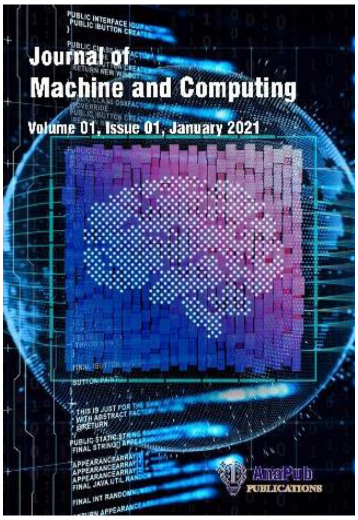
Journal Pre-proof



Detection and Recognition of Multi-Task Human Action Identification from Preloaded Videos Using CCTV Stationary Cameras

Pavankumar Naik and Srinivasa Rao Kunte R

DOI: 10.53759/7669/jmc202505120

Reference: JMC202505120

Journal: Journal of Machine and Computing.

Received 03 November 2024

Revised form 12 April 2025

Accepted 25 May 2025

AnaPub
PUBLICATIONS

Detection and Recognition of Multi-Task Human Action Identification from Preloaded Videos Using CCTV Stationary Cameras

[1]Pavankumar Naik and [2]Srinivasa Rao Kunte R
[1,2]Department of Computer Science and Engineering, Institute of Engineering and Technology, Srinivas University, Mangalore, Karnataka, India.
[1]pavanraj.cse@gmail.com, [2]kuntesrk@gmail.com

Correspondence should be addressed to Pavankumar Naik : pavanraj.cse@gmail.com

Abstract. Human activities include group activities, individual actions, and interactions between objects and people. In computer vision technology, recognizing and categorizing these activities is a vital process. This system will aim at developing a model that can recognize and detect such behavior and apply it in surveillance, healthcare, military operations, and patient monitoring. In the first place, videos were gathered in order to get a better understanding of the various human activities and interactions. Subsequently, we have converted Video frames into images and pre-processed each image. Characteristic features are extracted from video images by capturing spatial and temporal details. Spatio-temporal interest points using three descriptors Harris STIP, Gabor STIP, and HoG3D STIP are extracted as features. Extracted features are passed to a Heatmap generation process which gives confident key features related to human action. Support Vector Machine (SVM) Classifier is used to analyze these confident key features to label and classify human actions. Various classifier performance metrics, such as accuracy, sensitivity, and specificity, were used to evaluate the performance of the system. Classifier exhibiting accuracy of around 98.60% used as an indicator of the overall reliability of the proposed system in effectively recognizing human actions.

Keywords: Spatial-temporal Interest Point, Action Recognition, Multitask Human Action Recognition, HOG (Histogram of Oriented Gradients), Harris STIP, and Gabor STIP.

1 Introduction

Many applications, for example, video indexing and search, health care, sports visual systems, security surveillance, rely on the ability to interpret human actions in video inputs. The analysis of visual patterns in order to distinguish between different kinds of human actions is a complex task and demands consistent application of known techniques; it enhances computer visual perception abilities.

Three primary categories for a Human activity recognition are:

1. Single-user sensor-based action recognition: This technology uses sensors to identify and understand the actions or movements of a single individual. Applications include healthcare, sports analytics and security systems, where it helps monitor and analyze individual events using sensor data.

2. Multi-user sensor-based action recognition: It utilizes sensors to detect and interpret simultaneously the actions or movements of multiple people. The applications include smart environments, interactive games, crowd monitoring, and can be used for the recognition of multi-user actions and interactions through sensor data.

3. Action Recognition through Images: Here, the aim is to understand and recognize human gestures through images or even movie videos. This action of human computer interaction and sports analysis and tracking where computers learn through some visual data to recognize and understand gestures and actions. In order to segment video content effectively for indexing, search, and reliability, the actions that are present in the input video need to be identified. Unique mechanisms are required for complex human action recognition. HAR can be divided into two types [1, 2]:

(a) Low-level action recognition process: It involves action recognition based on extracted feature values. These processes are relatively easy to implement, but vary in reliability. They focus on identifying actions using feature points extracted from video clips of specific dimensions.
(b) High-level action recognition process: These processes are more robust and computationally intensive and require specific hardware such as high-resolution cameras. They offer increased reliability at the expense of increased computing requirements.

The final objective of HAR is to detect and classify activities performed by one or more individuals through the continuous monitoring of their activities and changes in environmental conditions.

2 Review of Related Works

The rapid development of technology has resulted in significant automation in most industries, which shows the increasing relevance of human face and facial expression detection in practical applications. Such applications range from subjects such as biometric identity, data privacy, information security, image and video security surveillance, human-computer interface (HCI), and human behavior interpretation (HBI) [3].

Taking into account the limitations of the HAR method, which relies on the STIP approach, improvements have been made concerning activity recognition through the inclusion of spatio-temporal (ST) interrelations involving various visual features of individuals. Researchers have attempted to look into the computer vision method that enhances Human-Robot Interaction (HRI). This idea involves developing systems that are able to extend the range of action capabilities [4]. A study was focused on the detection and recognition of activities using wearable sensors or mobile data collected by suitable sensors. The authors put forward the need for feature extraction to reduce the execution time and improve the accuracy of individual action recognition [5].

In [6], A technique that combines RGB and optical flow for HAR. Their work concentrated on the application of CNNs, which proved the feasibility of these networks in successfully detecting human visual actions in different input videos.

To identify and infer human actions from image or video data, current human action recognition systems make use of two predominant approaches: computer vision and machine learning. Generally, human action recognition systems can be roughly categorized into two approaches; 2D and 3D-based techniques. 2D-based methods classify human action by recognizing the 2D visual information. For example, pixel values and color distribution in an image or video frame. Optical flow analysis and histogram of oriented gradients are popular methodologies. While deep learning algorithms are Convolutional Neural Networks, CNN. All the 3D techniques involve the spatial and the temporal features of the actions that human beings undergo in the video. This comprises depths, movement, and series of activities that one uses overtime and usually record the images by utilizing 3-D sensors. Examples are 3D CNNs, motion histories volumes, skeletal point tracking.

Some of the most commonly used tools and libraries for developing human action recognition systems include OpenCV, TensorFlow, PyTorch, and Keras. Popular datasets such as UCF101, HMDB51, and Kinetics are being used to train and validate these systems.

These systems have applications in video surveillance, human-computer interaction, healthcare, and sports analysis. They are continuously evolving, with improvements in deep learning, sensor technologies, and larger annotated datasets that increase accuracy and robustness in recognizing a wide range of human activities.

