Privacy-Preserving Indoor Monitoring using Vision Sensors

Ph.D. Thesis

By
Ankit Kumar Jain

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE

December 2024

Privacy-Preserving Indoor Monitoring using Vision Sensors

A Thesis

Submitted in partial fulfillment of the
requirements for the award of the degrees
of
Doctor of Philosophy

by

Ankit Kumar Jain

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
December 2024

INDIAN INSTITUTE OF TECHNOLOGY INDORE

I hereby certify that the work which is being presented in the thesis entitled **Privacy-Preserving Indoor Monitoring using Vision Sensors** in the partial fulfillment of the requirements for the award of the degree of **DOCTOR OF PHILOSOPHY** and submitted in the **Department of Computer Science & Engineering**, Indian Institute of Technology Indore, is an authentic record of my own work carried out during the time period from **July 2018** to **June 2024** under the supervision of **Dr. Abhishek Srivastava**.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

31-12-2024

Signature of the student with date

(Ankit Kumar Jain)

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

31/12/24

Signature of Thesis Supervisor with date

(Dr. Abhishek Srivastava)

17-12-2024

Ankit Kumar Jain has successfully given his/her Ph.D. Oral Examination held on *Date of PhD Oral Examination*.

31/12/24

Signature of Thesis Supervisor with date

(Dr. Abhishek Srivastava)

ACKNOWLEDGEMENTS

Completing this PhD thesis has been a long journey and would not have been possible without the support and encouragement of many individuals. I would like to take this opportunity to express my deepest gratitude to all those who have contributed to this achievement.

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Abhishek Srivastava, for continuous support, guidance, and encouragement throughout PhD study and research. Dr. Abhishek's quest for excellence and his love for perfection have always inspired me to give my best. His insightful feedback and extensive knowledge have been invaluable, helping me navigate the complexities of my work and ensuring the highest standards of academic excellence. His confidence in me has fostered my motivation, helping to define my research and presentation skills, and to grow as an independent researcher. His guidance helped me in all the time of research and writing of this thesis. Above all, his positive attitude, calm nature, and his ability to stay optimistic even in most challenging situations always inspired me. I could not have imagined having a better supervisor and mentor for my PhD study.

Besides my supervisor, I would like to extend my heartfelt thanks to my PG Student's Progress Committee (PSPC) members, Dr. Somnath Dey and Dr. Amod Umrikar, for their keen observation, constructive feedback, and validation of the research work. Their input has been crucial in refining my work and broadening my perspective. I am also grateful to Dr. Rajendra Akerkar from Western Norway Research Institute Norway, for his constructive remarks on my research. I deeply thank my M.Tech supervisor Late Dr. S.K. Jena and all my teachers who have guided, inspired, and supported me throughout my educational journey.

My sincere thanks also go to my fellow labmates, Arun Kumar, Gyan Prakash, Rupendra Singh, Shekhar Tyagi, Prarthi, Anil, Utkarsh, Shibani, Drishti, and others for their unwavering support, stimulating discussions, and all the fun we have had in the last few years. Also, a special thanks to BTP student Mihir for implementing an algorithm, and other fellow colleagues for helping me in creating datasets. I also thank to my friends and fellow doctoral researchers at IIT Indore, Mahesh Joshi, Vikash Chouhan, Narendra Vishwakarma, Vivek Singh, Priyanka Joshi, Anuj Rai, Rahul Chourasia, Pratibha Khandait, among many others

for their motivation, fun times, and the much-needed breaks that kept me sane throughout this journey.

I am grateful to all my friends from different part of my life for being there with me and cheer me on. A heartfelt thanks to all my dear ones, whose constant presence lifted my spirits and celebrated every achievement of mine, no matter how small.

I would like to extend my deepest gratitude to my mother, Smt. Sunita Jain, whose unconditional love, encouragement, and sacrifices have been instrumental in my journey. She has been a primary source of strength for me, especially during challenging times. A special mention goes to my children, Shivank and Shreeja. Though they may not fully comprehend the magnitude of my academic journey at this young age, but their laughter, curiosity, and boundless energy have infused my days with joy and perspective. Their presence, even amidst the stacks of books and late-night study sessions, reminds me of the beauty and simplicity of life beyond academia.

Last but not the least, to my wife Shivani, whose endless love, patience, and constant encouragement have been my anchor. She has been my rock during the challenging times, and her belief in me has kept me going. I thank her for being my partner in this journey and for standing by me through the highs and lows. I am forever grateful for her presence in my life.

This thesis is dedicated to all of you. Thank you for being part of this incredible journey.

Ankit Jain

Dedicated

to

My Family, Friends, and Teachers

ABSTRACT

Indoor monitoring is the process of continuous observation and analysis of indoor spaces using sensors, cameras, and similar sensing deployments. These systems typically monitor: environmental parameters like temperature, air quality; disaster scenarios like fire, earthquakes; and human behaviour such as activities of daily living, unnatural and abnormal activities, occupancy of spaces, indoor localization, and marked behavioural changes. Indoor monitoring has a wide range of applications in elderly-care, health-care, and in systems facilitating smart-homes.

The monitoring devices used in indoor monitoring systems largely comprise ambient sensors, wearable sensors (useful mostly in human monitoring) and vision sensors (typically cameras). Ambient and wearable sensors for indoor monitoring have limited effectiveness and are constrained largely in terms of performance and convenience respectively. These constraints can largely be overcome by the use of vision sensors largely comprising visible cameras. The issue with visible cameras, however, is compromise of privacy in such indoor spaces. The main objective of this thesis, therefore, is to devise privacy-preserving mechanisms for indoor monitoring utilizing vision sensors. These mechanisms harness the benefits of vision sensors in indoor monitoring whilst overcoming the limitations of privacy preservation. The thesis also conceives the use of appropriate learning algorithms for analyzing privacy-preserving visual data to enable automated monitoring.

Initially, a vision based fire detection framework for monitoring private spaces whilst preserving the privacy of occupants is proposed. This is a novel endeavour as no other approach has looked at the issue of privacy preservation in fire detection with vision sensors. The framework utilises a Near Infra-Red (NIR) camera to capture images in a manner that the privacy of occupants is preserved. To confirm that images captured with this camera do preserve occupants' privacy, two random user surveys were conducted. For effective fire detection using these images, a novel system incorporating both spatial and temporal properties of fire is employed. Experiments were conducted and confirm the superiority of the proposed framework when compared with existing techniques in literature both in terms of performance and model size. In addition to this, the lightweight nature of the proposed

system enables its effective use over resource-constrained environments. This is validated through a real-world prototypical implementation.

Continuing from fire detection, the thesis progresses to monitoring human activities in indoor space whilst preserving occupants' privacy. This is of significant utility in applications like health-care and elderly-care. A robust framework for automatic human activity recognition is proposed that uses depth sensors that preserve privacy and are cost-effective. Depth sensors provide two data modalities, namely depth maps and skeleton sequences, used together for activity recognition. Two novel descriptors, Joint Position Descriptor (JPD) based on the position of joints; and Bone Angle Descriptor (BAD) based on bone inclination, are generated from the skeleton sequence data. The descriptors convey both spatial and temporal information and are scale and view-point invariant. The other set of data obtained from depth sensors, depth maps, are used along with the descriptors to deal with the issue of noisy and missing skeleton sequences. The data modalities and descriptors are fused using a two-level fusion strategy for a multi-channel Convolutional Neural Network (CNN) framework. The proposed system is validated and shown to be superior to the existing state of the art through comparisons over four widely used public datasets. A computational complexity analysis of the system confirms its efficacy in real time. A prototypical implementation of the proposed system further validates its practicability.

Depth sensors and their use in indoor monitoring are effective and are appropriate for concealing the identity and preserving of an individual. In certain circumstances, however, this is not enough. Considering a scenario where the occupant of a certain private space (a room, for instance) is common knowledge; concealing the identity of the individual is not enough. The need here is preserving the activity privacy of the individual i.e. information on the activity(ies) that the monitored individual is indulging in. This is a bigger challenge and requires even coarser granularity as far as the monitored individual's depiction is concerned. We endeavour to strike a balance between capturing depth images that do not betray the activity privacy of the monitored individuals whilst being good enough for deep learning models to assess their well being. The refinement of existing deep learning models to be up to this task is discussed in this thesi. In addition to this, a survey over a crowd-sourcing platform to assess the degree of privacy that people are comfortable with is also included.

Finally experiments are discussed that were conducted to confirm the utility of the proposed approach and validate its efficacy.

LIST OF PUBLICATIONS

(A) From PhD thesis work:

A1. Journal Articles:

Published/Accepted:

1. **A. Jain** and A. Srivastava, *"Privacy-Preserving Efficient Fire Detection System for Indoor Surveillance"*. IEEE Transactions on Industrial Informatics, vol. 18, no. 5, pp. 3043-3054, May 2022. **DOI: https://doi.org/10.1109/TII.2021.3110576.**

2. **A. Jain**, R. Akerkar and A. Srivastava, *"Privacy-Preserving Human Activity Recognition System for Assisted Living Environments"*. IEEE Transactions on Artificial Intelligence, vol. 5, no. 05, pp. 2342-2357, May 2024. **DOI: https://doi.org/10.1109/TAI.2023.3323272.**

3. Mihir Karandikar, **Ankit Jain**, Abhishek Srivastavaa, *"A Lightweight Human Activity Recognition System for Resource Constrained Environments"*. Journal of Electronic Imaging 33(4), 043025, July 2024. **https://doi.org/10.1117/1.JEI.33.4.043025.**

Under Review/Revision:

1. **Ankit Jain** and Abhishek Srivastava, *"Identity and Activity Privacy-Preserving Indoor Monitoring System for Assisted Living"*. ACM Transactions on Computing for Healthcare. ***(under review)***

(B) Other publications during PhD:

B2. Conference Articles:

Published/Accepted:

1. **Jain, A.**, Srivastava, A. *"A Comprehensive Framework for Detecting Behavioural Anomalies in the Elderly"*. In: AI, Data, and Digitalization. SAIDD 2023. Communications in Computer and Information Science, vol 1810. Springer, Cham. **DOI: https://doi.org/10.1007/978-3-031-53770-7_9**

Contents

List of Figures

List of Tables

List of Abbreviations & Acronyms

ADL	Activities of Daily Living
AMT	Amazon Mechanical Turk
AS	Action Set
BAD	Bone Angle Descriptor
CNN	Convolutional Neural Network
CONV	Convolutional Layer
CRC	Collaborative Representation Classifier
CS	Cross Subject
DMM	Depth Motion Map
DT	Decision Tree
DTW	Dynamic Time Warping
ELM	Extreme Learning Machine
ES	Evaluation Setting
FCN	Fully Convolutional Network
FLFS	Feature Level Fusion Strategy
FLOP	Floating Point OPeration
GAP	Global Average Pooling
GCN	Graph Convolutional Network
HAR	Human Activity Recognition
HOG	Histogram of Gradients
HP	Histogram Projection
I3D	Inflated Three Dimensional
INC	Inception Module
IR	Infrared
JPD	Joint Position Descriptor
LBP	Local Binary Pattern
LDPE	Low Density Poly Ethylene
LR	Logistic Regression

LSTM	Long Short Term Memory
MP	Max Pooling Layer
MSE	Mean Square Error
NIR	Near Infrared
NB	Naïve Bayes
R-CNN	Region-based Convolutional Neural Network
RF	Random Forest
RNN	Recurrent Neural Network
SA-FDS	Spatially Aware Fire Detection System
SIFT	Scale Invariant Feature Transform
SLFS	Score Level Fusion Strategy
SSD	Single Shot Detector
ST-FDS	Spatio-Temporal Fire Detection System
SVM	Support Vector Machine
TIR	Thermal Infrared
TOF	Time of Flight
YOLO	You Look Only Once

Chapter 1

Introduction

Monitoring of indoor spaces is imperative given the requirements of safety, security, and ensuring the well being of occupants of such spaces. The conventional approach to indoor monitoring is manual with individuals assigned and paid to monitor such spaces round the clock in shifts. This approach, while effective to an extent, is not commonly practicable today given the shortage of labour, rising salaries, and issues of intrusion into private spaces. Technology has come to the rescue and is being widely harnessed today for monitoring indoor spaces. Advancements in sensor technologies, computational infrastructure, and affordability of such systems is making them effective and easily accessible.

Such automated indoor monitoring systems typically involve the deployment of sensors to gather data, and surveillance systems with intelligent algorithms to analyse the data. The integration of sensing devices and artificial intelligence has enhanced the capabilities of indoor monitoring systems, allowing for real-time insights and proactive decision-making.

These systems form part of a wide range of applications, including and not limited to home safety and security (i.e., monitoring fire, earthquake, air-quality) [1–3], elderly-care through anomaly detection (i.e, detecting a fall of an elderly occupant) [4–6], health monitoring [7–9], activity recognition [10–12]. A typical use-case of indoor monitoring

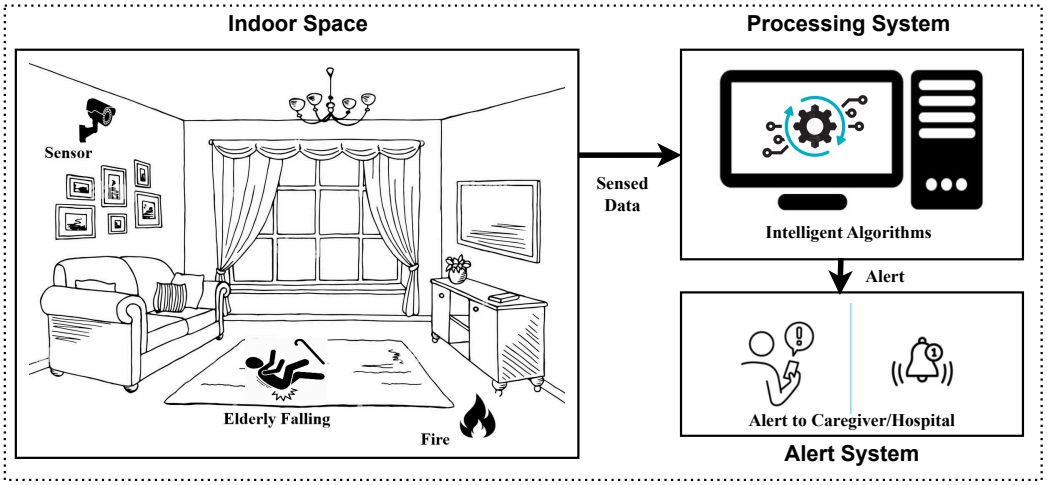


Figure 1.1: Illustration of Indoor Monitoring

system comprising sensor(s) and processing and alert system is shown in Figure 1.1.

An effective indoor monitoring system has two major components: a sensing component to gather data and an algorithmic component to analyse the sensed data. The sensing component comprises a variety of sensors, depending on the specific application and requirement. The sensors used in indoor monitoring are primarily categorized into ambient sensors, wearable sensors, and vision sensors. Figure 1.2 shows the categorization of the sensors used for indoor monitoring.

Ambient sensors are deployed all over the monitored space such as on walls, ceilings, floors. These mostly include, environmental sensors (to detect temperature, humidity, smoke, gas) [13], motion sensors (PIR, ultrasonic; to detect movements) [14], inertial sensors (i.e., accelerometer, gyroscope, pressure sensor; to detect relevant movement paramters) [15, 16], Wi-Fi, Bluetooth, microphones, magnetic reed switches, and so on. Wearable sensors, on the other hand, are sensors that are worn on the human body. These include, accelerometers, gyroscopes, pedometers, heart rate monitors, respiration rate sensors, GPS devices, and so on [6, 17, 18]. Wearable sensors are commonly integrated into smartwatches, fitness trackers, smart clothing, and other wearable devices. They provide

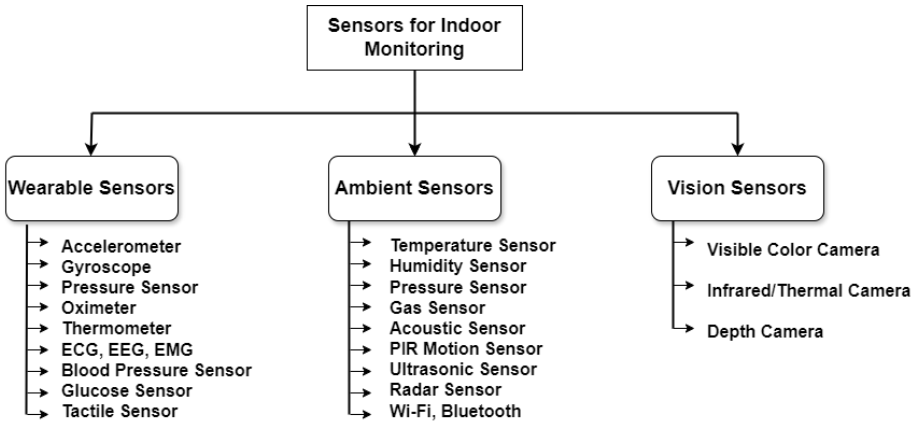


Figure 1.2: Categories of the sensors utilized in Indoor Monitoring

valuable data for monitoring the health and well-being of individuals.

Another category of sensors used for indoor monitoring are vision sensors and includes, RGB color cameras (mostly referred to as, visible sensor in this thesis), thermal infrared cameras, and depth cameras. Of these, color cameras [10, 19, 20] are extensively used in literature for monitoring as part of various applications and are proven to be superior in terms of accuracy and response time.

Thermal cameras work on the principle of temperature and capture images based on the amount of heat emitted by objects. Thermal infrared cameras are more commonly used in security applications such as monitoring military installation, home security, and others owing mainly to their 'night vision' capabilities [21]. Another important application of thermal cameras is in capturing remote sensing images using unmanned aerial vehicles (UAV) [22]. Depth cameras, on the other hand, work on the principle of distance and capture data based on the distance of the object from the camera. Depth cameras accurately model the third dimension (i.e., depth) of the object and are therefore used in applications like industrial automation, robotics, activity and gesture recognition, and several others [7, 12, 23, 24].

Each of these sensors have limitations, and the choice of the sensor becomes crucial for specific use-cases. The major limitation associated with ambient sensors are their sub-

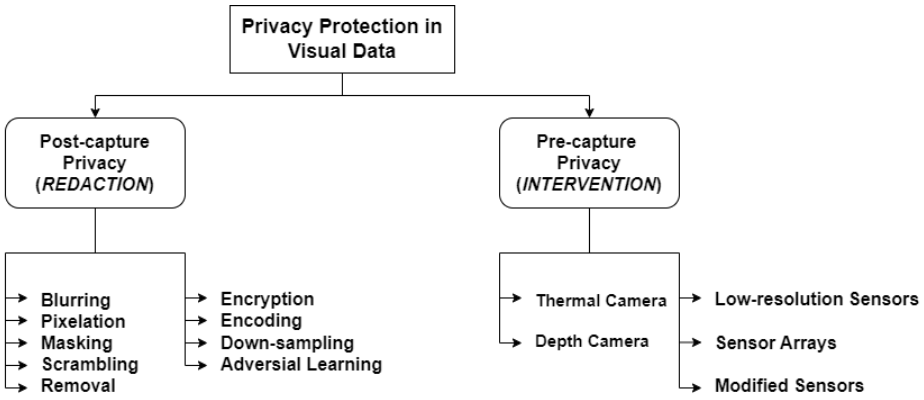


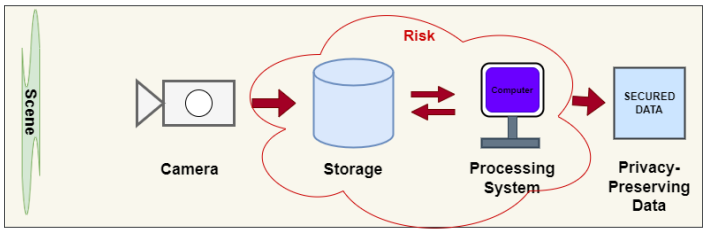
Figure 1.3: Approaches for privacy preservation in visual data

par precision owing largely to their dependence on environmental factors [25]. Wearable sensors on the other hand, are accurate but inconvenient as one needs to wear them all the time. It is also not uncommon for the monitored individual to have forgotten to wear the sensor [26]. Vision sensors, especially the visible color cameras, overcome the limitations of accuracy and inconvenience but are plagued by concerns of privacy in indoor locations [27, 28]. As color cameras capture rich visual representations of the monitored space, there become infeasible for indoor monitoring especially of private spaces such as bedrooms, living rooms, offices, toilets, and others. In addition to this, the other major limitation of color cameras is sensitivity to light; these cannot work effectively in low light conditions and hence are not suitable for monitoring in the night [29].

Privacy-preserving monitoring is the idea of collecting, sharing, and processing data for legitimate purpose whilst protecting the privacy (i.e., identity and activity) of the involved individual. Approaches for privacy-preservation in literature are divided into two categories: Post-capture privacy also known as *Redaction*; and Pre-capture privacy also known as *Intervention* [30]. Figure 1.3 shows the various ways of privacy preservation in visual data.

Post-capture privacy preservation, redaction, is the most common approach used in literature and captures the data first. Subsequently, algorithms process the data and remove

(a) Post-capture privacy



(b) Pre-capture privacy

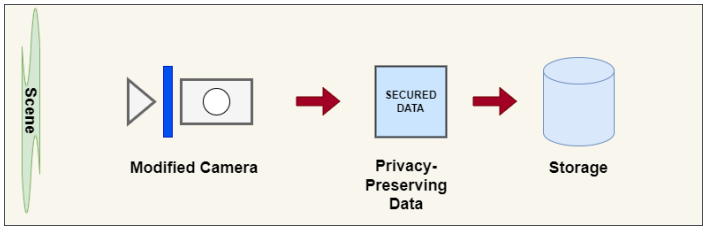


Figure 1.4: A high-level depiction of post-capture privacy and pre-capture privacy approaches

privacy sensitive information. Simple post-capture methods include, blurring, pixlation, masking, down-sampling, or removal of specific portions like human faces from the images [31, 32]. More complex post-capture methods include, image encryption, encoding, and adversarial learning to conceal privacy sensitive information [33, 34]. Post-capture privacy preservation approaches make the data vulnerable to intrusions during transit from the sensing device or during storage. Furthermore, most post-capture algorithms can be reverse engineered with modern technologies and infrastructure.

Pre-capture privacy preservation, intervention, on the other hand, is the approach wherein the sensor does not capture privacy-sensitive information. As sensitive information is never captured in this approach, it is completely privacy preserving. Common pre-capture privacy preserving approaches include: the use of sensors like thermal or depth cameras [35, 36], low resolution cameras, customised cameras [37, 38]. The images captured using depth or thermal do not preserve privacy completely as person can be identified from them. Moreover, the use of the images captured through low-resolution camera or down-sampled images is restricted to localization or motion sensing in the literature. Such sensors

are not suitable for complex applications like activity monitoring. A pictorial depiction of post-capture and pre-capture privacy preservation approaches is shown in Figure 1.4.

Privacy preservation is important and essential in assisted living and healthcare facilities because it upholds the dignity, autonomy, and well-being of residents, who may already face vulnerabilities. The breach of privacy in many cases may leads to emotional and physiological impact such as loss of security, anxiety, and loss of trust in the residents. The following case studies effectively convey the need of privacy preservation in assisted living and healthcare facilities.

[Case 1] Mrs. Elizabeth, an 82-year-old widow with early-stage dementia, moved into an assisted living facility to receive help with her daily needs. She was a retired teacher, valued her independence, and had always been private about her personal life. The facility implemented a new monitoring system with cameras in residents' rooms to enhance safety. One day, a video of Mrs. Elizabeth in her room, captured by the monitoring system, was accidentally accessed by an unauthorized individual. Shortly after, a caregiver inadvertently discussed Mrs. Elizabeth's routine loudly in a common area, where other residents and visitors overheard sensitive information. Mrs. Elizabeth felt humiliated and violated with this as her private moments were exposed. She became withdrawn, refusing care, and her trust in the staff and system eroded.

[Case 2] Mr. David, a 78-year-old retired engineer, lived alone in his family home. Due to his family's concerns about his safety, they installed a smart home surveillance system, including indoor and outdoor cameras, to monitor his well-being remotely. The cameras were intended to provide peace of mind and help the family respond in emergencies. Although, David valued his privacy and cherished his ability to live independently, he agreed to the installation, believing the cameras would be used only during emergencies. One day, David began noticing unusual interactions in his daily life. A neighbor mentioned details

about his activities, such as what he watched on television or what he ate in the dinner last night. David was confused because he had not shared this information. Upon investigation, he discovered that his video data had been accessed without his permission. The neighbor, a tech-savvy individual, had exploited a vulnerability in the devices to access David's activities.

Subsequent to capture of data, specialised algorithms are harnessed to make sense of the data and thus enable effective monitoring of the indoor space. The algorithms used for this purpose are divided into two categories: hand crafted feature based approaches; and deep learning based approaches. Recent literature mostly utilizes deep learning based approaches due to their automated feature extraction capabilities and superior accuracy [39, 40]. Algorithmic approaches for monitoring are further also categorized based on spatial and temporal representations of the images/videos. Spatial representations include extraction of class specific features from still images or frames of videos; whereas temporal representations consider the correlation of features in the temporal direction [41]. Most applications (i.e, activity recognition, fall detection) involving videos perform well when both spatial and temporal features are considered.

1.1 Motivation

With rapid increase in the number of elderly-care, health-care, and smart homes' facilities worldwide, demand for effective indoor monitoring systems are on the rise. These systems should be capable of monitoring indoor spaces continuously and sending an alert to concerned persons/organizations in case of a deviation from the normal sequence of events.

As discussed earlier, research in recent times prescribes the use of various types of sensors (i.e., ambient, wearable, and vision) for such indoor monitoring. Vision sensors, especially, have been extensively explored for data collection followed by the use of machine

learning and deep learning algorithms to make sense of the data. These have demonstrably proven to be effective and accurate. In spite of the superior performance of such vision sensor based systems, the major concern with them remains the compromise of occupants' privacy.

The work in this thesis is, therefore, towards devising mechanisms for indoor monitoring using vision sensors in a manner that the occupants' privacy is preserved whilst accurately and effectively monitoring the spaces. Our work mainly focuses on pre-capture privacy by preventing sensors from capturing sensitive information and using appropriate learning algorithms to analyze privacy-preserving data in an effective manner. The pre-capture privacy techniques used in our work are different from the existing ones in a way that we modified the existing sensors rather than just down-sampling the images.

1.2 Thesis Contributions

This section presents significant contributions of this research in effectively utilising the vision sensors for indoor monitoring in a privacy-preserving manner. In addition to this, the research focuses on the efficient use of learning algorithms for effective monitoring of indoor spaces.

1.2.1 Privacy-preserving efficient fire detection system

Indoor fires are becoming nowadays and often unfortunately lead to large causalities, property damage, and financial losses. The state of the art approaches for fire detection have major shortcomings related to privacy, cost, and performance. Keeping these factors in mind, we developed a lightweight, fast, and efficient fire detection system using Convolutional Neural Network and the temporal properties of fire. A strategically modified Near Infrared camera was utilized for privacy preserving monitoring. Since privacy is subjective

in nature, an acceptable level of privacy was identified through user surveys. In the absence of privacy preserving data, a large dataset of privacy preserving images of both fire and non-fire was created and a comparative analysis of the proposed system with existing state of the art methods was done. Finally, a prototypical deployment of the proposed system on a resource constraint device is done to show it‘s applicability in the real-world.

1.2.2 Privacy-preserving human activity recognition system

Monitoring daily activities is an important aspect in elderly-care environments that facilitates independent living for the elderly. This includes the detection of abnormal activities like fall, and can help in other behavioural monitoring. Considering the privacy, convenience, cost, and robustness, we developed an efficient and privacy preserving system for recognising human activities from depth data and skeleton sequences. A robust activity recognition system using a multi-stream convolutional neural network is proposed along with two level fusion strategy. Two novel descriptors are derived to enable scale and view invariant recognition from skeleton sequence data. A thorough evaluation and a comparative analysis on four public data sets in done to show the efficacy of the proposed system. Finally, a prototypical implementation and computational complexity analysis is done to show the suitability of the proposed system in the real-world.

1.2.3 Identity and activity privacy preserving posture recognition system

Most of the existing research in privacy-preserving indoor monitoring considers the identity of monitored individual as the only factor to be preserved. But, in the scenarios where only, only one or two individuals live in a house, the identity is not that crucial but the fine-grained activities need to be preserved form the intruders. In this work, we in-

troduced the concept of activity privacy in indoor monitoring and develop a system using Convolutional Neural Network and structural characteristics of human body to recognise body postures. The proposed system, utilizes strategically modified depth camera to capture the data that preserve both identity and fine-grained activities of the individuals while the course-grained activities can be classified efficiently. Privacy-accuracy trade-off of the system is calculated from both human and machine perception using user surveys and deep learning models, respectively.

1.3 Thesis Organization

The thesis is organized into six chapters including this introductory chapter. A flow diagram of the chapters is shown in Figure 1.5. Continuing from this chapter, the remainder of the thesis is structured as follows:

Chapter 2: Literature Review 2

This chapter provides a detailed overview of indoor monitoring and its applications in assisted living, health care, and smart homes. The review is done with an emphasis on privacy concerns and their solutions in vision based monitoring. The chapter also provides a summary of classification algorithms (both Machine Learning and Deep Learning) employed in indoor monitoring applications.

Chapter 3: Privacy-Preserving Fire Detection System 3

This chapter introduces a privacy-preserving system for detecting indoor fire whilst preserving occupants' privacy. This include creating a privacy-preserving fire dataset using modified near-infrared camera and developing a lightweight classification system using spatial and temporal properties of the fire.

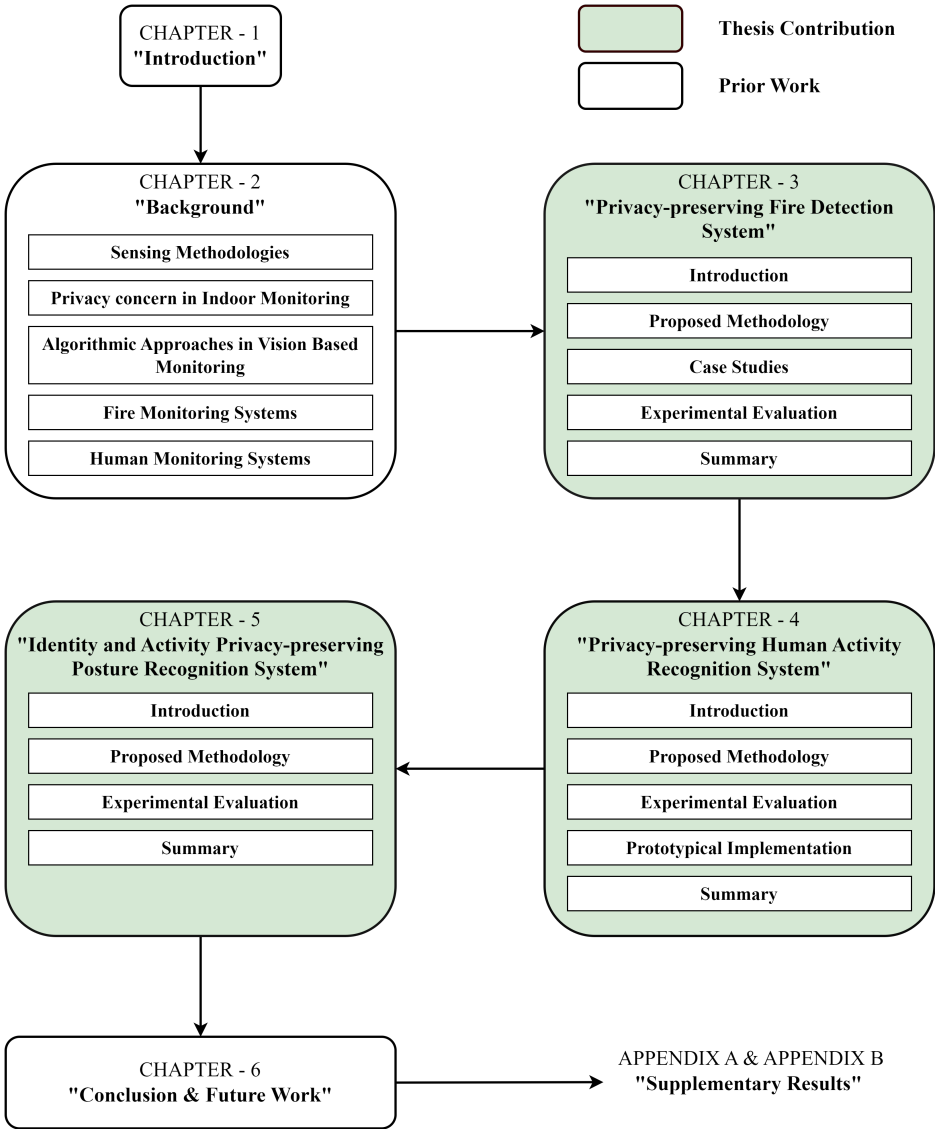


Figure 1.5: Flow diagram of the thesis

Chapter 4: Privacy-Preserving Human Activity Recognition System 4

This chapter introduces a robust human activity recognition system to recognise daily activities and detect a deviation from such activities (such a fall, for example) of the elderly in a manner that preserves their privacy. This includes the development of a multi-stream convolutional neural network to classify human activities based on the movements of different body part from depth data and skeleton sequences.

Chapter 5: Identity and Activity Privacy-Preserving Posture Recognition System 5

This chapter introduces the new concept of activity privacy in elderly care applications and develops an indoor monitoring system to recognise postures and coarse-grained activities (like, walk, sleep, fall). Fine-grained activities like (eating, drinking, talking on phone, and the like) are not discerned and thus the activity privacy of the individual is preserved.

Chapter 6: Conclusion & Future Work 6

This chapter comprises discussions and contributions of the thesis. It also includes possible future directions of work in privacy-preserving indoor monitoring systems, especially those meant for elderly-care and health-care applications.

Chapter 2

Literature Review

Indoor monitoring comprises tracking and observing conditions, activities, and events in indoor locations. This involves the monitoring of smart homes, elderly care, health care facilities, commercial buildings, to name just a few. Indoor monitoring has a wide range of applications including monitoring for security, environmental parameters, disasters, healthcare, daily activities, and behavioral analysis.

The literature related to indoor monitoring is reviewed along three dimensions, namely: sensing methodologies; privacy concerns; and algorithmic approaches for monitoring. The review of sensing methodologies explores studies utilizing various types of sensors, such as ambient, wearable, and vision-based sensors, in diverse indoor monitoring applications. Following this, research addressing privacy concerns and privacy-preserving techniques in indoor monitoring is analyzed, highlighting the importance of privacy and the strategies commonly employed to ensure it. Next, the commonly used algorithmic approaches including machine learning and deep learning algorithms are explored in the context of indoor monitoring. Finally, the literature on specific applications is reviewed, which include fire detection systems; human activity recognition systems; and posture recognition systems for indoor spaces in a privacy-preserving manner.

2.1 Sensing Methodologies for Indoor Monitoring

Researchers have explored the working of sensors along three main categories: wearable sensors; ambient sensors; and vision sensors.

Wearable sensors are typically lightweight sensors worn by individuals on their bodies to collect data on their activities and/or the environment. Wearable sensors may be deployed on various locations of the individuals body such as fingers, wrists, arms, thighs, neck, head, shoes, and clothing. Sensors of these kinds include accelerometers, gyroscopes, oximeters, thermometers, pressure sensors, electrocardiograms (ECG or EKG) sensors, electrodermal activity (EDA) sensors, among others. Authors in [18] review the use of wearable sensors like accelerometers and gyroscopes to detect human motion and as part of activity recognition systems. The review also include the use of EEG and EMG sensors for sleep monitoring and gesture recognition, respectively. Another study in [42] reviews the use of wearable sensors in health monitoring. This includes the use of pressure sensors for pulse and blood pressure monitoring, ECG sensors for heart disease monitoring, EEG sensors for brain diseases, GCM sensors for glucose monitoring, optical sensors for blood pressure monitoring, and so on. Recent studies in [43, 44] present the use of electro-mechanical sensors (like pressure sensor, strain sensor, tactile sensors) in human activity monitoring, health monitoring, tactile perception. [45] discusses the use of EEG and ECG sensors for mental health monitoring. Another study in [46] presents the use of wearable sensors in security applications like person identification. Studies in [6, 47] discuss the use of wearable sensors like accelerometer and gyroscopes in human safety applications like fall detection.

