Review of Literature on Human Activity Detection and Recognition

Pavankumar Naik [1]*, & R. Srinivasa Rao Kunte [2]
[1] Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India.
Orcid ID: 0000-0002-4295-0309; E-Mail ID: pavanraj.cse@gmail.com
[2] Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India.
Orcid ID: 0000-0002-5062-1505; E-Mail ID: kuntesrk@gmail.com

Area/Section: Computer Science.
Type of the Paper: Review paper.
Type of Review: Peer Reviewed as per |C|O|P|E| guidance.
Indexed in: OpenAIRE.
DOI: https://doi.org/10.5281/zenodo.10197162
Google Scholar Citation: IJMTS

How to Cite this Paper:
Naik, P., & Kunte, R. S. R. (2023). Review of Literature on Human Activity Detection and Recognition. *International Journal of Management, Technology, and Social Sciences (IJMTS)*, *8*(4), 196-212. DOI: https://doi.org/10.5281/zenodo.10197162

International Journal of Management, Technology, and Social Sciences (IJMTS)
A Refereed International Journal of Srinivas University, India.

CrossRef DOI: https://doi.org/10.47992/IJMTS.2581.6012.0318

Received on: 10/06/2023
Published on: 23/11/2023

Review of Literature on Human Activity Detection and Recognition

Pavankumar Naik [1]*, & R. Srinivasa Rao Kunte [2]
[1] Research Scholar, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India.
Orcid ID: 0000-0002-4295-0309; E-Mail ID: pavanraj.cse@gmail.com
[2] Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore, India.
Orcid ID: 0000-0002-5062-1505; E-Mail ID: kuntesrk@gmail.com

ABSTRACT

Purpose: *The objective of this research article is to methodically combine the existing literature on Human Activity Recognition (HAR) and provide an understanding of the present state of the HAR literature. Additionally, the article aims to suggest an appropriate HAR system that can be used for detecting real-time activities such as suspicious behavior, surveillance, and healthcare.*
Objective: *This review study intends to delve into the current state of human activity detection and recognition methods, while also pointing towards promising avenues for further research and development in the field, particularly with regards to complex and multi-task human activity recognition across different domains.*
Design/Methodology/Approach: *A systematic literature review methodology was adopted by collecting and analyzing the required literature available from international and national journals, conferences, databases and other resources searched through the Google Scholar and other search engines.*
Findings/Result: *The systematic review of literature uncovered the various approaches of Human activity detection and recognition. Even though the prevailing literature reports the investigations of several aspects of Human activity detection and recognition, there is still room for exploring the role of this technology in various domains to enhance its robustness in detecting and recognizing of multiple human actions from preloaded CCTV cameras, which can aid in detecting abnormal and suspicious activities and ultimately reduce aberrant human actions in society.*
Originality/Value: *This paper follows a systematic approach to examine the factors that impact the detection and recognition of Human activity and suggests a concept map. The study undertaken supplements the expanding literature on knowledge sharing highlighting its significance.*
Paper Type: *Review Paper.*
Keywords: Human Activity Recognition, Human Activity detection, Human Action Detection, Multi-Task Recognition, Spatio-Temporal Interest Points (STIPs).

1. INTRODUCTION :

In the field of computer vision, computers can now mimic human visual systems, thanks to the field of computer vision. It takes data from digital photographs or videos and processes it to determine the content, quality, and so on as a subset of artificial intelligence. The entire procedure entails the acquisition of images, screening, analysis, identification, and information extraction. Because of this deep processing, computers can interpret any visual input and respond correctly. Human action/interaction, MHAR (Multitask Human Action Recognition) is a key component in current research. MHAR using computer vision involves the understanding of motion and clutter in the given image or video, which is a complex and challenging issue in Multi-task based systems. For example, a long jump involves a series of physical motions that last a certain amount of time, such as sprinting, jumping, landing, and standing up. When an action interacts with one or more objects, the complexity

of the action increases. Hitting a cricket ball with a cricket bat, for example, entails a person utilizing one object (the cricket bat) to execute an action on another object (the cricket ball).

Basically a technique called Human Activity Recognition (HAR) can automatically identify what a person is doing in terms of the body's motion and function. Utilizing many technologies, including cameras, motion sensors, location sensors, and time, the fundamental objective in HAR is to identify a person's behavior. Recognizing human activity is crucial in a variety of fields, including social sciences, ubiquitous computing, artificial intelligence, human-computer interaction, healthcare, health outcomes, and rehabilitation engineering. There are various ubiquitous and pervasive computing systems in which user behaviors play a vital role. Human behavior conveys a wealth of context information and aids systems in achieving context awareness. It aids in functional diagnosis and measuring health outcomes in rehabilitation. Human activity recognition is a significant indication of involvement, quality of life, and lifestyle.

Human activities are divided into two categories based on movement of the body and function. The first class, which is a basic human activity, focuses on human body motion and posture while identifying various activities such as walking, running, and sitting. The second type is a complicated human activity or a multi-tasking human activity. Cooking, reading, and watching television, playing soccer, vacuuming, working at a desk or computer, brushing teeth, and eating a meal are all examples of multi-task human activity.For example, when a person is reading a book, he or she is most likely sitting somewhere (chair or sofa). The complicated action here is "reading a book," which is made up of the simple activity "sitting" and the function "reading." Another example is talking on the phone while watching television.

Human activity recognition has been an active multidisciplinary study subject for almost a decade. Significant study has been undertaken to identify human activities. However, there are other serious difficulties that must be solved. Addressing these challenges will result in considerable improvements in various elements of HAR applications in several fields. There has been a lot of study done on simple human activity identification, but very little research has been done on complicated human activity recognition. However, several essential elements (identification accuracy, computational cost, energy consumption, mobility) in these sectors must be addressed to increase their feasibility. Human activities can be categorized into different levels as indicated in Fig. 1.