The HAR mechanism helps in feature extraction of essential features from images for an improved understanding of the scene. The process is done seamlessly with the need for rearranging the current image and eventually leads to obtaining the result of the action along with its velocity in real time [7]. There have been many approaches proposed on variations of the Scale-Invariant Feature Transform, which detects a single action in a video, such as the extraction of samples and local descriptors along the motion trajectory based on SIFT [8].

A heatmap is a method of visualizing data using color gradients representing the intensity of data points so as to easily identify patterns and hotspots of high activity in human action recognition Heatmaps are handy as they can graphically point out useful information on notable body joints or movement regions in an image frame with which a computer model can classify the action more accurately.

Key points regarding heatmaps in human action recognition are:

Visualizing joint locations:
By overlaying the location of essential body joints on a heatmap, researchers can instantly notice which joints become most active while performing an action, helping with pose estimation and action recognition.

Identifying motion patterns:
The color intensity on the heatmap shows the extent of movement in each joint and enables the identification of fine motions that may otherwise be hard to detect.

Providing explanations of model predictions:
Used with machine learning models, heatmaps are able to display visual explanations for action classification by indicating which parts of the body it is paying attention to, while making a decision.

Benefits of Using Heatmap in HAR
Normalization across Key Points inside the frames: Conventional skeleton coordinates may differ substantially between key points of frames based on variations in measurement methods. Heatmap volumes eliminate this problem by creating a consistent representation that is easily computable and comparable between key points of frame. The research [9] proves that a pre-trained model is able to well extract shared motion features among human activities and achieve stable and accurate accuracy in all training conditions based on heatmap-based pseudo videos.
Rich feature representation: Heatmaps summarize spatial data regarding human poses without the requirement to distinguish between separate body parts explicitly. This feature enables capturing intricate actions that may be hard to categorize based on raw coordinate data alone. In human action recognition, heatmaps can be employed to describe human poses and movements in videos and calculates motion vectors in joint keypoints between consecutive frames [10].

The research paper [11] identifies human actions from pose estimation maps, something that has not been attempted in action recognition tasks previously. The method entails producing pose estimation maps from every frame of a video and then producing a heatmap and a pose to represent each frame.

HAR requires improved hybrid models for accuracy and anomaly detection. Paper [12] critically discusses the literature and proposes future development with hybrid optimization methods. This work introduces a deep learning HAR model based on a multiplicative 3D Convolutional Network. The four-stage model combines 3DCNN, LSTM, and Yolov6 to detect objects in real-time. It is more accurate than previous techniques, outperforming SOTA models on several datasets [13].

In [14], it presents a multi-sensor HAR system with improved action recognition through data refining and extracting critical features. Through a CNN-GRU classifier, it displays remarkable accuracy that surpasses the current models. Investigation of human action recognition with phase data from video rather than motion vectors is presented in [15]. A KNN classifier learns actions from phase correlations between frames. In [16], a HAR-LightCNN is introduced in a human activity recognition model based on Wi-Fi using CSI data. The model uses a lightweight CNN for real-time recognition and boosts performance with data augmentation.

The existing research on HAR has examined different scenarios, such as everyday life, group activities, and real-time activities. Most of the research has concentrated on simple activities pertaining to everyday life and user behavior, although there have been few studies on complex and real-time activity recognition in areas like healthcare, surveillance, and suspicious behavior. This scarcity of research is due to the challenges presented in recognition of real-time activities.

From the literature review we found that it is necessary to analyze the computational effectiveness of STIP-based approaches. Further, investigate alternative ways to represent STIP features to capture spatial and temporal information more effectively and explore the ways to increase the accuracy for the detection of simple as well as complex activities from a video.

3 Proposed HAR System and Methodology

The human visual system, comprising the human eye and brain, is an amazingly complex image processing system. Its ability to capture, interpret, and make sense of visual information inspires computer vision systems. We use a computer vision system in attempting to replicate this ability. This computer vision system takes a video input and breaks it down into individual frames. All these pictures are pre-processed to ensure quality, and this is enhanced by removing unwanted pixels in order to reduce noise. These processed images are preserved for future reference and re-use. An important concept of our methodology is information extraction or feature extraction of meaningful information from video pictures. This is accomplished with the use of STIP (Spatial Temporal Interest Points) descriptors of various sorts that are derived from preprocessed video frames. These descriptors allow for the identification and characterization of "distinguishing" features in images, making it easier to analyze and identify in the later stage. Fig. 1 depicts the detailed structure of the proposed HAR system.

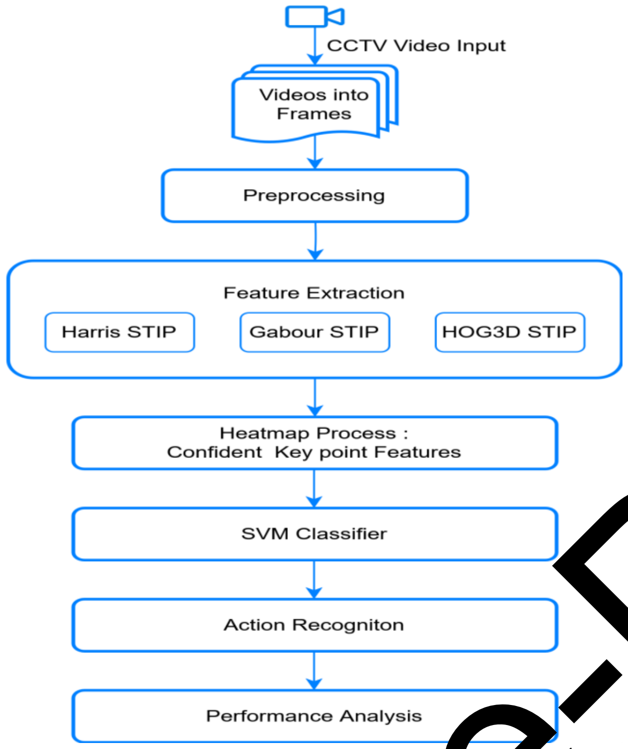


Fig. 1. The Structure of Proposed HAR System

The structure of the proposed HAR system has two phases one is the training phase and another is the test phase. Both the phases will undergo the same operations with a CCTV video given as input. The input video is converted into frames. Frames are preprocessed. Key points from each frame are captured using Harris STIP, Gabor STIP and HOG3D STIP descriptors. Extracted key points are further processed to find the Confident key points using Heatmap process. These confident key points are used for training our model using Support Vector Machine (SVM) classifier for categorizing the human actions. Similarly the testing phase will undergo all the processes as mentioned in Fig. 1 with the test video input containing the human action which will be detected and recognized by the trained classifier. Performance of the proposed HAR system is measured using accuracy, specificity and sensitivity.

Proposed system includes a GUI as shown in Fig. 2 that facilitates in selecting different operations involved in the HAR recognition process such as Load Video, Preprocessing, Feature Extraction, Action Recognition etc. These operations are explained in the following sections.