Ambient sensors are sensors deployed within the monitored space on walls, roof, gates, furniture, and so on. These include environmental sensors like temperature sensors, humidity sensors, gas sensors, sound sensors, motion sensors, among others. Studies in [14, 48]

discuss indoor localization approaches using various types of sensors such as wi-fi, bluetooth, RFID, LED light sensors, ultrasonic sensors, acoustic sensors, radar sensors, IR arrays. Similarly [44] discusses the use of sensors for human activity recognition in indoor spaces. Another interesting study presents the use of various gas sensors [13] for air quality monitoring and various sensors [49] for environmental monitoring in indoor spaces. [50, 51] presents the use of gas sensors in fire detection. Approaches for health monitoring using ambient sensors are explored in [52].

Vision sensors play an important role in indoor monitoring. These sensors capture visual data such as videos/images from the monitored space and provide information about the scene. Commonly used vision sensor are visible (color) camera, infrared thermal camera, and depth camera. A study in [53] reviews surveillance approaches based on vision sensors for varied applications such as person identification, activity recognition, abnormal/suspicious activity detection, safety applications like fall detection, security applications, and others. [54, 55] explores approaches for detecting abnormal events like violence, robbery, falls, both indoor and outdoor. Approaches for fire detection using vision sensors are extensively explored in [1, 56–58]. The authors in [59–63] review the approaches for human activity recognition using various kinds of vision sensors.

2.2 Privacy Concern in Indoor Monitoring

Privacy concerns in indoor monitoring revolve around the potential invasion of personal space, the collection of sensitive information, and the risk of unauthorized access to recorded data. Indoor monitoring systems, such as CCTV cameras or smart sensors, can intrude into an individuals' private spaces, such as homes, workplaces, or commercial establishments. It may inadvertently capture sensitive information about individuals, such as their identity information, activities, and/or personal interactions. This invasion of privacy

can lead to discomfort and may violate individuals' rights to solitude and confidentiality. Without proper safeguards, this information could potentially be used to track individuals and even cause harm to them.

Studies in [64, 65] consider privacy invasion as one of the prominent barriers in large-scale adoption of indoor monitoring systems in smart homes or indoor spaces. The study in [66] conducted a survey of the elderly and questioned them on their willingness to permit installion of surveillance systems in their homes. The survey concluded that along with financial cost, invasion of privacy was marked as a major concern with installing surveillance systems. Authors in [67] conducted surveys to understand the need of privacy in indoor monitoring. This survey discovered multiple reasons to justify the preservation of privacy of the elderly. [68] explores the legal compliance factors related to an individual's privacy. The findings of this work emphasize the importance of ensuring privacy of sensitive information in accordance with the relevant laws and regulations.

As compared to other sensing methodologies, vision sensor based (especially visible sensor) monitoring approaches raise serious privacy concerns due to their intrusive nature and the potential for collecting sensitive information about individuals and surrounding. The sensitive information may include facial features, body features, bio-metric data, personal activities, surrounding objects, and so on. Several researchers highlighted this issue of privacy invasion [26, 31, 37, 69, 70] in visible sensor based approaches.

Considering the importance of privacy in indoor monitoring, privacy preservation and studies related to it have gained prominence. Approaches for privacy preservation in literature fall into two categories, namely: post-capture privacy also known as redaction; and pre-capture privacy also known as intervention [30].

Post-capture privacy, as the name suggests, includes methods to hide sensitive information (such as images of person's face, body parts, and other identifiable information) after

capturing the data. The studies in [31, 32, 71, 72] review various approaches for post-capture privacy in the data captured with vision sensors (i.e. images or videos). A few common approaches include blurring [73], pixelation, masking, scrambling, removal of the human face or body parts from the captured images [**? ?**]. Encryption and down-sampling [74–76] of the images are other commonly used approaches in the literature to preserve privacy. Recent literature [33, 34, 77] explores the use of adversarial learning to conceal privacy sensitive information in images.

Pre-capture privacy, on the other hand, includes mechanisms to ensure that the sensing device captures data in such a way that sensitive information is excluded. Studies in this direction use various categories of vision sensors to preserve the privacy of individuals in monitored indoor spaces. The approaches in [35, 36, 78, 79] make use of depth or thermal sensors for monitoring indoor spaces in a privacy-preserving manner. This becomes possible owing to the relatively less intrusive nature of such vision sensors thanks mainly to the lack of color and texture information. Certain other approaches use low-resolution sensors [37, 80–82] or modified sensors [38] to capture privacy-preserving images/videos.

Enhanced privacy often comes at the cost of reduced accuracy, work in [83] emphasizes this fact and explores the effects of different privacy-preservation methods on the performance of the system.

2.3 Algorithmic Approaches for Vision Based Monitoring

Vision-based monitoring relies on a variety of algorithmic approaches to analyze and interpret visual data. These approaches can be categorized into traditional hand-crafted feature based techniques, and modern deep learning based methods [61]. Furthermore, the approaches are also categorized based on spatial representation in the case of static images and/or temporal representation in the case of videos, or a combination of both.

Hand-crafted feature based techniques comprise three modules: region of interest (ROI) identification; feature extraction; and classification [39, 40]. Background subtraction, temporal differencing, color-based segmentation, contour-based segmentation, and optical flow analysis are the commonly used techniques for ROI identification and tracking in the temporal domain [84–86]. Methods involved in feature extraction include Scale-Invariant Feature Transform (SIFT) [86], Histogram of Gradient (HOG) , Histogram of Flow (HOF) [63], and Local Binary Pattern (LBP) [23] from individual frames [84, 87] or the images representing the video clips such as Motion History Image (MHI), Motion Energy Image (MEI), Depth Motion Map (DMM) [88], and others. Feature representation methods such as dictionary learning or Bayesian networks are popularly used to combine the features extracted from individual frames of the videos [61]. Finally, the classification algorithms include Support Vector Machine (SVM) [89], Logistic regression [90], Extreme Learning Machine [91], K-means clustering [92], neural networks [93], and several others [61, 84, 87]. Commonly used methods for temporal modeling include Hidden Markov Model (HMM), wavelet transform, optical flow analysis in video sequences [1, 39, 87].

Hand-crafted feature based approaches used in vision based system are somewhat complicated and require significant human expertise and efforts. Hence, the selection of algorithms used in each module, especially feature extraction, becomes critical and is problem dependant [40, 41]. Deep learning based approaches reduce human intervention by taking the visual data directly as input and appropriately processing it.

Deep learning approaches used in vision-based monitoring systems commonly include Convolutional Neural Networks (CNN) [94–96], Recurrent Neural Networks (RNN), and Graph Neural Networks (GNN) [39, 61]. 2D-CNNs are most popularly used either as feature extractors or classifiers due to their automatic feature extraction capabilities [40, 41, 87]. Other works employ 3D-CNNs for classification in video clips considering the

temporal correlation of different frames of the videos [39, 40]. GNN based unsupervised learning approaches are explored for their use in pose estimation and activity classification in videos and/or skeleton data [29, 97]. RNN based models, such as Long Short Term Memory (LSTM) are capable of learning dependencies in sequential data and are extensively used for feature integration in the temporal domain [7, 41, 61, 98, 99].

2.4 Indoor Fire Monitoring System

Indoor fires, as discussed in the previous chapter, unfortunately lead to some of the worst disasters affecting human livelihood and economies. This leads to the need of automated fire detection systems and many researchers explored this. Literature in this area is reviewed along the following two dimensions: literature on sensing devices; and literature on algorithmic techniques to assess the sensed parameters.

The more common sensor-based fire detection systems' performance is adversely affected by limitations of the sensors [56] and these often result in false alarms. Vision-based fire detection systems, on the other hand, perform better in terms of accuracy and are widely researched. Visible color image cameras [85][1] are increasingly being used for fire detection in both indoor and outdoor spaces. Infrared based cameras have proven most suitable for fire detection as a large amount of infrared energy emanates from fire. IR cameras are extensively used in satellites and/or Unmanned Aerial Vehicles (UAVs) [22][58] for forest/outdoor fire detection. It is now a proven fact that infrared cameras are more accurate in fire detection. In spite of this, IR cameras are seldom used for fire detection in indoor spaces owing to the high cost of such cameras.

The performance of fire detection systems are commonly augmented with focused algorithms that make better sense of the collected data. Classification techniques used in traditional vision-based fire detection techniques mostly comprise three major steps: 1) fire re-

gion identification; 2) feature extraction; and 3) classification. There are several approaches to accomplish the first step: segmentation based on color or brightness [1][22][58] in still images, and background subtraction and/or frame differencing [85][100],[86] in videos are the popular approaches for fire region identification. Various kinds of features such as color features [85][1], motion features using optical flow analysis [93][22], flame centroid motion [1][86], and shape based features such as variations in flame shape, boundary roughness [85][1][58], key points features using SIFT descriptor [86], and the like are extracted. The last step comprises classifying the identified regions as fire or non-fire based on the features extracted in the previous step. Frameworks in [85][1][58] utilize the decision rules based approaches. Machine learning techniques such as Neural Network [93], Logistic Regression [100], Support Vector Machine [86] [92], K-mean clustering [92], and others are also employed in various endeavours.

Convolutional Neural Network (CNN) has emerged as a popular technique for object detection and classification in images mainly due to its automatic feature extraction capability that reduces human intervention to a minimum. The superiority of CNN in image classification led to the development of several fire detection models around it over the past few years. Approaches have been proposed by modifying existing CNN models or by developing new ones from scratch. [101] proposes a 9-layer CNN architecture from scratch for fire and smoke detection in videos. The model is trained on a small dataset and achieves superior performance. A CNN model based on the Xception network is proposed in [94] wherein the network is trained using an infrared image dataset of forest fires. The work also proposes a fire detection segmentation approach employing the UNet architecture. [102] proposes a fire detection framework based on deep learning models such as Faster-CNN, R-FCN, SSD and YOLO. These models are trained via transfer learning using relatively small datasets of around $15,000$ images and achieve good performance. An endeavour that uses

Faster R-CNN and SVM is proposed in [95] in which R-CNN is used to detect regions of fire within images and subsequently the VLAD representation vectors of the detected regions are classified using SVM. The model is trained on a very small dataset of 550 images and achieves comparable performance. Khan Muhammad *et al.* propose fire detection models [20] [103][104] based on MobileNet, SqueezeNet, and GoogleNet respectively. All these models employ pre-trained weights on the well-known Imagenet dataset, followed by fine-tuning of the fire dataset. The models are trained with a large number of labeled images and achieve better accuracy than other state of the art methods. Efforts towards reducing the size of the network are made in [105] and [106] to make them work in resource-constrained environments. In [107] and [98], fire detection frameworks are proposed using combinations of CNN and LSTM to utilize the temporal property of fire. A hybrid approach combining convolutional neural network with genetic algorithms is proposed in [108] to cater to a more generic scenario. The technique can readily be customised for fire detection.

2.5 Indoor Human Monitoring System

Monitoring of individual's activities especially in elderly care or health care is another prominent area where researchers explored automated human activity recognition or posture recognition systems using verity of sensors. As discussed in previous section, the vision sensor (especially visible camera) based systems raise a serious privacy concern, due to which many researchers explored ToF imaging depth sensor based systems due to their insensitivity to light and privacy unobtrusiveness.

2.5.1 Human Activity Recognition Using Depth Sensors

Research on Human Activity Recognition (HAR) using depth vision sensors in literature can be divided into three parts: HAR using depth data; HAR using skeleton sequence

data; and HAR using a combination of the two. Most recent techniques on human activity recognition that use depth data are based on Depth Motion Maps (DMM) [62, 88, 109] from multiple (mainly front, side, and top) views. Other approaches [110, 111] that use depth data are based on feature extraction from depth frames in sequence, followed by their concatenation in the time domain.

Approaches based on skeleton joint information primarily use the position of joints [112], angles between joints [91, 113], and/or a combination of the two [96, 114] for both spatial and temporal features. [112] uses the distance of joints from the floor. [91] utilizes the angles between selected joint triplets within frames and between pairs of joints in consecutive frames for spatial and temporal features. The distance between each pair of joints is combined with the joints' angles with the principle axes in [114], and the angle between joints triplets in [96]. Other approaches in [24, 115] involve arranging skeleton sequences in a 3D matrix and using these for feature extraction. Another approach proposes a solution based on graph theory by representing the skeleton joint information on graphs. A graph is created by concatenating the skeletons from all the frames in a skeleton sequence in [29].

Certain approaches use a combination of skeleton and depth data for human activity recognition. [12, 62] combine DMMs derived from depth data with angles obtained from spherical coordinates systems. Similarly, other approaches combine DMMs with the angle between joint segments in consecutive frames (joint segments are obtained by summing up joints in each body part) [63]. In [116], the surface normal vector for each body part is calculated in the depth maps and concatenated in the temporal domain. The skeleton joints are used to segment the body parts in the depth maps.

Algorithmic approaches used for human activity recognition can be categorized in two ways: hand-crafted feature based, and deep learning based. The techniques used in hand-crafted feature based approaches mostly comprise two major steps: 1) feature extraction and

representation; 2) classification. In systems based on depth maps, the approaches for feature extraction include Histogram of Gradients (HOG) [62, 63, 109], Local Binary Pattern (LBP) [23], Space-Time Auto Correlation of Gredients (STACOG) [88], and wavelet decomposition [117]. The features in skeleton based techniques comprise the distance between joints [96, 118], the angles between joints [12, 62, 63, 91, 96], and the original joint positions [114]. Classification algorithms like Support Vector Machine (SVM)[63, 89, 115, 116, 119], Extreme Learning Machine (ELM) [23, 62, 91], Logistic Regression (LR) [90, 111], and Collaborative Representation Classifier (CRC) [88] are commonly used for human activity recognition.

Convolutional Neural Networks (CNN), a class of deep learning, have emerged as an effective object detection and classification technique that are demonstrably superior in image classification. [12, 96, 114] utilize 2D CNN models for activity classification from the skeleton or depth based images. Another deep neural network class is Recurrent Neural Networks (RNN), which model information in the time domain. These networks, especially Long Short Term Memory (LSTM), are suitable for video activity recognition systems and work by processing information frame by frame. An approach in [7] employs LSTM based architectures for activity classification. Other approaches in [24, 99, 120] involve a combination of CNN and LSTM for exploiting spatial and temporal features. A special class of CNN, namely Graph Convolutional Network (GCN), is utilized in [29] involving skeleton based graphs.

2.5.2 Identity and activity privacy preserving posture recognition system

Although, the depth sensors are less privacy invasive as compared to the visible color sensors due to lack of color and texture information, but the depth sensors also invade the

privacy of individuals and are not completely suitable for privacy-preserving indoor monitoring.

2.5.2.1 Privacy in Depth Data

Most of the work in literature considered depth data privacy-preserving as the texture and color information in depth domain is missing which make it difficult to identify the individuals in depth images. Work in [121] generated Depth Motion Maps (DMM) from depth video clips and utilize them for elderly fall detection. Authors consider the depth videos privacy-preserving and develop a fall detection system using with CNN and ELM. Another work in [4] developed a fall detection system using top mounted depth sensors and considered top-view depth images as privacy preserving. Authors in [79, 122, 123] utilized depth sensors for action recognition in indoor spaces claiming the depth clips as privacy-preserving. Another interesting work in [124], utilized depth sensors for slipping detection of elderly in bathroom, which is extremely private space. In healthcare application, an in-bed pose monitoring system [35] is developed for monitoring patients activities using depth sensors considering them privacy preserving.

While majority approaches considered depth data privacy-preserving, there are studies questions the privacy of the depth images. A recent study in [125], claims that the depth images are not fully privacy-preserving and validated this claim by performing face recognition in depth images with 98% accuracy. Works in [76, 82] also question the privacy of depth images and proposed solutions using low-scale depth images for pose estimation and action recognition, respectively. These work assumed that down-sampling the depth images to significantly smaller dimensions can preserve the identity of individuals. Another work in [126] also made the same assumption and utilize severely down-sampled depth images for human pose detection and tracking inside smart room. Author in [38], considered depth im-

ages privacy-invasive and develop a modified depth sensor for capturing privacy-preserving images to preserve individual's identity. Some other works in [127–129] employ depth data based face recognition and each work achieved more than 85% accuracy. These observations concludes that the depth images do not preserve identity privacy and are not suitable for privacy-preserving indoor monitoring.

2.5.2.2 Privacy Preservation in Vision Sensor based Systems

The vision sensor based approaches those consider individual's privacy can be categorized into two categories; post-capture privacy and pre-capture privacy. In post-capture privacy, the idea is to hide the privacy sensitive information (i.e., human faces) using various methods after capturing the images. The simplest approaches for post-capture privacy includes down-sampling [76, 82], blurring human faces [73], and pixelation/removal of human region [130]. Another category of approaches used in post-capture privacy are based on encryption [74, 75, 75, 131] to change the visual representation of the images. The most recent approaches used in post-capture privacy use adversarial learning [33, 34, 77] to conceal identity oriented information from images.

Pre-capture privacy, as name suggests, enable sensor to capture data without privacy sensitive information. One of very few work in [126] utilize an sparse array of depth sensors for human tracking in indoor spaces. Another approach in [80] also utilize a low resolution depth camera to capture privacy-preserving data to locate elderly individual inside home. A very interesting study in [38] presents a modified depth sensor to capture privacy preserving data. Though, all the approaches discussed above consider privacy concern focus on preserving individual's identity based of facial recognition.

2.5.2.3 Posture Recognition

Human posture recognition is crucial part of coarse-grained daily activities (i.e., sleeping, walking, running, exercising) and the abnormal activities like fall. Many researchers developed posture recognition systems, we discuss the posture recognition systems using vision sensors in this section. A simplistic approach in [132] exploit the height of human body as a feature to classify different body postures using various decision rules. Some other approaches in [5, 133, 134] employ some complex machine learning techniques such as Neural Network and Support Vector Machine to classify human postures with LBP based features and projection histogram features, respectively. Other recent approach in [135] utilized CNN based classifier to recognised human postures from low scale infrared image. A work in [136] employ a multi-channel CNN model for posture recognition using a combination of RGB and depth images. Applications of posture recognition in yoga and sport activities are explored in [137, 138].

Chapter 3

Privacy-Preserving Indoor Fire Detection System

3.1 Introduction

Fire is unfortunately one of the primary sources of calamities on our planet. Casualties, both human and property, are large in fire based catastrophes. According to a report by the US Fire Administration [139], around 13 million fire incidents were reported per year in the five year period between 2013 and 2017 in the USA alone and resulted in 3316 deaths, 15370 injuries. To mitigate the damage from such incidents, several endeavours are directed towards developing effective frameworks for fire detection. A fire detection framework should ideally identify a fire in varied environments (such as residential, commercial, outdoor, forest) as quickly as possible to minimise damage.

The performance of a fire detection system is broadly dependent on two factors: a sensing device; and an algorithmic mechanism to assess the sensed parameters. Various kinds of sensors are employed in literature to detect fires effectively. These include and are not limited to heat, light, humidity, and gas sensors. The major limitations of such sensor based systems [56] include frequent false alarms and a high response time. To overcome these,

work in the more recent past is directed towards utilising vision-based sensors [93][85] for fire detection wherein images and/or videos across spectral bands such as visible, infrared, and multi-spectral are used. Vision-based sensors have superior classification accuracy and smaller response-time. Within the spectrum of vision-based sensors, infrared systems are the most effective for fire detection owing to their unique property of depiction wherein 'hotter objects are brighter'.

In spite of the advantages of vision-based fire detection systems, the big drawback of these systems that make them impracticable for most residential and office spaces is the intrusion on occupants' privacy [68]. While visible color image-based systems are not at all suitable for monitoring private spaces, even infrared image-based systems that do 'fuzzify' images to an extent are unacceptable. In addition to this, the high costs of thermal IR cameras is another important limiting factor in their use for regular monitoring. Two major directions of work towards privacy preservation are available in literature, these are: *Intervention* (preventing the camera from capturing private information); and *Image redaction* (hiding sensitive information such as human faces in captured images).

Complementing sensors in their bid to effectively detect fires are algorithms of various kinds that draw conclusions from the parameters sensed. Most algorithms used in vision-based fire detection systems, however, are complicated and require considerable human expertise and effort. In this context, Convolutional Neural Networks (CNN) are emerging as a powerful approach for object detection in images that significantly reduce human intervention by taking direct images as input instead of hand-engineered features. There are several approaches that utilise CNN [20][105] to detect fire using images/videos and are demonstrably superior to other machine learning approaches.

In this chapter, a vision-based privacy-preserving efficient fire detection system based on both spatial and temporal properties of fire is proposed for private spaces in residential

and commercial establishments. In addition to efficiently detecting fires whilst preserving privacy, the system reduces false alarms drastically as is characteristic of a vision-based approach. The approach to privacy preservation is of the *Intervention* kind and employs a Near Infra-Red (NIR) camera. The NIR camera is employed in a manner that it captures infrared images with severely reduced visibility and effectively conceals human presence and activities. The proposed fire detection technique harnesses the high intensity difference between regions of fire and normal objects in such images to detect fires and simultaneously preserve privacy. The proposed technique is able to efficiently distinguish such non-fire images from those of fire.

Another issue with privacy is its subjective nature. An image may be perfectly acceptable in terms of privacy preservation to one individual and not so much to another. To address this, a survey comprising individuals across demographics was conducted locally and globally through personal solicitations and using Amazon Mechanical Turk (AMT). The results of the survey enabled us to safely identify the 'level' of images to use that would be efficient in fire detection whilst simultaneously preserve occupant privacy.

Given the superiority of CNN in object detection and classification, we developed a light-weight CNN architecture for exploiting the spatial properties of the images. In addition to this, the temporal characteristics of fire motion (i.e. fire flame movement) were also harnessed using continuous frame differencing. An integrated model combining both spatial and temporal properties is proposed as an effective means of fire detection. In addition to fire detection capability, the framework is deliberately made light-weight so as to be seamlessly deployed in remote, resource-constrained locations.

Keeping all these factors: privacy, cost, performance, and resource constraints in mind; the key contributions of the work in this chapter are as follows:

1. Assessment of an acceptable level of privacy in images by surveying both local and

global audiences utilising crowd-sourcing services.

2. Development of a lightweight, fast, and efficient fire detection system using CNN and temporal features.

3. Demonstration of the efficacy of the proposed framework by comparing it with existing techniques and through a prototypical deployment in a real-world resource-constrained environment.

The content of this chapter is organised as follows. The proposed methodology including camera modification, privacy assessment, and the proposed architectures is elucidated in Section 3.2. A few case-studies demonstrating the utility of the proposed framework in the real world are included in Section 3.3. Experimental validation of the approach and deployment in a resource constrained environment are included in Section 3.4. Finally, Section 3.5 concludes the chapter.

3.2 Proposed Methodology

In this section, the approach proposed is discussed in detail. In an endeavour to capture images in private spaces without compromising on occupants' privacy, a vision sensor (a simple color camera) is modified appropriately. The details of this modification are discussed first. Subsequently, details of a survey, conducted to ensure that the images captured by the modified camera are indeed privacy preserving, are discussed. Finally, the efficient lightweight model proposed to detect fires within these diminished images/videos is described. The schematic diagram of the proposed framework is shown in Figure 3.1. In addition to depicting the steps of the framework, the diagram also shows deployment of the system in a resource-constrained environment.

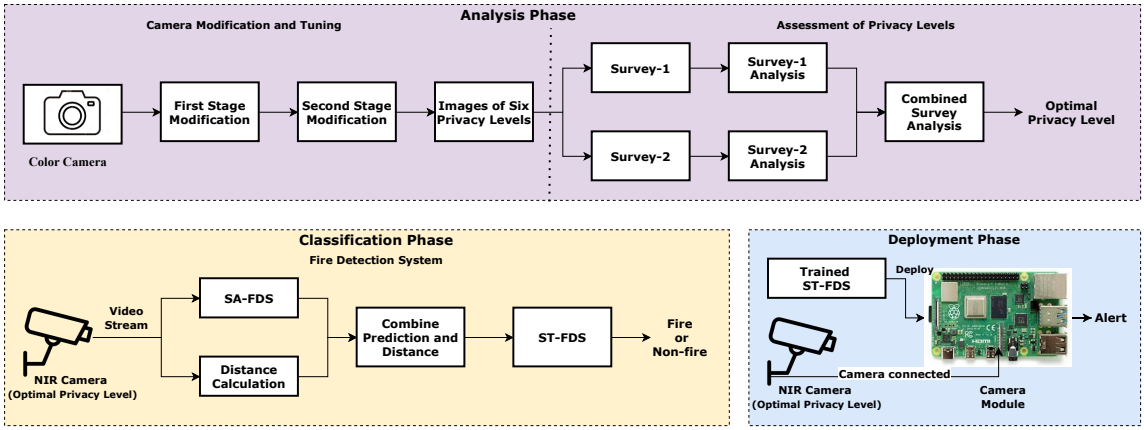
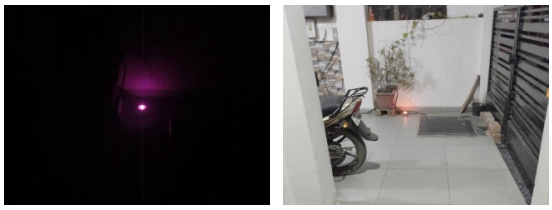


Figure 3.1: Workflow of the proposed Fire Detection System

3.2.1 Camera Modification and Tuning

The approach utilises an NIR camera in a manner that the images captured are of significantly reduced visibility and ensure privacy of inhabitants. In our work, we utilize a modified color camera to capture privacy-preserving images while retaining information that can distinguish fire flames from other non-fire objects. A *Canon SX430 PowerShot 20MP* digital camera is used whose cost is around $200. The camera modification was done in two stages to make it suitable for the proposed work: the first stage of modification involved the conversion of the given color camera into an NIR camera. This stage is required only if one chooses to use a color camera and may be skipped if one has an NIR camera to start with; subsequently modifications were done to further diminish the quality of images captured to effectively preserve occupants' privacy.

In the first stage of camera modification, the installed IR block filter was removed and replaced with an IR pass filter similar to [140] that allows radiations of wavelength above 850 nm (Near IR spectrum) to pass through and blocks light in the visible spectrum. The cost of this modification worked out to around $70 and included the cost of the filter. The camera modified in this manner became an NIR camera which can capture the reflected infrared radiations from objects and works much like a visible camera. The Sun and fire are two



(a) (b)

Figure 3.2: Differences in captured images: (a) Image using an NIR camera, (b) Image using a Color Camera

primary sources of infrared radiations, especially near infrared radiations. Tungsten bulbs, LEDs, quartz halogen lights also emit a small amount of NIR radiations. An NIR camera is sensitive to all NIR light sources such as fire, sunlight, light bulbs. As it is sensitive to NIR radiations and less sensitive to visible light, it is a suitable choice for fire detection whilst preserving occupants' privacy. This can be seen in the image in Figure 3.2(a) taken by the developed NIR camera. Figure 3.2(b) is an image taken by a regular color camera of the same setting at about the same time. The setting is of the evening hours with no sunlight and no other source of NIR radiations except a small candle. Therefore, nothing is visible in the first image except the light from the candle.

An NIR camera is sensitive to sunlight, hence objects in images captured by the NIR camera in daylight, are visible but are not very clear. To further reduce the visibility of such non-fire objects, the second stage of modification of the camera was done and comprised placing a thin translucent LDPE (Low-Density PolyEthylene) film before the camera lens. The image quality depends mainly on the amount of light that passes through the camera lens, hence the light reaching the lens in the camera is restricted by the LDPE film. According to a technical report [141], the wavelength of light is reciprocal to the refractive index of the material. Therefore, when light of different wavelengths (i.e., visible and NIR) passes through the translucent LDPE film, the low wavelength visible light experiences high refraction in comparison to NIR light. Most NIR light thus reaches the lens while most visible

(a) Color Image (b) Privacy Level-1 (c) Privacy Level-2 (d) Privacy Level-3 (e) Privacy Level-4 (f) Privacy Level-5 (g) Privacy Level-6

Figure 3.3: Images at the 6 privacy levels

light gets lost in refraction. This significantly reduces the visibility of objects. This reduced visibility is not applicable, however, to NIR light sources such as fire, sunlight (that comes through windows), light bulbs, LEDs, all of which are visible and easily distinguishable.

As we discuss the utility of NIR cameras in fire detection, it is important to note that Thermal Infra-Red (TIR) cameras that work by sensing the temperature of objects are also useful for fire detection tasks. The incapability of such cameras in preserving the privacy of occupants, however, make them unsuitable for indoor fire detection. TIR cameras are sensitive to the heat from fires and capture these very effectively. These cameras are, however, also sensitive to the heat emitted by human beings and effectively detect their presence and activities as well. While this may be a boon in certain circumstances, it is a disadvantage in indoor fire detection as the privacy of occupants is compromised. TIR cameras are thus not preferred for this application. In addition to this, TIR cameras in general are more expensive than NIR cameras and are not ideal for mass adoption and deployment. There are, however, contradicting claims in this regard with cheaper versions of TIR cameras also available.

3.2.2 Assessment of Privacy levels

As the proposed fire detection system is meant for private spaces, a privacy assessment exercise was conducted to understand the acceptable degree of privacy in images. Degrees of Privacy from least private to most private may be classified as: *everything is visible in the image*; to *nothing is visible*. *Nothing is visible* is the highest degree of privacy. From

the point of view of privacy preservation, this should be the degree of privacy maintained when capturing images. The limitation with this, though, is that the classifier being used to identify fire in such images may not effectively be able to differentiate fire from non-fire scenes. On the other hand, while working on an image where *everything is visible*, the performance of the classifier in fire detection is the best but the privacy of occupants is not preserved. The solution, therefore, is a level of privacy somewhere in between these two extremes that enables both precise fire detection and preserves occupants' privacy. To determine this optimal level of privacy in images, six privacy levels were identified comprising images captured by modifying the NIR camera to different extents.

The modifications were done by progressively changing the number of layers of the LDPE film placed before the camera lens and by varying the shutter speed. The six privacy levels in decreasing levels of visibility (and consequently increasing levels of privacy) are as follows:

- ***L1:*** Images taken with NIR camera at low shutter speed.
- ***L2:*** Images taken with NIR camera at high shutter speed.
- ***L3:*** Images taken with NIR camera with a layer of film before camera lens at low shutter speed.
- ***L4:*** Images taken with NIR camera with a layer of film before camera lens at high shutter speed.
- ***L5:*** Images taken with NIR camera with two layers of film before camera lens at low shutter speed.
- ***L6:*** Images taken with NIR camera with two layers of film before camera lens at high shutter speed.

Images at these six privacy levels for the same scene are shown in Figure 3.3 along with the reference color image. The decreasing visibility with increasing privacy levels is

Table 3.1: Survey Questionnaire

Q.No.	Questions
1	Where, in your opinion, is this picture taken (Kitchen, Bedroom, Living-room, Laboratory, Store-room, or any other location)?
2	What living or non-living objects do you see in the image? (e.g one or tow persons, one bed, two computers, or any thing else)?
3	If, according to you, the image has one or more human-being(s), what is his/her/their location? (e.g. Floor, Bed, Sofa, Chair).
4	If, according to you, the image has one or more human-being(s), what is his/her/their position? (e.g. Standing, Sitting, Lying).
5	If, according to you, the image has one or more human-being(s), what activity or activities do they seem to be doing? (e.g. Eating, Reading, Talking, Working, Sleeping).
6	Any other comments? (Optional)
Note: What can you see in this image? (If you cannot see anything, please write 'Not Clear' as the answer.) In cases that you feel there is more than one human-being in the image, please provide information for each person separately.	

apparent from the figure. To understand the acceptable levels of privacy in images, two surveys were conducted involving both local and global participants.

In the first survey, 70 participants from various geographical locations, age groups, educational backgrounds, and occupations were contacted of which we received responses from 50 participants. The second survey was published on Amazon Mechanical Turk (AMT) where we solicited 50 responses from participants across time zones, age groups, educational backgrounds, and occupations.

The survey involved showing the participants images captured at various locations such as Living room, Bedroom, Computer Lab, Storeroom, Hall, among others in varying light conditions. Fifteen images at different privacy levels (two at Level 1, two at Level 2, two at Level 3, six at Level 4, two at Level 5, and one at Level 6) were included in the survey. The images used in the survey were similar to those in Figure 3.3 and are available at following link: SurveyLink.

The questions asked in the survey are included in Table 3.1. Responses collected from both the surveys were analysed to identify the image level optimal for maintaining privacy as well as providing good discriminating properties for accurate fire detection. In the analysis, the collected responses were categorized as: *correct, incorrect, not clear*. The total number of responses in each category (*correct*, *incorrect*, and *not clear*) was counted for all images at that level corresponding to each question. The calculations were done as per Equation (3.1).

$\forall I_x \in L_i$

$$T_{L_i C_j Q_k} = \frac{\sum_{x=1}^{X} R_{C_j Q_k}(I_x)}{\sum_{C=1}^{j} \sum_{x=1}^{X} R_{C_j Q_k}(I_x)} \tag{3.1}$$

- i = Six privacy levels.
- j = Three categories (i.e. correct, incorrect, and not clear).
- k = Five questions (i.e. place, #humans, location, position, and activity).
- x = Number of images in survey of level L_i.

where L_i, C_j, Q_k represent the privacy-level$[i]$, category$[j]$, and question$[k]$, respectively. $R_{C_j Q_k}(I_x)$ is the response to question$[k]$ on image$[x]$ of category$[j]$. $T_{L_i C_j Q_k}$ is the percentage of total responses in category$[j]$ and privacy-level$[i]$ corresponding to question$[k]$.

3.2.3 Fire Detection System

The proposed system for fire detection based on deep learning and temporal properties is presented in this section. It is observed that combining both spatial and temporal features of images can significantly improve fire detection efficacy.

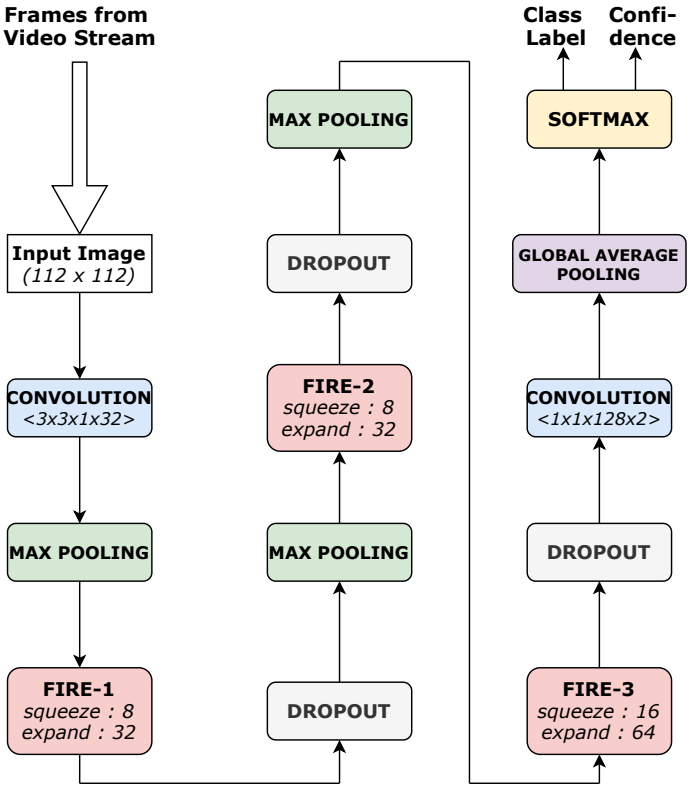


Figure 3.4: Spatially-Aware Fire Detection System (SA-FDS)

3.2.3.1 *System based on Spatial Properties*

To exploit the spatial properties of images, a lightweight convolutional neural network, namely SA-FDS (Spatially Aware Fire Detection System), is proposed that utilises the fire module of the well-known SqueezeNet architecture [142]. The Fire module comprises two layers: *squeeze* and *expand*; where the squeeze layer contains only 1x1 convolution filters; and the expand layer contains a combination of 1x1 and 3x3 convolution filters. The 1x1 filters play a vital role in parameter reduction as they have fewer parameters as compared to the high dimensional 3x3 filters.