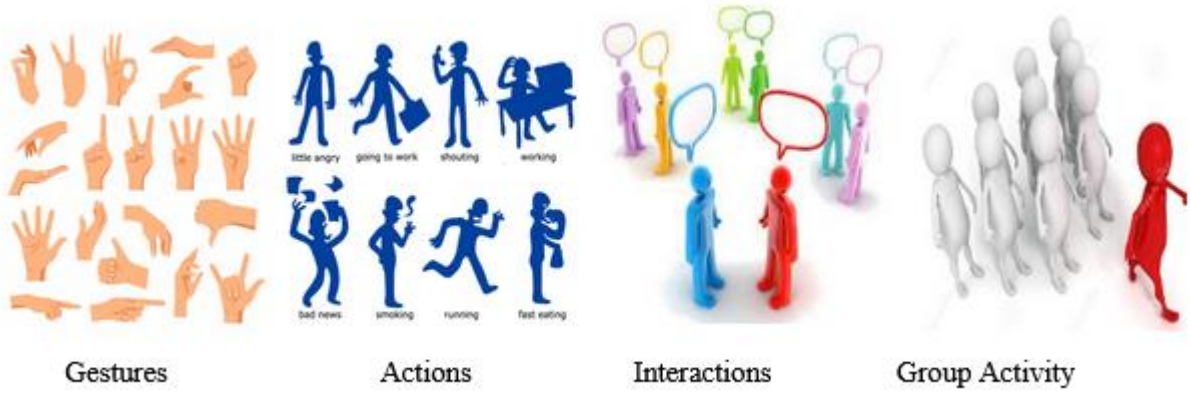

Fig. 1: Types of Human Activities.

Gestures – These are basic movements of a person or a person's bodily part that serve as atomic components in characterizing a person's meaningful motion.

- **Actions** – These are an object's actions that can consist of several movements organized in time, such as walking, running, boxing, waving, punching, and so on.
- **Interactions** - These are object activities that involve two or more people and/or objects, such as two people shaking hands, two people fighting, etc.
- **Group activities –**These are the actions carried out by conceptual groupings made up of many people and/or objects. A marching group, two or more groups fighting, and so on are common instances of group activity.
- **Emotions** – These are a person's good or negative mental states that are associated with a pattern of any physiological processes. Emotions characterize a person's mental state. There are six types

of basic emotions: happiness, anger, fear, disgust, surprise, and sorrow. These six fundamental emotions combine to generate complex emotions.

Images in a video involving human activities or actions will have considerable local fluctuations in both space and time. We require local interest point features or operators to identify local structures in space and time that give concise and abstract representations of patterns in the image or video for reorganization of any human activity or activities in image or video. Locally invariant features can withstand changes in rotation, scale, and perspective.

Spatio-Temporal Interest Points (STIPs) are a category of spatio-temporal invariant features used in video analysis. STIPs have advantages over global features, as they are more robust to video fluctuations like geometry transformation, perspective transformation, illumination variation, and convolution transformation. Moreover, STIPs can be directly detected from video to describe moving objects, avoiding the requirement for background modeling and foreground segmentation, Rasheed, M. B.et al.(2015) [1]. STIPs are useful in detecting human activity in environments with occlusions and a cluttered background.

The study of Spatio-Temporal Interest Points (STIP) has gained popularity in the field of video analysis due to its various applications, including HAR, video surveillance, video summarization, and content-based video retrieval. Several researchers have extensively studied STIP detection. The authors Mohana, H. S., & Mahanthesha, U. (2021, 2020) [2, 3] provide a comprehensive overview of available STIP detection techniques and their significance. They also discuss the current challenges in video STIP identification, including low time efficiency, weak robustness to camera movement, lighting change, perspective occlusion, and backdrop clutter. Many effective methods have been proposed for human activity recognition, Holte, M. B.etal.(2012) [4] and are widely used in various applications. Aggarwal J.K., & Park S (2004) [5], concentrate on four areas of high-level processing: (1) human body modelling, (2) the amount of information required to understand human activities, (3) methods for recognizing human actions, and (4) domain knowledge-based high-level recognition systems. The overview includes examples from each of the topics covered, as well as recent advances in human activity comprehension.

2. OBJECTIVES OF THE STUDY :

The objectives of our review study are:

(1) To study the prevailing methods and aspects used for human activity detection and recognition for simple and multiple/ complex human activities.

(2) To explore the future research directions for robust human activity detection and recognition methodology, especially for multi task human activity recognition in several domains.

3. METHODOLOGY :

The study utilized a systematic approach for conducting a historical literature review, which involved searching various resources from international and national journals, conferences, databases, and other sources of internet to collect and analyze relevant methods used for human activity recognition and to develop a robust HAR algorithm using a supervised learning framework, which can recognize many actions such as hand waving, running, jumping, bending, bowling, boxing, jogging, multiple human activity in crowded places, suspicious activity recognition in public etc. from the preloaded input video sequences.

4. LITERATURE SURVEY :

This section provides a summary of the HAR related articles obtained from various databases, including IEEE Xplore, SpringerLink, ACM, Google Scholar, ScienceDirect, and others. The articles were gathered using keywords such as Action Recognition, Action Detection, Activity Detection, Suspicious Activity Detection, Human Object Interaction, Multiple Actor Activity, Object Detection, Multi-Human Action Detection, Continuous Activity Recognition, Group Behaviour Analysis, Abnormal Behaviour Recognition, Violent Event Analysis, Event Detection, and Behaviour Detection. In this section, we will explore the methods and techniques that are most effective in recognizing human activities, as well as the challenges that still need to be addressed.