3.1 Video Input and Frame Extraction:
First Step in our proposed model is to load a CCTV video containing the human actions to be recognized as a source of input by selecting the 'Load Video' button in GUI. After loading the video, it is converted into the frames. Fig. 2 depicts the result of loading a sample video of Archery action, its size and its converted frames per second, both of which play a significant role in subsequent image analysis.

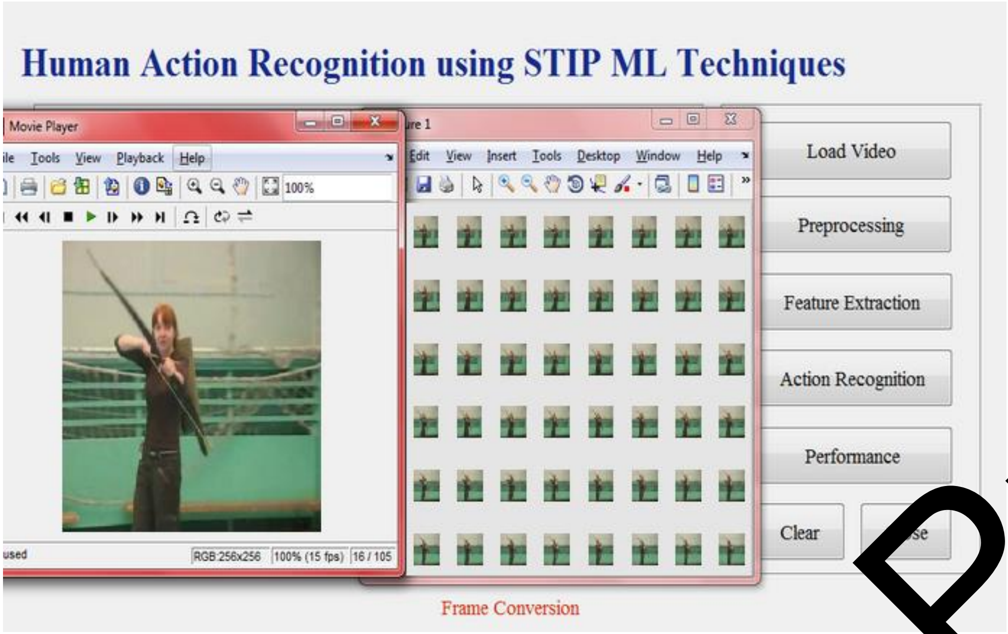


Fig. 2. Visualisation of the HAR system's input video loading and the video conversion process into individual frames.

Fig. 3 shows extracted images from frames of video. In order to gain a deeper understanding of activities, these images will be crucial in subsequent motion analysis during Scale-Invariant Feature Transform analysis.

Fig. 3. Extracted images from frames of an Archery action CCTV video

3.2 Preprocessing:

The goal of the preprocessing step is to eliminate various forms of noise present in the images, the most common being salt and pepper noise. This type of noise shows up as random white and black pixels scattered throughout the images. In addition, our preprocessing phase includes removing unwanted pixels and further enhancing image quality.

Videos with a high level of noise can be divided into three categories. The first type of noise is produced by very small fluctuations in the imaging device that are evenly distributed throughout the frame. A second form of noise is present near object boundaries, as indicated by regular jumps in values from background to foreground depth. The "holes" that appear in depth photographs are a third type of noise and are caused by random effects, fast movements, porous surfaces, and materials with unique reflectivity.

A combination of filtering techniques as described in [17, 18] is used to effectively remove noise from these images. Images are processed with a 2D Gaussian smoothing filter and then a temporal filter is used in all dimensions. Finally, the images are processed with a 1D complex Gabor filter.

3.3 Feature Extraction:

Once the noise is reduced, we proceed to extract key features from the preprocessed video frames. We use three different types of STIP descriptors, each offering a unique perspective. These descriptors calculate the basic information in the frame. We extract the remarkable features using Harris STIP, Gabor STIP, and HOG3D STIP.

The Harris STIP descriptor is used to identify corners in video frames. This algorithm not only identifies the corners but also takes into account the location of the corners using differential methods and directions. Additionally, it takes into account the sum of squared differences (SSD) to increase accuracy.

Gabor wavelets are used to identify corners at the exact position of an object using the Gabor STIP approach. The local spectral energy density provided by Gabor functions allows for two orthogonal directions of wavelet convolution with different scales.

Histograms of gradient values at different images are extracted using the HOG3D STIP approach. The image is split into small, connected blocks known as cells. Linked to each cell is a histogram of the orientations of gradients or edge detections for pixels within the cell. All these histograms are combined in order to create the descriptor. Accuracy can be enhanced by performing contrast normalisation on the local histograms. This is achieved by measuring the intensity in a block, which is a larger area of the image. All the cells in the block are then normalised.

3.3.1 Harris STIP Method

This method fulfills a key role in computer vision systems since it is designed to detect corners in images. The task of corner detection in images has extensive applications in various fields. Corners are the basic features of an image. They are usually at the intersection of two edges, which mark the points of sudden changes in brightness and draw the different elements of the image from them. The algorithm relies on a corner scoring mechanism that considers the corner scores variation with respect to direction.

For images in which all the pixels have nearly equal intensity, edges appearing in adjacent pixels are nearly indistinguishable. In contrast, images where most pixels have non-negligible differences in intensity do reveal significantly different regions; in this case, the edges. It is such differentiation between similar and non-similar regions that lies at the foundation of corner detection.

Many applications in computer vision need corner identification. It is widely applied in motion detection, image mosaicking (merging multiple images into a single composite image), image registration (aligning two or more images), video tracking, panorama stitching (creating panoramic views from multiple images), 3D modeling, and many other types of object recognition. Such versatility makes corner detection a well-founded and versatile tool in computer vision, enabling valuable information abstraction and image content inference.

Harris Corner Detection System

We locate points based on intensity changes in a nearby neighbourhood using the Harris mechanism. As described in [19, 20], this mechanism concentrates on the small portion of the element where the greatest shift in intensity levels is noticed in comparison to moving the windows in any direction. The following autocorrelation functions help to clarify this idea.