The SA-FDS architecture with ten layers, including three fire modules, is shown in Figure 3.4. A grayscale input image first goes through a convolution layer with a filter size of 3x3, followed by the rectified linear unit activation. Subsequent to this, three max-

pooling layers and fire modules are used in an alternating manner. A 1x1 convolution layer is applied with ReLU activation, followed by a global average pooling layer. To deal with model over-fitting, a dropout layer is added after each fire module with dropout ratios: 0.2, 0.2, and 0.5, respectively. Similar to SqueezeNet, the proposed model also does not include any fully connected layer, and this reduces the number of parameters significantly. Finally, a softmax layer is placed to produce the probability distribution for both the classes (i.e., Fire and Non-fire). The Softmax function computes the probability of each class as described in Equation (3.2).

$$P_j = \frac{\exp\left({C_j}^2\right)}{\sum_{j=1}^{2} \exp\left({C_j}^2\right)}, j = 1, 2 \tag{3.2}$$

C^2 is the output of the Global Average Pooling (GAP) layer and P_j is the probability of Class j. Here $j = 1$ denotes fire and $j = 2$ denotes non-fire. The final output is the class label corresponding to input image (I) along with the confidence score of the prediction. The class label is given based on the maximum probability index and is defined by Equation (3.3).

$$\begin{aligned} Class\_label[I] &= \arg\max_{j \in \{1,2\}} (P_j) \\ Confidence[I] &= \max_{j \in \{1,2\}} (P_j) \end{aligned} \tag{3.3}$$

The confidence score is the maximum probability corresponding to the class label. As fire detection is a binary classification problem, the binary cross entropy loss function is utilized to train the proposed SA-FDS. The binary cross entropy loss is defined in Equation (3.4).

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} y_i * log(p_i) + (1 - y_i) * log(1 - p_i) \tag{3.4}$$

where p_i is the probability of class label y_i and N is the number of samples.

The proposed SA-FDS framework is an adaption of the SqueezeNet [142] architecture

with certain modifications to make it light-weight and suitable for a resource constrained environment. The modifications involve reducing the number of fire modules and the number of filters within each fire module. This reduces the number of parameters significantly. In addition to this, pooling layers are added after each fire module to more quickly down-sample the feature maps and compensate for the reduced number of layers. A dropout layer is added after each fire module barring the last one to reduce over-fitting. The size of the input and output in SA-FDS are also different from the original SqueezeNet. A detailed description of the modifications is given in Appendix (A).

3.2.3.2 *Integrated System using Spatial and Temporal Properties*

A fire detection system based on convolutional neural networks takes individual images as input and predicts the image label based on spatial properties only. The continuous movement of a fire flame is an important feature and is used along with SA-FDS to further improve performance. Temporal considerations significantly reduce false alarms often triggered by fire shaped 'non-fire' objects. In a stationary camera setting, a fire scene changes very frequently when compared to a non-fire scene because of the movement of the flame(s). Figure 3.5 shows how the shape of the fire flame changes continuously with time.

Continuous frame differencing that calculates the difference (or similarity) between two frames is handy in this scenario. The distance between two similar frames is about zero and this value is high for dissimilar frames as given in Equation (3.6). The Mean Square Error (MSE) technique [143] is used for frame differencing because of its simplicity and low time-complexity. We also experimented with the Minkowski distance metric described in Equation (3.5) with different values of m. We found that Minkowski distance with $m = 2$ (similar to MSE) is better in terms of performance and/or speed. The results and a detailed

analysis of these experiments are included in Appendix (A).

$$Distance_m = (\sum_{i=1}^{n} |X_i - Y_i|^m)^{\frac{1}{m}} \tag{3.5}$$

where X and Y are two n-dimensional data points and m is the order.

The distances between subsequent frames in Figure 3.5 are 127.28, 63.88, and 73.73 respectively. Frame differencing alone is not sufficient for classification as there can be several other reasons (i.e., light variation, movement due to normal activities) that cause changes in the scene. The proposed integrated framework, ST-FDS (Spatio-Temporal Fire Detection System), therefore combines the spatial and temporal aspects respectively of the

Algorithm 3.1 Spatio-Temporal Fire Detection System (ST-FDS)

Input : Video Stream (VS)
Output: Prediction Labels ($prediction$)

Initialize: $F_{ref} \leftarrow None$

foreach k^{th} *frame* F_t *in VS* **do**
    CNN_Label, Confidence = SA-FDS(F_t)
    if F_{ref} *is None* **then**
        $prediction$=*CNN_Label*
    else
        if *Confidence* > *T1* **then**
            $prediction$=*CNN_Label*
        else

$$Distance = \frac{1}{w.h} \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} [F_t(i,j) - F_{ref}(i,j)]^2$$

            /* where w and h are the width and height of the frame, respectively. */
            if $Distance$ > *T2* **then**
                $prediction$ = *FIRE*
            else
                $prediction$ = *NON-FIRE*
    /* where *T1* and *T2* are the thresholds for SA-FDS confidence and distance between two images, respectively. */
    $F_{ref} = F_t$

two techniques for increased efficacy. The framework is described in Algorithm (3.1).

$$\begin{aligned} Distance &= 0, \quad \text{if } I_x \doteq I_y \\ Distance &> 0, \quad \text{if } I_x \neq I_y \end{aligned} \tag{3.6}$$

where I_x and I_y are the two images. Distance represents the difference between I_x and I_y.

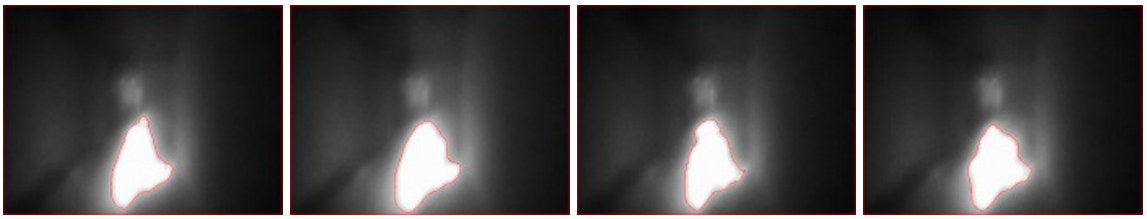
The video stream is the input to the algorithm. Each frame $F(t)$ of the video stream is extracted, and its distance $dist(t)$ from the frame captured k time units earlier, $F(t-k)$ is calculated. Frame $F(t)$ is then fed into a trained SA-FDS model to produce the predicted image label and confidence score. If the confidence score is below a threshold $T1$ indicating that the SA-FDS classifier is not so sure of its prediction; the distance $dist(t)$ is compared with another threshold $T2$. If $dist(t)$ is below $T2$, it implies that the image frames have not changed significantly and hence the classification is that of 'non-fire'. Conversely, if $dist(t)$ is larger than threshold $T2$, the implication is that the frames have changed as is the nature of fire and the classification is that of fire. The value of thresholds $T1$ and $T2$ are chosen according to Equation (3.7).

$$\begin{aligned} &\underset{T1 \in S_1, T2 \in S_2}{\text{maximize}} \quad A(T1, T2), \underset{T1 \in S_1, T2 \in S_2}{\text{maximize}} \quad R(T1, T2) \\ &\text{s.t.} \qquad\qquad T1, T2 > 0. \end{aligned} \tag{3.7}$$

where A and R are functions that denote Accuracy and Recall. These depend on the variables $T1$ and $T2$ that represent the thresholds. S_1=$\{s_{10}$,s_{11},s_{12},.....,$s_{1m}\}$ and S_2=$\{s_{20}$, s_{21}, s_{22}, ...,$s_{2n}\}$ are the domains of $T1$ and $T2$, respectively.

3.3 Case Studies

The following case studies effectively convey the utility of the proposed framework in the domain of fire detection in indoor spaces.



Frame at time *'t'* **F(t)** | Frame at time *'t+k'* **F(t+k)** | Frame at time *'t+2k'* **F(t+2k)** | Frame at time *'t+3k'* **F(t+3k)**

Figure 3.5: Movement of a fire flame in contiguous video frames

[Case I] *Jane is working in the kitchen of her 17th floor apartment. Her apartment is endowed with good quality smoke detectors. It has been a while and Jane decides to cook fried eggs today. While they are delicious, cooking them involves heating them for several minutes on a frying pan. She starts cooking them and the inevitable smoke starts escaping from the frying pan. Before she realises it, her smoke detector is triggered and sets off an alarm. Within a few seconds the alarm triggers other alarms throughout the sky-scraper and at least 500 people rush down the stairs because of what was unfortunately a false alarm.*

[Case 2] *John and Harry, as most others of their age, are easy-going, carefree teenagers. It is their lucky day as John's parents are travelling to another city. They decide to meet up at John's apartment and have some fun. They decide to build a small 'bonfire' in the balcony with some pizza, pop, and their favourite movie. They are having a good time. They build the fire and it burns smoothly with very little smoke of which most escapes outside from the balcony. The smoke detector in the apartment fails to detect anything, until, unfortunately a wooden chair accidentally falls on the fire and starts burning. The fire spreads and gets hold of the curtains. Only then are the smoke detectors triggered. It is too late.*

In the two cases above, a visual sensor would have been more effective. Visual sensors are not triggered by harmless kitchen smoke. In the second case, a visual sensor would have detected a fire even when it was very small and had not attained dangerous proportions.

The following case study is about issues with visual sensors.

[Case 3] *Chris and Molly are a happy couple and decide to install visual sensors in their apartment. These are cheaper, cleaner, and more effective. They, however, restrict these to only their living area, kitchen, and balcony. They understandably are not comfortable with such visual fire sensing systems in their bedrooms and washrooms.*

Taking into account, this rather important limitation of visual sensors, the proposed technique ensures the use of visual sensors in a manner that their benefits are harnessed whilst preserving the occupants' privacy.

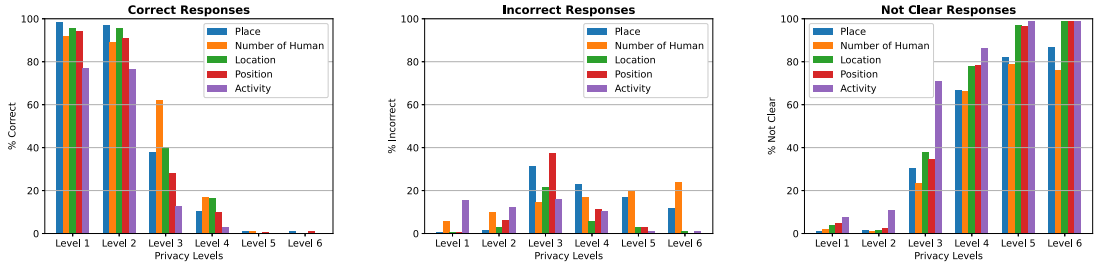
3.4 Experimental Evaluation

To identify a level of privacy acceptable to most people, a survey was conducted with images of different privacy levels and the survey results were analysed. Subsequently, images falling in the category of 'acceptable privacy level' were used to evaluate fire detection capabilities of the proposed integrated fire detection system.

3.4.1 Survey Analysis

Apart from the informal perception of the research team and verbal feedback from a number of people, two surveys with the random crowd were performed to identify the appropriate privacy-preserving level of images. Each survey comprised fifty participants from different geographic locations, with different levels of education, professions, and age groups. The survey where participants were contacted by the research team comprised mostly people in country India; whereas in the survey conducted over AMT, the participants were mostly from countries USA and India. Furthermore, most participants were in the age group of 20-40 years and had a Bachelors' or Masters' degree. The number of salaried employees in both the surveys was larger than other occupations.

Images corresponding to the six levels of privacy (described in detail in Section III B) i.e.



(a) (b) (c)

Figure 3.6: Number of 'correct', 'incorrect', and 'not clear' responses received at each privacy level in the surveys

from Privacy Level 1 (Highest Visibility, Least Privacy) to Privacy Level 6 (Least Visibility, Highest Privacy) were included in the surveys. Images in the survey were similar to those in Figure 3.3.

A wide range of responses was received for each question corresponding to each image. Responses were categorized as: *correct*, *incorrect*, and *not clear*. For each level of privacy, the total number of responses in each category (*correct*, *incorrect*, and *not clear*) was calculated for all images at that level corresponding to each question using Equation (3.1). Figure 3.6 (a)-(c) includes plots of *correct*, *incorrect*, and *not clear* responses at the six privacy levels. All 100 survey responses from the two surveys (50 from each survey) were used to calculate the percentage values plotted in Figure 3.6. Percentage values for correct responses calculated separately for the two surveys are shown in Table 3.2.

Table 3.2: Analysis of the correct responses received in two surveys (values are given in %)

	Amazon Mechanical Turk Survey						**Survey with Random Participants**					
	L1	**L2**	**L3**	**L4**	**L5**	**L6**	**L1**	**L2**	**L3**	**L4**	**L5**	**L6**
Place	98	98	32	5	1	2	99	96	44	16	1	0
#Human	93	87	61	13	1	0	91	91	63	20	1	0
Location	96	95	57	12	0	0	95	96	24	21	0	0
Position	97	91	45	10	0	0	92	91	11	11	1	2
Activity	81	77	9	4	0	0	73	76	17	2	0	0

The majority of respondents of the two surveys gave *incorrect* or *not clear* responses to questions corresponding to images at Privacy Levels 5 and 6. Given this fact, images at Privacy Levels 5 and 6 would be perfect for privacy preservation. These images, however, do not have discriminating properties that are adequate for effectively distinguishing between fire and non-fire objects. This is because objects in these images are not apparent, and edges are highly blurred. At the other extreme, around 90% respondents gave *correct* responses to questions corresponding to images in Privacy Levels 1 and 2, implying that most objects in these images are visible thus making these image unsuitable for privacy preservation.

The number of correct responses received for images at Privacy Level 3 was smaller than those for Privacy Levels 1 and 2. At Privacy Level 3, question related to human presence was correctly responded to by about 60% of the respondents. Questions about the place and location of the human were correctly responded to by about 40% of the respondents. Even though the *correct* responses pertaining to the activities of the human-beings in the images were significantly low at Privacy Level 3, it is still not good enough to be used for a privacy preserving system. Interestingly, images at Privacy Level 3 received the largest number of *incorrect* responses as compared to other levels indicating that the level saw a good number of random guessing, owing perhaps to the degree of visibility in the images which tempted a guess but was not good enough for the guesses to be correct.

The images at Privacy Level 4 received very few *correct* responses, less than 20% across questions. Over 70% of the responses were *not clear* at this level. The questions on the location of humans in the images and the number of humans in the images received between 15-20% *correct* responses, and possibly even these correct responses were due to there being fewer choices and thus increased chances of a correct guess. The number of *correct* responses to questions on human presence and location at Privacy Level 4 is smaller than what could be achieved by random guessing (i.e., for five choices, the probability of

a *correct* response is 20% in random guessing). For questions pertaining to other crucial information like place, human position and activities, the *correct* response rate was not even 10%, which is again much worse than random guessing. This indicates that images at Privacy Level 4 are appropriate for privacy preservation. In addition to this, images at this level have discriminating properties and are adequate for correctly distinguishing fire from other bright non-fire objects. We also did some preliminary investigations on the efficacy of images at Levels 5 and 6 in detecting fires and these were found to be inferior to Level 4 images. Level 5 and 6 images are, however, better than Level 4 images in privacy preservation. Intuitively also, it is clear that the blurred edges and low intensity of Level 5 and Level 6 images make objects indistinguishable and result in poor classification performance. The images at Privacy Level 4 are, therefore, ideal for a privacy preserving and efficient fire detection system.

To more concretely analyze the image clarity at various privacy levels, we calculate a distance of the image at that level from the image at Level 1 (the clearest of all images) using the MSE technique [143]. A large distance indicates a less clear image and a low 'clarity index'. A small value of the distance indicates a more clear image and a large clarity index. The distances of the images at different privacy levels (given in Figure 3.3) along with their clarity indices are included in Table 3.3.

Table 3.3: Clarity Index of the images of different privacy levels

Image-1	**Image-2**	**Distance**	**Clarity Index**
Privacy Level-1	Privacy Level-2	183.48	6
Privacy Level-1	Privacy Level-3	203.34	5
Privacy Level-1	Privacy Level-4	229.24	4
Privacy Level-1	Privacy Level-5	300.79	3
Privacy Level-1	Privacy Level-6	328.60	2

The results in Table 3.3 indicate that the distance of images at Privacy Level 2 are less

than the corresponding images at higher privacy levels. The distances of images at Level 6 are highest, thus indicating lower clarity of images at higher levels of privacy.

3.4.2 Fire Detection System

This section comprises details on the experiments conducted, dataset created, and training of the proposed CNN architecture SA-FDS and the integrated system ST-FDS, for fire detection. All experiments were conducted using a dataset containing images/videos at Privacy Level 4, collected using the modified NIR camera. Comparisons of the proposed framework with other state of the art techniques are also included in this section.

3.4.2.1 Dataset

Research on fire detection using vision-based sensors has mostly been done using color images/videos. Therefore, datasets of images in the infrared or near-infrared spectrum are not available (except aerial infrared images). Also, vision based fire monitoring has only been explored for outdoor and/or public spaces until now. The focus in this chapter is at exploring the use of vision-based systems for fire detection in personal spaces (such as living rooms, bedrooms, kitchens, office spaces) whilst simultaneously preserving the occupants' privacy. In the absence of an appropriate dataset of images in the infrared and near-infrared regions, especially those of private spaces, we created a fire and non-fire dataset using the modified camera described earlier. This dataset includes images at Privacy Level 4 in line with the conclusions drawn in Section V(A) about Privacy Level 4 being the most appropriate level for privacy preservation and fire detection.

The fire dataset, images of actual fires, was collected from eight different locations inside under-construction buildings and storerooms at our institute, after due approvals. Similarly, the non-fire dataset was collected from different locations comprising private spaces inside

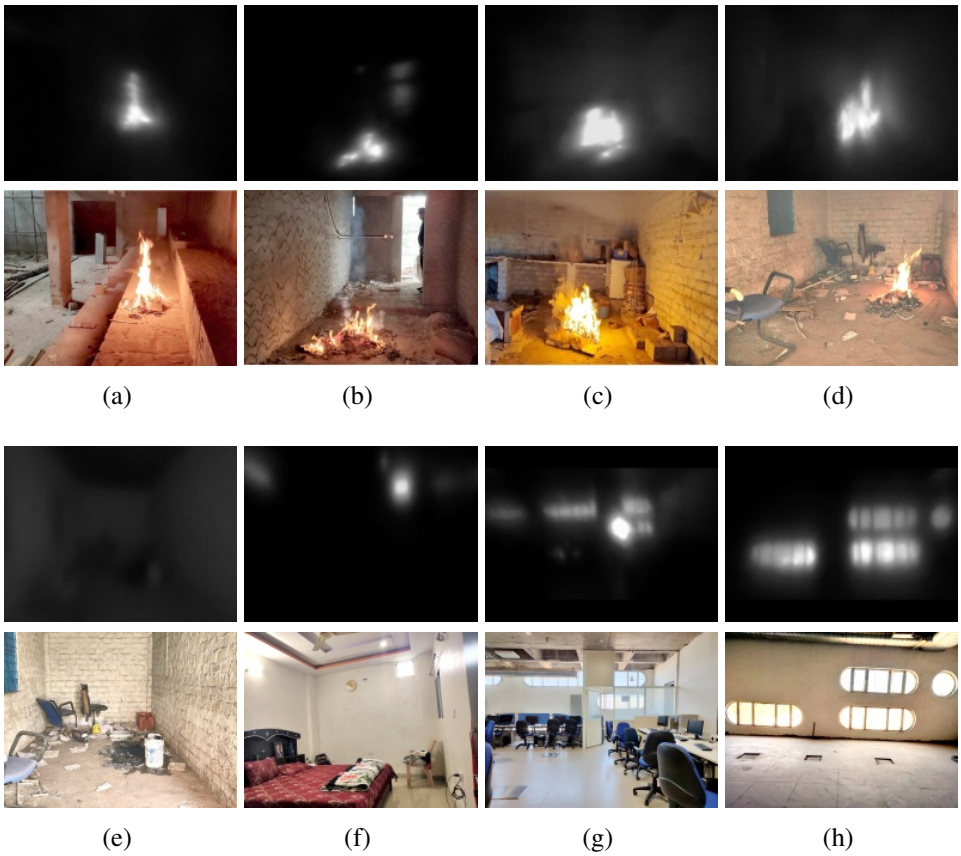


Figure 3.7: Sample images from the dataset created, (a)-(d) are fire images, and (e)-(h) are non-fire images. (Images from the created dataset are in the first & third row and the corresponding color images are in the second & fourth row, respectively)

a home, an office space, computer laboratories, store rooms, and other spaces inside under-construction buildings. To introduce variation, the dataset was collected by placing the camera along all four directions (East, West, North, and South) for each captured scene and at varying distances in the range from 2 meters to 6 meters. The fire dataset contains images captured mostly at daytime (90%) and a few at night (10%).

All images in the non-fire dataset were taken during the day. Both the datasets were collected in the form of short videos of 30 to 40 seconds each to start with. To make the dataset diverse, several non-redundant small video clips of 1 to 5 seconds duration were extracted from random points in the original videos. This was done to reduce the number

of similar images extracted from the long videos. This resulted in around 1000 small video clips (136,056 frames), comprising an equal number of fire and non-fire clips. Sample images from the created dataset are shown in Figure 3.7. The dimensions of the original images are 640x480 and are converted into grayscale. Our dataset contains a large number of non-fire images comprising scenarios with sunlight (coming through the window), light sources such as LED, bulbs and other such NIR sources. The dataset is made publicly available and can be downloaded from the following link: DATASET.

3.4.2.2 Training

Training of the SA-FDS and ST-FDS algorithms was done separately. The dataset was randomly divided into training and test sets in a 70:30 ratio. The training and test set splits were done randomly from 1000 small video clips captured from different settings. Also, clips in the test set were not seen by the model during training. We, therefore, ensured that the training and test sets contained sufficiently different samples. In training the SA-FDS model, 80% of the training data was used to train the model and the remaining 20% of the training data was used for validation. Adam optimization was employed with an initial learning rate of 0.001 and decay parameters: $\beta 1 = 0.90$ and $\beta 2 = 0.99$. The model weights were initialized randomly and the model was trained for 100 epochs using a binary cross entropy loss function. A regularization parameter weight decay of 0.001 was used to avoid over-fitting of the model. Apart from this, data augmentation of various types (i.e. horizontal flip, width-shift, height-shift, shearing, variation in brightness and size) were harnessed to effectively handle model over-fitting. To compare the proposed models with prominent ones in literature, the latter were implemented and trained using the same dataset with similar augmentation strategies. All the model structures and their respective parameters were kept same as mentioned in original papers. The input image shapes of the various architectures

were different, and hence images were re-scaled to the required dimensions before feeding them to the respective models.

ST-FDS combines the SA-FDS framework with the temporal factor. For training the ST-FDS framework, two parameters $T1$ and $T2$ used in Algorithm (3.1) need to be tuned. $T1$ is the optimal confidence score of the prediction made by SA-FDS, and $T2$ is the optimal distance of the current frame from the previous one. The values of the thresholds $T1$ and $T2$ are decided through a grid search on a large range of values. For threshold $T1$, values ranging from 0.8 to 1.0 were tried at fixed intervals of 0.01. Similarly, for threshold $T2$, values ranging from 0 to 100 were tested (the range was decided based on the mean distance of non-fire data). It was observed that both $T1$ and $T2$ are dependent upon each other and need to be chosen in pairs. In fire detection, false negatives are more dangerous than false positives and hence there should be greater emphasis on eliminating these. Thresholds $T1$ and $T2$ were, therefore, chosen so as to achieve the highest possible accuracy whilst keeping the false negatives as low as possible. At threshold $T1 = 0.92$, and $T2 = 36$, the accuracy with the training set was the best, and false negatives were the least.

Another parameter k used in Algorithm (3.1) determines the gap between two frames. A value of $k = 5$ implies that every 5^{th} frame is used for classification, which works out to approximately $1/6^{th}$ of a second. A value of $k = 5$ was chosen because with a value of k smaller than 5, the distance between two frames was not discriminating enough (the scene does not change significantly from one frame to the next); whereas with a large k, only a small number of frames reach SA-FDS and the classification was delayed.

3.4.2.3 Results

Every 5^{th} frame of the clips in the test-set was used for testing. Comparison of the performance of SA-FDS (Spatially Aware Fire Detection System) and ST-FDS (Spatio-

Table 3.4: Performance comparison of fire-detection techniques

Technique	Precision (%)	Recall (%)	F-Score (%)	Accuracy (%)
Proposed ST-FDS	**96.38**	**100**	**98.16**	**98.13**
Proposed SA-FDS	90.52	99.99	95.01	94.76
[104]	92.67	99.79	96.10	95.95
[20]	91.92	99.71	95.66	95.50
[103]	89.24	99.37	94.03	93.62
[106]	88.30	98.75	93.23	92.85
[105]	95.88	92.06	93.93	94.08
[101]	89.48	99.79	94.35	94.05

Temporal Fire Detection System) with existing models is included in Table 3.4 and comparisons of the different model sizes are shown in Table 3.5. It is evident from Table 3.4 that the proposed ST-FDS comfortably outperforms all existing techniques.

Four evaluation metrics (i.e. *Accuracy*, *Precision*, *Recall*, and *F-Score*) in a manner similar to [20] are employed to evaluate the system performance. These metrics are calculated as shown in Equation (3.8).

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + TN + FP + FN} \\ Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ F - Score &= \frac{2 * Precision * Recall}{Precision + Recall} \end{aligned} \tag{3.8}$$

- TP: True Positive (Positive examples classified as positive).
- TN: True Negative (Negative examples classified as negative).
- FP: False Positive (Negative examples classified as positive).
- FN: False Negative (Positive example classified as negative).

Here TP and TN are correct classifications whereas FP and FN are mis-classifications. Accuracy indicates the percentage of correctly classified examples. Precision is a measure of

Table 3.5: Model size comparison of fire-detection techniques

Technique	Input Shape (WxH)	FLOPs (in millions)	Number of Parameters	Size on Disk (in KB)
Proposed ST-FDS	112x112	**8.21**	**18018**	**324**
Proposed SA-FDS	112x112	**8.21**	**18018**	**324**
[104]	224x224	530.98	723522	4560
[20]	224x224	612.73	2260546	12768
[103]	299x299	11469.10	5601954	21370
[106]	299x299	4705.67	986370	3982
[105]	64x64	17.81	646818	7631
[101]	64x64	23.75	30035	411

the classified positive examples that are actually positive whereas Recall is a measure of positive examples that are correctly classified. The F-Score is the harmonic mean of the Precision and Recall and is a good indicator even in the case of unbalanced data.

The proposed techniques perform well not just in terms of Accuracy but also other standard performance metrics like Precision, Recall, and F-score. ST-FDS has an Accuracy of 98.13% . It has a 100% Recall, which implies that all the fire images were classified correctly. It has a Precision of 96.38% indicating that only 3.6% non-fire images are classified as fire. The F-score is 98.16% which indicates an overall superior performance.

The results included in Table 3.5 demonstrate the superiority of the proposed models in terms of size, requiring markedly less space on disk. Also, the proposed models are the fastest among existing techniques as they perform the least number of FLOPs (Floating point Operations). The small size and speed of SA-FDS and ST-FDS makes them useful for constrained environments like those of a *Raspberry pi* or mobile phones commonly used in IoT deployments.

We now compare the proposed SA-FDS model with the SqueezeNet [142] architecture especially because SA-FDS is based on Squeezenet with certain modifications. Table 3.6 compares the performance of the two models. The table also compares the proposed ST-

Table 3.6: Comparison of proposed technique with SqueezeNet

Technique	Accuracy (%)	Parameters (K)	FLOPs (M)
Proposed ST-FDS	**98.13**	**18.02**	**8.21**
Proposed SA-FDS	94.76	**18.02**	**8.21**
SqueezeNet [142]	95.95	723.52	530.98
SqueezeNet with TmP	97.54	723.52	530.98

FDS with SqueezeNet augmented with temporal properties (in a manner that SA-FDS is combined with temporal properties to give ST-FDS). Here, SqueezeNet is trained for two class classifications with transfer learning to avoid over-fitting given the large network size.

Table 3.6 shows that the performance of Squeezenet is better than the proposed SA-FDS without taking the temporal properties into consideration. Subsequent to inclusion of the temporal properties, the proposed ST-FDS is marginally superior to SqueezeNet with temporal properties. The important point to note here is that SA-FDS and ST-FDS, while being comparable to SqueezeNet in classification efficacy, are much superior to the SqueezeNet architecture in terms of speed and size. This is shown in Table 3.6 by the Number of Flops (speed), and Number of Parameters (size), respectively.

3.4.3 Analysis of the System

The common approach to analyse algorithms in terms of computational complexity is by representing their time and space requirements asymptotically. Usually computational analysis is performed to determine the time and space requirements of an algorithm in the worst-case scenario. This is known as the worst case complexity and is denoted by the 'big oh' notation: $\mathcal{O}$.

In the case of CNN models, however, it is uncommon to do their asymptotic complexity analysis. The time complexity of a CNN model is estimated by the measure of computations it devours. The larger the computations, higher is the time complexity. In this chapter,

the time complexity of the proposed model is estimated by the number of FLOPs (Floating Point Operations) performed during inference. Similarly, the space complexity of the model is measured by the number of its parameters. The number of FLOPs performed and the number of parameters of the proposed models along with those of existing models are shown in Table 3.5.

A quick analysis of the performance, size, and speed of the proposed models when compared with existing state-of-the-art models is as follows: of the existing models, [20] and [104] perform the best in terms of Accuracy. The proposed ST-FDS comfortably outperforms the same while SA-FDS has Accuracy results close to those of [20] and [104]. The model in [101] is of the smallest size (on disk) and [105] is the fastest amongst the existing techniques in literature. Both SA-FDS, and ST-FDS are superior in terms of size requiring markedly less space on disk. Both models are also the best in terms of speed performing least number of FLOPs.

3.4.4 Deployment in a Resource Constrained Environment

The effectiveness of the proposed framework in terms of speed was tested by deploying it over a real world resource-constrained environment comprising a *Raspberry Pi 3B* device. *Raspberry Pi 3B* is a small, single board computer that is composed of a 1.2 GHz Quad core processor, and a 1 GB RAM. Both the proposed frameworks, SA-FDS and ST-FDS, were tested on a *Raspberry Pi* and were found to be light enough to run efficiently with an impressive frame rate. The Frame rate was calculated in two ways: 1) classification time only; and 2) frame reading time along with classification time. Table 3.7 shows the frame rates for the two models.

The classification time includes the time required for frame prepossessing that involves conversion to gray-scale, resizing to 112x112, and pixel normalization to the $[0 - 1]$ range

Table 3.7: Frame Rates on Raspberry Pi

Model	Classification Only	Frame Reading + Classification
SA-FDS	115 fps	66 fps
ST-FDS	72 fps	**50 fps**

along with the time taken by the model for actual classification. The frame reading time includes the time taken for extracting every 5^{th} frame of the video from the time that the video capturing process starts. The effective frame rate of the ST-FDS model is 50 frames/second which enables it to comfortably process the streaming video frames (a regular camera captures videos at a frame rates of 30 frames/second of which only 6 frames/second need to be processed in the ST-FDS model).

3.5 Summary of the chapter

In this chapter, a privacy-preserving system for fire detection using vision-based monitoring was proposed. The main contribution of this chapter is the development of a privacy-preserving, lightweight, and efficient fire detection system for indoor spaces. To ensure privacy, a strategically modified vision camera was utilized to capture videos followed by identification of the appropriate privacy level using user surveys. In addition, a lightweight and efficient fire detection system was developed to work with these privacy-preserving images without compromising accuracy. The use of a modified vision sensor, the privacy assessment and the combination of the motion characteristics of the fire flame with the shape of the flame for fire detection was a novel idea which was not explored in the literature.

Systematic surveys were conducted to assess the privacy requirements of people and an appropriate 'level' of images was identified that was able to preserve privacy whilst providing enough discrimination for accurate fire detection. In lieu of the absence of an

appropriate dataset, a new dataset was created and used to effectively validate the system. The proposed fire detection system ST-FDS based on both spatial and temporal properties of fire, was able to detect fire accurately and found to outperform existing state of the art techniques in both detection accuracy and model size. The proposed model was shown to be lightweight by running it efficiently on a resource constrained environment at an acceptable frame rate.

Chapter 4

Privacy-Preserving Human Activity Recognition System

4.1 Introduction

With a rapidly aging population and a depleting workforce, automated human activity recognition systems are slated to become the norm in societies worldwide. These systems facilitate, among other things, monitoring of the daily activities of the elderly, healthcare interventions, indoor surveillance, and public safety [7, 10, 11, 15, 19, 120]. The approach is to continuously monitor indoor locations and raise a flag on detecting an event out of the ordinary. The working of these systems are also based on two components: a sensing component; and an algorithmic component. The sensing component comprises a single or heterogeneous combination of sensing devices of varied kinds. Sensing devices employed for human activity recognition can broadly be categorized into three types: wearable sensors (accelerometers, gyroscopes, EEG sensors, GPS trackers) [11, 17]; ambient sensors (passive infrared sensors, pressure sensors, contact switches, radar sensors, wi-fi routers) [15, 16]; and vision sensors (color cameras, infrared cameras, depth cameras) [7, 10, 12, 19, 23, 62, 89, 109–111, 114, 115, 144].

As discussed in Chapter 1, ambient and wearable sensors have limitations related to accuracy and convenience, respectively. Vision sensors, on the other hand are privacy invasive and can not be used for monitoring private indoor spaces. To overcome these limitations, an effective solution for human activity recognition in private spaces is one that harnesses depth information obtained from depth sensors. Depth sensors work on the principle of an object's distance from the sensor. Depth sensors are, therefore, not sensitive to light and work well in dimly lit spaces. These are also relatively inexpensive compared to thermal infrared cameras. The images captured also appear distorted to the human eye and, therefore, serve the purpose of preserving privacy [4, 35].

With such advantages, several studies [23, 62, 109–111] employ depth data to recognize human activities effectively. These methods are either based on analyzing hand-crafted features extracted directly from depth frames [110, 111] or by creating a depth motion map (DMM) [23, 62, 109]. A DMM is an image corresponding to a depth video clip generated by computing the differences between consecutive frames of the clip. Systems based on processing individual frames of depth videos do an outstanding job of recognizing simple actions (like standing, running) but fail to recognize complex/similar-looking human activities (like drinking, eating etc.) owing to a lack of strong temporal correlation. DMM based methods possess temporal correlation but lack in terms of modeling speed variations and changes in the order of movements. They, therefore, suffer from intra-class variations. Another issue with DMMs is that two activities appear similar from one view (like the front view) and are totally different from another (like a side or top view). This potentially leads to incorrect classification when only one DMM is used. Furthermore, the data captured using depth sensors is noisy and prone to occlusion, degrading the activity recognition efficacy.

An interesting approach that utilizes depth videos for human activity recognition extracts skeleton sequence data from these videos. Real-time skeleton tracking algorithms [145]

extract such skeleton sequence data, which can unambiguously provide human presence and movement information. The easy availability of skeleton data led several researchers to work on skeleton based human activity recognition systems[7, 89, 114, 115].

Activity recognition using skeleton sequence data is effective to an extent but somewhat restricted by the absence of skeleton joint information, as most skeleton tracking algorithms fail to extract joint information correctly. Lack of joint information adversely affects recognition. Furthermore, skeleton sequence data based approaches using joint positions features [112] are susceptible to scale variation (i.e., each individual's height and size in the image can differ due to distance from the camera). Also, approaches using joint angles based features [89, 91] are susceptible to change in orientation but remain the same with varying scales. Studies in [12, 62] use a combination of depth and skeleton sequence data for better activity recognition. Still, these also suffer from a lack of temporal features in DMM and scale/orientation in skeleton sequences.

In this work, we use a combination of depth videos and skeleton sequences to develop a robust, cost-effective activity recognition system for private spaces. The datasets used in this work are captured using a depth camera embedded in the Microsoft Kinect device. The depth camera captures information in a scene using the distance of objects from itself and maps the same into gray-scale images. The use of depth data enables privacy preservation as the individual's identity is preserved. Use of raw depth video instead of DMMs in the analysis preserves the temporal aspects of the data. Additionally, we utilise joint information from skeleton sequence data and incorporate this in our analysis. To do this, we devise two novel descriptors: Joint Position Descriptor (JPD), which records the variations in the body joint positions with time; and Bone Angle Descriptor (BAD), which records the variations in the inclination of the bones with time.