In the study by Dr. Mohana H. S. et.al. [2], the focus was on person activities, which were classified as person acts, interactions, and group actions. The ability to identify actions in input video is highly valuable in computer vision technologies and enables the development of models that can detect and recognize activities. The interactions between electronic devices and people are important in various surveillance environment systems, healthcare systems, military, patient monitoring systems (PMS), and other HAR applications. The researchers manually collected motion photos with motions or interactions, which were then divided into frames and preprocessed using a median filter to reduce noise in the input frames. Three STIP approaches, namely Harris SPIT, Gabor SPIT, and HOG SPIT, were used to extract features from the video frames. The SVM algorithm was used to classify the extracted features. The classifier's identification of the colored label was used to recognize actions. System performance was measured using the classifier's accuracy, sensitivity, and specificity. The classifier's accuracy demonstrated its reliability, while specificity and sensitivity explained how the classifier allocated its characteristics to each correct category and rejected aspects that did not belong to a specific correct category.

A new algorithm was developed by Mr. Mahanthesh U. et.al. for smart attendance tracking systems, which utilizes facial detection and identification methods to automatically take attendance in classrooms or laboratories. The algorithm is based on the SURF (Speed Up Robust Feature) algorithm [6]. The process involves capturing an image of a single student or a group of students in class and comparing it to previously saved images. If the captured image matches a previously saved image of a student, their attendance will be marked. However, there are still challenges in recognizing faces in group photos.

Hong-Bo Zhang and their collaborators [7] conducted an exhaustive review of the existing literature on human action recognition and proposed future research directions in this domain. Despite the substantial body of work dedicated to human action recognition, it remains a formidable challenge in real-world scenarios due to factors like intricate body poses, occlusions, and background interference. In their analysis, the authors assessed various methodologies for recognizing human actions, encompassing manually crafted action features in both RGB and depth data, action feature representation techniques based on deep learning, approaches for recognizing human-object interactions, and action detection methods. They presented the most notable and effective approaches within these research paths, providing a concise overview for researchers interested in these fields. Their study yielded several key insights into the realm of human action recognition research: (1) the careful selection of appropriate data for capturing actions and the use of robust algorithms for action recognition are foundational prerequisites for fruitful research in this area. (2) Deep learning-based solutions demonstrate superior performance when it comes to addressing challenges related to action feature learning. (3) Beyond the fundamental task of single-person action classification, the domains of interaction recognition and action detection are emerging as critical research frontiers in this field.

The primary objective of the study conducted by Nicki Efthymiou and their colleagues [8] was to enhance the Human-Robot Interaction (HRI) experience by incorporating computer vision techniques. Their research was dedicated to advancing the capabilities of an action recognition system in challenging HRI scenarios, with a specific focus on situations involving special user groups such as children. These scenarios often present limitations in training data and pose challenges to state-of-the-art system techniques. To address these challenges, the researchers devised a multi-view action recognition system. This system involved the integration and evaluation of various feature extraction methods, encoding techniques, and fusion approaches, all aimed at creating a system capable of recognizing pantomime motions performed by children. To assess the system's effectiveness, they integrated it into a robotic platform and conducted testing within an engaging Children-Robot Interaction scenario.

Nweke Henry Friday and their research team [9] undertook an investigation into the utilization of mobile or wearable sensors for the purpose of activity recognition. Initially, they provided a concise overview of contemporary deep learning methodologies employed in the context of human activity recognition. Subsequently, they introduced a conceptual deep learning framework designed to leverage

Gated Recurrent Units for the extraction of global features that capture temporal relationships. This system encompasses seven convolutional layers, two Gated Recurrent Units, and a Support Vector Machine (SVM) layer for the classification of activity details. Although this proposed approach is currently in the developmental phase, it will undergo evaluation using established benchmark datasets, allowing for comparisons with existing baseline deep learning algorithms.

Van-Minh Khong et al. [10] introduced a new method for detecting human activity using both RGB and optical flow information extracted from input video. The proposed approach uses a two-stream convolutional neural network, with each stream consisting of the architecture of a previously developed 3D convolutional neural network (C3D), which is known for its efficiency in action recognition. The two streams operate independently and the recognition results are generated through early or late fusion. The article provides evidence to show that 2stream C3D is better in accuracy compared to single-stream C3D on two benchmark datasets: UCF101 (from 82.37% to 88.79%) and HMDB51 (from 48.43% to 62.54%).

Yifeng He and their research team [11] explored the utilization of multisets generated by multiple sensors for the purpose of human action recognition. They introduced two innovative techniques for amalgamating data from diverse sets: BisetGlobality Locality Preserving Canonical Correlation Analysis (BGLPCCA) and MultisetGlobality Locality Preserving Canonical Correlation Analysis (MGLPCCA). These methods are capable of acquiring a lower-dimensional shared space that retains both local and global structural characteristics of the data samples. Furthermore, they put forward two distinct descriptors for depth and skeleton information and introduced a novel framework for human action recognition. This framework leverages either BGLPCCA or MGLPCCA to learn a common subspace from a range of data sets, including skeleton, depth, and optical flow data. The effectiveness of this proposed framework was extensively demonstrated through rigorous testing on five publicly accessible datasets, which encompassed the MSR Action3D, UTD multimodal human action dataset, multimodal action database, Kinect activity recognition dataset, and SBU Kinect interaction dataset.

Chen Chen and their research team [12] emphasized that prior surveys in the field of human action recognition have traditionally concentrated on either vision sensors or inertial sensors in isolation. However, recognizing the inherent limitations of each sensor modality, recent research has increasingly demonstrated that combining data from visual and inertial sensors can significantly enhance recognition accuracy. This comprehensive review article serves as a concise compilation of recent experiments that are dedicated to improving human action recognition through the integration of visual and inertial sensors. The authors place particular emphasis on depth cameras and inertial sensors due to their widespread availability and affordability, both of which contribute to their appeal. Furthermore, these sensor types are especially valuable as they provide 3D human action data, offering a richer dataset for analysis.

Rawya Al-Akam and Dietrich Paulus investigated a new method for detecting human activities in 3D movies using RGB and depth data. The method suggested in [13] involves using Bag-of-Information techniques to extract local-spatial temporal features from all frames of a video and to distinguish between human activities. To achieve this, K-means clustering and multi-class Support Vector Machines are utilized for classification, and the system is designed to be invariant to scale, rotation, and lighting. The novel approach of combining these features results in higher recognition rates compared to other existing techniques for recognizing human activities.