A scalar function P, represented as P(R), characterizes P as a scalar function as shown in equation (1). The symbol ▲, denoted as h, means a slight increment at any point in the domain. The corners are identified as points of x that give remarkable values of the function shown for infinitesimal h.

$$E(h) = \Sigma w(a)[P(a + h) - P(a)] \qquad (1)$$

This indicates a significant change in various directions. The function w(a) allows the selection of a support region, commonly known as a Gaussian function. To linearize the expression P(a + q), Taylor expansions will be used as follows:

$$P(a + q) \approx P(a) + \nabla(a)Tq$$

Hence, the right hand of equation (1) gives

$$E(q) \approx \sum w(a)(\nabla P(a) \cdot q)^2 \, da = \Sigma w(a)(q^T \, \nabla P(a) \, \nabla P(a)^T q) \quad (2)$$

The final equation (2), is dependent on the image gradient, included in the autocorrelation matrix or tensor structure, expressed as

$$Z = \sum w(a)(\nabla P(a) \, \nabla P(a)^T) \qquad (3)$$

The initial Z in equation (3) is an eigenvalue indicating the orientation of the most significant intensity change while the slave eigenvalue is aligned with the perpendicular direction of the intensity change.

Fig. 4 illustrates the Harris STIP features extraction points of frame 21 of the Archery action video. Extracted key points are shown as circles in the frames and its respective values are preserved.

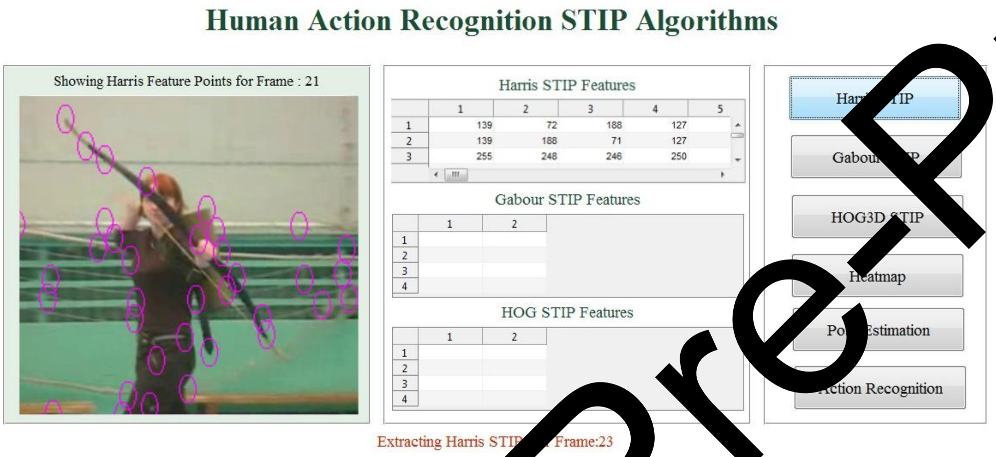


Fig. 4. Harris STIP features extraction points of frame 21.

3.3.2 Gabor Method

The Gabor function plays a key role in capturing the energy distribution of local spectral values situated at a certain location and frequency orientation. When used in two-dimensional convolution in the circular domain, Gabor functions display unique attributes different from their one-dimensional counterparts. In the context of corner detection, we use Gabor wavelets, serving as second-order partial derivative (PD) operators. Gabor functions find significant applications in edge detection and are named in recognition of Dennis Gabor.

Both the orientation description and the frequency attributes of Gabor filters have strong similarities to human component analysis techniques. They have proven to be extremely suitable for tasks such as texture description and differentiation. A Gabor filter can be thought of as a spatial-domain 2D filter that modulates the plane wave of a sine signal using a Gaussian kernel function.

The Gabor filter has a wide range of applications, including pattern recognition, optical character identification, fingerprint recognition, and facial expression recognition. These filters are particularly valuable for their ability to capture and represent complex visual patterns, making them a versatile tool in a wide range of image and signal processing tasks.

Gabor filter features: The basic feature extraction of a two-dimensional Gabor filter also known as multi-resolution Gabor filter is created by combining the outputs of Gabor filters applied at different frequencies (fa) and orientations, resulting in different representations. Equation (4) describes these frequency representations.

$$f_a = h - a f_{\max} \quad a = \{0, \ldots, A-1\} \qquad (4)$$

In this context, fa symbolizes the ath frequency, where, fmax ≥ 0 represents the maximum generated frequency, and h > 1 serves as the frequency scaling factor. θn represented in equation (5) are the filter orientations, and these orientations are determined as follows:

$$\theta n = 2\pi n/N = 0,\ldots,N-1 \qquad (5)$$

Here, n represents the nth orientation, and N is the maximum number of orientations and Fig. 5 illustrates the features extraction points of frame 21 of Archery video using Gabor method.

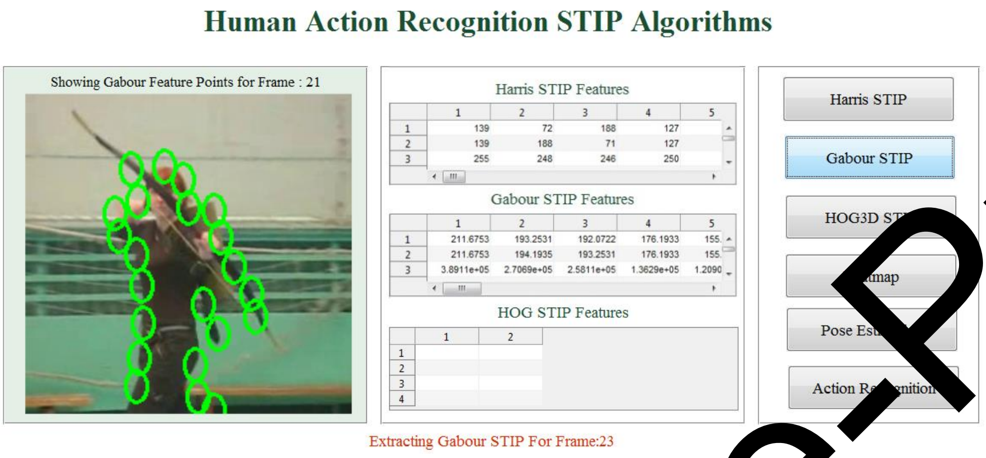


Fig 5. Gabor STIP features extraction points of frame 21.

3.3.3 HOG3D STIP

Histogram of Oriented Gradients (HOG3D) is a widely used method in computer vision and image processing for object detection that relies heavily on feature descriptors. The process involves dividing the image into interconnected segments called cells. In each cell, we calculate the HOG3D directions or edge orientations for all pixels. The gradient weights for each angular bin are determined based on the contributions of each pixel in the cell. To improve the analysis, we group neighboring cells into blocks that form spatial regions. This grouping of cells into blocks forms the basis for histogram categorization and normalization [18].

Histogram of Oriented Gradients Calculation:

In the early stage of descriptor construction in HOG, the process involves the calculation of one-dimensional derivative points in a and b directions, denoted G_a and G_b. This is achieved by convolving the gradient masks M_a and M_b with the source image I as shown in equation (6) and equation (7).

$$G_a = M_a * I \quad where \quad M_a = -(1\,0\,1) \qquad (6)$$

$$G_b = M_b * I \quad where \quad M_b = -(1\,0\,1)^T \qquad (7)$$

The basis functions G_a and G_b use derivatives to calculate the size of |G(a, b)| and the angle in the F(a, b) direction for each pixel in the HOG3D descriptor.