In addition to using a combination of depth data and skeletal sequence data effectively

for activity recognition, the other significant contribution of this work is in using 3D CNN models for extracting spatial and temporal features from depth videos. The use of 3D CNN is limited in literature owing to the unavailability of large datasets of depth videos. We overcome this limitation by utilising transfer learning approaches to reuse pre-trained weights for models like ResNet [146] inspired 2D-CNN and I3D [147] inspired 3D-CNN. ResNet (Residual Network) is a popular 2D Convolutional Neural Network (CNN) used for image classification by applying 2D convolution filters successively on the image data. Similarly, I3D (Inflated 3D) is a popular 3D CNN used for video classification by applying 3D convolution filters on video data, with the third dimension being time. ResNet and I3D are extensively used in applications employing RGB color data.

Finally, a two-level fusion scheme is proposed: fusion of features extracted using the two descriptors (JPD and BAD) called Feature Level Fusion Strategy (FLFS). Subsequently, the classification scores using the depth data and skeletal data, respectively, are also fused, and this fusion is called Score Level Fusion Strategy (SLFS). Combining the two descriptors of skeleton data, JPD and BAD, makes the system robust and shields it from variations in scale and orientation of the human body. Further, combining the two modalities, depth and skeleton data, protects the system from noisy depth maps and missing skeleton joints.

Keeping the following factors in mind: privacy, convenience, cost, and robustness; the key contributions of this work include:

1. A robust system for activity monitoring using a two-stream CNN architecture is proposed for depth and skeleton data, respectively.

2. Two novel descriptors from skeleton sequence data, JPD and BAD, are proposed to model the movement of body parts irrespective of the scale and orientation of an individual.

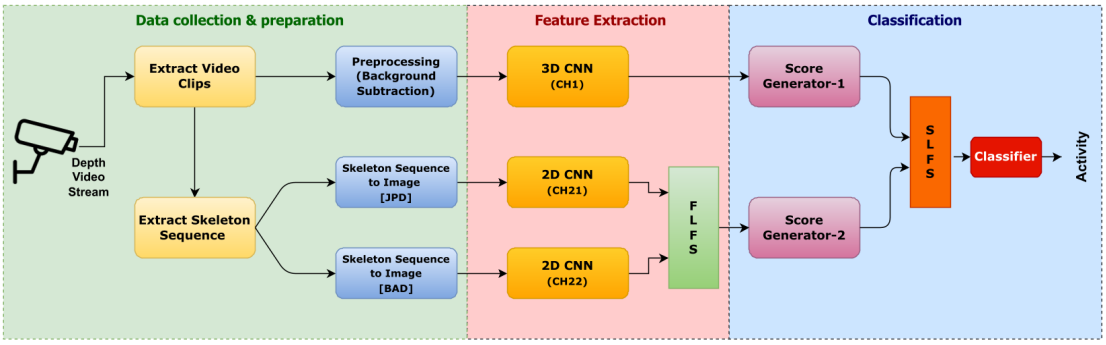


Figure 4.1: Workflow of the proposed Human Activity Recognition System

3. A two-level fusion strategy is proposed to effectively combine the different inputs.

4. Two modified deep CNN architectures are proposed to deal with the overfitting problem in the absence of large datasets.

5. A thorough evaluation of the proposed system is done on four public datasets, and it is shown to outperform most existing work. Also, computational complexity analysis and a prototypical implementation of the system suggest that it can work in real-time.

4.2 Proposed Methodology

In this section, we discuss the proposed approach for human activity recognition in detail. We use two data modalities for human action recognition: 1) depth data; and 2) skeleton sequence data. The first input (i.e., depth data) is used directly in its raw form after necessary pre-processing. The second input (i.e., skeleton sequence) is converted into two descriptors, namely JPD and BAD, followed by their mapping into color images. The pre-processing of depth data and the generation of the descriptors, JPD and BAD, are first discussed in this section. Subsequently, the proposed deep learning based framework comprising two primary channels (one for each data modality, depth, and skeleton sequence) is discussed. Finally, we discuss the fusion strategies adopted for combining the analyses of the two data

modalities. The schematic diagram of the proposed framework is included in Figure 4.1.

4.2.1 Data Preparation

4.2.1.1 Depth Data

In this work, we utilise depth data for each activity largely in its raw form, with minimum pre-processing. The minor pre-processing required includes: 1) background removal; and 2) temporal normalization. Video captured using a depth camera contains gray-scale values and often suffers from a phenomenon called 'depth camouflage' wherein the background and foreground have the same gray-scale values. Therefore, it becomes difficult to distinguish between the two, especially when the foreground is close to the background. To rectify this, we adopt the well-known 'background subtraction' approach [148] that removes the background from such frames following Equation (4.1). A background model $B(x, y, t)$ is first computed from the initial frames of the video in the absence of not-stationary foreground objects. Subsequently, each frame of the video is subtracted from the background model. This highlights the foreground objects by removing the stationary background details. A threshold is imposed that suppresses tiny noises arising due to the environmental changes and produces clear foreground objects in the background subtracted image. Post-processing exercises like morphological operations are undertaken if the resultant images are still noisy. Figure 4.2(a) shows a depth frame after background subtraction.

$$S(x, y, t) = \begin{cases} 1, & \text{if } |F(x, y, t) - B(x, y, t)| \geq \mathrm{T} \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

Where, $S(x, y, t)$ is the pixel after background subtraction, $F(x, y, t)$ is the frame at time t, $B(x, y, t)$ is the background model, and T is a threshold. Threshold T varies for each dataset, therefore we utilized Otsu method [149] to calculate appropriate value of the

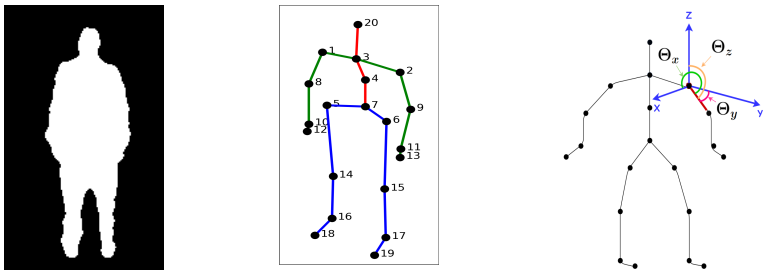
threshold T. The threshold value obtained using Otsu method for our dataset is T=40.

The second pre-processing task, temporal normalization, comprises converting varying length depth video clips to a fixed temporal length. The optimal temporal length of clips for a dataset depends on the lengths (i.e., time taken to perform an activity) of all the clips in the dataset and is close to the most frequent temporal length (M_f). This gives a rough idea of the lengths of most clips. To determine M_f, we categorize the lengths of clips as those falling in the intervals 0-10, 10-20, 20-30, and so on. We consider such intervals when the value of the range $R = 10$. Based on their lengths, the clips fall into one of these intervals. The number of clips that fall in the interval with the largest number of clips is M_f. The number of clips falling in the interval immediately preceding the interval with the largest number is M_p. The number of clips in the interval immediately succeeding the one with the largest number is M_s. These values are used with Equation (4.2), and we get an estimate of the optimal length (T_e) of the video clips in a dataset. This broadly gives us the statistical mode of the temporal lengths of the clips. In addition, there is a minor spatial normalization that the depth video clips are subjected to and undergo simple image resizing to a smaller 224×224 size. This is mainly done for easier accommodation in the framework and to reduce computational costs.

$$T_e = L + \left(\frac{M_f - M_p}{2M_f - M_p - M_s} \right) * R \tag{4.2}$$

where:

- T_e is the estimated temporal length for a dataset.
- M_f: Number of the most frequent temporal lengths in any interval.
- M_p, M_s: Number of temporal lengths preceding and succeeding the interval of M_f.



(a) (b) (c)

Figure 4.2: Depiction of Human Body: a) Depth image; b) Skeleton joints and bones; c) Bone angles.

- L: Lower limit of the interval with M_f.
- R: Range of intervals.

4.2.1.2 Skeleton Joint Data

Human activity recognition from videos containing skeleton joint information is based on two aspects: the spatial posture of the body in each frame of the video; and the movement of body parts in the temporal domain. The spatial and temporal information from the skeleton sequences are encoded in two descriptors: Joint Position Descriptor (JPD); and Bone Angle Descriptor (BAD), based respectively on joint positions and inclination of bones.

Datasets comprising skeleton joint information usually describe the human body with a certain number of joints. The datasets used in our experiments, in most cases, have 20 or 25 joints in the human body. Sample skeleton joints in the MSR Action3D dataset are shown in Figure 4.2(b).

In skeleton joint information, the body is usually divided into five parts: the spine, the two hands, and the two legs. Each body part is represented by a certain number of joints. In the MSR Action3D dataset, each body part has four joints. Human activities involve the movement of one or more body parts, and hence the movement of the joints of those parts. An activity is distinguished from another based on the dominant movement of one set of joints compared to others and vice-versa. The joint groupings for the MSR Action3D

dataset are as follows (Figure 4.2(b)): right-hand {1, 8, 10, 12}; left-hand {2, 9, 11, 13}; right-leg {5, 14, 16, 18}, left-leg {6, 15, 17, 19}, and spine {20, 3, 4, 7}.

A skeleton joint sequence comprises several frames of joint information. A joint sequence JS is represented as: $JS\{F_1, F_2, F_3, ...F_n\}$, where F_i is the i^{th} frame of the sequence. Each frame with m joints is represented as $F_i\{J_1, J_2, J_3, ...J_m\}$, where J_k is the k^{th} joint of F_i. Each joint is represented by its position $J(x, y, z)$ in the 3D Cartesian coordinate system. The Joint Position Descriptor (JPD) image is generated based on this information described in Algorithm (4.1). The movement of spine joints is minimal, which does not contribute much to activity classification. We, therefore, only consider joints of the hands and legs while generating the JPD.

The joint sequence corresponding to an activity is the input to Algorithm (4.1) along with the grouping of joints (i.e., joints of the left hand are grouped together, followed by the joints of the right hand, and so on). Joints in each frame are first arranged according to the grouping sequence to create a 2D matrix AJ containing three column vectors (one for each coordinate). The matrices AJ of consecutive frames are further concatenated to form an *action cube of position* (P), a matrix with dimensions: Width x Height x Coordinates ($W \times H \times C$). Figure 4.3 is a high-level depiction of the formation of the *action cube*.

The ranges of the three coordinates are usually quite different, with the large range dominating the smaller ones. To overcome this discrepancy, normalization of the *action cube of positions* (P) is done with respect to each coordinate (i.e., c = x, y, z) within the fixed range of [0-255]. Finally, the normalized *action cube of positions* (N) along three coordinates are concatenated to form JPD, which is mapped to an RGB image. The RGB image is further used for human activity recognition.

In addition to joint positions, the inclination of bones of the human body also encodes human activities' spatial and temporal dynamics. The hands and legs contain four bones

Algorithm 4.1 Joint Position Descriptor (JPD) based image generation.

Input : Skeleton Joint Sequence JS $\{F_1, F_2, ..., F_n\}$
Input : Joints grouping (JG) according to body parts
Output: RGB image (W x H x C)

P=EmptySequence
foreach `frame in JS` **do**
  AJ = *ArrageJoints*($frame$,JG)
  P=AppendColumn(P,AJ)

$N_c = 255 * \frac{P_c - \min\{P_c\}}{\max\{P_c\} - \min\{P_c\}}, \qquad c = 1, 2, 3$

JPD = $[N_{c=1} | N_{c=2} | N_{c=3}]$

I = ConvertToImage(JPD)

Procedure *ArrageJoints*(F, Seq)**:**
  AJ = $EmptyFrame$
  for `s in Seq` **do**
    AJ = AppendRow(AJ, $F[s]$)
  return AJ
End Procedure

```
// '|' is concatenation over 3rd axis
      // AppendRow(): method to concatenate m joints to create a column
vector.
  // AppendColumn(): method to concatenate n column vectors, where n is
number of frames.
// ConvertToImage(): method to map a 3D array into an RGB image.
```

each, of which three are more susceptible to movement, as shown in Figure 4.2(b). We select three bones, each from the two hands and legs, making a total of 12. The angles between each bone and the X, Y, and Z axes, namely, Θ_x, Θ_y, and Θ_z respectively, are calculated. Figure 4.2(c) shows the three angles of one of the bones.

The complete calculation of angles corresponding to each bone in each frame and their subsequent concatenation in the temporal direction to form the Bone Angle Descriptor (BAD) is described in Algorithm (4.2). The algorithm takes the joint sequence corresponding to an activity as input along with the sequence in which the bones are arranged in a

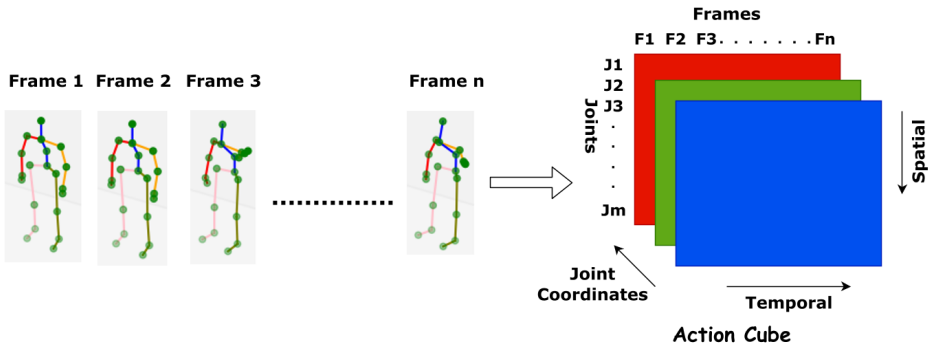


Figure 4.3: Image generation from the skeleton joint sequence of consecutive frames of an activity.

body part (e.g., bones $B_1(J_2, J_9)$, $B2(J_9, J_{11})$, and $B3(J_{11}, J_{13})$ represent the left hand, and similarly bones $B4$, $B5$, and $B6$ represent the right hand in the human body). The three angles for each bone, calculated against the three principle axes, are appended into a vector forming a 12×3 matrix (F). This process is repeated for each frame and concatenated one after the other to form an *action cube of angles* (I). The bones have a different range of angles for the three axes that lead to different features for the same activity when the human-camera angle changes. To overcome this, normalization of each angle is done in a fixed range of [0-255]. Finally, the normalized *action cube of angles* (N) along three axes are concatenated to form BAD. BAD is further mapped to an RGB image which is used for activity recognition.

The images generated from both the JPD and BAD have the following dimensions: the height of the image; the width of the image; and the depth of the image. The height of the image represents the number of joints in the image in the case of JPD and the number of bones in the case of BAD. The width of the images is representative of the number of frames generated from the video. Finally, the depth of the images is the number of coordinates used to depict the joint position in the case of JPD and the number of bone angles in the case of BAD.

For any dataset, the number of joints or bones is fixed, and therefore the height of the

Algorithm 4.2 Bone Angle Descriptor (BAD) based image generation

Input : Skeleton Joint Sequence JS $\{F_1, F_2, ..., F_n\}$
Input : Sequence of bones ($BONES$) according to body parts as tuple of joints (J_a, J_b)
Output: RGB image (W x H x C)

I=EmptySequence
foreach *frame in JS* **do**
  F=EmptyFrame
  foreach *bone in BONES* **do**
    $angles$ = *CalculateAngles(bone)*
    F= AppendRow(F,$angles$)
  end
  I=AppendColumn(I,F)
end

$N_c = 255 * \frac{I_c - \min\{I_c\}}{\max\{I_c\} - \min\{I_c\}}, \qquad c = 1, 2, 3$

BAD = [$N_{c=1}|N_{c=2}|N_{c=3}$]

I = *ConvertToImage(BAD)*

Procedure *CalculateAngles*(bn):
  AJ = $EmptyFrame$
  if *($bn[J_b]$!= 0) and ($bn[J_a]$!= 0)* **then**
    $b \leftarrow bn[J_b] - bn[J_a]$
  else
    $b \leftarrow 0$
  end
  b = b / $\|b\|$
  $\Theta_x = acos(b.AX)$
  $\Theta_y = acos(b.AY)$
  $\Theta_z = acos(b.AZ)$
  return [$\Theta_x, \Theta_y, \Theta_z$]
End Procedure

```
// AX, AY, and AZ are X, Y, and Z axes, respectively.
```

images is the same. The depth, as mentioned, depends on the number of coordinates/angles and is also fixed at 3. The width, however, is variable as the length of the video clips can vary, and correspondingly so do the number of frames. As a 2D CNN framework requires a fixed size, we make the images undergo an image resizing operation and fix the image size at

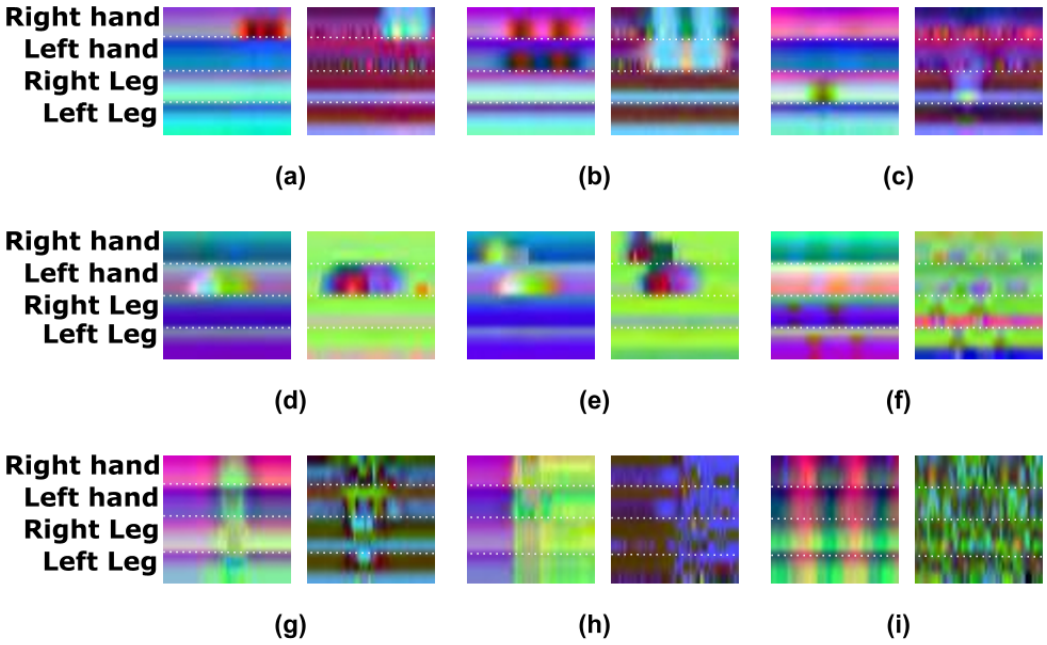
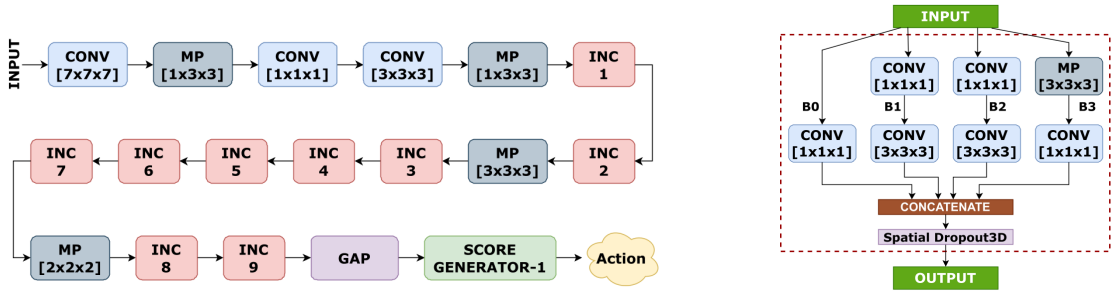


Figure 4.4: Sample JPD (left) and BAD (right) based images for various activities in the three datasets. MSR Action3D dataset (a-c); UTD-MHAD dataset (d-f); and TST Fall dataset(g-i). a) High arm wave; b) Hand clap; c) Forward kick; d) High arm throw; e) Tennis serve; f) Walk; g) Grasp object; h) Front fall; and i) Walk

$224 \times 224 \times 3$. Sample JPD and BAD of nine different activities from three datasets, MSR Action3D, UTD-MHAD, and TST Fall, are shown in Figure 4.4. The activities depicted include one hand movement (a & d); two hands movement (b & e); and leg movement (c & f); whereas activities (g-i) involve the movement of all body parts.

4.2.2 Human Activity Recognition System

As shown in Figure 4.1, the proposed method comprises two primary channels for human activity recognition: a 3D CNN based on I3D [147] and a 2D CNN based on ResNet [146].



(a) (b)

Figure 4.5: (a) Modified I3D based 3D-CNN architecture, (b) Modified Inflated Inception Module employed in our 3D-CNN

4.2.2.1 3D CNN

Inflated 3D (I3D) [147] is a popular 3D CNN architecture for video classification in the color domain. As the name suggests, I3D is based on inflated convolutional filters and pooling kernels (inflated filters are 3D filters obtained by adding an additional dimension to the usual 2D filters). The I3D architecture utilizes an Inception Module based design with inflated 3D filters to process video data. Figure 4.5(a) shows the complete I3D architecture. I3D takes an input video of size (Time (T) x Width (W) x Height (H)) as input and gives a activity category as output.

The original I3D architecture is modified in two ways: the last average pooling layer is replaced by a Global Average Pooling (GAP) layer that averages over the entire feature map instead of just over the size filters. Next, a module named *Score Generator-1* is added that generates scores corresponding to each activity. The *Score Generator-1* module contains a dense layer with *'A'* neurons, where *'A'* is the number of activity categories. A softmax layer is added at the end, giving a probability distribution for the *'A'* classes; next, a spatial dropout layer [150] is included in the 3D inception module after the concatenation layer. This helps reduce the dependence among values in the motion maps and reduces over-fitting by dropping the entire motion map with a drop probability of p_d. Equation (4.3) expresses the feed-forward process with a spatial dropout. The structure of the modified

inflated inception module is shown in Figure 4.5(b).

$$X^l = \Phi\left(\sum_{c=1}^{C} K_c^l * (X_c^{l-1}.m_c^l)\right) \tag{4.3}$$

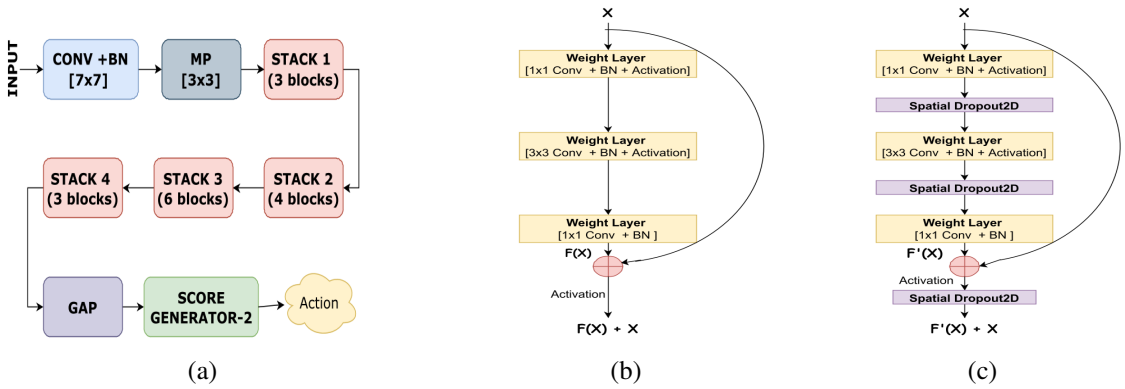
Where, X^{l-1} and X^l are the outputs of the previous and current layers, respectively. m_c^l is Bernoulli's random variable that takes a value of '0' or '1' depending on the drop probability p_d; whenever $m_c^l = 0$, the entire feature map becomes '0' and does not contribute to the weight updation of the model. Other feature maps where $m_c^l = 1$ update the weights in that iteration. This process repeats in every iteration during training and avoids over-fitting. $K_c^l \in \mathrm{R}^{K_w \times K_h \times K_T}$ is the convolution kernel and Φ is the non-linearity activation function. '.' denotes scalar multiplication, and '*' denotes the correlation operation using the given filter.

Unlike standard dropout, which drops the input features randomly without considering correlations between nearby pixels, spatial dropout drops either the entire feature map or none. This results in adjacent values in the feature map that are either all '0' or all active. The spatial dropout at the input layer helps handle noise or missing data. Through experiments, it was observed that spatial dropout was more effective at higher layers and hence used in the last three inception modules and in the input layer.

4.2.2.2 2D CNN

ResNet [146] is a reasonably successful CNN architecture for image classification. It comprises a modularized architecture composed of residual blocks and skip connections that help solve the vanishing gradient problem in deep CNN architectures. The ResNet50 architecture and the structure of single residual module is shown in Figure 4.6(a) and Figure 4.6(b), respectively.

In our work, the original ResNet50 architecture is modified in two ways: the residual



(a) (b) (c)

Figure 4.6: (a) Modified ResNet50 based 2D-CNN architecture; (b) Residual Block used in original ResNet50; (c) Modified Residual Block employed in our 2D-CNN

block is first modified by inserting a spatial dropout layer after each processing layer (i.e., Conv + BN + Activation). Figure 4.6(c) shows the modified residual block with the bottleneck design. As mentioned earlier, spatial dropout helps deal with data independence and over-fitting at higher layers. We use modified residual blocks in the last two stacks. As mentioned, spatial dropout is added to the input layer to deal with missing or noisy data. In addition to this, spatial dropout also augments the data by randomly removing a feature map at each iteration. The second modification to the ResNet50 architecture involves the removal of the dense layer at the end and its use in feature extraction. After feature fusion (discussed in the next section), a module named *'Score Generator-2'* is added to generate scores from fused feature vectors. *'Score Generator-2'* includes three dense layers with 512, 128, and *'A'* neurons respectively. A softmax layer is added at the end to generate probability scores corresponding to each activity *'A'*.

4.2.3 Fusion Strategies

As discussed earlier, the depth and skeleton data modalities are effective and mostly give precise results. They do, however, suffer from shortcomings and are occasionally erroneous. We seek to harness the redundancy advantage that the availability of two modalities offers

and compensate for the errors of one with the correctness of the other and vice-versa. Also, the features based on only joint positions often suffer from scale variation. We adopt a two-level fusion approach and combine the classification results of the channels.

4.2.3.1 Feature Level Fusion Strategy

The Feature Level Fusion Strategy (FLFS) combines the features extracted from the skeleton data's sub-channels (CH21 and CH22) for better performance. In this strategy, features are combined in the following four ways: 1) element-wise average (FLFS-avg); 2) element-wise product (FLFS-prod); 3) concatenation of two feature vectors (FLFS-cc); and 4) element-wise maximum (FLFS-mx). In FLFS, features from sub-channels CH21 and CH22 are combined following Equation (4.4) after the Global Average Pooling layer.

$$
\begin{aligned}
X_{FLFS-avg_i} &= \frac{X_{JPD_i} + X_{BAD_i}}{2} \\
X_{FLFS-prod_i} &= X_{JPD_i} * X_{BAD_i} \\
X_{FLFS-mx_i} &= max(X_{JPD_i}, X_{BAD_i}) \\
X_{FLFS-cc} &= (X_{JPD_1}, X_{JPD_2}, ..., X_{JPD_l}, \\
&\qquad X_{BAD_1}, X_{BAD_2}, ..., X_{BAD_l})
\end{aligned}
\tag{4.4}
$$

Where $X_{JPD} \in R^l$ is the vector comprising JPD based features; $X_{BAD} \in R^l$ is the vector of BAD based features; $X_{FLFS-avg} \in R^l$, $X_{FLFS-prod} \in R^l$, $X_{FLFS-mx} \in R^l$, and $X_{FLFS-cc} \in R^{2l}$ are respectively the fused feature vectors following the four methods mentioned above. Feature fusion in $X_{FLFS-avg}$ is done by calculating the element-wise average of the values in the two feature vectors. A similar approach is used in $X_{FLFS-prod}$ and $X_{FLFS-mx}$ to find the element-wise product and maximum. In $X_{FLFS-cc}$, two input feature vectors are appended, resulting in a larger vector. The fused feature vectors are fed to the *Score Generator-2* to generate scores for each activity.

4.2.3.2 Score Level Fusion Strategy

In Score Level Fusion Strategy (SLFS), the score vectors of the two Score Generators, *Score Generator-1* and *Score Generator-2*, are combined. The softmax function in each of the Score Generators generates a vector S$\{s_1, s_2, ..., s_A\}$ that comprises the probability distributions of the various activities according to Equation (4.5).

$$S_j = \frac{\exp\left({F_j}^2\right)}{\sum_{j=1}^{A} \exp\left({F_j}^2\right)}, j = 1, 2, ..., A. \tag{4.5}$$

Where F_j is the j^{th} value in the output of the last dense layer of the score generator; S_j is the probability of class j; and A denotes the total number of activity categories. For the skeleton data channel (CH2), *Score Generator-2* takes the feature vector X_{FLFS} as an input, and processes it in the dense layers, thus leading to a feature vector F with a size equal to the number of activities 'A'. A softmax layer, at the end, takes F as an input and generates the probability score vector S containing the score of each activity j as per Equation (4.5). Similarly, *Score Generator-1* takes the feature vector obtained from the 3D-CNN and generates the score vector S for the depth data channel (CH1).

Score level fusion happens in the following six ways: 1) weighted sum of scores (SLFS-ws); 2) weighted product of scores (SLFS-wp); 3) maximum of scores (SLFS-mx); 4) logistic regression (SLFS-lr); 5) Random Forest (SLFS-rf); and 6) Naive Bayes (SLFS-nb). The first three are based on simple score fusion operations described in Equation (4.6), whereas the latter three utilise supervised learning algorithms on concatenated score vectors.

$$
\begin{aligned}
X_{SLFS-ws} &= \sum_{m=1,2} W_m * S_m \\
X_{SLFS-wp} &= \prod_{m=1,2} S_m^{W_m} \\
X_{SLFS-mx} &= \forall i \; max_i\big(||_m(S_m)\big) \\
s.t. \sum_{m=1,2} W_m &= 1
\end{aligned}
\tag{4.6}
$$

Where, $SLFS$'s are the fused score vectors; S_m's are the score vectors from each trait; W_m's are the fusion weights assigned to each classifier that are tuned empirically. In $X_{SLFS-ws}$ and $X_{SLFS-wp}$, the element-wise weighted sum and the weighted product of the score vectors from the depth channel (CH1) and the skeleton (CH2) channel are calculated. Similarly, the element-wise maximum of the corresponding values in the two score vectors is calculated in $X_{SLFS-mx}$. The class label is determined by the index of the maximum value in the fused score vector from Equation (4.6).

4.3 Experimental Evaluation

The proposed human activity recognition system is evaluated by training and testing it with popular and publicly available datasets comprising depth and skeleton data. The system's performance is subsequently compared with existing state-of-the-art techniques and shown to be superior. Four public datasets are harnessed in our experiments: MSR Action3D[110]; UTD-MHAD[151]; TST Fall V2[152]; and MSR Daily Activity[118]. The first two datasets comprise general activities, the third dataset contains data on fall activities, and the fourth dataset includes data on daily activities. The third and fourth datasets are specific to indoor activities and are related to elderly care applications. A detailed description and a few visual examples from each of the above datasets are included in the Appendix (B).

4.3.1 Experimental Setup

In our experiments, a combination of a 2D-CNN and a 3D-CNN is used for feature extraction from skeleton and depth based inputs. For the 2D-CNN, pre-trained weights from the well-known Imagenet dataset are used, comprising over 14 million annotated color images distributed across 1000 categories. The approach for reusing pre-trained weights, known as transfer learning, is adopted. This is followed by fine-tuning the relatively small set of images generated using algorithms (4.1) and (4.2). For the 3D-CNN, pre-trained weights from a large-scale Kinetics dataset containing 65000 annotated video clips of around 400 activities are used. Here also, transfer learning is adopted with fine-tuning with the pre-processed depth clips, as discussed in Section III(A.1).

The three chosen datasets [110, 118, 151] are already clean and do not have a background. The fourth dataset [152] does include background information in the depth maps, but these are used as it is. The only pre-processing, therefore, for the depth video clips needed is temporal and spatial normalization. The optimal values of the parameter, temporal length (T_e), for each dataset are estimated using Equation (4.2), and their effectiveness is verified empirically. The length of activities in the MSR DailyAct dataset and the TST Fall dataset is large, and the activities are performed multiple times in a depth clip. For this reason, every 3rd and 4th frame, respectively, are used for these two datasets. T_i is the input temporal length used in 3D CNN after frame selection. The estimated temporal lengths (T_e) and the input temporal lengths (T_i) for different datasets are included in Table 4.1.

The approach taken is data augmentation to further address the issue of over-fitting, especially with less training data. Image data augmentation commonly involves rotation, translation, scaling, brightness, cropping, flipping, shifting, etc. As images in this work are generated from the skeleton joint information where each joint/angle is mapped to a pixel, the techniques mentioned are not helpful as they alter the pixel position. The augmentation

Table 4.1: Optimal Values of the Temporal Lengths and Score Fusion Weights in Various Datasets.

Dataset	T_e	T_i	**Weights** (W_1, W_2)
MSR Action	38	40	(0.43,0.57)
UTD MHAD	68	68	(0.68,0.32)
MSR DailyAct	194	65	(0.41,0.59)
TST Fall v2	144	36	(0.46,0.54)

T_e: Estimated Temporal Length; T_i: Input Temporal Length used in the 3DCNN.

approach adopted, therefore, is width-shift augmentation that varies an activity's start and end time. Similarly, for video clips, the augmentation approaches include *speed sampling* that changes the speed of performing activities; *random temporal cropping* that varies the starting point of activities; *random rotation*, *random translation*, *random resizing*, *spatial random cropping*, and *horizontal flipping* for camera angle, human position, distance from the camera, and left/right-handed activities.

The learning rate is initialized at 0.001 and decreased several times with a decay of 0.004 until it becomes $1e{-}4$. The training takes different epochs for the three channels and subsequent fusions. A weight decay of 0.001 with the L_2 regularizer is used, and the training is done using the ADAM optimization algorithm. During training, the batch sizes respectively for 2D-CNN and 3D-CNN are 32 and 16. The framework in Figure 4.1 is implemented, trained, and tested using Keras with a TensorFlow backend on a PC with Ubuntu 18.04 and 16GB Tesla V100 GPU. To improve the reliability of the results, all the fusion experiments are conducted thrice, and the average performance is reported.

4.3.2 Model Selection and Ablation Study

To select the most suitable 2D-CNN architecture for our work, we experimented with eight well-known models with transfer learning and fine-tuning, as discussed earlier.

Table 4.2: Impact of Input Temporal Length (T_i) on the Performance of 3DCNN.

Length(T_i)	24	32	40	64	96
Accuracy(%)	87.64	92.00	**93.45**	92.73	92.73

ResNet50 gave the best results of these eight models with the two descriptors and hence was selected for our 2D-CNN. Similarly, for appropriately configuring our 3D-CNN, we experimented with different temporal lengths and augmentation methods. Although the temporal length is estimated by Equation (4.2) and is validated in Table 4.2. The best results were obtained with a certain amount of data augmentation. Augmentation beyond this point resulted in the degradation of results. Details related to our choice of models for both 2D-CNN and 3D-CNN are included in the Appendix (B).

The weights used in SLFS as given in Equation (4.6) are obtained empirically using grid search on the training data for each dataset. These are shown in Table 4.1. Numeric data augmentation is employed on the scores for better generalization, where 10% noise is randomly imputed (added/subtracted) in the training data scores.

We conducted an ablation study to analyze the effectiveness of the spatial dropout module. The experiments were conducted on the MSR Action3D dataset. Including spatial dropout modules in both 2D-CNN and 3D-CNN leads to an improvement in performance while avoiding overfitting. The accuracy improved from 92.00% to 93.61% in the JPD based channel (CH21), from 85.45% to 88.73% in the BAD based channel (CH22), and from 92.53% to 94.21% after FLFS (CH2). In the depth based channel (CH1), the accuracy improved from 93.45% to 94.18%. The overall performance improvement was from 96.72% to 98.18% when spatial dropout was used.

Table 4.3: Classification Performance of the Individual and Fused Streams on Four Public Datasets.