The main objective of Henry Friday Nweke and his colleagues [14] was to conduct a thorough evaluation of various data fusion and multiple classifier system techniques for human activity detection, especially for mobile and wearable sensors. They initially explained the techniques and modalities of data fusion, and then examined feature fusion, including deep learning fusion for human activity recognition, along with their benefits and drawbacks. The paper also explores several recently proposed design and fusion strategies for multiple classifier systems discussed in the literature. Finally, they identify and discuss some of the unresolved research issues that require further investigation and improvement.

V. D. Ambeth Kumar and colleagues [15] proposed a facial recognition model that aims to detect and notify the system when a specific person is identified in a designated area monitored by CCTV cameras. The system is composed of a centralized server that receives live streaming footage from multiple camera feeds and maintains a database of individuals to be found. The proposed method uses image processing techniques to match the real-time facial images with previously stored images of the individual in question. By focusing on the individual's most distinctive feature, i.e. their facial image, the system only requires the person's face image to be stored in the database for identification. As a result, the search for a person is simplified to detecting human faces in the video feeds and matching them with the stored images in the database. Once a match is found, the system tracks the person's location and sends an alert to the appropriate authorities.

Luca Turchet, Roberto Bresin, and their team conducted two studies [16] to investigate how the use of acoustically mimicked ground materials can affect the production and recognition of emotional walking. In the first experiment, the researchers found that the auditory feedback influenced emotional walking patterns in various ways, but the impact was not consistent. The second experiment showed that sound conditions did not have any effect on emotion recognition based on acoustic information alone. Both studies yielded similar results for participants with and without musical training. The researchers concluded that tempo and sound intensity are two crucial acoustic features in both the production and recognition of emotions in walking.

Some more related works reported in the literature are summarized in Table 1.

Table 1: Scholarly literature on Classifiers and Machine Learning algorithms used for Human Activity Detection & Recognition, Source [17-29].

S. No	Area and Focus of Research	Outcome of Research	Remarks	Reference
1.	Recognizing abnormal human behaviour with a Bayes classifier and a convolutional neural network	The Kalman filter is utilized in each frame to detect the moving human target. This method compares the Bayes Classifier and CNN algorithms for walking, running, punching, and tripping.	The results of the experiments reveal that the CNN approach outperforms the Bayes classifier. It is possible to raise the value of false positives.	Congcong Liu et al. (2018). [17]
2.	Human Emotion Recognition in Static Action Sequences Using Tree Based Classifiers	They used a method to extract motion data using a region of interest (ROI), and then used a tree-based classifier to identify human action.	Some emotions cannot be recognized with great precision.	R Santhosh Kumar et al. (2018). [18]
3.	Using the STIP Feature to Recognize Emotions from Human Activity.	The proposed emotion recognition method based on body movement is introduced. Harrier's method was used to detect the cornet. To classify the activity in terms of emotions, the STIP features are fed into the KNN and SVM classifiers.	Anger and joy in activities are interconnected and difficult to separate.	R Santhosh Kumar et al. (2018). [19]

4.	An overview of current machine learning trends for human activity identification.	This article looks into several data mining and machine learning techniques.	When compared to the data mining method, machine learning will produce better results. The combined ML method will provide better performance.	R Sreenivasan et al. (2019). [20]
5.	Emotion recognition from skeletal movements	The proposed methodology employs a variety of deep neural network algorithms on skeleton-based images.	The recognition of activity in terms of emotions can be improved.	T Sapinski et al., (2019). [21]
6.	A hybrid technique for recognising human activities using a support vector machine and a 1D convolutional neural network.	The Random forest method is used to detect static and moving activity. For activity classification, SVM and 1D CNN methods are used.	This method can be used to identify various activities in real time.	Md Maruf Hossain et al. (2020). [22]
7.	Anomaly detection in real time via CCTV using neural networks	In this study, real-time CCTV object detection is used. RNN and CNN are used for classification.	Real-time activity detection in this work is more difficult and can be made better.	Virender Singh et al. (2020). [23]
8.	A review of recent developments, datasets, challenges, and applications in video-based human action detection.	In particular, this study focused on content-based video outline, human-PC interaction, instruction, medical services, video observation, abnormal action discovery, sports, and entertainment along with various activity recognition techniques and HAR applications.	A multimodal understanding for activity recognition might be viewed as the future bearing. Later on, multi-person recognition could be included. The method for classifying videos with overlapped actions should be considered. Apply activity detection techniques to an online situation and also make action predictions.	Preksha Pareek et al. (2020). [24]
9.	Human activity recognition based on a CNN-SVM learning technique	The hybrid strategy to detect human activity is introduced in this method. CNN will extract the feature from the frames and assign it to SVM.	Although this method was tried on a short dataset, CNN has been shown to be beneficial when used to huge datasets.	Hend Basly et al. (2021). [25]
10.	Activity normalization in surveillance videos for activity detection	In order to distinguish between people and moving objects before categorization, the system used activity normalization.	The multi-person action recognition could be made better.	Takashi Hosono et al. (2021). [26]

11.	An investigation on vision-based human activity recognition	The many HAR techniques and difficulties are examined in this article.	It can be challenging to understand daily activities over a longer period of time. It is also difficult to classify the same action when it is performed in different ways. These are both difficult tasks that require further development.	Beddiar, D. R et al. (2020). [27]
12.	Survey of human activity recognition systems, problems, and applications using computer vision	This article compared the various vision-based HAR systems that are currently in use.	The precision and speed of the system's recognition have been found to be impacted by the use of cheap, low-quality cameras. In real time, HAR that is based on deep learning can be used to identify emotional behaviour such as happy sitting and angry running.	F Abdul Manaf, (2021). [28]
13.	Human activity recognition using convolutional neural network tools: A state-of-the-art overview, data sets, issues, and future prospects	The CNN-based deep learning technique was described, and the investigation was separated into four different input categories: multimodal sensing devices, telephones, radar, and vision devices. On several generic datasets, it was discussed.	With some enhancements, CNN can distinguish many human actions in a single frame and can make predictions about future behaviour via activity tracking.	Md. Milon Islam, (2022). [29]