As shown in equation (8), the degree of HOG indicates its power in pixels.

$$|G(a,b)| = \sqrt{Ga(a,b)2 + Gb(a,b)2} \qquad (8)$$

Fig. 6 shows the HOG3D features extraction points of frame 21 of Archery video and its respective key points. The dense color spectrum in the frame 21 displays gradient intensities (HOG feature magnitudes) at various locations in space and time. More dense intensity indicated using Thick blue color in HOG3D STIP shows that action present in the frame.

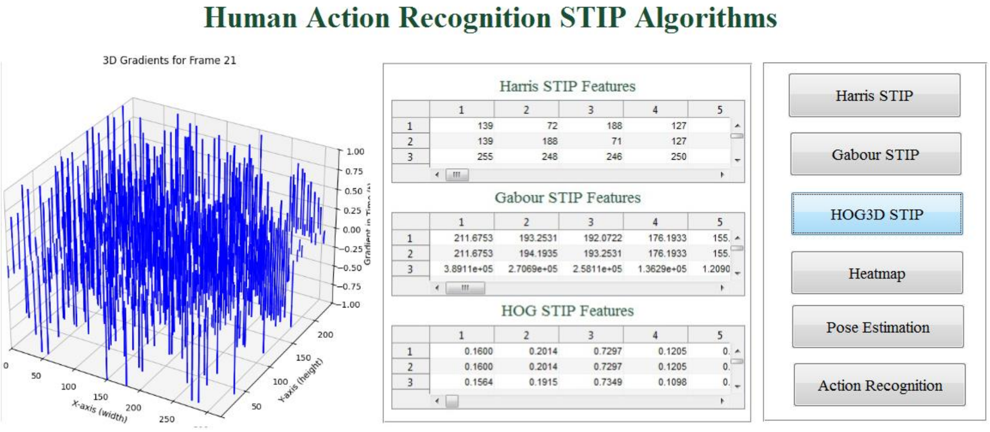


Fig. 6. HOG3D STIP features extraction points of frame 21.

In Fig. 7 it shows that when we overlay the color on extracted feature points, we can check that action has been traced in frame 21.

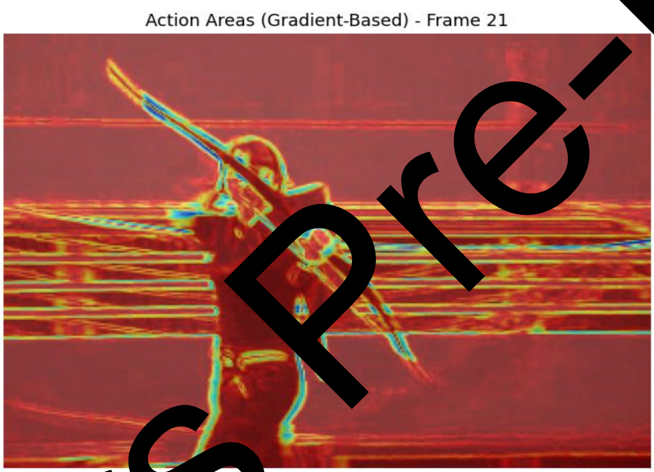


Fig. 7. Color overlay on HOG3D STIP feature extraction points of frame 21.

After processing all the frames using all the three methods of STIP, the extracted feature points are used to generate Heatmap and Pose Estimation.

3.3.4 Heatmap Generation and Pose Estimation

Heatmaps are visual representations of data intensity or density over a provided space. They are frequently utilized in a number of applications such as, HAR and Heatmap offer a method to denote human poses and movement over time, facilitating understanding and classification of actions.

Each individual keypoint of the frames extracted from all three STIP methods are given as input for generating heatmap for each frame. Heatmaps give a confidence of the key point's presence at that location in each frame. After generating a heatmap for all the frames, the next step is to extract coordinates of the key points (detecting the accurate coordinates) from these frames. High values in heatmap correspond to high confidence of key points. For each heatmap, we find the maximum confidence point, which corresponds to the location of the confident key points as shown in Fig. 8.

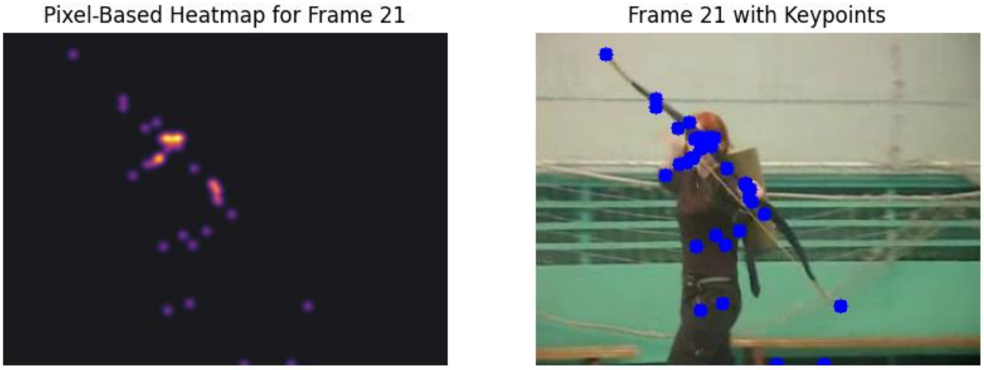
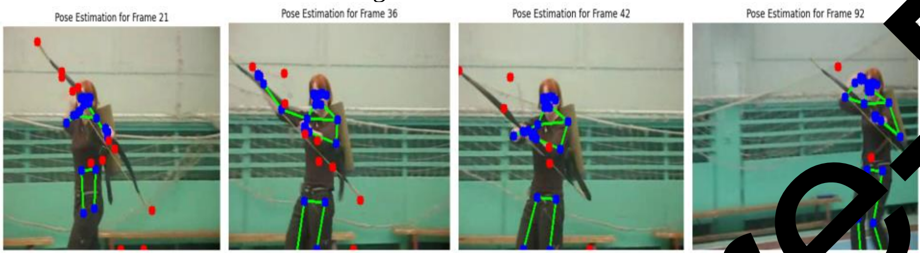


Fig. 8. Pixel based Heatmap for frame 21 and detected confident key points.

Pose Estimation:
If lines are drawn between the confident key points to form the human skeleton, then this gives a pictorial Pose Estimation of the human action as shown in Fig. 9 for different frames of the same video.



(a) (b) (c) (d)

Fig. 9. Pose Estimation by drawing lines between the confident key points of different frames, (a) represents the pose estimation of frame 21, (b) represents the pose estimation of frame 36, (c) represents the pose estimation of frame 42 and (d) represents the pose estimation of frame 92.