Streams	MSR Action3D	MSR Action3D (ES2)				UTD-MHAD	TST Fall	MSR DailyAct
	(ES-1)	AS1	AS2	AS3	Avg	(CS)	(CS)	(CS)
CH1	94.18	94.33	95.57	99.10	96.33	93.26	98.33	78.75
CH21	93.61	87.73	93.75	92.85	91.44	95.34	94.16	73.12
CH22	88.73	87.73	91.96	89.29	89.66	94.42	89.99	65.63
FLFS-cc	**94.21**	90.88	96.13	93.45	93.49	97.52	95.00	**76.25**
FLFS-avg	93.60	**91.51**	**96.43**	**93.75**	**93.90**	**98.32**	**95.00**	74.99
FLFS-prod	93.55	88.68	94.64	92.26	91.86	95.58	93.89	72.29
FLFS-mx	92.88	90.25	95.53	92.26	92.68	96.74	94.44	73.33
FLFS (best)	94.21	91.51	96.43	93.75	93.90	98.32	95.00	76.25
SLFS-ws	97.45	99.06	96.46	99.06	98.19	98.60	100.00	81.87
SLFS-wp	**98.18**	**99.06**	**96.46**	**99.10**	**98.21**	**98.83**	**100.00**	**81.87**
SLFS-mx	95.67	94.34	92.03	99.10	95.16	93.02	99.16	81.25
SLFS-lr	96.00	98.11	96.46	99.10	97.89	97.20	100.00	80.00
SLFS-rf	94.90	95.28	95.57	95.53	95.46	96.51	99.16	78.75
SLFS-nb	96.00	98.11	96.46	99.10	97.89	96.51	100.00	79.37
SLFS (best)	98.18	99.06	96.46	99.10	98.21	98.83	100.00	81.87

4.3.3 Performance Evaluation

A concise summary of the classification performance of all the channels and fusions on the four datasets is included in Table 4.3. The first three rows provide information on classification based on depth data, classification based on skeleton data using JPD, and classification based on skeleton data using BAD. The subsequent four rows provide information on the performance of FLFS using different methods, with the best FLFS performance included in the following row. Similarly, the performance of SLFS, using six different methods, is also included, followed by the best overall accuracy.

Of the various methods used with FLFS, the average operation method performs well in most cases, along with the concatenate operation. The average method performs best as it is a linear operation where the gradient flows nicely, leading to the better tuning the training weights considering both inputs. The concatenate operation, on the other hand, works on any kind of input irrespective of the possible correlation. With SLFS, the weighted product

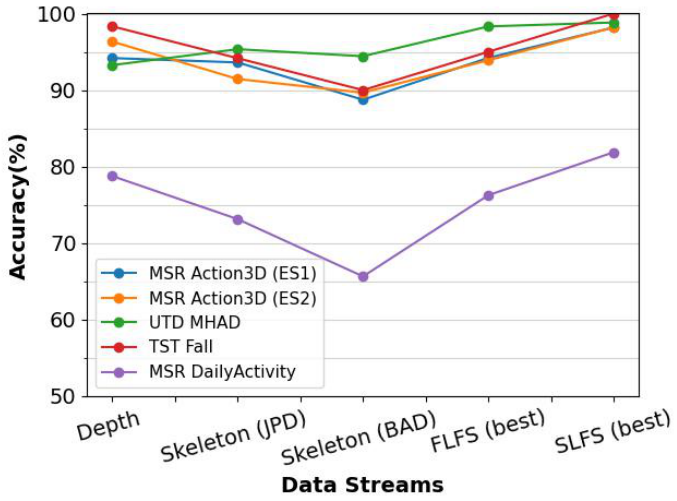


Figure 4.7: Effect of multilevel fusion on the overall performance on different datasets

operation performs best in all cases. This is because it maximizes the fused score of the correct classes (and the misclassified correct classes in one of the traits). Figure 4.7 shows the effect of multilevel fusion depicting the performance improvement.

For comparison with state-of-the-art, we only consider existing work using either skeleton or depth data. Techniques that use RGB color data or sensor data are excluded as these are privacy-invasive and discomforting. Also, we do not consider literature where the evaluation settings are not properly described or are inconsistent.

4.3.4 Results on MSR Action3D dataset

The MSR Action3D dataset comprises depth maps and skeleton sequences of 567 action instances of 20 action categories. The MSR Action3D dataset is widely used in literature for evaluation in the following three settings: 1) cross-subject evaluation on the entire dataset [109]; 2) cross-subject evaluation on three subsets of the data (i.e., AS1, AS2, AS3) [110]; and 3) evaluation by randomly dividing the dataset into training and test set [110]. In this work, evaluation settings (1) and (2) (namely ES-1 and ES-2, respectively) are used for assessing the proposed system. Setting (2) is especially popular and has been used in most

Table 4.4: Performance Comparison on MSR Action3D Dataset

Technique	**Year**	**Data**	**Method**	**Acc (%)**
Nunez et al. [24]	2018	S	DL	95.70
Qi et al. [115]	2018	S	ML	86.81
Huynh-The et al. [114]	2019	S	DL	97.90
Liu and Zhao [89]	2020	S	ML	94.60
Sima et al. [91]	2022	S	ML	93.68
Zhang et al. [29]	2022	S	DL	94.81
Li et al. [110]	2010	D	ML	74.70
Farooq et al. [109]	2018	D	ML	96.50
Weiyao et al. [23]	2019	D	ML	98.20
Trelinski and Kwolek [111]	2021	D	DL	95.64
Bulbul et al. [88]	2022	D	ML	93.00
Ji et al. [116]	2018	D+S	ML	90.30
Kamel et al. [12]	2018	D+S	DL	94.51
Chao et al. [63]	2020	D+S	ML	91.58
Li et al. [62]	2021	D+S	ML	95.60
Proposed (IncludingAll Joints)	2022	D+S	DL	98.21
Proposed (ExcludingErroneous Joints)	2022	D+S	DL	98.16

D: Depth data; S: Skeleton data ; D+S: Both Depth and Skeleton data.
ML: Use Machine Learning Algorithm; DL: Use Deep Learning Algorithm

recent endeavors. Both these settings are cross-subject evaluations where odd subjects are used for training and even subjects for testing.

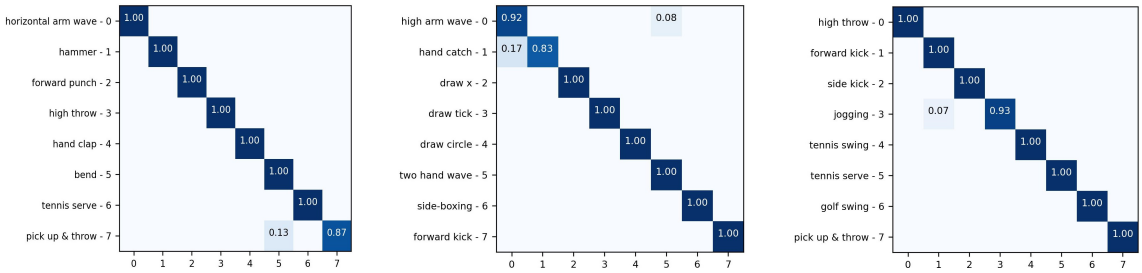
Table 4.4 compares the proposed framework's performance with existing state-of-the-art techniques. The MSR Action3D dataset has missing joints in a few of its sequences. Therefore, most existing skeleton based monitoring methods ignore these erroneous sequences (that are 10 in number) and use 557 out of the total 567 sequences. Our model, on the other hand, utilizes all 567 sequences and gives good results. This displays its robustness even in the face of missing data. Therefore, the accuracy of the skeleton based channel of the proposed framework is 93.90% as given in FLFS(best) row of Table 4.3 which is a

little less than that of existing methods, but it makes up for it through fusion strategies and reports an overall accuracy of 98.21%. The accuracy using data with only non-erroneous joints (i.e., 557) is also included in Table 4.4 to show fair comparison with the existing methods those excluded erroneous joints. The best accuracies using skeleton data, depth data, and a combination of both, reported in the literature, are 97.90% [114], 98.20%[23], and 95.60%[62], respectively. The proposed framework outperforms these techniques by a comfortable margin.

Skeleton data based approaches in [89, 91, 115] extract hand-crafted features and utilize machine learning algorithms. Their performance is mostly inferior to the proposed method, which uses deep learning. The approach in [24] extracts features from each frame separately and combines them in the temporal dimension using LSTM. Although LSTMs are meant for modeling sequential data, they have trouble recalling details about lengthy sequences with a large number of time steps. [114] uses a combination of positions and angles of joints to model an activity but suffers from missing joint data in practical use-case scenarios. In contrast, the proposed approach uses depth data and is able to effectively handle missing joints.

Depth data based approaches [110] exhibit weak temporal correlation due to the concatenation of individual frame features. DMM [23, 88, 109], on the other hand, lacks modeling speed and order variation. In [111], features extracted from each depth frame are individually combined using an ensemble of a multi-channel 1D-CNN and Dynamic Time Warping (DTW). The complexity of DTW limits its use, and it is cannot be used for large datasets with lengthy sequences. The proposed approach uses 3D-CNN, which processes the temporal information through the whole network. Moreover, only depth data based methods experience degraded performance due to noise and occlusion in depth maps.

Approaches using both depth and skeleton data [12, 62, 63] mostly generate DMMs from

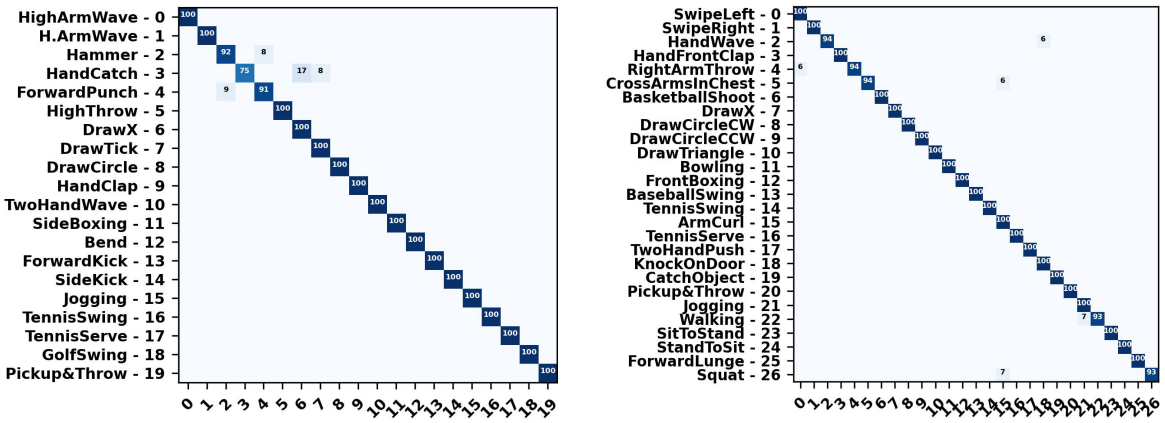


(a) (b) (c)

Figure 4.8: Confusion matrices for three action subsets (AS) of MSR Action3D dataset using Evaluation Setting (2): a) AS1; b) AS2; c) AS3

depth data which has inherent drawbacks, as discussed earlier. [12] and [62] use features based on the spherical coordinate of joints and are susceptible to viewpoint and rotation. Approaches in [63, 89, 116] use hand-crafted features and machine learning algorithms. In contrast, the proposed approach models temporal features in a better way using 3D-CNN and utilizes joint positions and bone inclinations from skeleton joints to make the system robust and unaffected by scale and viewpoint variations. Although, the methods in [114] and [23] have comparable performance to the proposed method on this dataset but has a huge difference in the performance on UTD-MHAD dataset.

It is evident from Table 4.3 that the accuracies of the FLFS operations are superior to those of individual skeleton based channels. The accuracy of the SLFS operations is also better than the combinations of skeleton and depth channels. Two operations, SLFS-mx and SLFS-rf, give accuracy slightly inferior but still better than those of depth and skeleton channels. The confusion matrices in Figure 4.8(a-c) show the precision of each activity in (AS1, AS2, and AS3). The matrices in Figure 4.9(a) show the precision of activities in the entire dataset. Only a few instances of activities are misclassified. These include $PickUp\&Throw$, which is confused with $Bend$ in AS1, $HandCatch$, which is confused with $HighArmWave$ in AS2, and $Jogging$ is confused with $ForwardKick$ owing to somewhat similar movement of body parts.



(a) (b)

Figure 4.9: Confusion matrices: a) MSR Action3D dataset (ES1); b) UTD MHAD dataset

4.3.5 Results on UTD-MHAD Dataset

The UTD MHAD dataset is captured along four data modalities: RGB clips, depth maps, skeleton sequences, and sensor readings. The dataset comprises 861 instances of 27 action categories. This work uses the evaluation setting followed in [151], where odd subjects are used for training and even subjects for testing. The UTD-MHAD dataset is comparatively large and includes more activities. The depth clips in this dataset are clean, and the skeleton sequences are less noisy. Therefore the performance of individual skeleton based channels is much better, as seen in Table 4.3. Integrating JPD and BAD in FLFS significantly improves performance even when using just skeleton data.

A comparison of the proposed framework with the state-of-the-art is included in Table 4.5, and the former comfortably outperforms the latter. A skeleton based method in [96] that uses multi-stream CNN with decision-level fusion has a performance that is somewhat close to the proposed. The proposed framework is, however, more robust due to the integration of the depth data with skeleton sequences. The approach in [91] suffers from temporal correlation due to concatenation and LSTM. Another approach in [29] represents skeleton joints as a large graph and uses a Graph Convolution Network (GCN) that is sensitive

Table 4.5: Performance Comparison on UTD-MHAD dataset.

Technique	**Year**	**Data**	**Method**	**Acc (%)**
Huynh-The et al. [114]	2019	S	DL	90.90
Banerjee et al. [96]	2020	S	DL	97.91
Sima et al. [91]	2022	S	ML	86.37
Zhang et al. [29]	2022	S	DL	94.19
Weiyao et al. [23]	2019	D	ML	88.70
Trelinski and Kwolek [111]	2021	D	DL	88.14
Bulbul et al. [88]	2022	D	ML	93.30
Kamel et al. [12]	2018	D+S	DL	88.14
Chao et al. [63]	2020	D+S	ML	89.53
Li et al. [62]	2021	D+S	ML	94.20
Proposed	2022	D+S	DL	98.83

D: Use Depth data only; S: Use Skeleton data only; D+S: Use Depth and Skeleton data.
ML: Use Machine Learning Algorithm; DL: Use Deep Learning Algorithm

to noise and small variations in joints. The proposed approach exhibits a strong temporal correlation in skeleton data by combining frames in an image and is comparatively less sensitive to noise.

The proposed approach is significantly better in terms of performance when compared to existing depth data based approaches [23, 88, 111] due mainly to the strong temporal correlation provided by the 3D-CNN, and the fusion of data modalities. Approaches [12, 62, 63] that utilise both depth and skeleton data are also notably inferior to the proposed approach. This is largely owing to weak temporal correlations in DMM and the use of only one of the two features: joint position or joint angle. In contrast, the proposed approach utilizes both the joint position and the novel bone inclination based feature to represent even a small movement of the bones effectively.

The confusion matrix in Figure 4.9(b) shows the precision of each activity. There are a few misclassifications here as well, owing to similarities in action, such as

$RightArmThrow$ being confused with $SwipeLeft$, $Squat$ being confused with $ArmCurl$, and $Walking$ being confused with $Jogging$.

4.3.6 Results on TST Fall V2 dataset

The TST fall detection dataset is specific to indoor activities, and the fall instances in this dataset seem useful for studying elderly care. The dataset is generated along three data modalities: depth maps, skeleton sequences, and acceleration data. The dataset comprises 264 action instances depicting 8 actions. The actions in the dataset are grouped into two categories: Falls and Activities of Daily Living (ADL). This work also uses a cross-subject evaluation setting similar to [112]. Table 4.6 compares the proposed work with the existing ones. A few existing techniques in literature classify Fall vs. ADL (i.e., 2 class classification) while others classify all the 8 activities (i.e., 8 class classification). We compare our work with both techniques using the 8 class classification, where our approach comfortably outperforms existing techniques. All the activities are correctly classified in the test dataset.

Table 4.6: Performance Comparison on TST Fall Dataset.

Technique	**Year**	**Class**	**Method**	**Acc (%)**
Ghojogh et al. [113]	2017	2	ML	90.15
Xu and Zhou [153]	2018	2	DL	95.84
Hristov [120]	2021	2	DL	91.00
Maldonado et al. [112]	2022	2	ML	92.20
Ghojogh et al. [113]	2017	8	ML	88.64
Ghodsi et al. [117]	2018	8	ML	92.30
Akyash et al. [119]	2020	8	ML	98.80
Yin et al. [7]	2021	8	DL	93.90
Proposed	2022	8	DL	100

ML: Use Machine Learning Algorithm; DL: Use Deep Learning Algorithm.
2: Binary classification (ADL vs Fall); 8: Eight class classification.

The existing approaches utilize only the skeleton joint data from this dataset as the depth frames contain rich background information and are unsuitable for methods like DMM. As 3D-CNN is adept at extracting useful information from images with a background, we were able to utilize depth data as well. Approaches in [112, 113] extract features from each frame, concatenate the same in the temporal dimension and classify using machine learning algorithms. Approaches in [7, 120, 153] harness LSTM for temporal modeling, leading to inferior performance compared to the proposed approach. Approaches in [117, 119] utilize dynamic time warping for temporal modeling. All approaches use raw skeleton joints and are susceptible to variations in viewpoint and scale and the variations arising from missing joints.

Table 4.3 includes comparisons of accuracies achieved with individual modalities and fusion strategies on the TST Fall detection dataset. The performance is 95% when tested on the skeleton data alone (after FLFS) and 98.33% using depth data alone. Most methods in SLFS return 100% accuracy except SLFS-mx and SLFS-rf.

4.3.7 Results on MSR DailyActivity Dataset

As the name suggests, the MSR Daily Activity dataset contains 320 instances of 16 daily routine activities performed in each sitting and standing position. In this work, the evaluation setting followed in [118] is used. Herein the evaluations are done in a cross-subject (CS) manner while considering all activities.

The MSR Daily Activity dataset contains heavy occlusions in the depth clips as the actors often sit or stand in front of the sofa. This results in very noisy skeleton sequences owing to erroneous estimations by the tracking algorithm. Also, the fact that the actors are in both sitting and standing positions makes this dataset more complex than others. Due to these reasons, like other approaches, our model is also imprecise on this dataset. Another

Table 4.7: Performance Comparison MSR Daily Activity Dataset.

Technique	**Year**	**Data**	**Method**	**Acc(%)**
Nunez et al. [24]	2018	S	DL	63.10
Qi et al. [115]	2018	S	ML	68.75
Reily et al. [90]	2020	S	ML	82.00 [1]
Liu and Zhao [89]	2020	S	ML	91.20
Debnath et al. [99]	2021	S	DL	76.30 [1]
Farooq et al. [109]	2018	D	ML	76.30
Wang et al. [118]	2012	D+S	ML	85.75
Ji et al. [116]	2018	D+S	ML	81.30
Proposed	2022	D+S	DL	81.87

D: Use Depth data only; S: Use Skeleton data only; D+S: Use Depth and Skeleton data.
ML: Use Machine Learning Algorithm; DL: Use Deep Learning Algorithm

crucial factor behind the model's poor performance is the human-object interactions in this dataset. Actions like eating and drinking; or using the laptop, reading a book, and just sitting still are semantically similar unless the objects (i.e., plate or glass; laptop or book) are considered. In literature, too, this dataset performs well when used for RGB data (alone or in combination with depth/skeleton data) or the objects are taken into account. Despite this, we validate our proposed framework on this dataset using skeleton and depth data only without object consideration.

Table 4.7 compares the proposed framework with the existing ones. The former mostly outperforms the existing techniques except [89] and [118]. In skeleton based methods, [89] uses the angle features of the bones by representing joints in Riemannian Geometry which is well suited for curved spaces. This work uses an additional pre-processing (i.e., interpolation technique) on skeleton joints to reduce noise, which leads to better performance. [118], on the other hand, Local Occupancy Patterns (LOP) from the point cloud data of depth maps to represent objects involved in each activity led to improved performance. The proposed

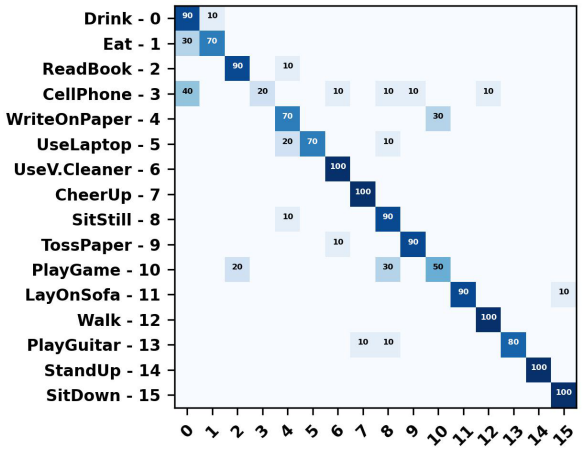


Figure 4.10: Confusion matrices of MSR Daily Activity Dataset

approach significantly improves from [24, 115], which combines features from individual frames in the temporal domain. Approaches in [109] and [116] utilizes machine learning algorithms on hand-crafted features, resulting in inferior performance. Approaches in [90, 99] utilize RGB data along with skeleton joints; the results shown here only use skeleton data.

Table 4.3 includes comparisons between various input channels and fusions. As skeleton sequences are pretty noisy and irregular, the framework's performance on both the skeleton based channels is poor. The FLFS, however, somewhat improves the performance by combining the joint positions and bone inclinations. SLFS further improves the performance as compared to the depth and skeleton traits. The confusion matrix in Figure 4.10 shows the precision of each activity. This dataset observes several misclassifications owing to human-object interaction and noisy skeleton and depth data. For example, the $playinggame$ activity is mostly misclassified as the $readingbook$ or $sitstill$ activity; $callingcellphone$ activities are misclassified as the $drinking$ or $sitstill$ or $tosspaper$; and eat activity misclassified as $drink$. These activities have similar hand movements and are difficult to distinguish, especially with noisy data.

[1]This also used RGB data; accuracy using skeleton data only is reported here.

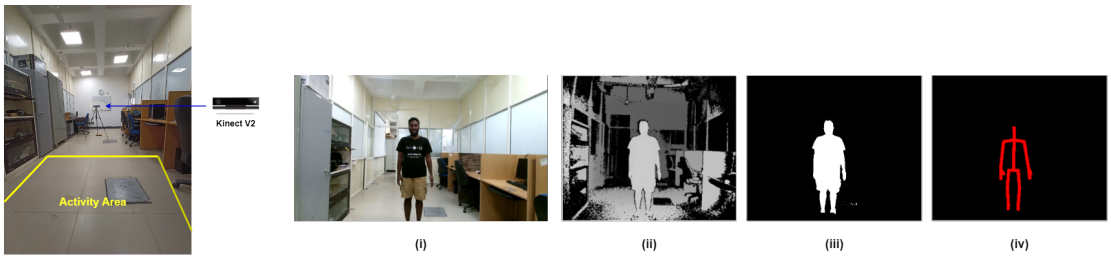
4.3.8 Computational Complexity

The usual approach to analyse algorithms in terms of computational complexity is by asymptotically representing their time and space requirements. For CNN models, however, asymptotic complexity analysis is usually not done. Alternatively, to demonstrate the effectiveness of the proposed framework in real-time, we calculate the proposed framework's inference time, which is a rough estimate and is dependent on the available resources. Table 4.8 includes the inference time of the framework on different datasets. The total inference time is divided as follows: the preprocessing time of the depth data; the score generation time taken by the 3D-CNN; the JPD and BAD generation time from the skeleton sequences; the score generation time taken by the 2D-CNN (including FLFS); and the time taken by the SLFS module. The preprocessing of depth data includes temporal & spatial normalization and normalization of the input pixel values in the range of [-1,1]. As the average temporal lengths of the datasets are different, the processing time for each dataset is also different. Similarly, the JPD and BAD generation time varies for different datasets due

Table 4.8: Average Computation Time per Activity Instance (in Milliseconds) of the Proposed Framework on Different Datasets

Operation	**MSR Action3D**	**UTD MHAD**	**TST FallV2**	**MSR DailyAct**
Depth Preprocessing	29.91	53.89	58.77	61.44
3D-CNN Score Generation	35.46	60.39	32.47	58.76
CH1 (total time) (P2)	65.37	114.28	91.24	120.20
JPD Image Generation (P1)	1.12	2.03	3.19	3.67
BAD Image Generation (P1)	9.98	14.21	35.96	36.43
2D-CNN + FLFS + Score Generation	7.77	7.77	7.77	7.77
CH2 (total time) (P2)	17.75	21.98	43.73	44.19
SLFS	0.42	0.46	0.28	0.37
Total Time (considering P1 & P2)	**65.79**	**114.74**	**91.52**	**120.57**

Two (P1) operations can execute in parallel with each other; Similarly, two (P2) operations can also execute in parallel with each other.



(a) (b)

Figure 4.11: Experimental setup for data collection; a) Placement of KinectV2 device; b) Multiple data streams: (i) Color image (ii) Depth image (iii) Depth image after background removal (iv) Human body skeleton

to different average temporal lengths. The score generation time of the 3D CNN depends on the input Temporal Lengths, T_i, for each dataset. Higher the temporal length, the longer the score generation time. The calculation of the total inference time also considers the parallel execution of segments: like JPD & BAD can be generated in parallel (P1), and CH1 & CH2 can execute in parallel (P2). The inference time is calculated as the ratio of the total time taken by all the inputs and the total number of inputs in the test set. The maximum inference time per input is $\approx$ 121 ms for the MSR Daily Activity dataset, which is well within the acceptable range in real-time applications.

4.4 Prototypical Implementation of Proposed System

The feasibility of the proposed framework in the real world is validated through a prototypical implementation. A KinectV2 depth sensor is utilized for capturing the depth and skeleton data in a laboratory setup, as shown in Figure 4.11(a). Multiple streams captured by the device are shown in Figure 4.11(b)[i-iv]. The Kinect device was placed on a tripod around 4 feet from the floor, and activities were performed in the activity region between 4 feet and 10 feet from the device. The depth data was captured at approximately 30 fps.

A small dataset of five activities, namely, HandWave, HandClap, ForwardKick, Walk, and Fall, is created to demonstrate the effectiveness of the proposed framework in a practical

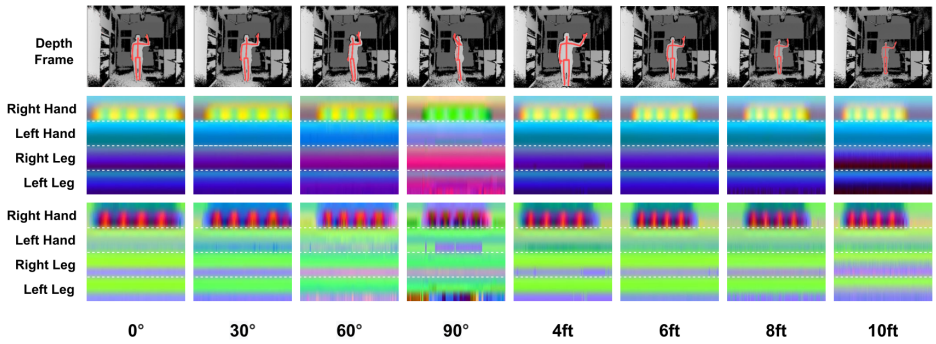


Figure 4.12: Hand Wave activity performed from four different angles and four different distances in our prototype dataset; depth frame (first row); JPD based images (second row), BAD based images (third row)

scenario. Six actors (four male, two female) with varying physiques performed the activities. Each actor performed an activity two to three times. To demonstrate the capability of the proposed JPD and BAD in effectively representing activities at varying distances and angles, each actor performed at least one activity at four different angles (i.e., 0°, 30°, 60°, 90°, where 0° implies that actor is facing the camera) and four different positions (i.e., 4 ft, 6ft, 8ft, 10ft from the device). In this way, the five activities were performed with eight variations at least once.

The JPD and BAD based images shown in Figure 4.12 confirm that the descriptors at different angles and distances for an activity are similar and can be distinguished from other activities. For temporal variation, the activity clip length was varied roughly between 2.5 and 6 seconds. The dataset altogether contains 104 instances of 5 activities. A few visual examples and the sample JPD and BAD based images for five activities are included in the Appendix (B). The dataset and the codes for reading the data, can be downloaded from the link: Dataset.

We utilized our pre-trained model for the TST Fall dataset (as this dataset was also captured using a similar Kinect device) after appropriate fine-tuning and achieved 100% classification accuracy. The fine-tuning involved modification and retraining both the score

generators using the activities performed by the first five actors. Subsequently, the validation was done using the activities performed by the sixth actor to conform with cross-subject evaluation.

4.5 Summary of the chapter

In this chapter, a framework for human activity recognition using privacy-preserving depth sensors was proposed. The main contribution of this chapter is the development of a robust two-channel CNN-based architecture for classifying human activities using depth clips and skeleton sequences. To ensure privacy, a depth sensor was employed, which inherently preserves privacy. The proposed architecture utilizes 3DCNNs and 2DCNNs for feature extraction from depth clips and images generated from skeleton data, respectively. Two novel descriptors, JPD and BAD, were introduced for scale- and view-invariant activity recognition from skeleton data. Furthermore, a novel two-level fusion strategy, FLFS and SLFS, was employed to effectively combine the JPD and BAD descriptors as well as the depth and skeleton modalities. The integration of 3DCNN and 2DCNN for two data modality, generation of JPD and BAD, and the two-level fusion strategy is a novel idea which was not explored in the literature.

The depth camera provided depth data clips that were directly used for activity recognition over an appropriate CNN framework after necessary preprocessing. The skeleton sequences obtained from depth data were mapped to two descriptors based on joint positions (JPD) and inclination of bones (BAD), respectively. The descriptors were further encoded into color images and provided both spatial and temporal information. These were used over a second channel of the CNN framework for activity recognition. Limited data was available for most analyses, and this was offset by transfer learning and data augmentation. To benefit from the multiple channels of analysis, fusion strategies were employed. A feature-level fu-

sion strategy, FLFS, was used to combine JPD and BAD based features, and subsequently, SLFS was used to combine the scores of the skeleton and depth based channels. The results of the proposed framework were shown to outperform existing techniques on four public datasets. The proposed framework was shown to be feasible in real-time through computational complexity analysis and prototypical implementation. The proposed system can recognize activities from partially occluded depth videos through skeleton tracking algorithms that enable the estimation of the joint information. However, occlusion beyond a point prevents the tracking algorithm from effectively estimating joints. To overcome this, a strategically planned multi-camera setup can be explored in future research.

Chapter 5

Identity and Activity Privacy-Preserving Posture Recognition System

5.1 Introduction

The global population is rapidly aging and the number of individuals aged 65 and more is anticipated to become 1.6 billion by 2050 [154]. This demographic shift is leading countries to seriously consider providing a larger number of independent living environments for the elderly. In this respect, the high cost and shortage of labor for elderly care, is spurring significant investments in the development of automated monitoring systems for such environments. Automated monitoring systems continuously observe living spaces and raise an alarm on detecting unusual activities, such as falls, of the elderly occupants. The system, on detecting an adverse medical condition, notifies caregivers or the next of kin of the victim to minimize potential harm. According to the Center for Disease Control and Prevention [155], around 3 million elderly are treated for injuries from falls each year of which around 800,000 are seriously injured and need to be hospitalized.

Most automated monitoring systems for assisted living environments function through readings received from sensors of various kinds. Of these, as discussed in Chapter 1, wear-

able and ambient sensor based systems are quite common. Although effective to an extent, these have limitations related to convenience and accuracy. In this respect, more effective systems are based on vision sensors. These are both convenient and accurate but are plagued the severe limitation of the compromise of the monitored individuals' privacy [27, 28, 73]. A good solution for overcoming privacy compromise amongst vision sensors is to use depth sensor based systems [121, 124], which is also utilized in our work proposed in chapter 4. Depth sensors are largely vision sensors that work on the idea of the distance of the object from the sensor and capture images that are quite unclear, comfortably concealing the identity of the monitored individual.

Most research that uses depth sensors for monitoring spaces conveniently assumes that their use automatically guarantees privacy [121, 124]. A few studies, however, argue that the images captured by depth sensors provide sufficient information for facial recognition especially when the number of individuals involved is small (up to around 30 individuals) [76, 125].

Moreover, for privacy preservation in assisted living environments, concealing an individual's identity, as claimed by most depth sensor based systems, is not enough and also not important. This is because the living space allocated to individuals in assisted living centers is fixed and known publicly and hence even if the depth camera based systems conceal the identity of the monitored individual, it does not serve much purpose. In such cases, therefore, a much higher degree of information concealment is required such that fine-grained activities performed by the monitored individuals, such as talking on the phone, consuming alcohol, keeping money, etc. are indiscernible. This is another level of privacy that we call 'activity privacy'. We, therefore, define two levels of privacy: identity privacy (that restricts the detection of an individual's identity from the captured images); and activity privacy (that restricts the detection of fine-grained activities that an individual performs).

It is important to understand, though, that depth sensor based systems that monitor assisted living environments need to work with images that preserve activity privacy but at the same time need to be coarse-grained enough for the algorithmic component of the monitoring mechanisms to assess the well being of the individual. This means that the images should conceal information on the specific tasks being done by the individual (like drinking, eating, smoking) whilst providing enough information to detect actions like standing, sitting, walking, lying, falling. It is critical that this balance, in the images captured, be maintained to be able assess the well being of the individual along with preserving privacy.

This chapter introduces a novel vision-based privacy-preserving system that utilizes customized depth sensors for monitoring indoor spaces. The modified depth sensors capture images that effectively preserve both identity and activity privacy, while also efficiently recognizing the coarse-grained postures of the individuals enabling an assessment of their well being.

The other challenge in devising such monitoring systems for private spaces is owing to the fact that privacy is a subjective concept. One individual may consider a certain level of privacy acceptable whereas another may find it to be an intrusion. To tackle this, we endeavoured to understand and identify the level of privacy widely accepted by most people. We conduced a survey involving individuals from diverse demographic backgrounds over a crowd-sourcing platform, Amazon Mechanical Turk (AMT). The survey's findings enabled us to determine the degree of privacy in images that most people are comfortable with in terms of both identify and activity privacy. In addition to this, two deep learning based classifiers (person identification system [156] & human activity recognition system proposed in chapter 4) were also employed to verify that the images, also preserved privacy from a machine learning perspective again in terms of both identity and activity privacy.

With the privacy of the monitored individuals ensured, the other important task of the

proposed system is correctly assessing their well being. The well being of monitored individuals whilst preserving their identify and activity privacy, is done through the well-established effectiveness of Convolutional Neural Networks (CNNs) in object classification. A CNN architecture based on VGG-16 [157] was employed for recognizing human postures within privacy-preserving depth images.

Furthermore, to enhance the performance of posture recognition in privacy-preserving images, structural characteristics of the human body in different postures were exploited using horizontal projections and geometric information of the human body. A horizontal projection map comprises images generated from the projection histogram of the horizontal pixels in the binary image. Projection maps are especially popular and effective in handwritten text recognition [158, 159] despite the noise and variations in handwritten text. We utilized projection maps because our privacy-preserving depth images, much like handwritten text, have variations in shapes owing to noise.

We also explored the statistical features of the shapes in the images such as the ratio of height to width of the lower and upper body, and other such factors that provide high discrimination in different postures. An integrated model was developed that combines the CNN-based features from depth images and projection maps with statistical features.

Keeping privacy-preservation and performance in mind, the key contributions of this chapter are as follows:

1. Modification and tuning of depth sensors for capturing depth images that preserve both identity and activity privacy.

2. Validation of acceptable privacy level of depth images by surveying global audiences utilizing crowd-sourcing services.

3. Validation of appropriate privacy levels for preserving identity and activity privacy

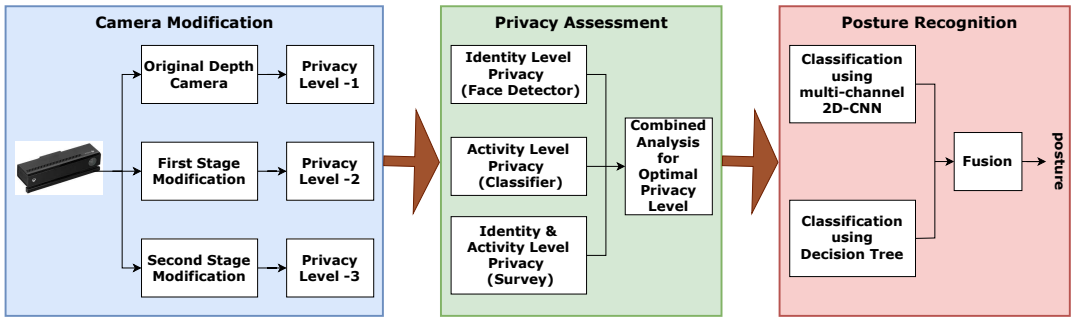


Figure 5.1: Workflow of the proposed framework

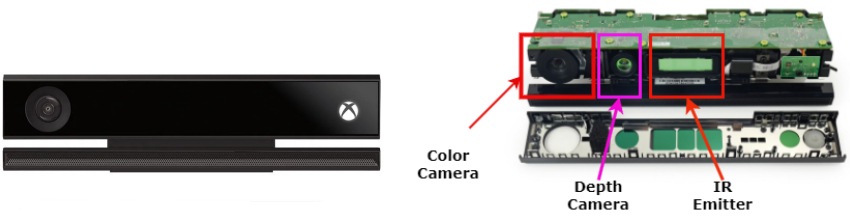
using machine learning methods.