5. SUMMARY OF THE LITERATURE SURVEY :

The current reseach on HAR has investigated various situations, including daily life, group activities, and real-time events. However, the bulk of the research has focused on simple activities related to daily life and user behavior, while there have been limited studies on complex & real-time activity recognition in domains such as healthcare, surveillance, and suspicious behavior. This scarcity of research is due to the challenges presented by real-time activities, hardware and technological limitations, and the lack of available data. In particular, real-time activities in crowded environments require robust processing capabilities to adapt to changing contexts, but there is a shortage of real-time data. It is worth noting that the activities described in the literature are not the only ones that occur in real-time, and researchers have recently used HAR to monitor individuals' behavior during pandemics and online exams.

The literature has compared different types of sensors, including Kinect, vision-based, and sensor-based, each with its own strengths and weaknesses. The ownership of smartphones has been a significant boost to HAR, as the sensors on these devices have greatly aided the field. The literature has also employed various machine learning and deep learning methodologies, with researchers suggesting innovative hybrid approaches to enhance performance. However, training numerous algorithms on a limited number of activities may lead to underfitting, while training on the entire dataset may result in overfitting. Additionally, the need for specialized hardware for training and testing is a challenge that could be addressed by transfer learning.

Although this review may not cover every article published on HAR, it provides valuable insights into current trends and obstacles in this area.

6. CURRENT STATUS AND ISSUES :

Based on the literature survey, there have been significant advances in HAR research, particularly in the development of computer vision techniques and deep learning-based solutions. However, several new challenges and issues have also emerged, including:

a. Recognizing human actions in complex and cluttered environments, such as those with occlusion, background clutter, and complex body postures.
b. Recognizing human-object interactions, which involves identifying the actions that humans perform on or with objects in their surroundings.
c. Recognizing human activity for surveillance environments such as to detect abnormal and suspicious activities in public places like airports, railway stations etc.
d. Detecting actions in group scenarios, which is a challenging task due to the large number of individuals and the variability in their actions.
e. Enhancing the Human Robot Interaction (HRI) experience by employing computer vision techniques that can improve the capabilities of an action recognition system in challenging HRI situations, particularly those involving special users like children.
f. Developing new algorithms and techniques for HAR using multisets from multiple sensors, such as depth, skeleton, and optical flow data.
g. Addressing privacy concerns related to HAR applications that involve the collection and processing of personal data, such as facial recognition for attendance tracking.
h. Ensuring the reliability and accuracy of HAR systems in real-world scenarios by improving the performance metrics of classifiers, such as accuracy, sensitivity, and specificity.

Overall, there is a growing need for more robust and reliable HAR systems that can recognize human actions in complex and cluttered environments and address the emerging challenges and issues in HAR research.

7. RESEARCH GAP :

- There is a limited literature reported on detection of real-time human activities in real time environments, such as detection of suspicious activities, monitoring of human behavior, human computer interaction etc.
- Existing models are limited to recognize only the simple activities like standing, sitting, walking, jumping etc. from specific stored video samples and they lack in recognition of many more human activities.
- There is a limited amount of research publications focused on recognizing complex activities in real-time, which includes both complex activities and real-time detection, as reported by several sources.
- There is a need to investigate the computational efficiency of STIP-based methods and explore ways to improve their speed without compromising accuracy for recognizing simple as well as more complex activities from a video.
- Although human activity recognition (HAR) systems have several advantages for various application domains, there are still significant limitations and unresolved issues in areas such as data collection, data preparation, complex activity recognition, activity misalignment, and hardware limitations.
- Various studies have indicated that collecting data for activity recognition and prediction poses several challenges, such as unannotated datasets, insufficient temporal information, recognition of unknown classes, and limitations on data availability that must be overcome.
- The preprocessing of data and the extraction of meaningful information are crucial for human activity recognition (HAR). According to several studies, challenges in data preprocessing include "appearance and feature extraction" as well as "background reduction."
- When a single action frame is divided into multiple frames, important information may be lost during frame segmentation, leading to incorrect action detection. Misaligned activities can also contribute to inaccurate action detection.

- The surveillance and healthcare industries have seen limited research on face and emotional information alongside human action recognition.

8. RESEARCH AGENDA :

Here's a research agenda to explore this area:

(1) Improved STIP Detection and Representation:
- Develop more robust algorithms for detecting STIPs in videos, especially in challenging conditions (e.g., low-light, occlusions, cluttered backgrounds).
- Investigate alternative ways to represent STIPs, including 3D structures, to capture spatial and temporal information more effectively.

(2) Feature Extraction and Encoding:
- Explore novel feature extraction techniques for STIPs to capture more discriminative information.
- Investigate different encoding methods (e.g., bag-of-words, deep learning-based) for generating representations from detected STIPs.

(3) Spatio-Temporal Analysis:
- Develop methods to analyze the spatio-temporal relationships between STIPs to better understand complex activities.
- Explore techniques for modeling the temporal evolution of STIPs within activities.

(4) Multi-modal Fusion:
- Investigate the integration of multiple data sources, such as audio and depth information, with STIPs to improve activity recognition accuracy.
- Study techniques for fusing information from different sensors/modalities effectively.

(5) Benchmark Datasets and Evaluation Metrics:
- Create and curate benchmark datasets specifically designed for STIP-based human activity recognition.
- Define standardized evaluation metrics that consider both accuracy and computational efficiency.

(6) Real-time Processing:
- Develop real-time or low-latency STIP-based activity recognition systems suitable for applications like surveillance and robotics.
- Optimize algorithms for efficient hardware deployment, including edge devices.