4 HAR Classifier System for Recognition of Actions

The proposed HAR recognition system uses a SVM as a classifier to accurately identify human actions in the given video. A probabilistic binary linear classifier is known as a SVM. It captures key motion features, transforming complex movements into identifiable patterns for classification. The SVM ensures reliable performance by optimizing decision boundaries, making it effective especially for real-time applications like surveillance, healthcare, and human-computer interaction. Its ability to handle large datasets and adapt to different activities enhances accuracy while keeping computational demands low .

Training and Testing of SVM Classifier

We have used the UCF 50 dataset containing videos for different human actions to train our SVM classifier. The videos of Archery, Baby Crawling, Bowling, Clapping, Boxing Punching Bag, Hair drying, Front kick, Playing Cricket, Bending, Sideway kick, Horse riding and Surfing actions were used for training the system. Around 75% of the dataset of each human action was used to train the system and the remaining dataset for testing the recognition results.

From each human action video, key points of the frames are extracted using the proposed three STIP methods. They are given as input for generating the heatmap to obtain the confident key points for each frame. These confident key points are used as the input for the classifier to train it to recognize the particular human action.

After training the classifier, the system was tested for the recognition results of the human actions. The proposed HAR system accurately recognized the different human actions. We could obtain an average recognition accuracy of 98.61%.

Sample recognition results for the Archery, Baby Crawling, Bowling, Boxing Punching Bag and Horse Riding actions are shown in Fig. 10.

Fig. 10. Sample recognition results for the human action of (a) Archery (b) Baby Crawling (c) Bowling (d) Boxing Punching Bag and (e) Horse Riding.

5 HAR System Performance Indicators

Evaluation of the HAR mechanism involves the assessment of the performance parameters of the classifier, including accuracy, sensitivity, and specificity. The accuracy of a classifier measures its success in correctly identifying an image based on a provided label. The true hit rate, or recall rate, measures how accurately the classifier assigns data to specified categories [22]. Conversely, specificity assesses the classifier's ability to reject data that does not belong to any category, also known as the true-negative rate.

The precise computations for fundamental performance measures, including sensitivity, specificity, and accuracy, with respect to a particular input action video are shown in the following equations.

Sensitivity, also referred to as True Positive Rate, is determined by the ratio of True Positives to the sum of True Positives and False Negatives.

$$Sensitivity = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative (FN)} \qquad (9)$$

Specificity is calculated as the ratio of True Negatives to the sum of False Positives and True Negatives.

$$Specificity = \frac{True\ Positive\ (TP)}{False\ Positive\ (FP) + True\ Negative (TN)} \qquad (10)$$

Accuracy is determined by the ratio of the sum of True Positives and False Negatives to the sum of False Positives and True Negatives.

$$Accuracy = \frac{True\ Positive\ (TP) + False\ Negative\ (FN)}{False\ Positive\ (FP) + True\ Negative (TN) + TP + FN} \qquad (11)$$

Equations (9), (10), and (11) express the mathematical relationships for Sensitivity, Specificity, and Accuracy, respectively.

The proposed model was evaluated using the UCF-50 database. Table 1 depicts the performance indicators of the proposed HAR system in recognising the different human actions. The average accuracy attained was 98.61% with 100% specificity and 97.38% sensitivity.

Table 1. Performance analysis of the proposed HAR model on UCF100 database

Human Actions	**Accuracy**	**Specificity**	**Sensitivity**
Archery	99.6	100	97.9
Baby Crawling	99.1	100	98.2
Bowling	98.6	100	97.4
Clapping	99.4	100	99.1
Boxing Punching Bag	98.2	100	97.5
Hair drying	97.7	100	96.4
Front kick	98.9	100	97.6
Playing Cricket	97.6	100	95.65
Bending	97.0	100	95.4
Sideway kick	98.9	100	98.4
Horse riding	99.7	100	97.3
Surfing	98.6	100	97.7

Further, the proposed system was evaluated by testing its recognition results for the similar human actions data available in a different dataset. Table 2 shows the results obtained when the system was tested with human actions available in Kinetics 400 dataset.

Table 2. Performance analysis of the proposed HAR model on Kinetic 400 database

Human Actions	**Accuracy**	**Specificity**	**Sensitivity**
Archery	99.7	100	97.9
Baby Crawling	99.3	100	98.4
Bowling	98.6	100	97.5
Clapping	99.5	100	99.3
Boxing Punching Bag	98.5	100	97.6
Playing Cricket	97.7	100	95.8
Bending	97.2	100	95.7
Sideway kick	98.9	100	98.4
Horse riding	99.7	100	97.3
Surfing	98.4	100	97.4

Features Selection:
Before finalizing the features to be used to train the proposed model, we did experiments in training and testing the system by selecting the combination of the different STIP features. The recognition results were better when we used all the three STIP features with Heatmap and Pose estimation. Hence, we finalized to use the combination of all the three STIP features with Heatmap and Pose estimation to train our HAR recognition model. The results obtained for the Archery and Baby crawling human actions for different combinations of features are shown in Table 3 as sample results. From the table results it is evident that the accuracy was maximum when all the three STIP features were used along with the heatmap and pose estimation.

Table 3. Result Analysis of different Hybrid STIP Model for Archery and Baby crawling human actions

Activity	Features	Accuracy	Sensitivity	Specificity
Archery	Harris + Gabor	83.95	88.71	63.67
	Gabor +HoG3D	69.19	80.63	81.07
	Harris + HoG3D	82.54	63.97	93.76
	Harris + Gabor + HoG3D	98.20	100.00	95.00
	(Harris + Gabor + HoG3D) + Heatmap	99.70	100.00	97.9
Baby Crawling	Harris + Gabor	84.12	96.80	97.03
	Gabor +HoG3D	97.45	69.10	87.79
	Harris + HoG3D	82.81	88.45	88.93
	Harris + Gabor + HoG3D	97.30	100.00	95.00
	(Harris + Gabor + HoG3D)+Heatmap	99.30	100.00	98.40

Fig. 12 depicts the performance analysis results of the proposed model for a set of activities. The findings of the experiment of testing the system for different human actions shows that the proposed model performs better at recognizing the human actions. The model exhibits higher specificity, meaning that it is good at not mistakenly identifying actions when they aren't actually happening. Fig. 13 shows a graphic depiction of the performance of the proposed model.