4. Developing an efficient posture recognition system using CNN with depth images, horizontal projection maps, and geometric information of the human body.

5. Demonstration of the efficacy and feasibility of the proposed work through a prototypical implementation.

5.2 Proposed Methodology

In this section, the proposed approach for monitoring occupants' well being in indoor locations whilst preserving their privacy is discussed. As mentioned earlier we endeavour to capture depth images that preserve the privacy of the monitored individual(s). To do this, we modify standard depth sensors appropriately. We dwell upon these modifications first in this section. Subsequently, we describe three levels of privacy of the depth images captured and attempt to assess and validate the extent of privacy preservation ensured at these three levels. We also assess the acceptability and effectiveness of these privacy levels to the human eye and to appropriately trained machine learning algorithms. Finally, in this section, we describe a posture recognition system that is meant to correctly identify the posture of the monitored individuals using depth images that are validated to be privacy preserving. The idea is that accurate posture recognition of the monitored individual provides informa-



(a) (b)

Figure 5.2: Kinect Device: (a) Original Kinect Device, (b) Sensors in Kinect Device

tion such as falls, sitting at one place for extended periods of time, sleeping for extended periods of time, and so on, that are helpful in assessing the well being of the monitored individual(s). All this using images that are privacy preserving. A schematic diagram depicting the workflow of the proposed system is shown in Figure 5.1.

5.2.1 Modification & Tuning of Depth Sensor

We utilize a modified depth sensor to capture the privacy-preserving images that preserve both identity and activity privacy. Identity privacy, as discussed earlier, prevents the recognition of an individual through the images; whereas activity privacy prevents the identification of the activity being performed. For our work, we use the depth sensor embedded in the KinectV2 device. Before discussing the modifications to the sensor, a brief introduction of the KinectV2 device and Time Of Flight (TOF) depth-sensing technology is provided. A KinectV2 device and the sensors available in KinectV2 are shown in Figure 5.2.

A Kinect device is equipped with a color camera, a depth camera, and an IR emitter. The color camera works along the principles of light and captures high-resolution (1920 × 1080) color images. The depth camera works on the principle of the distance of the objects from the camera and captures depth and infrared images with a pixel resolution of 512 × 424. The distance between the object and camera is calculated using the IR wave's TOF. The IR emitter projects IR light waves that are reflected back to the depth sensor. The phase

shift between the projected and captured wave is used to calculate the distance between the object and the sensor as per Equation (5.1).

$$d = \frac{c}{2f}\frac{\phi_d}{2\pi} \tag{5.1}$$

Where ϕ_d denotes the phase shift, f is the frequency of the light wave, and c is the speed of light.

The distance d is then mapped to a pixel intensity value to generate a depth image. Similarly, the amplitude of the captured wave is mapped to generate an infrared image. The depth, infrared, segmented depth image, and the reference color images captured using an original depth sensor are shown in Figure 5.3 (first row).

To capture images that are privacy preserving, the modifications to the kinect sensor is done in three steps. The first step involves placing an opaque black sheet before the color camera sensor to restrict it from capturing color images. We call the images captured after this first step of modification as being at privacy level 1, P1. The first row of Figure 5.3, excluding the colour image comprises images at level PI. The second step of modification is done by placing a plano-convex lens before the depth camera sensor. This modification is inspired by the work in [38], where a plano-convex IR lens is used to capture de-focused images. Images captured at this level are more privacy preserving than P1 and we call these images at privacy level 2, P2 (second row of Figure 5.3). Finally, an LDPE film with a rough surface is included before the plano-convex lens to further diminish the information in the captured images, and augment the privacy. This modification is inspired by the work in [160], and is referred as privacy level 3, P3, shown in the third row of Figure 5.3.

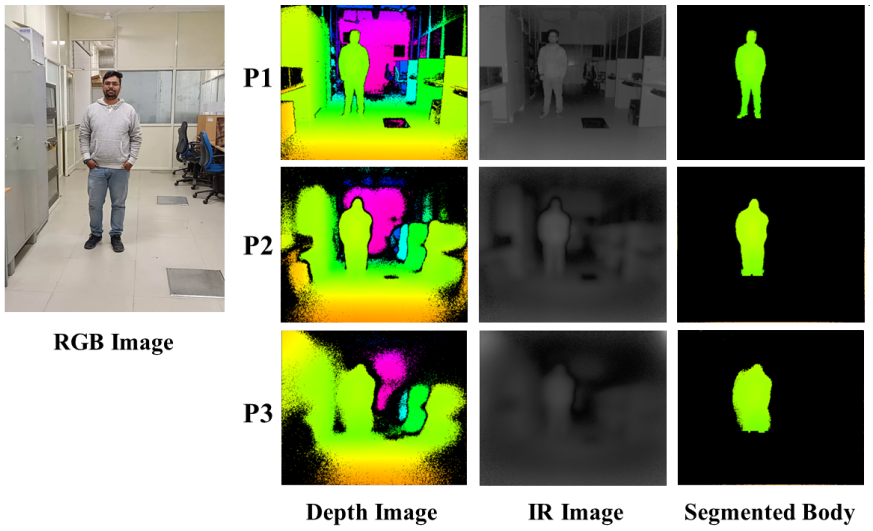


Figure 5.3: Captured data in three privacy levels

5.2.2 Validation of Privacy Preservation

The images captured by the modified depth sensor need to be validated for their degree of privacy, to ensure that they are indeed privacy preserving. We look at ensuring privacy preservation as perceived by human beings in general and from the point of view of trained machine learning algorithms. In other words, the privacy preserving images should preserve identity and activity privacy, when seen by a human being as well as when seen by an appropriate machine learning algorithm. The validation of the privacy-preserving nature of each privacy level is therefore done in three ways: 1) identity and activity privacy assessment through a user survey to ensure privacy when seen by humans in general; 2) identity privacy assessment using an automated face recognizing system; and 3) activity privacy assessment using a standard activity classifier. A detailed description of the approach taken for privacy validation follows.

5.2.2.1 Identity and activity privacy assessment through a user survey

To validate the privacy-preserving nature of depth images from the perception of humans in general, a user survey was conducted involving global participants. The survey was

Table 5.1: Survey questionnaire

Sr. No.	Questionnaire	Options	Keyword
Q1	Where, in your opinion, is this video taken? (choose the most suitable option)	Living Room, bed-room, computer-lab, kitchen, store-room, office, some other place, unable to recognize	Location
Q2	What living or non-living objects do you see in the video? Please name the objects. (select all appropriate options)	Person, bed, computer, fan, door, almirah, refrigerator, unable not recognize	Detection
Q3	If, according to you, the video includes a human being, what is their posture? (Choose any one option)	Standing & facing the camera, standing with back towards the camera, sitting & facing the camera, sitting with back towards the camera, some other posture, unable recognize	Posture
Q4	If, according to you, the video includes a human being, what activity or activities do they seem to be carrying out. (Choose any one option)	Drinking, eating, reading book, using mobile phone, sitting/standing still, some other activity, unable to recognize	Activity
Q5	Depth images of six persons are shown in first row and color images are shown in the second row (in random order). Match each person in depth image to the corresponding person in color image. (Match each depth images to it's correct color image)	First column shows the depth image numbering (D1, D2, D3, D4, D5, D6) and the second column shows the color (RGB) image numbering (R1, R2, R3, R4, R5, R6, Can't identify).	Identity

published on Amazon Mechanical Turk (AMT), where we considered responses from the first 100 participants representing a range of age groups, occupations, educational levels, and time zones.

The survey involved showing images and videos at different privacy levels (i.e., P1, P2, and P3) to participants. For each privacy level, two video clips comprising random activities were included. Along with the clips, questions around the clips were asked. These questions are included in Table 5.1.

In addition to this, the survey also comprised a matching exercise wherein a set of images for each privacy level with different actors was included. Also included was a set of visible

color images of the actors. The privacy-preserving images and the visible color images were arranged in a random order and participants were asked to match the actors in the privacy-preserving depth images to the actors in the visible color images. This exercise was yet another approach in the survey to assess the preservation of identity privacy in the images. The videos and images used in the survey were similar to those in Figure 5.3 and the survey is available at following link: SurveyLink.

The participants' observations for location, person detection, posture, and activity were noted in the survey and were analysed to assess the privacy preservation capability of the images at the three levels. The detection and identification of actors in the videos and images respectively enabled the assessment of identity privacy of the images at various levels. Similarly, the identification of activities in the videos in the survey enabled assessment of activity privacy of the images at various privacy levels.

The collected responses of the survey are categorized as *correct*, *incorrect*, and *not clear*. For a given privacy level, responses in each category are counted for all videos and images separately. The category-wise scores for each privacy level are calculated as per Equation (5.2) and Equation (5.3) for videos and images, respectively.

$$\forall V_x \in L_i, i = 1, 2, 3$$
$$T_{L_i,Q_j,C_k} = \frac{\sum_{x=1}^{X} R_{Q_j,C_j}(V_x)}{\sum_{j=1}^{C}\sum_{x=1}^{X} R_{Q_j,C_j}(V_x)} \tag{5.2}$$

Where, L_i denotes the privacy level (i.e., P1, P2, and P3), Q_j denotes the question number (i.e., Q1, Q2, Q3, and Q4), and C_k denotes the response category (i.e. 1 for Correct, 2 for incorrect, and 3 for Not-clear). T_{L_i,Q_j,C_k} represents the percentage of responses falling in category C_k of question Q_j for videos at privacy level L_i. $R_{Q_j,C_j}(V_x)$ is the response to Question Q_k falling in category C_j. V_x denotes the set of video clips included in the survey

for a given privacy level.

$$\forall D_x \in L_i, i = 1, 2, 3$$
$$N_{L_i,C_k} = \frac{\sum_{y=1}^{Y} M_{C_j}(D_x)}{\sum_{j=1}^{C}\sum_{x=1}^{X} M_{C_j}(D_x)} \tag{5.3}$$

Where, N_{L_i,C_k} represents the percentage of responses falling in the C_k category of privacy level L_i. $M_{C_j}(D_x)$ is a response wherein the respondent was able to correctly match the person/object in the depth image D_x to that in the corresponding color image; and the response falls in the C_k category.

In Equation (5.2), T_{L_1,Q_1,C_1}, for example, indicates the percentage of correct (C_1) responses to Question 1 (Q_1) at privacy level 1 (L_1). T_{L_1,Q_1,C_1} is calculated by dividing the correct responses to Question 1 (Q_1) at privacy level 1 (L_1) by all the responses to Question 1 at privacy level 1. The numerator is calculated by adding all the correct (C_1) responses to Question 1 (Q_1) received across all videos at privacy level 1 (L_1). The denominator is calculated by adding all the responses (C_1- correct, C_2- incorrect, and C_3- not clear) to Question 1 (Q_1) across videos at privacy level 1 (L_1).

Similarly, in Equation (5.3), N_{L_1,C_1} indicates the percentage of correct (C_1) responses at privacy level 1 (L_1). N_{L_1,C_1} is calculated by dividing the total number of responses correctly (C_1) matching a person in depth image to the person in color image, by all the responses at privacy level 1 (L_1). In the numerator ($M_{C_1}(D_x)$, the responses that correctly (C_1) match a person in the depth images with those in the color images are added across all images at privacy level 1 (L_1). In the denominator, all responses (C_1- correct, C_2- incorrect, and C_3- not clear) for privacy level 1 (L_1) are added.

In this manner, the identity and activity privacy of the images at the three privacy levels are assessed as perceived by human beings in general.

5.2.2.2 Identity privacy in a face recognition system

In addition to assessment of privacy as perceived by humans, we also endeavour to validate privacy of the images as perceived by appropriate machine learning algorithms. We first look at identity privacy assessed by a standard face recognition system. Low accuracy in the face recognition system indicates a high degree of identity privacy and vice versa.

We use the well recognised, YOLOV3 (You Only Look Once) [156] based architecture for face detection and recognition. YOLOV3 is a popular object detection algorithm that utilizes a deep convolutional neural network to detect and identify objects in images. YOLOV3 divides an input image into small regions and predicts the bounding boxes by checking the 'objectness' score (the probability of having an object in a given region). The Objectness score is calculated based on the similarity of the given region with one of the predefined classes. The high scoring regions are considered to be correctly classified (as a member of the closest class).

The YOLOV3 architecture comprises two parts, a feature extractor and a multi-scale detector. The feature extractor used in YOLOV3 is DarkNet53 [156] which a 53-layer convolutional neural network that extracts meaningful information from an image. DarkNet53 contains combinations of 3x3 and 1x1 convolution filters in the form of residual blocks with skip connections. Residual blocks allow the gradients to flow from a high layer to a lower layer and help in convergence of deeper networks. Another 53 layers are added after the feature extractor that serves as a detector thus making the architecture comprise 106 layers. The object identification is done at three scales from feature maps of size 13×13, 26×26, and 52×52 (taken from the 82^{nd}, 94^{th}, and 106^{th} layer). Detection at three scales helps preserve fine-grained details and thus detect both small and large objects. The feature map of size 13×13 detects larger objects, whereas the feature map of size 52×52 detects smaller objects in the image.

We modified the original YOLOV3 architecture to make it suitable for our face detection/recognition task. The original YOLOV3 is trained for 80 different classes; we first changed the number of 1×1 filters to utilize it for our work (2 for face detection and 6 for face recognition, similar to the number of actors). Further, the number of layers is increased in the first two residual blocks to enable the extraction of extra fine details, which usually get lost due to down-sampling. Experiments show that these modifications make the model suitable for face detection and recognition.

5.2.2.3 Activity Privacy in an activity recognition system

In addition to validation of activity privacy by humans in general through the user survey, we also validate that the activities being conducted in the privacy preserving images are indeed private when an established activity recognition system is used. The activity recognition system employed here is a fairly successful one and was proposed in an earlier work of ours [123]. The system comprises a 3DCNN model that takes a depth video clip as input and classifies the activity in the video clip into one of the defined categories. The activity recognition system is based on a 3D-CNN architecture *I3D* [147], that is popular for video classifications in a color domain.

I3D, as its name implies, is built around inflated convolutional and pooling filters (inflated filters are 3D filters created by giving standard 2D filters an extra dimension). The third dimension in a 3D filter learns the correlations in the temporal direction of the video, while the other two dimensions learn the spatial correlation in the frames. An inception module based design is utilized in the I3D architecture to process the video data. The Inception module contains a combination of 3×3 and 1×1 convolutional filters in parallel that make the network progressively wide but not deep. An inception module also uses $(1\times1\times1)$ convolutions for dimensionality reduction. The use of the inception module makes I3D, effi-

cient and faster. Also, asymmetric kernels ($2\times7\times7$) are used in the first max pooling layers of I3D to handle the frame rate of videos (i.e., the speed of performing activities vary).

The I3D architecture comprises a series of convolutional layers, max pooling layers and inflated inception modules. At the end, an average pooling layer followed by a convolution layer with filters of size ($1\times1\times1$) are used. As discussed in our work [123], two modifications are made to the original I3D architecture. These include, modifying the classifier module and including a spatial dropout [150] layer in the inception module of I3D. Modification in the classifier module is done by replacing the last average pooling layer with a Global Average Pooling (GAP) layer followed by adding a dense layer with a number of neurons similar to the activity categories. Finally, a soft-max layer is added at the end of the classifier module to compute the score probability distribution of each activity category.

A spatial dropout layer is incorporated in the 3D inception module due to its efficacy in handling missing values and mitigating over-fitting. Spatial dropout drops the complete motion map with a drop probability of P_d and thus reduces the dependence among the values in motion maps. Equation (4.3) shows the feed-forward process with a spatial dropout layer.

5.2.3 Posture Recognition

The aim of the proposed system ultimately is to ensure the well-being of the monitored elderly individuals. Therefore, it is imperative that the images that are unclear enough to be privacy preserving, have some characteristics remaining that enable accurate posture recognition. Posture recognition permits assessing the well being of the monitored elderly through recognising incidents like falls, sitting idle at one place for an unnaturally long time, lying down for an unnaturally long time at one place, and so on. In this chapter, for simplicity we recognise two postures only: standing and sitting down. The work can easily be extended to include other postures as well. The posture recognition approach in

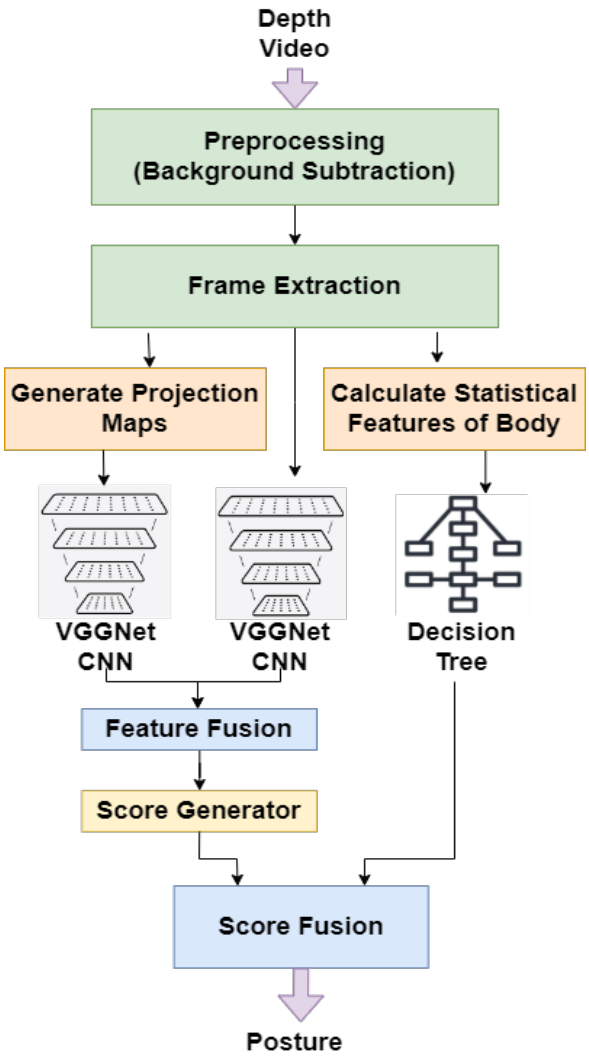


Figure 5.4: Integrated posture recognition system

the proposed system is based on deep learning and the structural properties of the human body in different postures. It is observed that combining the structural properties with deep learning methods significantly improves the posture recognition capability of the system from the privacy-preserving images.

The depth data is pre-processed first before feeding it to the posture recognition system. The pre-processing mainly includes the human body segmentation by removing the background details. A depth image containing gray-scale values often suffers from 'depth camouflage', a phenomenon in which the foreground and background pixels have identical

values. Due to depth camouflage, it becomes challenging to distinguish the human being from the background, especially when the human being is close to background. In order to address this, a well-known 'background subtraction' method [148] is utilized in this work, which segment the human body from the depth frame by eliminating the background details.

A background frame $B(x, y)$ is first computed from the background frames of the video in the absence of human being. Subsequently, the background frame is subtracted from each frame of the video. This highlights the human body by removing the stationary background details. The tiny noise arising due to environmental changes is suppressed by imposing a threshold. Equation (4.1) expresses the background subtraction with threshold. Finally, the post-processing exercises like morphological operations are undertaken if the resultant images are still noisy. This results into a depth image with segmented human body as shown in the last column of Figure 5.3.

As shown in Figure 5.4, the proposed posture recognition system comprises two CNN channels, one for depth images and the other for projection maps. Projection histogram maps are popularly used in handwritten text recognition [158, 159] and despite large intra-class variations in handwritten letters, such methods have proven to be successful. The privacy preserving images at privacy level P3, also have large shape variations due to noise, much like handwritten letters, and thus we employed horizontal projection maps of the human body for posture recognition in privacy preserving images.

A horizontal projection map is a histogram of the number of white pixels (pixels representing human body) accumulated along the rows of a binary image of size $H \times W$, where H is the number of rows and W the number of columns. A mathematical representation of the horizontal projection map is shown in Equation (5.4).

$$HPM(r) = \sum_{0<c<W} P(r, c) \tag{5.4}$$

A VGG16 [157] based architecture is exploited for feature extraction in both the channels. VGG16 is a simple yet fairly successful convolutional neural network model for object classification in images. VGG16 consists of thirteen convolutional layers to extract useful features from the image, five pooling layers for preserving features while decreasing the size of the feature maps. Finally, there are three fully connected layers to classify the image based on the features extracted by the preceding layers. The fully connected layers act as a linear transformation function to classify the input into one of the categories.

We modified the existing VGG16 architecture by removing all the fully connected layers, as these are dependent on specific classification tasks. The rest of the VGG16 architecture is utilized for feature extraction from the depth images and projection maps, separately. The features extracted from the two inputs (i.e., depth image and projection map) are combined into a large feature vector; this process is termed *feature-level fusion*.

Subsequently, a score generator module containing three dense layers and a softmax layer is added at the end. The score generator generates a score vector containing the confidence score of each posture category. Finally, the posture label is assigned using the maximum value index in the score vector.

Although convolutional neural network based approaches are very popular mostly due to their automatic feature extraction capabilities and the need for less human intervention, CNN considers spatial relations by pooling the local features into a global representation that sometimes performs sub-optimally when learning certain patterns. This fact is exhaustively tested in [161, 162], where authors emphasize that CNNs are sensitive to an object's local features but have no access to global shapes. This motivated us to consider some of the structural properties of the human body postures as additional features to counter the fuzziness in the privacy-preserving images.

The ratio of the height of the body to the width of the body is very important when

Algorithm 5.1 Posture Recognition using Weighted Average Score Fusion

Inputs: DI-Depth Image

HPM- Horizontal Projection Map

SP- Structural Features

Output: Prediction

CS = CNN(DI , HPM) `// CNN Score`

DS = DT(SF) `// Decision Tree Score`

SCORE$_c = W_1 * CS_c + W_2 * DS_c,$ $c = c_1, c_2$

$s.t., W1 + W2 = 1$

`/*` CS_c `and` DS_c `are the scores of CNN and DT for a posture class 'c'. */`

$$\text{Prediction} = \begin{cases} c_1, & \text{If } (SCORE_{c_1} > SCORE_{c_2}) \\ c_2, & \text{Otherwise} \end{cases}$$

return Prediction

considering the body postures, especially in the case of standing and sitting postures. Furthermore, the human body is also divided into the upper body and lower body using the center of mass. The ratios of height to width for the upper and lower body are also strong indications of the body posture. These three features are quite different for different postures and are used with a Decision Tree (DT) based approach to classify the body postures. Finally, the decisions of both CNN and DT are combined to get more accurate classifications of body postures. The complete process of combining the decisions of the two classifiers (i.e., CNN and DT) is described in algorithm (5.1).

5.3 Experimental Evaluation

Validation of the privacy-preserving nature of the depth images for identity privacy and activity privacy is done first and is conducted in three ways; namely, identity privacy using a face recognition system, activity privacy using an activity recognition system, and

both identity and activity privacy using a user survey. The results of these experiments are presented first. Subsequently, the validation of the utility of the depth images in posture recognition is done using a CNN based posture recognition system. Finally, the performance of the integrated posture recognition system on privacy preserving images (P3) is presented. In this way, the effectiveness of the proposed system for ensuring the well being of the elderly (through posture recognition) whilst preserving their privacy is validated.

Research on indoor monitoring using vision sensors is mostly done using color/depth images, and most existing works in this direction do not look into the aspect of privacy preservation. The few works that do consider privacy preservation utilize depth images, that mostly conceal only the identity of the monitored individuals. As discussed earlier, this is not true in the context of indoor monitoring systems as living spaces in most indoor locations are small in number and so are the occupants, thus making identity privacy ineffective and redundant. The other aspect of privacy, activity privacy, where images conceal the fine grained activities being performed by the individual, is therefore more important from a privacy point of view. Very little depth images data, that preserve both identity and activity level privacy, is available. This work focuses on exploring privacy-preserving monitoring of elderly with a vision-based monitoring system. In the absence of such privacy-preserving images' datasets, we created datasets of the three different privacy levels (i.e., P1, P2, and P3) using modified depth cameras, as described in Section III(A). The datasets were collected in a laboratory setup, as shown in Figure 5.5.

5.3.1 Dataset Description

The dataset for each privacy level includes depth video clips of five activities performed by six different actors. The activities were performed in two postures (i.e., standing and sitting) and at four different positions in the activity area. All the datasets were collected

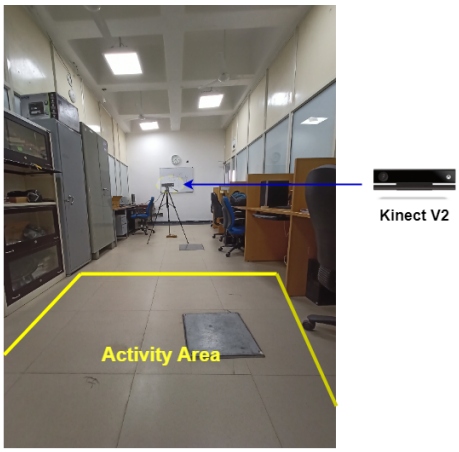


Figure 5.5: Laboratory setup for data collection (a modified depth sensor placed on a tripod)

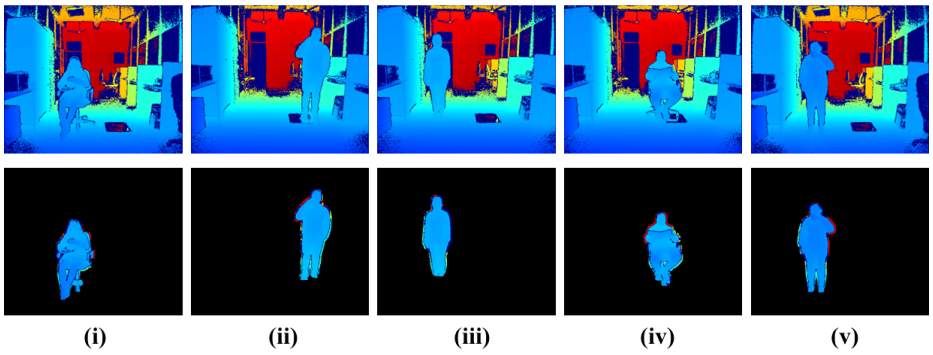
in the form of short videos of 5-10 seconds each to start with. For activity recognition, several small clips were extracted from the original videos with almost 10-20% overlap. This resulted in around 1000 small clips (40000 frames), comprising an equal number of clips for each of the daily activities. A detailed description of the dataset for each privacy level is included in Table 5.2.

Sample images from the created datasets are shown in Figure 5.6. The figure includes sample images at each privacy level (i.e., P1, P2, and P3). Background subtracted images corresponding to the depth images are also included in the figure.

5.3.2 Survey Analysis

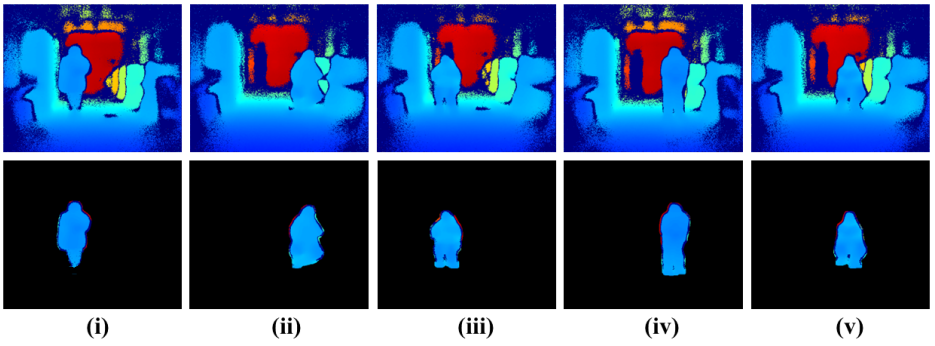
A user survey was conducted with global participants to validate the privacy of the depth images at the three privacy levels. The survey comprised 100 random participants from diverse demographics, educational backgrounds, and age groups. Most participants had bachelors' level education and were in the age group of 20-40 years. The participants were shown videos and a set of images for each privacy level and were asked questions related to the location, presence of persons in the video/image, body posture of the persons, activity being indulged in, and person identification as discussed in Section 5.2.2.1.

(a) Images of privacy level P1



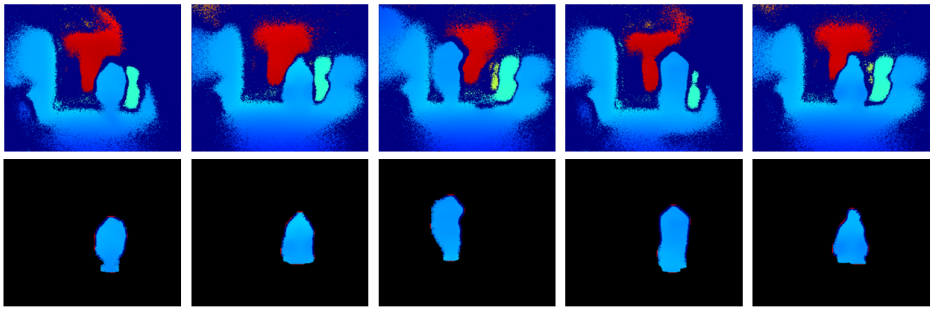
(i) (ii) (iii) (iv) (v)

(b) Images of privacy level P2



(i) (ii) (iii) (iv) (v)

(c) Images of privacy level P3



(i) (ii) (iii) (iv) (v)

Figure 5.6: Sample images from the datasets created (original depth images and background subtracted images). a) Images at privacy level P1, b) Images at privacy level P2, c) Images at privacy level P3

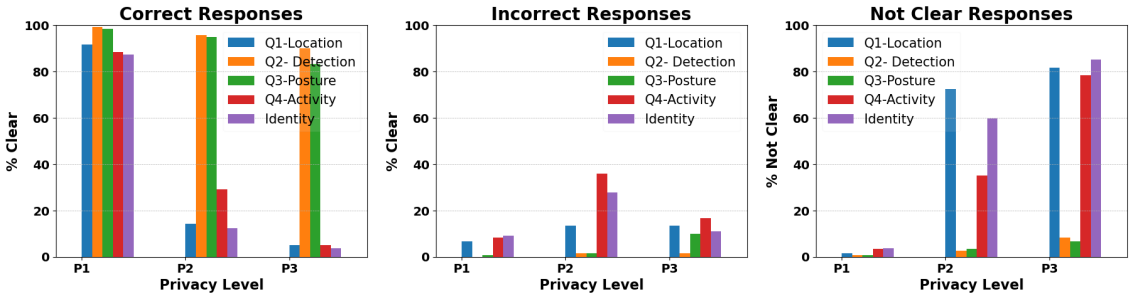
The responses to the questions were classified into three categories: *correct*, *incorrect*, and *not clear*. Correct response, as is obvious, implies that the participant gave the correct answer to the given question, whereas incorrect response indicates that the participant's response was not correct. The 'not clear' response means the participant could not see the required details in the video/image and chose the "cannot recognize" option.

Table 5.2: Description of the datasets

Description	
No. of activities	5
No. of Actors	6
Variations	4 locations, 2 postures
Clips Recorded	240
Smaller Clips	1000
Images	40000
Clips/activity	200
Images/posture	20000
Images/person	6600

Figure 5.7 shows the question-wise responses in the three categories at each privacy level. The majority of respondents gave correct responses for videos/images at privacy level P1, whereas the least number of correct responses were received at privacy level P3. The questions involving privacy sensitive information like location, activity, and identity of the individual were answered correctly by more than 85% of participants at privacy level P1. This indicates that the videos/images at privacy level P1 do not preserve identity and activity privacy. On the contrary, only around 5% of participants were able to answer questions related to images/videos at privacy level P3 correctly, indicating the suitability of images at level P3 for privacy-preserving monitoring.

Moreover, at privacy level P3, a large number of 'not clear' responses were received for questions involving privacy sensitive information, information like facial features that gives away the identity of the individual. Participants were clearly unable to recognize privacy-sensitive information at level P3 and this further bolsters the belief of level P3 being privacy-preserving. Similarly, a large number of incorrect responses at level P2 indicates that participants were not sure about this information but were a little confident and were at least trying to respond. This gives credence to out hypothesis that privacy preservation at



(a) (b) (c)

Figure 5.7: Survey analysis; percentage of a) Correct; b) Incorrect; and c) Not Clear responses at three privacy levels

level P3 is better than that at level P2.

However, the questions related to person/object detection and posture recognition were answered correctly by 95% and 83% respondents respectively at privacy level P3 which indicates the utility of these images in recognition of posture and other coarse grained activities.

5.3.3 Validation of Identity Privacy

For validating the claim that the depth images at the higher levels of privacy are indeed privacy preserving with respect to the person's identity even when established face detection/recognition systems are harnessed, the system was trained for each privacy level separately. As mentioned earlier, the dataset was collected at four locations (L1, L2, L3, and L4) in the activity area. A cross-location evaluation was employed, where the data from odd locations (i.e., L1 and L3) was used for training and data from even locations (i.e., L2, L4) for testing. This is named EvenTest. Similarly, the process was repeated with training with data from even locations and testing with data from odd locations, named OddTest. The final results were an average of the two tests.

We used pre-trained weights from the popular Imagenet dataset, which contains more than 14 million color images with annotations. Further, the model trained on color images

Table 5.3: Face detection (person detection) at three privacy levels.

Privacy Level	OddTest	EvenTest	Average
P1	98.64	99.63	99.13
P2	80.75	77.52	79.14
P3	50.15	48.65	49.40

was retrained on the Pandora dataset [163], which is a publicly available dataset of human faces in the depth domain. Finally, the retrained model was fine-tuned with our privacy-preserving datasets. To avoid over-fitting, data augmentation techniques like blurring, flipping, rotation, and change in brightness were utilized. The Adam optimization algorithm was used for training the model with an initial learning rate of 0.0001 and decay of 0.0005.

For face detection , the number of classes was set to one, as a face only needed to be detected in the depth image. In the face recognition task, on the other hand, the number of classes was set to six, corresponding to the number of actors. Table 5.3 and Table 5.4 show the accuracy of face detection and face recognition in images at the three privacy levels. The results are shown for cases where the confidence of the detection was at least 0.5.

The performance of the face detection and face recognition system is poor for images at privacy levels P3 and P2, which indicates that images at both these privacy levels preserve identity privacy. The performance of the face detection system for images at privacy level P3 is around 50%, which means that half of the human faces were not even detected in the images. Furthermore, the recognition accuracy for privacy level P3 is just 4%, which indicates that these images are significantly privacy-preserving in the context of identity privacy. On the contrary, the face recognition accuracy for system with images at privacy level P1 was quite high (86%) and these, therefore, do not preserve the identity privacy of the individuals.

Table 5.4: Face recognition (person identification) accuracy at three privacy levels.

Privacy Level	OddTest	EvenTest	Average
P1	88.18	83.77	85.98
P2	16.06	16.22	16.14
P3	3.71	3.90	3.80

5.3.4 Validation of Activity Privacy

Similarly, to ensure that the depth images at higher levels of privacy indeed preserve activity privacy, activity recognition systems were trained for each privacy level separately. Cross-subject evaluation was employed here as well where the images of odd-numbered (i.e., 1,3,5) actors were used for training and even-numbered (i.e., 2,4,6) actors for testing. This was named EvenSubTest. Similarly, images of even-numbered actors were used for training, and odd-numbered actors for testing in OddSubTest. Cross-subject evaluation of this kind is quite popular in activity recognition tasks.