(7) Robustness and Generalization:
- Investigate techniques to improve the robustness of STIP-based models to variations in lighting, camera perspectives, and background clutter.
- Focus on domain adaptation methods to enhance model generalization across different environments.

(8) Privacy and Ethical Considerations:
- Address privacy concerns associated with video-based human activity recognition, particularly in public spaces.
- Explore methods for anonymizing or protecting sensitive information in video data.

(9) Human Interaction and Behavior Understanding:
- Extend research to not only recognize actions but also understand human behavior, intentions, and emotions from STIP-based representations.
- Explore applications in human-computer interaction and affective computing.

(10) Real-world Applications:

- Apply STIP-based activity recognition to various practical domains, such as healthcare (fall detection), sports analysis, and smart environments.
- Collaborate with industry partners to deploy and evaluate STIP-based systems in real-world scenarios.

(11) Explainability and Interpretability:
- Develop methods to make STIP-based models more interpretable, allowing users to understand why certain activity predictions are made.

(12) Long-term Activity Understanding:
- Investigate techniques for recognizing and understanding long-term activities or events that span extended periods, beyond short-term actions.

(13) Cross-modal Learning:
- Explore cross-modal learning techniques that leverage STIP data in combination with textual descriptions or natural language to enhance activity recognition.

(14) Adversarial Robustness:
- Study adversarial attacks and defenses in the context of STIP-based activity recognition systems.

(15) User-Centric Design:
- Engage with end-users to understand their needs and preferences in STIP-based activity recognition systems to design more user-friendly applications.

This research agenda will guide investigations in the field of human activity recognition using STIP, fostering innovation, and advancing the capabilities of computer vision systems in understanding human actions and behaviors. Researchers can choose specific directions from this agenda based on their expertise and the evolving needs of the field.

9. FINDINGS :

- Even there are many work related to human activity recognition are found in literature but Improving Robustness with respect to Multitask of an object using stationary camera with image segregation is still remains challenging.
- Activity recognition has connections to many fields of study such as human computer interaction, sociology, security systems, surveillance environments, entertainment environments and healthcare systems.
- Multi-tasking recognition and detection is potential to identify the complex object activities from videos enables the construction of several important applications. Where majority of cases in all the human activity recognition applications or models need skilled human involvement for smooth management of such application as there is a need of automated surveillance systems in public places like airports and railway stations are required to detect abnormal and suspicious activities.
- The ability to recognize human activities also has significant applications in real-time monitoring within medical contexts for patients, children, and the elderly. It opens the door to developing gesture-based human-computer interfaces and vision-based intelligent environments, made feasible through activity recognition systems.

10. SUGGESTION :

As there is a limited amount of research publications focused on recognizing complex activities in real-time, which includes both complex activities and real-time detection. So we suggest that there is a need of more robust system for recognizing of simple and complex human activity using more advanced classification methods for versatile real time dataset where we can concentrate on following parameters:
(1) Effective recognition of frames which hold more information's of present objects at instance from a given input video.
(2) Concentrating on effective preprocessing methods which suits for preserving more informatics data present in the frames.

(3) Finding an effective Classifier which help in recognizing the complex human activity for a given input video.
(4) Performance is a key parameter which helps us to measure efficiency of the system due to which following will be the key parameter used to measure performance such as Sensitivity, Specificity & Accuracy.

11. PROPOSED EFFICIENT MULTI-TASK ACTION RECOGNITION MODEL :

From the inference of our literature review it is evident that there is a need for a robust system for the recognition of simple and complex human activity. In this section we are proposing such a Multi-task action recognition (MAR) system as shown in Fig. 2 and we carry out our further research in developing such an efficient system.