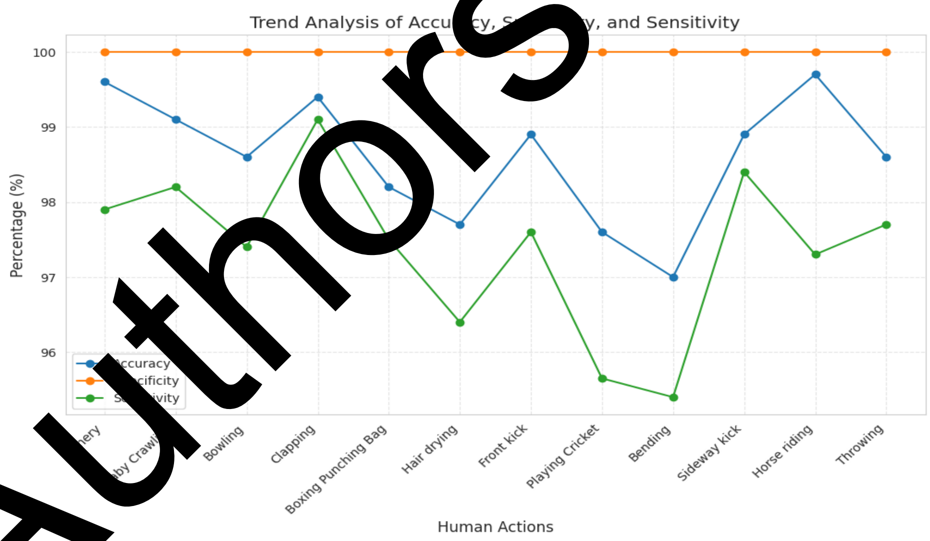


Fig. 12. Performance analysis result of proposed HAR model

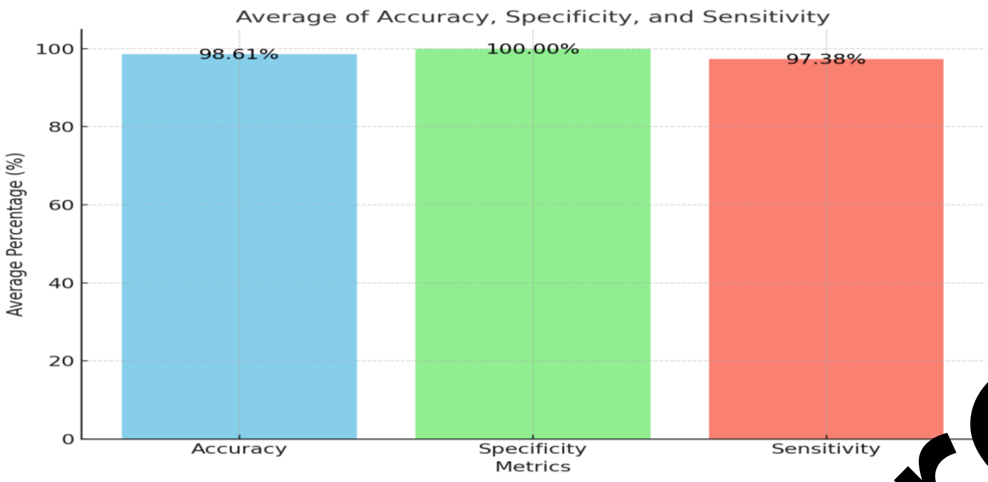


Fig. 13. Graphical evaluation of proposed HAR model

Table 4 compares the performance of our HAR recognition model with other recognition models as reported in the literature. It is evident that our proposed model has better recognition accuracy.

Table 4. Performance Analysis Matrix.

Model	**Accuracy (%)**
Traditional HAR (SVM + HOG) [23]	78.4
Deep Learning (RNN + LSTM) [23]	86.7
CNN Model [23]	92.5
CVRL (Contrastive Video Representation Learning Model) [24]	92.2
HAR STIP Method [25]	95.2
Proposed HAR Recognition Model	98.6

Table 5 gives the confusion matrix results of our HAR recognition system for the Surfing activity. It is worth noting that the diagonal parts of the confusion matrix have higher values than the ones in the top and lower triangular portions, indicating that this activity is recognized 100% accurately.

Table 5. Confusion matrix of Proposed HAR system for surfing action

37	0	0	0	0	0	0	0
3	38	0	0	0	0	0	0
0	2	39	0	0	0	0	0
0	0	1	39	0	0	0	0
0	0	0	1	37	0	0	0
0	0	0	0	3	40	0	0

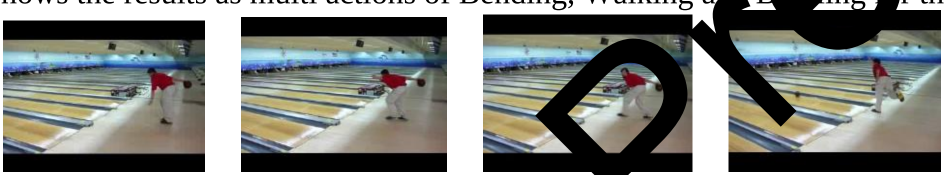
0	0	0	0	0	0	36	1
0	0	0	0	0	0	4	39

Multi-task Recognition:

The proposed system can recognize more than one human action; multi-task present in a given human action video by suitably modifying the training of the classifier. For example, for the human action video containing the Bowling action, it has Bending, Walking and Bowling activities together. In Table 1 for that video while training the classifier we have labelled that final action as a single action of Bowling. Instead, we can also train it to include the Bending, Walking and Bowling actions together, in which case the classifier generates the detection and recognition result as Bending, Walking and Bowling.

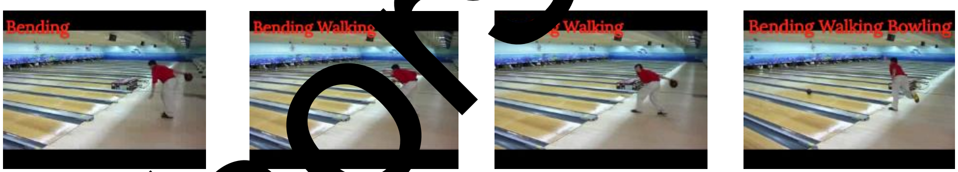
Fig. 14 shows the different frames of a Bowling sample video that contains the combination of Bending, Walking and Bowling actions. Fig. 14 (a) shows the Frame 7 of the sample video that has the action of Bending, (b) and (c) are the Frames 10 and 13 of Walking action and (d) is Frame 25 showing the action of Bowling.

The results of the action recognition of our modified multi-task recognition system for the selected sample video are shown in Fig. 15. The recognition results are shown on the top of the frames. Recognition result output of Frame 7 is Bending as shown in Fig 15 (a). The results of Frame 10 and 13 are multi actions of Bending and Walking as shown in Fig. 15 (b) and (c). Whereas, Fig 15 (d) shows the results as multi actions of Bending, Walking and Bowling for the Frame 25.