For activity recognition, pre-trained weights from a large-scale kinetics dataset with 60,000 annotated video clips divided into 400 categories were employed. Transfer learning with fine-tuning was adopted by training the classifier module while leaving the weights of lower layers unchanged. The depth clips were made to undergo temporal and spatial normalization to make them suitable for the CNN-based classifier. Temporal normalization includes converting the variable length depth video clips to a fixed length. Spatial normalization involves down-sampling of the depth frames according to the CNN architecture.

To further reduce over-fitting, especially with less training data, the approach taken was data augmentation. The augmentation approaches included speed sampling to vary the speed of performing the activities; temporal random cropping, to vary the starting point of the activities. Other image augmentation approaches like random rotation, translation, horizontal flipping, resizing, shifting were also utilized to make the model scale and loca-

Table 5.5: Activity recognition accuracy at three privacy levels.

Privacy Level	OddSubTest	EvenSubTest	Average
P1	86.27	89.43	87.85
P2	56.81	57.28	57.05
P3	32.40	29.65	31.03

tion invariant. The Adam optimization algorithm with an initial learning rate of 0.001 and a decay of 0.004 was utilized.

Table 5.5 shows the performance of the activity recognition system with images at the three privacy levels. At privacy level P1, around 88% of activity instances were classified correctly, which means that privacy level P1 does not preserve activity privacy. The activity recognition system performed poorly for images at privacy level P3 and only around 30% instances were classified correctly for a five-class classification which is little higher than the random guessing. Images at privacy level P3, therefore, seem to preserve activity privacy properly and can be used for privacy preservation in indoor monitoring.

5.3.5 Posture Recognition System

Having validated the efficacy of the modified depth images in preserving identity and activity privacy, it is important that the proposed system be able to recognise the posture of the monitored individuals in the images at the higher levels of privacy. Effective posture detection can be used for drawing conclusions on the well-being of the monitored individual.

CNN architectures based on VGGNet were trained separately for images at different privacy levels using a transfer learning approach. The models were trained for binary classification as the datasets contain humans in two different postures namely, sitting and standing (this can easily be extended for a few other postures). The accuracy of the posture recognition for different privacy levels is shown in Table 5.6.

Table 5.6: Posture recognition accuracy at different privacy levels.

Privacy Level	OddTest	EvenTest	Average
P1	99.24	99.70	99.47
P2	95.85	95.28	95.57
P3	93.92	84.05	88.99

The accuracy of posture recognition using images at privacy level (P1) is very high which is obvious as these images contain rich information on the monitored individual. On the other hand, the accuracy of posture recognition with images at privacy level (P3) is 89% which is also quite good and indicates the utility of these images in posture and coarse-grained activity recognition.

In order to improve the performance of posture recognition with images at privacy level P3, an integrated system as shown in Figure 5.4 using original depth images, projection map images, and structural features of the human body is explored. The performance of the individual streams/features is included in Table 5.7.

Table 5.7 shows that combining the projections and structural features with depth images significantly improves the posture recognition performance. The accuracy of posture recognition using the proposed integrated system is 96.28% which is very good and suitable for real-time applications.

Table 5.7: Performance of integrated posture recognition system at privacy level P3.

Method	EvenTest	OddTest	Average
Depth Images (DI)	84.05	93.92	88.99
Horizontal Projections (HP)	84.83	85.95	85.39
Multi-channel CNN [DI + HP](MC-CNN)	89.03	94.10	91.57
Decision Tree (DT)	92.89	91.03	91.96
MC-CNN + DT	**94.16**	**98.40**	**96.28**

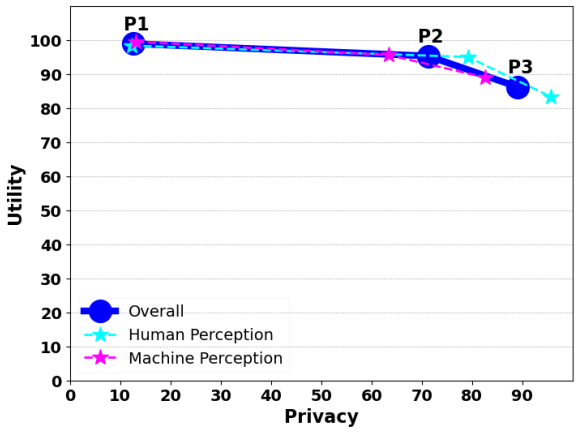


Figure 5.8: Utility vs Privacy trade-off of three privacy levels

5.3.6 Utility vs. Privacy trade-off

A trade-off between utility vs. privacy is an indicator of the utility of privacy preserving images. Our work discusses two measures (i.e., user survey and ML algorithms) to validate the privacy at different privacy levels. Similarly, a posture recognition system is used to validate the utility of the images in terms of posture recognition at different privacy levels. The results of all three measures are discussed in the previous sections. Figure 5.8 shows the trade-off of utility vs. privacy. The X-axis represents the privacy and the Y-axis represents the utility of the system at different privacy levels. Privacy of the data is inversely proportional to the accuracy of the recognition/detection systems. For example, if the accuracy of the face recognition system is high (i.e., 90%) then the data has less privacy (i.e, 10%), and vice-versa. We utilize this fact and calculate the privacy by subtracting the accuracy from 100.

The line shown in light blue indicates the trade-off of utility vs privacy from a human perception of the images and is calculated from the user survey. The line in magenta indicates the trade-off of utility vs. privacy from a machine perception and is calculated from the results of learning models given in Table 5.4, Table 5.5, and Table 5.6. Finally, the dark blue line shows the trade-off of utility vs. privacy using both the human and machine

perception and is calculated by taking an average of the results.

Figure 5.8 clearly indicates that the privacy of images at level P3 is low (approx 13%) but the accuracy is very high (almost 100%). On the other hand, the privacy of images at level P3 is high (i.e., around 90%) and the accuracy is significant (i.e., 86%), considering only CNN based classification. The accuracy at privacy level P3 is however improved (to 96%) by including additional features in integrated posture recognition system as shown in Table 5.7. This concludes that the images at privacy level P3 have significant utility and preserve both identity and activity privacy.

5.3.7 Computational Complexity

Usually the computational complexity of an algorithm is analyzed asymptotically. However, in CNN based approaches, it is uncommon to perform asymptotic complexity analysis. Mostly, the time consumption for a given input or the number of floating point operations (FLOPs) are calculated for CNN models. To illustrate the efficacy of the proposed approach in real-world applications, we also conduct a time-complexity analysis by calculating the inference time of the input instances (i.e., a depth image).

Table 5.8 shows the inference time of the proposed framework for a given input. The total inference time for an input is calculated by adding the time taken by each module shown in Figure 5.4. The modules include, depth pre-processing, projection map generation, statistical feature calculation, CNN score generation, decision tree score generation, and classification. The depth data pre-processing includes background removal and spatial normalization of the depth images. The depth pre-processing module runs the process in a manner that the subsequent frame runs in parallel with the other modules processing the current frame. This is, therefore, marked as (P0). Projection maps and statistical feature are generated from the background subtracted images and these two modules run in parallel.

Table 5.8: Average Inference Time Per Frame (in milliseconds) of the Proposed Integrated Posture Recognition System (P0, P1, and P2 indicates the parallel execution of the modules)

Operation	Time (ms)
Depth Preprocessing (P0)	22.22
Projection Map Generation (P1)	18.24
Statistical Feature Calculation (P1)	14.86
CNN Score Generation (P2)	12.03
Decision Tree Score Generation (P2)	0.006
Score Fusion & Classification	0.16
Total Inference Time	**30.44**

They are, therefore, marked as (P1). Furthermore, the CNN score generation time is the time taken by the two-stream CNN architecture to generate scores, and the decision tree score generation time is the time taken by the decision tree classifier to generate the classification scores. These two modules also run in parallel and are, therefore, marked as (P2). Finally, the score fusion & classification time is the time taken for classifying the posture after fusing the scores of the CNN and decision tree. For the modules running in parallel, the largest time taken is considered in the calculation of total inference time.

The inference time is calculated as the ratio of the total time taken for all inputs to the number of inputs. The average time per input is 30.44 millisecond that is 33 FPS, which is well within the range of real-time systems.

5.4 Summary of the chapter

In this chapter, an indoor monitoring system was proposed for assisted living environments that preserves the identity and the activity privacy of the residents. The main contribution of this chapter is the development of an indoor human monitoring system that preserves both identity and activity privacy of residents. To ensure privacy, a strategically

modified depth camera was used to capture privacy-preserving data. The privacy and utility of the captured data were validated through user surveys and deep learning methods. Additionally, an integrated posture recognition system was developed by combining the structural and statistical characteristics of the human body. *The use of a modified depth sensor, privacy validation from both human and machine perspectives, and the introduction of the concept of activity privacy represent a novel contribution not previously explored in the literature.* The proposed system employs a vision-based monitoring system comprising modified depth sensors. In the absence of privacy-preserving data, datasets containing privacy-preserving videos/images were created by us. The privacy and utility of the system was validated through a user survey and over appropriate deep learning frameworks. Finally, an integrated posture recognition system was developed and validated with privacy preserving data. The privacy validation and posture recognition results conform with the usability of the proposed framework in privacy-preserving indoor monitoring, especially in elderly care. In future, this work can be extended for coarse-grained activity recognition and pose estimation from privacy-preserving images.

Chapter 6

Conclusions and Future Works

Indoor monitoring is a multifaceted approach that enhances health, safety, comfort, and facilitate independent living in indoor environments. The limitations related to convenience in wearable sensors, performance in ambient sensors, and privacy in visible color sensors makes them unsuitable for indoor monitoring in private spaces such as smart home, elderly care, and healthcare. The primary challenges for effective indoor monitoring system thus include privacy, cost, convenience, and feasibility in real-world applications due to speed and resource requirements. Most approaches in literature focus on the performance of monitoring systems and neglect the issue of an individual's privacy and the system's feasibility for real-world applications. The main objective of the research in this thesis includes privacy-preserving indoor monitoring using vision sensors. Considering the major limitation of post-capture privacy, wherein data (videos/images) captured runs the risk of being compromised or reverse-engineered, this thesis explores the use of vision sensors modified appropriately to capture images that are privacy preserving and hence suitable for indoor spaces. Along with the use of appropriate vision sensor, the thesis also develops and effectively uses classification frameworks utilizing deep learning and machine learning approaches for privacy-preserving data. This chapter provides the concluding remarks on the

work carried out in this thesis. Section 6.1 presents a summary of the contributing chapters in this thesis. Subsequently, Section 6.2 highlights the possibilities to further extend the work in future.

6.1 Summary of Contributions

This section presents the contribution made through the research work in this thesis. A summary of each of the contributions is as follows:

6.1.1 Privacy-preserving fire detection system

A privacy-preserving efficient fire detection system using modified near-infrared cameras (NIR) is proposed in this work. The images of different privacy levels are captured using a progressively modified NIR camera. The camera captures images, based on the extent of its modification, across multiple levels of privacy ranging from least private to most private. Given the subjective nature of privacy, two user surveys are conducted and analyzed to assess and identify a level of privacy acceptable to most people. Finally, a lightweight fire detection system is developed by utilizing the spatial and temporal properties of fire using a CNN model and the idea of frame differencing, respectively. The experimental results demonstrate that while the images preserve the privacy of occupants, the proposed framework is capable of comfortably detecting fire from these privacy-preserving images. A comparative analysis with the state-of-the-art techniques demonstrates the superiority of the proposed framework. A prototypical implementation on a Raspberry Pi device shows its applicability over a resource-constrained environment.

6.1.2 Privacy-preserving human activity recognition system

A privacy-preserving efficient human activity recognition system using depth sensor is proposed in this work. Two data modalities, depth clips and skeleton sequences extracted from depth clips are utilized for human activity recognition. Two novel descriptors based on the position of joint (JPD) and the angle between bones (BAD) are generated to model the spatial and temporal dynamics and activity. A multi-channel CNN architecture comprising a 3D-CNN for feature extraction from depth data and a two-channel 2D-CNN for feature extraction from skeleton data are employed and fused through a multi-level fusion strategy. The proposed framework is evaluated on four public datasets and found to be superior than the state of the art. Finally, the computational complexity analysis and a prototypical implementation of the proposed framework shows its applicability in the real-world.

6.1.3 Identity and activity privacy preserving posture recognition system

An identity and activity privacy preserving human posture recognition system using modified depth sensor is proposed in this part of the thesis. The privacy-preserving data of three different privacy levels are collected and the privacy is validated frob both the human and machine perspective. To establish the level of privacy that people, in general, are comfortable with, a user survey was conducted and analyzed to decide optimal level. To validate the privacy from the machine perspective, deep learning based models like a face recognition system for identity privacy and activity recognition systems for activity privacy are utilized. Moreover, to validate the utility of the privacy preserving data at all privacy levels, a posture recognition system is employed. Finally, an integrated posture recognition system based on CNN and decision tree is developed for posture recognition from the images captured at the identified optimal privacy level. The experimental results and privacy

vs. utility analysis demonstrate that while the captured data preserves both identity & activity privacy, it is also useful and effective in detecting postures. A computational complexity analysis confirms the applicability of the proposed system in the real-world.

6.2 Future Research Directions

Research is a never ending process, the contributions made in this thesis offer various avenues for further exploration. The following points provide the future directions to extend the outcomes of this thesis.

1. A privacy-preserving human activity recognition system using skeleton and depth data is proposed in chapter 4. The proposed system recognizes the activities from the segmented clips, which can further be extended to the detection and recognition of activities from streaming data. The activity recognition from non-segmented streaming, data will enable continuous monitoring.

2. The proposed human activity recognition system can also be further extended to behavioral analysis of the elderly. Behavioural analysis includes the detection of behavioural abnormalities in the individual over both the short term and the long term. Behavioural analysis may also be extended with the development of recommendation systems that predict and communicate the expected next activity to the elderly and/or dementia patients.

3. A privacy-preserving posture recognition system preserving both identity and activity privacy is proposed in chapter 5. The proposed posture recognition system can be further extended for the detection of coarse-grained activities (instead of postures) in a privacy-preserving manner. A course-grained activity recognition system that also preserves activity privacy will be very useful in elderly care and health care facilities.

4. A pose-estimation framework can be developed to estimate the prominent joints of the human body relevant for the posture/coarse-grained activity recognition from the privacy-preserving data. The key joints extracted from the privacy-preserving data captured using the modified depth sensor would enable the coarse-grained activity recognition using skeleton data in a resource-constrained environment.

The core contributions made in this thesis include development of privacy preserving indoor monitoring systems by leveraging characteristics of pre-capture privacy concepts. To ensure pre-capture privacy, various types of vision sensors were strategically modified and utilized in our work. The acceptable privacy level for residents, balancing privacy concerns and monitoring utility, was determined through user surveys and learning algorithms. Since the quality of privacy-preserving data captured by modified sensors differs from that of conventional vision sensors, particularly in terms of visibility, optimized learning algorithms are essential for effective analysis. We developed multiple machine learning algorithms to analyze privacy-preserving data without compromising accuracy.

The broader impact of this thesis lies in demonstrating the utility of pre-capture privacy mechanisms to safeguard privacy in indoor spaces, particularly in assisted living and healthcare facilities. Future research on privacy preservation in indoor monitoring can be extended to other contexts. Pose estimation from privacy-preserving data is a promising area to explore, as it could enable the effective classification of a broader range of coarse-grained activities. Additionally, the activity detection framework could be further developed to support behavioral anomaly detection, particularly for elderly individuals, enhancing its utility in assisted living and healthcare settings.

Bibliography

[1] T. Wang, L. Bu, Z. Yang, P. Yuan, and J. Ouyang, "A new fire detection method using a multi-expert system based on color dispersion, similarity and centroid motion in indoor environment," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 1, pp. 263–275, 2019.

[2] J. Saini, M. Dutta, and G. Marques, "A comprehensive review on indoor air quality monitoring systems for enhanced public health," *Sustainable environment research*, vol. 30, pp. 1–12, 2020.

[3] H.-C. Chang, Y.-L. Hsu, C.-Y. Hsiao, and Y.-F. Chen, "Design and implementation of an intelligent autonomous surveillance system for indoor environments," *IEEE Sensors Journal*, vol. 21, no. 15, pp. 17 335–17 349, 2021.

[4] M. Ricciuti, S. Spinsante, and E. Gambi, "Accurate fall detection in a top view privacy preserving configuration," *Sensors*, vol. 18, no. 6, p. 1754, 2018.

[5] A. Iazzi, M. Rziza, and R. Oulad Haj Thami, "Fall detection system-based posture-recognition for indoor environments," *Journal of imaging*, vol. 7, no. 3, p. 42, 2021.

[6] S. Nooruddin, M. M. Islam, F. A. Sharna, H. Alhetari, and M. N. Kabir, "Sensor-based fall detection systems: a review," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 5, pp. 2735–2751, 2022.

[7] J. Yin, J. Han, R. Xie, C. Wang, X. Duan, Y. Rong, X. Zeng, and J. Tao, "Mc-lstm: Real-time 3d human action detection system for intelligent healthcare applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 2, pp. 259–269, 2021.

[8] K. Chetry, H. Nath, A. Ahmed, S. Hazarika, and M. K. Muchahari, "A reliable patient indoor monitoring system based on iot in ambient-assisted living," in *Smart Intelligent Computing and Applications, Volume 2: Proceedings of Fifth International Conference on Smart Computing and Informatics (SCI 2021)*. Springer, 2022, pp. 271–279.

[9] J. Lin, R. Fu, X. Zhong, P. Yu, G. Tan, W. Li, H. Zhang, Y. Li, L. Zhou, and C. Ning, "Wearable sensors and devices for real-time cardiovascular disease monitoring," *Cell reports physical science*, vol. 2, no. 8, 2021.

[10] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5281–5292, 2022.

[11] W. Huang, L. Zhang, W. Gao, F. Min, and J. He, "Shallow convolutional neural networks for human activity recognition using wearable sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[12] A. Kamel, B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 9, pp. 1806–1819, 2018.

[13] J. P. Sá, M. C. M. Alvim-Ferraz, F. G. Martins, and S. I. Sousa, "Application of the

low-cost sensing technology for indoor air quality monitoring: A review," *Environmental Technology & Innovation*, vol. 28, p. 102551, 2022.

[14] P. S. Farahsari, A. Farahzadi, J. Rezazadeh, and A. Bagheri, "A survey on indoor positioning systems for iot-based applications," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7680–7699, 2022.

[15] H. Y. Yatbaz, S. Eraslan, Y. Yesilada, and E. Ever, "Activity recognition using binary sensors for elderly people living alone: Scanpath trend analysis approach," *IEEE Sensors Journal*, vol. 19, no. 17, pp. 7575–7582, 2019.

[16] L. Zhang, W. Cui, B. Li, Z. Chen, M. Wu, and T. S. Gee, "Privacy-preserving cross-environment human activity recognition," *IEEE Transactions on Cybernetics*, vol. 53, no. 3, pp. 1765–1775, 2023.

[17] Y. Dong, X. Li, J. Dezert, M. O. Khyam, M. Noor-A-Rahim, and S. S. Ge, "Dezert-smarandache theory-based fusion for human activity recognition in body sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 11, pp. 7138–7149, 2020.

[18] X. Wang, H. Yu, S. Kold, O. Rahbek, and S. Bai, "Wearable sensors for activity monitoring and motion control: A review," *Biomimetic Intelligence and Robotics*, vol. 3, no. 1, p. 100089, 2023.

[19] A. S. Dileep, S. Nabilah, S. Sreeju, K. Farhana, and S. Surumy, "Suspicious human activity recognition using 2d pose estimation and convolutional neural network," in *2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*. IEEE, 2022, pp. 19–23.

[20] K. Muhammad, S. Khan, M. Elhoseny, S. H. Ahmed, and S. W. Baik, "Efficient fire detection for uncertain surveillance environment," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 3113–3122, 2019.

[21] A. N. Wilson, K. A. Gupta, B. H. Koduru, A. Kumar, A. Jha, and L. R. Cenkeramaddi, "Recent advances in thermal imaging and its applications using machine learning: A review," *IEEE Sensors Journal*, vol. 23, no. 4, pp. 3395–3407, 2023.

[22] C. Yuan, Z. Liu, and Y. Zhang, "Fire detection using infrared images for uav-based forest fire surveillance," in *2017 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2017, pp. 567–572.

[23] X. Weiyao, W. Muqing, Z. Min, L. Yifeng, L. Bo, and X. Ting, "Human action recognition using multilevel depth motion maps," *IEEE Access*, vol. 7, pp. 41 811–41 822, 2019.

[24] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognition*, vol. 76, pp. 80–94, 2018.

[25] Z. Hussain, Q. Z. Sheng, and W. E. Zhang, "A review and categorization of techniques on device-free human activity recognition," *Journal of Network and Computer Applications*, vol. 167, p. 102738, 2020.

[26] S. Bian, M. Liu, B. Zhou, and P. Lukowicz, "The state-of-the-art sensing techniques in human activity recognition: A survey," *Sensors*, vol. 22, no. 12, p. 4596, 2022.

[27] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, "Human action recognition: A taxonomy-based survey, updates, and opportunities," *Sensors*, vol. 23, no. 4, p. 2182, 2023.

[28] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3200–3225, 2022.

[29] C. Zhang, J. Liang, X. Li, Y. Xia, L. Di, Z. Hou, and Z. Huan, "Human action recognition based on enhanced data guidance and key node spatial temporal graph convolution," *Multimedia Tools and Applications*, vol. 81, no. 6, pp. 8349–8366, 2022.

[30] A. A. Chaaraoui, J. R. Padilla-López, F. J. Ferrández-Pastor, M. Nieto-Hidalgo, and F. Flórez-Revuelta, "A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context," *Sensors*, vol. 14, no. 5, pp. 8895–8925, 2014.

[31] J. Liu, K. Wang, H. Yang, and N. Sun, "Visual privacy-preserving coding for video intelligence applications: A compressed sensing mechanism via bee-eye bionic vision," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[32] R. Zhao, Y. Zhang, T. Wang, W. Wen, Y. Xiang, and X. Cao, "Visual content privacy protection: A survey," *arXiv preprint arXiv:2303.16552*, 2023.

[33] F. Hellmann, S. Mertes, M. Benouis, A. Hustinx, T.-C. Hsieh, C. Conati, P. Krawitz, and E. André, "Ganonymization: A gan-based face anonymization framework for preserving emotional expressions," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[34] I. R. Dave, C. Chen, and M. Shah, "Spact: Self-supervised privacy preservation for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 164–20 173.

[35] T. Dayarathna, T. Muthukumarana, Y. Rathnayaka, S. Denman, C. de Silva, A. Pemasiri, and D. Ahmedt-Aristizabal, "Privacy-preserving in-bed pose monitoring: A fusion and reconstruction study," *Expert Systems with Applications*, vol. 213, p. 119139, 2023.

[36] J. Lee, D.-O. Woo, J. Jang, L. Junghans, and S.-B. Leigh, "Collection and utilization of indoor environmental quality information using affordable image sensing technology," *Energies*, vol. 15, no. 3, p. 921, 2022.

[37] A. Naser, A. Lotfi, M. D. Mwanje, and J. Zhong, "Privacy-preserving, thermal vision with human in the loop fall detection alert system," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 1, pp. 164–175, 2022.

[38] F. Pittaluga and S. J. Koppal, "Pre-capture privacy for small vision sensors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2215–2226, 2016.

[39] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021.

[40] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[41] A. Gaur, A. Singh, A. Kumar, A. Kumar, and K. Kapoor, "Video flame and smoke based fire detection algorithms: A literature review," *Fire technology*, vol. 56, pp. 1943–1980, 2020.

[42] X. Lin, J. Luo, M. Liao, Y. Su, M. Lv, Q. Li, S. Xiao, and J. Xiang, "Wearable sensor-based monitoring of environmental exposures and the associated health effects: a review," *Biosensors*, vol. 12, no. 12, p. 1131, 2022.

[43] N. Dai, I. M. Lei, Z. Li, Y. Li, P. Fang, and J. Zhong, "Recent advances in wearable electromechanical sensors—moving towards machine learning-assisted wearable sensing systems," *Nano Energy*, vol. 105, p. 108041, 2023.

[44] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–40, 2021.

[45] S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," *IEEE Access*, vol. 9, pp. 84 045–84 066, 2021.

[46] R. Yin, D. Wang, S. Zhao, Z. Lou, and G. Shen, "Wearable sensors-enabled human–machine interaction systems: from design to application," *Advanced Functional Materials*, vol. 31, no. 11, p. 2008936, 2021.

[47] M. A. Al-Qaness, A. M. Helmi, A. Dahou, and M. A. Elaziz, "The applications of metaheuristics for human activity recognition and fall detection using wearable sensors: A comprehensive analysis," *Biosensors*, vol. 12, no. 10, p. 821, 2022.

[48] S. Hayward, K. van Lopik, C. Hinde, and A. A. West, "A survey of indoor location technologies, techniques and applications in industry," *Internet of Things*, vol. 20, p. 100608, 2022.

[49] M. Carminati, G. R. Sinha, S. Mohdiwale, and S. L. Ullo, "Miniaturized pervasive sensors for indoor health monitoring in smart cities," *Smart Cities*, vol. 4, no. 1, pp. 146–155, 2021.

[50] F. Khan, Z. Xu, J. Sun, F. M. Khan, A. Ahmed, and Y. Zhao, "Recent advances in sensors for fire detection," *Sensors*, vol. 22, no. 9, p. 3310, 2022.

[51] A. Solórzano, J. Eichmann, L. Fernández, B. Ziems, J. M. Jiménez-Soto, S. Marco, and J. Fonollosa, "Early fire detection based on gas sensor arrays: Multivariate calibration and validation," *Sensors and Actuators B: Chemical*, vol. 352, p. 130961, 2022.

[52] M. H. Kashani, M. Madanipour, M. Nikravan, P. Asghari, and E. Mahdipour, "A systematic review of iot in healthcare: Applications, techniques, and trends," *Journal of Network and Computer Applications*, vol. 192, p. 103164, 2021.

[53] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," *Journal of Visual Communication and Image Representation*, vol. 77, p. 103116, 2021.

[54] Y. Liu, D. Yang, Y. Wang, J. Liu, J. Liu, A. Boukerche, P. Sun, and L. Song, "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," *ACM Computing Surveys*, 2023.

[55] J. Ren, F. Xia, Y. Liu, and I. Lee, "Deep video anomaly detection: Opportunities and challenges," in *2021 international conference on data mining workshops (ICDMW)*. IEEE, 2021, pp. 959–966.

[56] A. Gaur, A. Singh, A. Kumar, K. S. Kulkarni, S. Lala, K. Kapoor, V. Srivastava, A. Kumar, and S. C. Mukhopadhyay, "Fire sensing technologies: a review," *IEEE Sensors Journal*, vol. 19, no. 9, pp. 3191–3202, 2019.

[57] J. Pincott, P. W. Tien, S. Wei, and J. K. Calautit, "Indoor fire detection utilizing com-

puter vision-based strategies," *Journal of Building Engineering*, vol. 61, p. 105154, 2022.

[58] M. Valero, O. Rios, E. Pastor, and E. Planas, "Automated location of active fire perimeters in aerial infrared imaging using unsupervised edge detectors," *International journal of wildland fire*, vol. 27, no. 4, pp. 241–256, 2018.

[59] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–34, 2021.

[60] M. S. Momin, A. Sufian, D. Barman, P. Dutta, M. Dong, and M. Leo, "In-home older adults' activity pattern monitoring using depth sensors: A review," *Sensors*, vol. 22, no. 23, p. 9067, 2022.

[61] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30 509–30 555, 2020.

[62] C. Li, Q. Huang, X. Li, and Q. Wu, "Human action recognition based on multi-scale feature maps from depth video sequences," *Multimedia Tools and Applications*, vol. 80, no. 21, pp. 32 111–32 130, 2021.

[63] X. Chao, Z. Hou, J. Liang, and T. Yang, "Integrally cooperative spatio-temporal feature representation of motion joints for action recognition," *Sensors*, vol. 20, no. 18, p. 5180, 2020.

[64] S. Zheng, N. Apthorpe, M. Chetty, and N. Feamster, "User perceptions of smart home iot privacy," *Proceedings of the ACM on human-computer interaction*, vol. 2, no. CSCW, pp. 1–20, 2018.

[65] W. Li, T. Yigitcanlar, I. Erol, and A. Liu, "Motivations, barriers and risks of smart home adoption: From systematic literature review to conceptual framework," *Energy Research & Social Science*, vol. 80, p. 102211, 2021.

[66] C.-Y. Wang and F.-S. Lin, "Exploring older adults' willingness to install home surveil-lance systems in taiwan: Factors and privacy concerns," in *Healthcare*, vol. 11, no. 11. MDPI, 2023, p. 1616.

[67] A. McNeill, P. Briggs, J. Pywell, and L. Coventry, "Functional privacy concerns of older adults about pervasive health-monitoring systems," in *Proceedings of the 10th international conference on pervasive technologies related to assistive environments*, 2017, pp. 96–102.

[68] Q. M. Rajpoot and C. D. Jensen, "Video surveillance: Privacy issues and legal compliance," in *Promoting Social Change and Democracy through Information Technology*. IGI global, 2015, pp. 69–92.

[69] M. Amir, "The visual side of privacy: State-incriminating, coproduced archives," *Public Culture*, vol. 32, no. 1, pp. 185–213, 2020.

[70] S. S. Prasad, N. K. Mehta, A. Banerjee, H. Kumar, S. Saurav, and S. Singh, "Real-time privacy-preserving fall detection using dynamic vision sensors," in *2022 IEEE 19th India Council International Conference (INDICON)*. IEEE, 2022, pp. 1–6.

[71] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, "Visual privacy protection methods: A survey," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4177–4195, 2015.

[72] L. Rakhmawati *et al.*, "Image privacy protection techniques: A survey," in *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE, 2018, pp. 0076–0080.

[73] K.-S. Wong, N. A. Tu, A. Maratkhan, and M. F. Demirci, "A privacy-preserving framework for surveillance systems," in *2020 the 10th International Conference on Communication and Network Security*, 2020, pp. 91–98.

[74] X. Kong, Z. Meng, L. Meng, and H. Tomiyama, "A privacy protected fall detection iot system for elderly persons using depth camera," in *2018 International Conference on Advanced Mechatronic Systems (ICAMechS)*. IEEE, 2018, pp. 31–35.

[75] S. Nicolazzo, A. Nocera, and D. Ursino, "Anonymous access monitoring of indoor areas," *IEEE Access*, vol. 9, pp. 56 664–56 682, 2021.

[76] V. Srivastav, A. Gangi, and N. Padoy, "Human pose estimation on privacy-preserving low-resolution depth images," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 583–591.

[77] T. Xiao, Y.-H. Tsai, K. Sohn, M. Chandraker, and M.-H. Yang, "Adversarial learning of privacy-preserving and task-oriented representations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 434–12 441.

[78] S. Sathyanarayana, R. K. Satzoda, S. Sathyanarayana, and S. Thambipillai, "Vision-based patient monitoring: a comprehensive review of algorithms and technologies," *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, pp. 225–251, 2018.

[79] A. S. Rajput, B. Raman, and J. Imran, "Privacy-preserving human action recognition as a remote cloud service using rgb-d sensors and deep cnn," *Expert Systems with Applications*, vol. 152, p. 113349, 2020.

[80] S. Simonsson, F. D. Casagrande, and E. Zouganeli, "Location prediction in real homes of older adults based on k-means in low-resolution depth videos," in *2020*

25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 9046–9053.

[81] A. Rezaei, M. C. Stevens, A. Argha, A. Mascheroni, A. Puiatti, and N. H. Lovell, "An unobtrusive human activity recognition system using low resolution thermal sensors, machine and deep learning," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 115–124, 2023.

[82] E. Chou, M. Tan, C. Zou, M. Guo, A. Haque, A. Milstein, and L. Fei-Fei, "Privacy-preserving action recognition for smart hospitals using low-resolution depth images," *arXiv preprint arXiv:1811.09950*, 2018.

[83] A. Erdélyi, T. Winkler, and B. Rinner, "Privacy protection vs. utility in visual data: An objective evaluation framework," *Multimedia tools and applications*, vol. 77, pp. 2285–2312, 2018.

[84] J. Qi, L. Ma, Z. Cui, and Y. Yu, "Computer vision-based hand gesture recognition for human-robot interaction: a review," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 1581–1606, 2024.

[85] P. Foggia, A. Saggese, and M. Vento, "Real-time fire detection for video-surveillance applications using a combination of experts based on color, shape, and motion," *IEEE TRANSACTIONS on circuits and systems for video technology*, vol. 25, no. 9, pp. 1545–1556, 2015.

[86] F. Gong, C. Li, W. Gong, X. Li, X. Yuan, Y. Ma, and T. Song, "A real-time fire detection method from video with multifeature fusion," *Computational intelligence and neuroscience*, vol. 2019, 2019.

[87] S. Geetha, C. Abhishek, and C. Akshayanat, "Machine vision based fire detection techniques: A survey," *Fire technology*, vol. 57, no. 2, pp. 591–623, 2021.

[88] M. F. Bulbul, S. Islam, Z. Azme, P. Pareek, M. Kabir, H. Ali *et al.*, "Enhancing the performance of 3d auto-correlation gradient features in depth action classification," *International Journal of Multimedia Information Retrieval*, vol. 11, pp. 1–16, 2022.

[89] X. Liu and G. Zhao, "3d skeletal gesture recognition via discriminative coding on time-warping invariant riemannian trajectories," *IEEE Transactions on Multimedia*, vol. 23, pp. 1841–1854, 2020.

[90] B. Reily, Q. Zhu, C. Reardon, and H. Zhang, "Simultaneous learning from human pose and object cues for real-time activity recognition," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8006–8012.

[91] M. Sima, M. Hou, X. Zhang, J. Ding, and Z. Feng, "Action recognition algorithm based on skeletal joint data and adaptive time pyramid," *Signal, Image and Video Processing*, vol. 16, no. 6, pp. 1615–1622, 2022.

[92] N. Ghassempour, J. J. Zou, and Y. He, "A sift-based forest fire detection framework using static images," in *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*. IEEE, 2018, pp. 1–7.

[93] M. Mueller, P. Karasev, I. Kolesov, and A. Tannenbaum, "Optical flow estimation for flame detection in videos," *IEEE Transactions on image processing*, vol. 22, no. 7, pp. 2786–2797, 2013.

[94] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P. Z. Fulé, and E. Blasch, "Aerial imagery pile burn detection using deep learning: the flame dataset," *Computer Networks*, vol. 193, p. 108001, 2021.

[95] P. Barmpoutis, K. Dimitropoulos, K. Kaza, and N. Grammalidis, "Fire detection from images using faster r-cnn and multidimensional texture analysis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8301–8305.

[96] A. Banerjee, P. K. Singh, and R. Sarkar, "Fuzzy integral-based cnn classifier fusion for 3d skeleton action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2206–2216, 2020.

[97] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[98] B. Kim and J. Lee, "A video-based fire detection using deep learning models," *Applied Sciences*, vol. 9, no. 14, p. 2862, 2019.

[99] B. Debnath, M. O'Brient, S. Kumar, and A. Behera, "Attention-driven body pose encoding for human activity recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5897–5904.

[100] S. G. Kong, D. Jin, S. Li, and H. Kim, "Fast fire flame detection in surveillance video using logistic regression and temporal smoothing," *Fire Safety Journal*, vol. 79, pp. 37–43, 2016.

[101] S. Frizzi, R. Kaabi, M. Bouchouicha, J.-M. Ginoux, E. Moreau, and F. Fnaiech, "Convolutional neural network for video fire and smoke detection," in *IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2016, pp. 877–882.

[102] P. Li and W. Zhao, "Image fire detection algorithms based on convolutional neural networks," *Case Studies in Thermal Engineering*, vol. 19, p. 100625, 2020.

[103] K. Muhammad, J. Ahmad, I. Mehmood, S. Rho, and S. W. Baik, "Convolutional neural networks based fire detection in surveillance videos," *IEEE Access*, vol. 6, pp. 18 174–18 183, 2018.

[104] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Efficient deep cnn-based fire detection and localization in video surveillance applications," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 7, pp. 1419–1434, 2018.

[105] A. Jadon, M. Omama, A. Varshney, M. S. Ansari, and R. Sharma, "Firenet: A specialized lightweight fire & smoke detection model for real-time iot applications," *arXiv preprint arXiv:1905.11922*, 2019.

[106] A. J. Dunnings and T. P. Breckon, "Experimentally defined convolutional neural network architecture variants for non-temporal real-time fire detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 1558–1562.