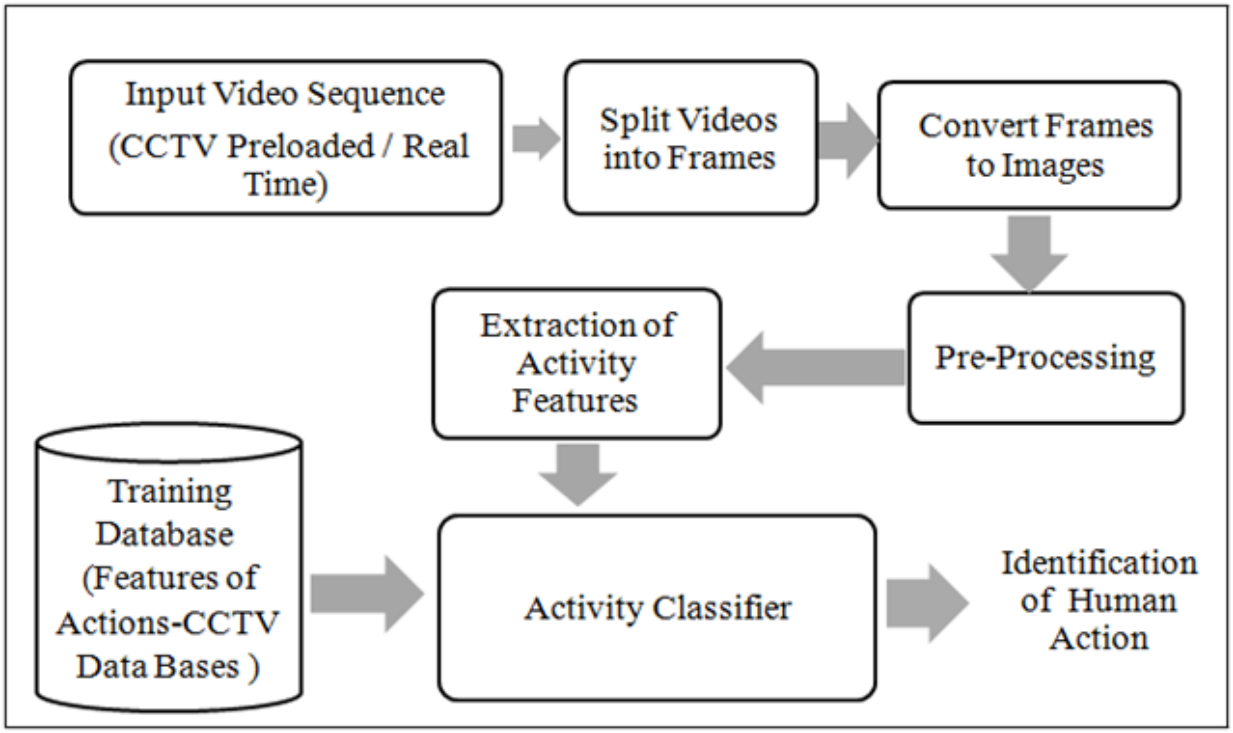

Fig. 2: Proposed Efficient Multi-Task Human Activity Recognition System
Source: Authors

Brief details of each block of the proposed system is given below.

Input Video Sequence:
A CCTV video or any real-time video will be used as the input video, since a video clip contains far more visual information than static photographs.

Split Videos into Frames:
To extract object features such as shape and motion characteristics, video should be analyzed frame by frame. As a result, the video sequence is transformed into a series of frames.

Convert Frames to Images:
Images are created from the frames retrieved from the video sequence. Specific images with the Region of interest are chosen from among these images for further processing.

Pre-Processing:
In general, while recording the input video, it may be contaminated with undesired noise. It could be due to changes in lighting, background clutter such as trees, and so on. As a result, the necessary details

may not be extracted satisfactorily. Hence, the input image is subjected to pre-processing to remove undesirable noise prior to the feature extraction stage.

Extraction of Activity Features:
The features related to the motion activities present in the images such as hand movement, leg movement and any other movements are extracted in this stage. Using the STIP algorithm, many different types of characteristics may be retrieved to characterize the specifics of human movement motions.

The STIP algorithm works by identifying and tracking interest points in a video sequence over time. These interest points correspond to locations in the video where there is a significant change in the intensity or color of pixels over time, which typically correspond to human joints or other key points of movement.

Activity Classifier:
Activity Classifier recognizes the different activities present in the given input sequence by using the extracted activities features. Also it classifies the recognized activities by labeling them as the set of known human activities. We are proposing to use Support Vector Machine (SVM) or Random Forest or a suitable Machine Learning algorithm or integrate all of them for recognition and classifying the human activity from the processed input video sequences.

Training Database:
In order to recognize the activities present in the given input video sequence from its features, first the proposed classifier system will be trained using the training database under supervised learning. The training database consists of a set of features that are extracted from the set of known possible pre-loaded human activities video sequence frames that represent the different possible human actions that the system is supposed to recognize. It encompasses a person's individual stance, atomic activity, person-person interaction, person-group interaction, group-group interaction, and so on. After training, the classifier is used to recognize and classify human actions in video sequences.

12. CONCLUSION :

There has been significant research in the area of HAR using computer vision technologies. This research has explored various approaches in identifying and recognizing human actions, including manual collection of motion photos, facial detection and identification methods, and the use of multiple sensors to recognize human actions. Deep learning-based solutions have been found to outperform other methods for action feature learning challenges. However, recognizing actions in real-life scenarios remains a difficult task due to several factors such as complex body postures, occlusion, and background clutter. Emerging research challenges include interaction recognition and action detection. Despite these challenges, researchers continue to work on developing HAR systems that can recognize and identify actions in a variety of situations, including those involving special users like children and also the literature survey reveals that there have been significant advancements in HAR research But there are also several challenges and gaps that need to be addressed.
The challenges include recognizing human actions in complex and cluttered environments, detecting actions in group scenarios, addressing privacy concerns related to HAR applications, and ensuring the reliability and accuracy of HAR systems in real-world scenarios. The research gaps include limited literature on detecting real-time human activities in real-time environments, limited recognition of many human activities (only simple human activities are recognised like walking, sitting, jumping etc), limited research on recognizing complex activities in real-time, and challenges in data collection and preprocessing. Addressing these challenges and research gaps is crucial in developing more robust and reliable HAR systems for various application domains. The study undertaken supplements the expanding literature on HAR highlighting its significance and need for further research. Based on the inference of the study we have also proposed an efficient model for the Multitask human activity recognition system and we propose to carry out our further research on its development.

REFERENCES :

[1] Rasheed, M. B., Javaid, N., Alghamdi, T. A., Mukhtar, S., Qasim, U., Khan, Z. A., & Raja, M. H. B. (2015). Evaluation of human activity recognition and fall detection using android phone. *In 2015 IEEE 29th International Conference on Advanced Information Networking and Applications*, *1* (1), 163-170. Google Scholar↗ CrossRef/DOI↗

[2] Mohana, H. S., & Mahanthesha, U. (2021). Human action recognition using STIP evaluation techniques. *Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE, 1* (1), 399-411. Springer Singapore. Google Scholar↗ CrossRef/DOI↗

[3] Mohana, H. S., & Mahanthesha, U. (2020). Human action Recognition using STIP Techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN*, 2278-3075, *9* (7), 878-883. Google Scholar↗ CrossRef/DOI↗

[4] Holte, M. B., Tran, C., Trivedi, M. M., & Moeslund, T. B. (2012). Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of selected topics in signal processing*, *6*(5), 538-552. Google Scholar↗ CrossRef/DOI↗

[5] Aggarwal, J. K., & Park, S. (2004, September). Human motion: Modeling and recognition of actions and interactions. *In Proceedings 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004, 1* (1), 640-647. Google Scholar↗ CrossRef/DOI↗

[6] Mohana, H. S., & Mahanthesha, U. (2018, July). Smart digital monitoring for attendance system. *In 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), 1* (1), 612-616. Google Scholar↗ CrossRef/DOI↗