(a) (b) (c) (d)

Fig. 14 Different frames of a Bowling sample video (a) Frame 7 (b) Frame 10 (c) Frame 13 (d) Frame 25



(a) (b) (c) (d)

Fig. 15. Recognition Results for the different frames of Fig. 14.
(a) Frame 7 Action result - Bending (b) Frame 10 Action result - Bending Walking (c) Frame 13 Action result - Bending Walking (d) Frame 25 Action result - Bending Walking Bowling

4 Conclusion

The proposed HAR recognition system identifies the actions performed by an individual in the video using features gathered from STIP methods. In addition to STIP features, heatmap gives more prominent key points which gives much more accurate results. Action recognition is performed by applying the kernel function within the SVM classifier. Proposed system demonstrates enhanced accuracy compared to existing methodologies due to a reduction in categorization variances. STIP detectors and descriptors have been adapted to accommodate diverse photometric channels and image intensities, resulting in the utilization of STIPs. The proposed method precisely identifies the actions carried out by the individual in the video. Despite challenges such as lighting variations, contrast disparities, swift movements, and changes in the scale of the individual in the video, the results were accurate and the average accuracy attained was 98.61% with 100% specificity and 97.38% sensitivity.

5 Future Scope

The optimization of feature extraction and classification would further upgrade action recognition in complicated real-life situations. By integrating spoken actions and facial expressions, this framework favors intelligent monitoring systems to boost realistic performance of HAR Systems.

References

1. Pavankumar Naik et.al. (2023) Review of Literature on Human Activity Detection and Recognition. International Journal of Management, Technology, and Social Sciences (IJMTS) 8(4) Nov 2023
2. Lalitha K, Deepika TV, Sowjanya MN, Michahial S (2016) Human identification based on iris recognition using support vector machines. Int J Eng Res Electr Electron Eng (IJEREEE) 2(5) May 2016
3. Mahanthesh U, Mohana HS (2016) Identification of human facial expression signal classification using spatial temporal algorithm. Int. J. Eng. Res. Electr. Electron. Eng. (IJEREEE) 2(5)
4. Efthymiou N, Koutras P, Filntisis PP, Potamianos G, Maragos P (2018) Multi-view fusion for action recognition in child-robot interaction. 978-1-4799-7061-2/18/$31.00 ©2018 IEEE
5. Friday NH, Mujtaba G, Al-garadi MA, Alo UR (2018) Deep learning fusion conceptual frameworks for complex human activity recognition using mobile and wearable sensors: 978-1-5386-1370-2/18/$31.00 ©2018 IEEE
6. Khong V, Tran T (2018) Improving human action recognition with two-stream 3D convolu- tional neural network. 978-1-5386-4180-4/18/$31.00 ©2018 IEEE
7. El Din Elmadany N (2018) Student Member, IEEE, Yifeng He, Member, IEEE, and Ling Guan, Fellow, IEEE. Information fusion for human action recognition via biset/multiset globality locality preserving canonical correlation analysis. IEEE Trans Image Process 27(11) Nov 2018
8. Pavithra S, Mahanthesh U, Michahial S, Shivakumar M (2016) Human motion detection and tracking for real-time security system. Int J Adv Res Comput Commun Eng ISO 3297:2007 Certified 5(12) Dec 2016
9. Yuan, L., He, Z., Wang, Q., Xu, L., & Ma, X. (2023). Improving small-scale human action recognition performance using a 3D heatmap volume. *Sensors*, *23*(14), 6364.
10. Song, I., Lee, J., Ryu, M., & Lee, J. (2024, August). Motion-aware heatmap regression for human pose estimation in videos. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (pp. 1245-1253).
11. Liu, M., & Yuan, J. (2018). Recognizing human actions as the evolution of pose estimation maps. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1159-1168).
12. D. Nagpal and S. Gupta, "Human Activity Recognition and Prediction: Overview and Research Gaps," *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, Lonavla, India, 2023, pp. 1-5, doi: 10.1109/I2CT57861.2023.10126458.
13. C. Gupta *et al*., "A Real-Time 3-Dimensional Object Detection Based Human Action Recognition Model," in *IEEE Open Journal of the Computer Society*, vol. 5, pp. 14-26, 2024, doi: 10.1109/OJCS.2023.3334528.
14. Batool, M., Alotaibi, M., Alotaibi, S. R., AlHammadi, D. A., Jamal, M. A., Jalal, A., & Lee, B. (2024). Multimodal Human Action Recognition Framework using an Improved CNNGRU Classifier. IEEE Access
15. Hejazi, S. M., & Abhayaratne, C. (2022). Handcrafted localized phase features for human action recognition. Image and Vision Computing, 123, 104465.
16. El Zein, H., Mourad-Chehade, F., & Amoud, H. (2024). CSI-based Human Activity Recognition via Lightweight CNN Model and Data Augmentation. IEEE Sensors Journal.
17. Ryoo MS, Aggarwal JK Stochastic representation and recognition of high-level group activ- ities. Robot Research Department, Electronics and Telecommunications Research Institute, Korea, e-mail: mryoo@etri.re.krs
18. Xia L, Aggarwal JK Spatio-temporal depth cuboid similarity feature for activity recogni- tion using depth camera. Computer and Vision Research Center/Department of ECE, The University of Texas at Austin, aggarwaljk@mail.utexas.edu
19. Holte MB (2012) Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments. IEEE J Sel Top Sign Process 6(5) Sept 2012
20. Aggarwal JK, Ryoo MS Human motion: modeling and recognition of actions and interactions. In Proceedings of the 2nd international symposium on 3D data processing, visualization, and transmission (3DPVT'04) 0-7695-2223-8/04 $ 20.00 IEEE
21. P. B S and N. Thillaiarasu, "Personality Prediction From Handwriting Using Adaptive Deep Convolutional Recurrent Neural Network," 2025 Fifth International Conference on Advances in Electrical, Computing, Communication and

Sustainable Technologies (ICAECT), Bhilai, India, 2025, pp. 1-8, doi: 10.1109/ICAECT63952.2025.10958973
22. Jafari R, Kehtarnavaz N (2018) A survey of depth and inertial sensor fusion for humanactionrecognition.https://link.springer.com/article/10.1007/s11042-015-3177-1. 07/12/2018
23. Harshavardhan Patil et.al. (2024) Precise Human Activity Recognition using Convolutional Neural Network and Deep Learning Models. International Journal of Computer Sciences and Engineering Vol.12, Issue.7, pp.24-32. DOI: https://doi.org/10.26438/ijcse/v12i7.2432
24. R. Qian et al., "Spatiotemporal Contrastive Video Representation Learning," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 6960-6970, doi: 10.1109/CVPR46437.2021.00689.
25. Mohana, D. H., & Mahanthesha, U. (2020). Human action Recognition using STIP Techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN*, 2278-3075.