[107] Y. Cao, F. Yang, Q. Tang, and X. Lu, "An attention enhanced bidirectional lstm for early forest fire smoke recognition," *IEEE Access*, vol. 7, pp. 154 732–154 742, 2019.

[108] D. Połap, "An adaptive genetic algorithm as a supporting mechanism for microscopy image analysis in a cascade of convolution neural networks," *Applied Soft Computing*, vol. 97, p. 106824, 2020.

[109] A. Farooq, F. Farooq, and A. V. Le, "Human action recognition via depth maps

body parts of action," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 12, no. 5, pp. 2327–2347, 2018.

[110] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 9–14.

[111] J. Trelinski and B. Kwolek, "Cnn-based and dtw features for human activity recognition on depth maps," *Neural Computing and Applications*, vol. 33, no. 21, pp. 14 551–14 563, 2021.

[112] C. Maldonado, S. Hernandez-Mendez, D. Torres-Muñoz, and C. Hernandez-Mejia, "Fall detection using features extracted from skeletal joints and SVM: Preliminary results," *Multimedia Tools and Applications*, vol. 81, no. 19, pp. 27 657–27 681, 2022.

[113] B. Ghojogh, H. Mohammadzade, and M. Mokari, "Fisherposes for human action recognition using kinect sensor data," *IEEE Sensors Journal*, vol. 18, no. 4, pp. 1612–1627, 2017.

[114] T. Huynh-The, C.-H. Hua, and D.-S. Kim, "Encoding pose features to images with data augmentation for 3-d action recognition," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3100–3111, 2019.

[115] J. Qi, Z. Wang, X. Lin, and C. Li, "Learning complex spatio-temporal configurations of body joints for online activity recognition," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 6, pp. 637–647, 2018.

[116] X. Ji, J. Cheng, W. Feng, and D. Tao, "Skeleton embedded motion body partition for human action recognition using depth sequences," *Signal Processing*, vol. 143, pp. 56–68, 2018.

[117] S. Ghodsi, H. Mohammadzade, and E. Korki, "Simultaneous joint and object trajectory templates for human activity recognition from 3-d data," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 729–741, 2018.

[118] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1290–1297.

[119] M. Akyash, H. Mohammadzade, and H. Behroozi, "A dynamic time warping based kernel for 3d action recognition using kinect depth sensor," in *2020 28th Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 2020, pp. 1–5.

[120] P. Hristov, "Real-time abnormal human activity detection using 1dcnn-lstm for 3d skeleton data," in *2021 12th National Conference with International Participation (ELECTRONICA)*, 2021, pp. 1–4.

[121] G. Arulselvi, D. Poornima, and S. J. Anand, "Privacy preserving elderly fall detection using kinect depth images based on deep convolutional neural networks," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 3, pp. 5492–5510, 2020.

[122] C. Liang, D. Liu, L. Qi, and L. Guan, "Multi-modal human action recognition with sub-action exploiting and class-privacy preserved collaborative representation learning," *IEEE Access*, vol. 8, pp. 39 920–39 933, 2020.

[123] A. Jain, R. Akerkar, and A. Srivastava, "Privacy-preserving human activity recognition system for assisted living environments," *IEEE Transactions on Artificial Intelligence*, 2023.

[124] H. Zong, H. Lei, Z. Jiao, and Z. Zhong, "Privacy-preserving automatic slipping de-

tection method for elderly in bathroom using depth sensors," in *2021 33rd Chinese control and decision conference (CCDC)*. IEEE, 2021, pp. 1990–1994.

[125] W. Mucha and M. Kampel, "Beyond privacy of depth sensors in active and assisted living devices," in *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, 2022, pp. 425–429.

[126] L. Jia and R. J. Radke, "Using time-of-flight measurements for privacy-preserving tracking in a smart room," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 689–696, 2013.

[127] Z. Cheng, T. Shi, W. Cui, Y. Dong, and X. Fang, "3d face recognition based on kinect depth data," in *2017 4th International Conference on Systems and Informatics (ICSAI)*. IEEE, 2017, pp. 555–559.

[128] Z. Feng and Q. Zhao, "Robust face recognition with deeply normalized depth images," in *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*. Springer, 2018, pp. 418–427.

[129] S.-k. Kwon, "Face recognition using depth and infrared pictures," *Nonlinear Theory and Its Applications, IEICE*, vol. 10, no. 1, pp. 2–15, 2019.

[130] J. Yan, F. Angelini, and S. M. Naqvi, "Image segmentation based privacy-preserving human action recognition for anomaly detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8931–8935.

[131] Z. Ma, Y. Liu, X. Liu, J. Ma, and K. Ren, "Lightweight privacy-preserving ensemble classification for face recognition," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5778–5790, 2019.

[132] M. A. Mousse and B. Atohoun, "Saliency based human fall detection in smart home environments using posture recognition," *Health Informatics Journal*, vol. 27, no. 3, p. 14604582211030954, 2021.

[133] W.-J. Wang, J.-W. Chang, S.-F. Haung, and R.-J. Wang, "Human posture recognition based on images captured by the kinect sensor," *International Journal of Advanced Robotic Systems*, vol. 13, no. 2, p. 54, 2016.

[134] M. Hamdi, H. Bouhamed, A. AlGarni, H. Elmannai, and S. Meshoul, "Deep learning and uniform lbp histograms for position recognition of elderly people with privacy preservation." *International Journal of Computers, Communications & Control*, vol. 16, no. 5, 2021.

[135] M. Gochoo, T.-H. Tan, F. Alnajjar, J.-W. Hsieh, and P.-Y. Chen, "Lownet: Privacy preserved ultra-low resolution posture image classification," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 663–667.

[136] K. Adhikari, H. Bouchachia, and H. Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network," in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, 2017, pp. 81–84.

[137] M. Gochoo, T.-H. Tan, S.-C. Huang, T. Batjargal, J.-W. Hsieh, F. S. Alnajjar, and Y.-F. Chen, "Novel iot-based privacy-preserving yoga posture recognition system using low-resolution infrared sensors and deep learning," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 7192–7200, 2019.

[138] C. Zhi-chao and L. Zhang, "Key pose recognition toward sports scene using deeply-learned model," *Journal of Visual Communication and Image Representation*, vol. 63, p. 102571, 2019.

[139] M. Ahrens, "Home structure fires," 2019. [Online]. Available: https://www.nfpa.org/

[140] J. D. Burnett and M. G. Wing, "A low-cost near-infrared digital camera for fire detection and monitoring," *International journal of remote sensing*, vol. 39, no. 3, pp. 741–753, 2018.

[141] SCHOTT Technical Information TIE-29, "Refractive index and dispersion," SCHOTT North America, Inc., Tech. Rep., Feb 2016. [Online]. Available: https://www.us.schott.com/$advanced_optics/$

[142] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[143] S. Winkler, *Digital video quality: vision models and metrics*. John Wiley & Sons, 2005.

[144] G. Batchuluun, D. T. Nguyen, T. D. Pham, C. Park, and K. R. Park, "Action recognition from thermal videos," *IEEE Access*, vol. 7, pp. 103 893–103 917, 2019.

[145] G. T. Papadopoulos, A. Axenopoulos, and P. Daras, "Real-time skeleton-tracking-based human action recognition using kinect data," in *International Conference on Multimedia Modeling*. Springer, 2014, pp. 473–483.

[146] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2016, pp. 770–778.

[147] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 6299–6308.

[148] S. C. Sen-Ching and C. Kamath, "Robust techniques for background subtraction in urban traffic video," in *Visual Communications and Image Processing 2004*, vol. 5308. SPIE, 2004, pp. 881–892.

[149] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[150] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2015, pp. 648–656.

[151] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International conference on image processing (ICIP)*. IEEE, 2015, pp. 168–172.

[152] S. Gasparrini, E. Cippitelli, E. Gambi, S. Spinsante, J. Wåhslén, I. Orhan, and T. Lindh, "Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion," in *International conference on ICT innovations*. Springer, 2015, pp. 99–108.

[153] T. Xu and Y. Zhou, "Elders' fall detection based on biomechanical features using depth camera," *International journal of wavelets, multiresolution and information processing*, vol. 16, no. 02, p. 1840005, 2018.

[154] "World population prospects 2022," 2022. [Online]. Available: https://desapublications.un.org/

[155] (2023) Older adult fall prevention. Accessed: 2024-03-01. [Online]. Available: https://www.cdc.gov/falls/data/

[156] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[157] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[158] T. T. Zin, S. Thant, M. Z. Pwint, and T. Ogino, "Handwritten character recognition on android for basic education using convolutional neural network," *Electronics*, vol. 10, no. 8, p. 904, 2021.

[159] S. Shamim, M. B. A. Miah, A. Sarker, M. Rana, and A. Al Jobair, "Handwritten digit recognition using machine learning algorithms," *Indonesian Journal of Science and Technology*, vol. 3, no. 1, pp. 29–39, 2018.

[160] A. Jain and A. Srivastava, "Privacy-preserving efficient fire detection system for indoor surveillance," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3043–3054, 2021.

[161] C. Bisogni, L. Cimmino, M. De Marsico, F. Hao, and F. Narducci, "Emotion recognition at a distance: The robustness of machine learning based on hand-crafted facial features vs deep learning models," *Image and Vision Computing*, vol. 136, p. 104724, 2023.

[162] N. Bento, J. Rebelo, M. Barandas, A. V. Carreiro, A. Campagner, F. Cabitza, and H. Gamboa, "Comparing handcrafted features and deep neural representations for domain generalization in human activity recognition," *Sensors*, vol. 22, no. 19, p. 7324, 2022.

[163] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth

for driver pose estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5494–5503.

Appendix A

Supplementary Results (Fire Detection System

A.1 Analysis of Minkowski Distance Metric

We analyzed the proposed ST-FDS using the Minkowski distance metric as given in Equation (3.6). The results with different values of parameter m are included in Table A.1.

Table A.1: Performance and speed comparison for different values of m in Minkowski Metric

Parameter (m)	**Threshold**	**Accuracy %**	**Precision %**	**Recall %**	**F-Score %**	**Processing Speed (fps)**
MSE	36	98.13	96.38	100	98.16	70
m=1	3	95.78	92.20	100	95.94	53
m=2	36	98.13	96.38	100	98.16	52
m=3	800	98.23	96.56	100	98.25	23
m=4	25000	98.28	96.66	100	98.30	23
m=5	632410	98.29	96.68	100	98.31	23
m=10	9.6 E+16	98.71	97.62	99.85	98.72	23

The summary of the findings in Table A.1 are summarized in the following points.

1. The Accuracy, Precision, and F-Score increase with increasing values of m but the increase is not significant.

2. The Recall decreases when the value of m is very high (i.e. $m = 10$).

3. The processing speed is the number of distances calculated per second. The processing speed is less than half (i.e. 23 fps) when the value of $m > 2$.

4. The MSE and Minkowski with $m = 2$ are the same. The processing speeds, however, are different owing to the absolute function used in the Minkowski method which is not the case in MSE, thanks to the even power.

5. We do not calculate the m^{th} root of the distance (similar to MSE) as it is a time saving approach. For analysis purposes, we compute the time taken to calculate the MSE with square root which is much higher than the original MSE without square root. The processing speed for MSE with square root is 65 fps which is slower than the original MSE.

6. Processing speeds were calculated on a PC with an i7 Processor and 16 GB memory.

7. According to the data given in Table A.1, MSE/Minkowski with $m = 2$ is the most suitable metric for distance while keeping performance and speed in mind.

In accordance with the above analysis, $m = 2$ is used for distance calculation in Equation (3.6).

A.2 Comparison with Original SqueezeNet

The proposed SA-FDS is an adaption of the SqueezeNet architecture, with the significant modifications. A diagram highlighting the key differences between the original SqueezeNet architecture and the proposed SA-FDS is shown in Figure A.1. A comprehensive discussion on the differences between the two architectures is as follow:

1. An important aspect of this work is the development of a light-weight system that can be easily deployed over resource constrained environments. To achieve this, we use only three fire modules as compared to eight fire modules in the original SqueezeNet.

2. We use fewer filters in each fire module as compared to the original SqueezeNet, again to reduce the model size.

3. We use a max pooling layer after each fire module except the last one whereas SqueezeNet uses the pooling layer after every 3 or 4 fire modules. As the architecture in our case is small, hence a pooling layer after each fire module is useful to down-sample the image faster as compared to a large model.

4. We use a dropout after each fire module which is not the case in the original SqueezeNet. Dropout forces the model to learn and avoid over-fitting . This is important as the model is trained on relatively small datasets as compared to the very large ImageNet dataset.

5. The input for SA-FDS is a single channel grayscale image, whereas SqueezeNet takes a three channel color image as input.

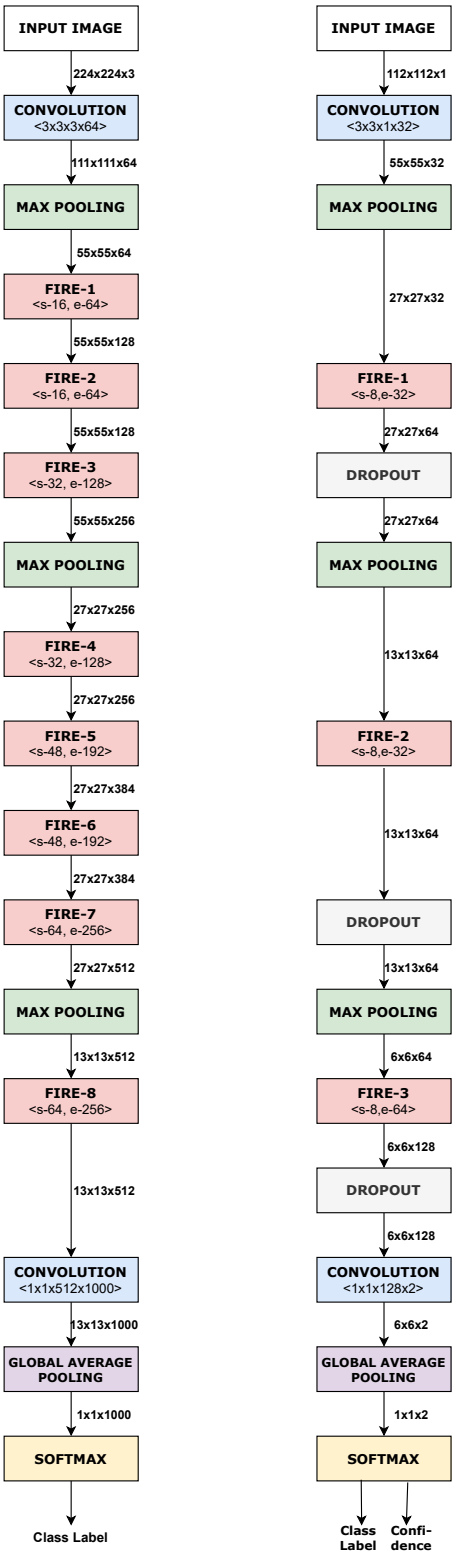


Figure A.1: Comparison of the proposed SA-FDS architecture with the original SqueezeNet (SqueezeNet on the left; SA-FDS on the right)

Appendix B

Supplementary Results (Human Activity Recognition)

B.1 Dataset Description

A detailed description and visual examples of a few activity samples from various datasets are included in this section. The depth maps in the visual examples are converted to color-maps for better visualization. The skeleton joints are displayed in 2D geometry.

B.1.1 MSR Action3D Dataset

The MSR Action3D dataset is captured using a depth camera similar to the Microsoft Kinect device along two data modalities: depth maps and skeleton sequences. The data set comprises 20 actions performed by ten subjects in 2-3 repetitions, resulting in 567 depth video clips and an equal number of skeleton sequences. The dataset is divided into three subsets, namely AS1, AS2, and AS3. Actions with similar movements are grouped into AS1 and AS2, and complex activities are included in AS3. Table B.1 gives a complete list of activities in each action set. Visual examples of three activities containing depth frames and skeleton joints are shown in Fig B.1.

Table B.1: List of activities in MSR Action3D dataset.

Action Set 1 (AS1)	**Action Set 2(AS2)**	**Action Set 3 (AS3)**
Horizontal Arm Wave	High Arm Wave	High Throw
Hammer	Hand Catch	Forward Kick
Forward Punch	Draw X	Side Kick
High Throw	Draw Tick	Jogging
Hand Clap	Draw Circle	Tennis Swing
Bend	Two Hand Wave	Tennis Serve
Tennis Serve	Forward Kick	Golf Swing
Pickup & Throw	Side Boxing	Pickup & Throw

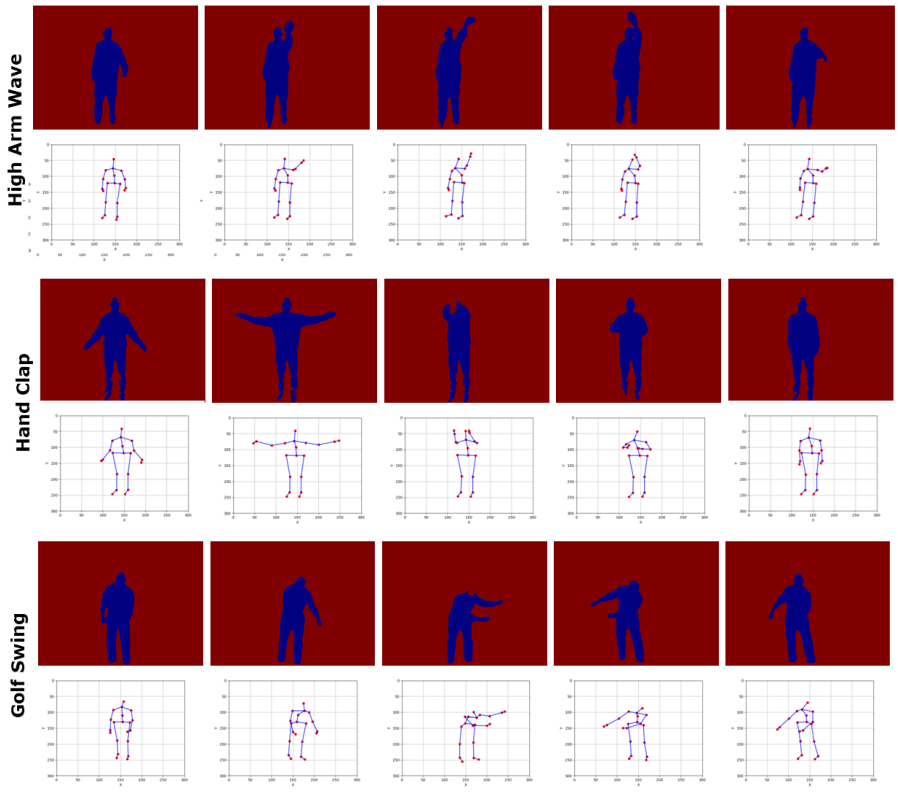


Figure B.1: Visual examples (Depth frames and Skeleton joints) of three activities: High Arm Wave; Hand Clap; and Golf Swing; from MSR Action3D dataset.

B.1.2 UTD MHAD Dataset

The UTD MHAD (Multimodel Human Action Dataset) is captured using a Microsoft Kinect device and an inertial sensor along four data modalities: RGB clips, depth maps, skeleton sequences, and sensor readings. The dataset comprises 27 actions performed by eight subjects in four repetitions, resulting in 861 action instances. A complete list of actions included in the dataset is given in Table B.2. Visual examples of three activities containing depth images and skeleton joints are shown in Figure B.2.

Table B.2: List of activities in UTD MHAD dataset

Activities	
1. Right Arm Swipe to Left	15. Tennis Right Hand Forehand Swing
2. Right Arm Swipe to Right	16. Arm Curl (Two Arms)
3. Right Hand Wave	17. Tennis Serve
4. Two Hand Front Clap	18. Two Hand Push
5. Right Arm Throw	19. Right Hand Knock on Door
6. Cross Arm in the Chest	20. Right Hand Catch an Object
7. Basketball Shoot	21. Right Hand Pickup & Throw
8. Right Hand Draw X	22. Jogging in Place
9. Right Hand Draw Circle (Clockwise)	23. Walking in Place
10. Right Hand Draw Circle (Counter-clockwise)	24. Sit to Stand
11. Draw Triangel	25. Stand to Sit
12. Bowling (Right Hand)	26. Forward Lunge (Left Foot Forward)
13. Front Boxing	27. Squat (Two Arm Stretch Out)
14. Baseball Swing From Right	

B.1.3 TST Fall Dataset

The TST Fall detection dataset is generated using a Microsoft Kinect device and a wearable inertial sensor along three data modalities: depth maps, skeleton sequences, and acceleration data. The dataset comprises 8 actions performed by 11 actors in three repetitions, resulting in 264 action instances. The size of the depth maps is 512×424, and the skeleton sequence contains 25 joints in each frame. The actions in the dataset are grouped into two categories: Fall and Activities of Daily Living (ADL). A complete list of activities in

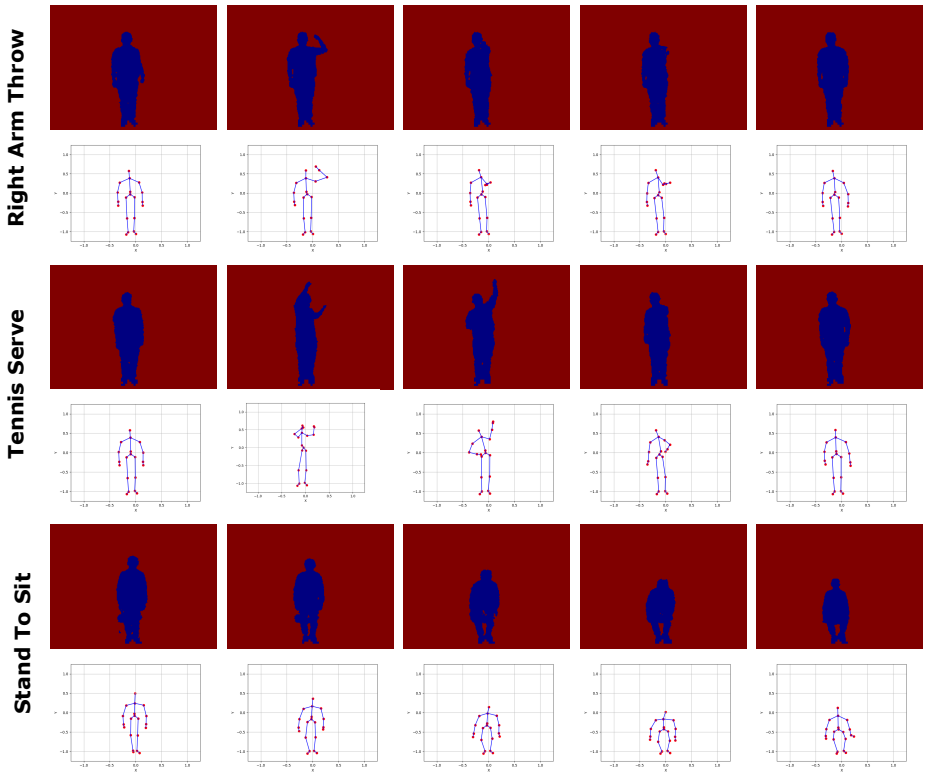


Figure B.2: Visual examples (Depth frames and Skeleton joints) of three activities: Right Arm Throw; Tennis Serve; and Stand to Sit; from UTD MHAD dataset.

both categories is given in Table B.3. Visual examples for three activities containing depth frames and skeleton joints are shown in Fig B.3. In contrast to the above datasets, the depth frames in this dataset contain background information.

Table B.3: List of activities in TST Fall dataset.

ADLs	**Fall**
Grasp Object from Floor	Fall from Front
Sit on a Chair	Fall Backward
Walk	Fall from Side
Lay Down	Fall backward & End up Sit

B.1.4 MSR Daily Activity Dataset

The MSR Daily Activity dataset is also captured using a Kinect device along three modalities: RGB clips, depth maps, and skeleton sequence information. The dataset con-

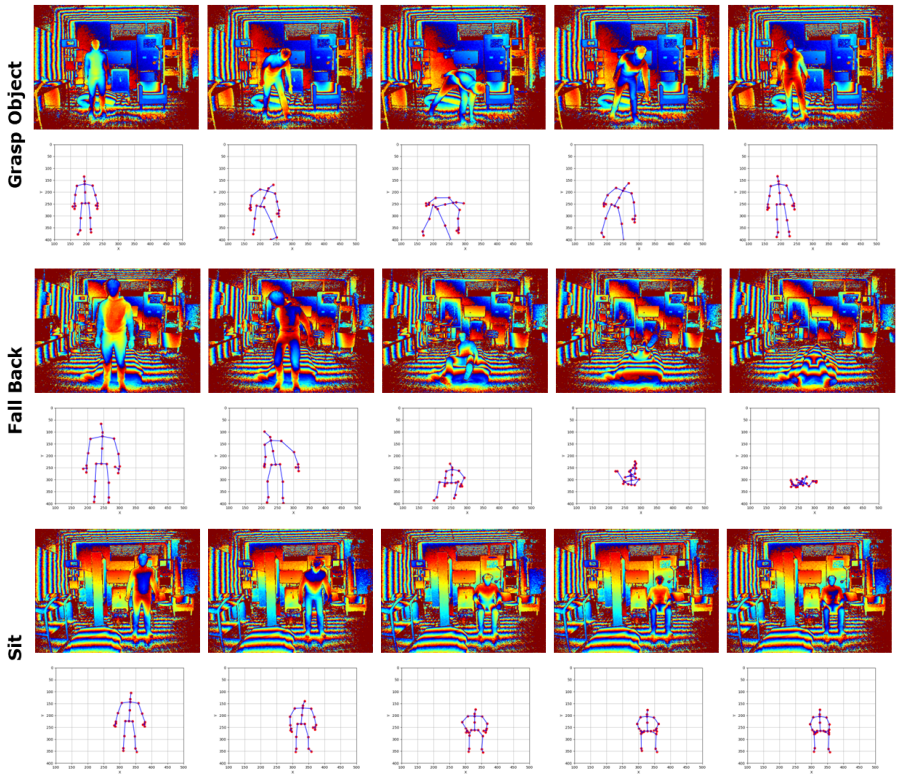


Figure B.3: Visual examples (Depth frames and Skeleton joints) of three activities: Grasp Object; Fall Backward; and Sit on a Chair; from TST Fall dataset.

tains 320 clips of 16 actions performed by ten individuals, twice each (one in a standing position and the other in a sitting position). A complete list of activities included in this dataset is included in Table B.4. The dataset is quite noisy, and the depth clips are not fully clean. The background of the depth maps is partially subtracted, and close objects like the sofa, side table, and other small objects are still visible. A few visual examples showing the movement of body parts for the given activity are shown in Fig B.4.

Table B.4: List of daily activities in MSR Daily Activity dataset.

Activities	
1. Drink	9. Sit Still
2. Read	10. Toss Paper
3. Read Book	11. Play Game
4. Call Cellphone	12. Lay Down on Sofa
5. Write on a Paper	13. Walk
6. Use Laptop	14. Play Guitar
7. Use Vaccume Cleaner	15. Stand Up
8. Cheer Up	16. Sit Down

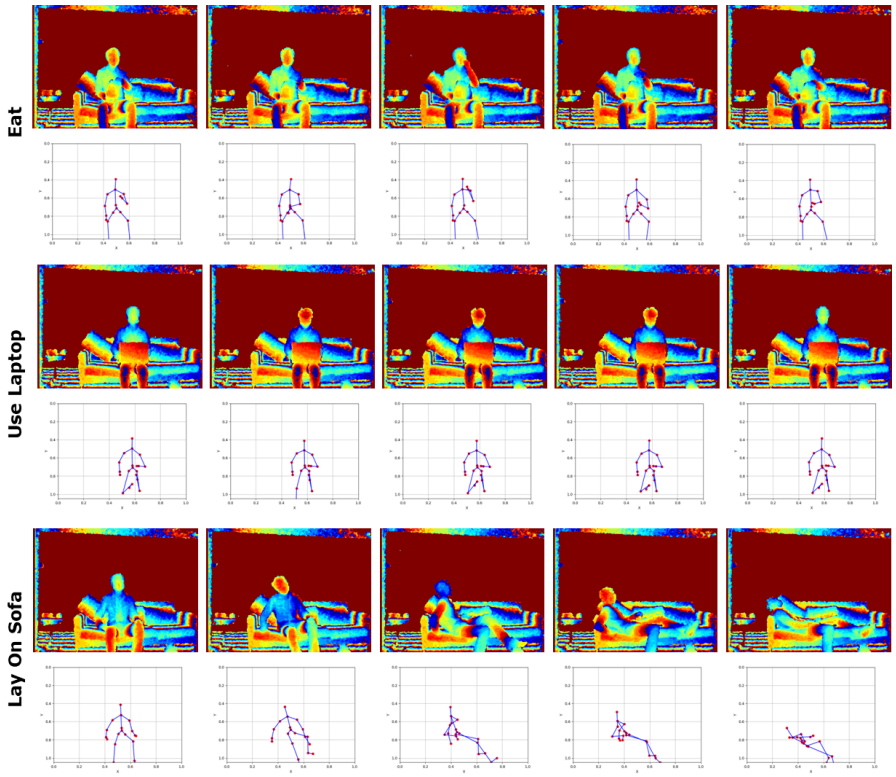


Figure B.4: Visual examples (Depth frames and Skeleton joints) of three activities: Eat; Use Laptop; and Lay on Sofa; from MSR Daily Activity dataset.

B.2 Model Selection

A selection of the most suitable models and parameters is crucial for the optimal performance. We experimented with different architectures and set of parameters for selecting optimal models. The results of these experiments are included in this section.

B.2.1 2DCNN

To select the best 2DCNN model for our work, we experimented with eight well-known 2D-CNN models in similar settings. Table B.5 provides a performance comparison of the CNN models using both the JPD based images and the BAD based images. It can be seen that the Resnet50 model outperforms the rest on both types of images, whereas DenseNet201 has the next best results. We further modified the ResNet50 and DenseNet201 models as discussed in Section 3.2.2(in the paper). The modification leads to significant performance improvement of ResNet50. We therefore use the modified ResNet50 model for the 2D-CNN.

CNN Model	JPD Accuracy(%)	BAD Accuracy (%)
MobileNet	85.45	80.72
VGG16	86.10	81.45
InceptionV3	87.27	78.54
Xception	89.45	79.63
DenseNet121	87.99	79.63
DenseNet201	89.45	82.18
ResNet50	92.00	85.45
Mod-DenseNet201	90.17	83.27
Mod-Resnet50	**93.61**	**88.73**

Table B.5: Performance comparison of 2DCNN models on MSR Action3D dataset using Evaluation Setting(1).

B.2.2 3DCNN

An Inflated 3D (I3D) CNN, which has demonstrably superior performance in video classification, is utilized in our work. To select the optimal configuration for our 3D-CNN, we experiment with different combinations of augmentation methods and temporal lengths. The performance comparison of some of the combinations of augmentation methods is shown in

Table B.6. We observed significant performance improvement with an adequate amount of augmentation as compared to training without augmentation. At the same time, augmentation beyond a point degraded the performance. Use was made, therefore, of the combination shown in the third row of Table B.6.

Augmentation	**Accuracy (%)**
None	88.00
Rotate (10%), Speed (20%), HFlip	90.55
Rotate (15%), Speed (30%), Resize (6%), Translate (20%), Random Crop, HFlip	**93.45**
Rotate (20%), Speed (40%), Resize (10%), Translate (20%), Random Crop, HFlip	92.00

Table B.6: Performance comparison of 3DCNN with different combinations of augmentation methods

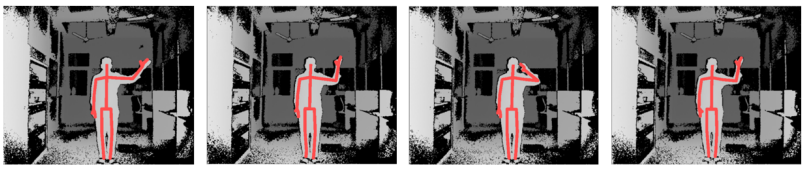
We also modify the I3D model, as discussed in Section 3.2.1 (in the paper). The modification led to some improvement in performance whilst also avoiding over-fitting. The performance on the original I3D and the modified I3D is shown in Table B.7.

Model	**Accuracy (%)**
I3D	93.45
Mod-I3D	**94.18**

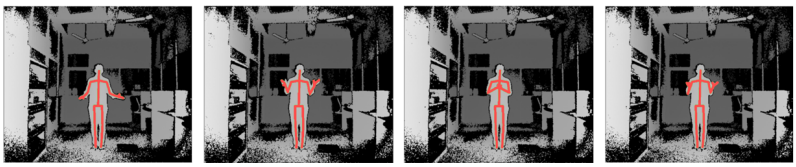
Table B.7: Performance comparison on 3DCNN before and after modification

B.3 Prototype Dataset

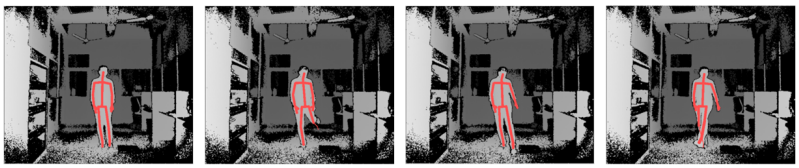
As part of a prototypical implementation, we created a small dataset of 104 activity instances containing 5 different activities. Each activity is performed by 6 actors in two/three iterations. Figure B.5 shows the depth frames from activity clips of five activities in prototype dataset. The red lines in on the bodies in the depth frames show the lines connecting the skeleton joints.



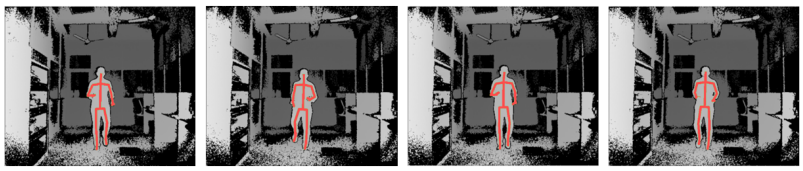
(a) Arm Wave activity



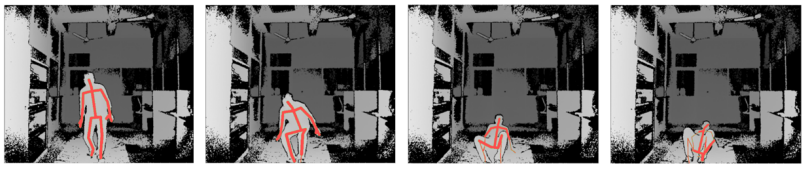
(b) Hand Clap activity



(c) Forward Kick activity



(d) Walking activity



(e) Falling activity

Figure B.5: Sample depth frames (skeleton are marked with red color in depth frame) of five activities from prototype dataset.

Figure B.6 shows the sample JPD and BAD based images for different activities from our

dataset generated using the algorithms mentioned in our paper. It is evident from Figure B.6 that both JPD and BAD based images are sufficiently distinguishable and would facilitate the classification of different activities.

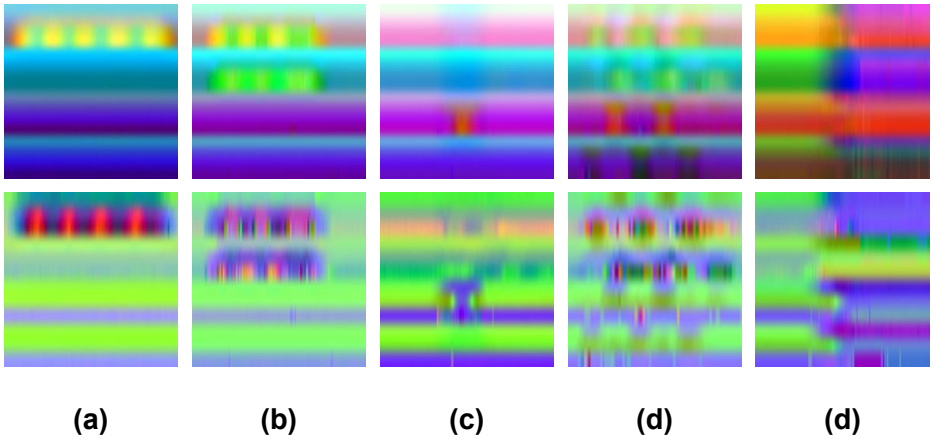


Figure B.6: Sample images for five activities (JPD-top, BAD-bottom). a) Arm Wave; b) Hand Clap; c) Forward Kick; d) Walking; e) Falling