[7] Zhang, H. B., Zhang, Y. X., Zhong, B., Lei, Q., Yang, L., Du, J. X., & Chen, D. S. (2019). A comprehensive survey of vision-based human action recognition methods. *Sensors, 19(5),* 1005. Google Scholar↗ CrossRef/DOI↗

[8] Efthymiou, N., Koutras, P., Filntisis, P. P., Potamianos, G., & Maragos, P. (2018). Multi-view fusion for action recognition in child-robot interaction. *In 2018 25th IEEE International Conference on Image Processing (ICIP), 1* (1), 455-459. Google Scholar↗ CrossRef/DOI↗

[9] Friday, N. H., Al-garadi, M. A., Mujtaba, G., Alo, U. R., &Waqas, A. (2018). Deep learning fusion conceptual frameworks for complex human activity recognition using mobile and wearable sensors. *In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 1* (1), 1-7. Google Scholar↗ CrossRef/DOI↗

[10] Khong, V. M., & Tran, T. H. (2018, April). Improving human action recognition with two-stream 3D convolutional neural network. *In 2018 1st international conference on multimedia analysis and pattern recognition (MAPR), 1* (1), 1-6. Google Scholar↗ CrossRef/DOI↗

[11] Elmadany, N. E. D., He, Y., & Guan, L. (2018). Information fusion for human action recognition via biset / multiset globality locality preserving canonical correlation analysis. *IEEE Transactions on Image Processing*, *27*(11), 5275-5287. Google Scholar↗ CrossRef/DOI↗

[12] Chen, C., Jafari, R., & Kehtarnavaz, N. (2017). A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools and Applications*, *76*(1), 4405-4425. Google Scholar↗ CrossRef/DOI↗

[13] Al-Akam, R., & Paulus, D. (2018). Local feature extraction from RGB and depth videos for human action recognition. *International Journal of Machine Learning and Computing*, *8*(3), 274-279. Google Scholar↗ CrossRef/DOI↗

[14] Nweke, H. F., Teh, Y. W., Mujtaba, G., & Al-Garadi, M. A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion*, *46*(1), 147-170. Google Scholar↗ CrossRef/DOI↗

[15] Kumar, V. A., Kumar, V. A., Malathi, S., Vengatesan, K., &Ramakrishnan, M. (2018). Facial recognition system for suspect identification using a surveillance camera. *Pattern Recognition and Image Analysis*, *28*(1), 410-420. Google Scholar↗ CrossRef/DOI↗

[16] Turchet, L., &Bresin, R. (2015). Effects of interactive sonification on emotionally expressive walking styles. *IEEE Transactions on affective computing*, *6*(2), 152-164. Google Scholar↗ CrossRef/DOI↗

[17] Liu, C., Ying, J., Han, F., &Ruan, M. (2018). Abnormal human activity recognition using bayes classifier and convolutional neural network. In *2018 IEEE 3rd international conference on signal and image processing (ICSIP)*, *1* (1), 33-37. Google Scholar↗ CrossRef/DOI↗

[18] Santhoshkumar, R., & Geetha, M. K. (2018). Human Emotion Recognition in Static Action Sequences based on Tree Based Classifiers, *International Journal of Scientific Research in Computer Science Applications and Management Studies, 7(3)*. Google Scholar↗ CrossRef/DOI↗

[19] Santhoshkumar, R., & Geetha, M. K. (2018). Recognition of Emotions from Human Activity Using STIP Feature, *International Journal of Engineering Science Invention (IJESI)*, *1* (1), 88-97 Google Scholar↗ CrossRef/DOI↗

[20] Ramasamy Ramamurthy, S., & Roy, N. (2018). Recent trends in machine learning for human activity recognition—A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), 125-132. Google Scholar↗ CrossRef/DOI↗

[21] Sapiński, T., Kamińska, D., Pelikant, A., &Anbarjafari, G. (2019). Emotion recognition from skeletal movements. *Entropy*, *21*(7), 646-652. Google Scholar↗ CrossRef/DOI↗

[22] Shuvo, M. M. H., Ahmed, N., Nouduri, K., & Palaniappan, K. (2020). A hybrid approach for human activity recognition with support vector machine and 1D convolutional neural network. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, *1* (1), 1-5. GoogleScholar↗ CrossRef/DOI↗

[23] Singh, V., Singh, S., & Gupta, P. (2020). Real-time anomaly recognition through CCTV using neural networks. *Procedia Computer Science*, *173*, 254-263. Google Scholar↗ CrossRef/DOI↗

[24] Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, *54*(1), 2259-2322. Google Scholar↗ CrossRef/DOI↗

[25] Basly, H., Ouarda, W., Sayadi, F. E., Ouni, B., &Alimi, A. M. (2020). CNN-SVM learning approach based human activity recognition. In *Image and Signal Processing: 9th International Conference, ICISP 2020, Marrakesh, Morocco, June 4–6, 2020, Proceedings* *9*(1), 271-281. Google Scholar↗ CrossRef/DOI↗

[26] Hosono, T., Sawada, K., Sun, Y., Hayase, K., &Shimamura, J. (2020, October). Activity normalization for activity detection in surveillance videos. *In 2020 IEEE International Conference on Image Processing (ICIP)*, *1* (1), 1386-1390. Google Scholar↗ CrossRef/DOI↗

[27] Beddiar, D. R., Nini, B., Sabokrou, M., &Hadid, A. (2020). Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, *79*(1), 30509-30555. GoogleScholar↗ CrossRef/DOI↗

[28] Manaf, A., & Singh, S. (2021, May). Computer vision-based survey on human activity recognition system, challenges and applications. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, *1* (1), 110-114. Google Scholar↗ CrossRef/DOI↗

[29] Islam, M. M., Nooruddin, S., Karray, F., & Muhammad, G. (2022). Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects. *Computers in Biology and Medicine, 149*(1), 106060. Google Scholar↗ CrossRef/DOI↗

****