



Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

An efficient and lightweight multiperson activity recognition framework for robot-assisted healthcare applications

Syed Hammad Hussain Shah ^{a,*}, Anniken Susanne T. Karlsen ^a, Mads Solberg ^b, Ibrahim A. Hameed ^a

^a Department of ICT and Natural Sciences, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Larsgårdsvegen 2, Ålesund, 6009, Norway

^b Department of Health Sciences, Faculty of Medicine and Health Science, Norwegian University of Science and Technology (NTNU), Larsgårdsvegen 2, Ålesund, 6009, Norway

ARTICLE INFO

Keywords:

Human activity recognition
Multiperson activity recognition
Exercise recognition
Robot-assisted rehabilitation
Virtual coaches
Eldercare

ABSTRACT

Aging is inevitably associated with a decline in physical abilities and can pose challenges to the social lives of elderly individuals. In long-term care facilities, group exercise is instrumental for keeping elderly residents physically and socially healthy. Accommodating these needs in elderly care can be challenging due to staff shortages and other lacking resources. A robotic exercise coach could be helpful in such contexts. Intelligent human-robot interaction requires accurate and efficient human activity recognition. Several solutions focusing on human activity recognition in healthcare robotics have been proposed. However, multiperson activity recognition remains a challenging task in case of using vision-based or wearable sensors data, and past research has mainly focused on single-person rather than multiperson or group activity recognition. Moreover, the existing state-of-the-art methods for activity recognition mainly use heavyweight Convolutional Neural Network (CNN) models to achieve good accuracy. However, these models have certain drawbacks, such as requiring significant computational resources, higher memory and storage needs, and slower inference times. Another challenge is the limited number of publicly available datasets containing few activities for physical activity recognition. In this work, we propose a lightweight, deep learning-based, multiperson activity recognition system for group exercise training of elderly persons. Considering the limited publicly available datasets, we curated a new dataset named the Routine Exercise Dataset (RED), comprising 19 routine exercise activities recommended for elderly persons. The RED dataset has 14,440 samples collected from 19 participants and is one of the most extensive datasets of its kind. We evaluated our proposed activity recognition method based on proposed feature extraction modules and a one-dimensional multilayer long short-term memory network on 16 datasets, including 10 publicly available benchmark activity recognition datasets, an RED dataset, a publicly available dataset combined with RED dataset, and four noise-corrupted RED datasets. The results indicate the efficiency of the proposed method for real-time activity recognition compared to the state-of-the-art methods. The proposed method achieved F1-scores of 98.64%, 97.95%, and 99% on large-scale datasets named UESTC RGB-D, NTU RGB+D, and RED, respectively. We also developed a Robot Operating System (ROS)-based application to deploy our proposed system in a social robot and test it in real-life scenarios.

1. Introduction

Intelligent robotic systems have the potential to offer assistance in many domains, including healthcare (Jamil et al., 2022a; Zaabar et al., 2021), education, entertainment, manufacturing, and other industries (Jamil et al., 2022b). Globally, the proportion of people above 65 in the general population is increasing in many countries and is predicted to rise further in the following decades. Inadequate social

engagement and physical inactivity are common among older adults, negatively affecting their health (Jaarsma et al., 2015). Many research studies suggest that elderly individuals may feel more motivated to exercise in a social group rather than exercising alone (Ahmad et al., 2022; Shah et al., 2022a, 2022b). Therefore, group exercise is usually recommended by healthcare professionals and often administered as

* Corresponding author.

E-mail addresses: syed.h.h.shah@ntnu.no (S.H.H. Shah), anniken.t.karlsen@ntnu.no (A.S.T. Karlsen), mads.solberg@ntnu.no (M. Solberg), ibib@ntnu.no (I.A. Hameed).

<https://doi.org/10.1016/j.eswa.2023.122482>

Received 11 June 2023; Received in revised form 4 November 2023; Accepted 5 November 2023

Available online 22 November 2023

0957-4174/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

a joint activity in care facilities. However, daily group activity sessions can be challenging for healthcare professionals since time and other resources are limited. Integrating intelligent robotic systems in healthcare to perform such jobs could mitigate these challenges in the future (Blindheim et al., 2023). To be useful, these systems must be able to recognize and understand specific and relevant human behaviors for successful human–robot interaction. In this context, Human Activity Recognition (HAR) has emerged as a promising development. HAR can be used in healthcare to monitor and detect changes in the activity of patients (Schrader et al., 2020; Taylor et al., 2020) and be integrated with personalized smart devices and smart-home solutions based on individual needs (Mekruksavanich & Jitpattanakul, 2021). Many existing systems use powerful wearable sensors to acquire time series data for HAR (Andrade-Ambriz et al., 2022). Among these sensor systems, the Motion Capture (MoCap) system can capture and transmit motion data from human users. MoCap has a high level of accuracy when reproducing intricate movements. However, it is an expensive technology and requires specialized training to be used precisely (Stumpf, 2010). Furthermore, sensors such as gyroscopes, accelerometers, magnetometers, and Global Positioning Systems (GPS) embedded in smartphones or Inertial Measurement Units (IMU) are also used to acquire time series data for HAR (Ihianle et al., 2020; Qiu et al., 2022). The data are collected using multiple wearable sensors attached to the wrist, ankle, and chest. HAR systems relying on wearable or smartphone sensor data are popular because of their light weight and ease of collection. However, achieving better accuracy, efficiency, and effectiveness with less computational and financial cost in multiperson activity recognition remains challenging. Using wearable or smartphone-embedded sensors for multiperson activity recognition is computationally and economically costly. Moreover, carrying these sensors can be cumbersome for older people, making invasive HAR methods challenging, especially for multiperson activity recognition.

Vision-based HAR using RGBD images and depth information acquired by an RGB camera and an infrared projector is also a popular and cost-effective approach. Several cameras, such as Intel RealSense Depth Camera and Microsoft Kinect, are available on the market to collect RGB-D images. However, processing color and depth images for HAR is computationally expensive. Researchers have therefore proposed several cost-effective vision-based solutions for HAR. For instance, Andrade-Ambriz et al. (2022) proposed a method for HAR using 3D convolutions and convolutional Long Short-Term Memory (LSTM) on short video sequences.

The work presented in this paper is motivated by pressing challenges faced by healthcare professionals working in long-term care facilities for elderly individuals in Norway. In our interviews with the healthcare professionals, we learned that conducting group activities such as exercise or fun activities can become challenging to offer due to a lack of resources. Group activities are often only provided once or twice weekly for 15 to 30 min, which is not unique in the Norwegian context. A US-based survey (CareerStaff Unlimited, 2023) suggests that 98% of long-term care providers are experiencing difficulties in hiring staff, with this shortage in labor having consequences for the level of services staff are able to provide. Moreover, continuously monitoring residents located in different areas of care facilities can be challenging. Accordingly, there is a growing need for intelligent systems to assist staff in care environments. Past research has focused on noninvasive or vision-based solutions (Andrade-Ambriz et al., 2022; Khan et al., 2022) for single-person activity recognition. On the other hand, invasive methods (Cippitelli et al., 2016; Gaglio et al., 2014; García-de Villa et al., 2022) based on sensors attached to the body are not feasible for multiperson activity recognition due to the cost and hardware complexity. The complexity of invasive methods for multiperson activity recognition extends beyond cost and hardware complexity. Another reason these methods may not be feasible is the practical challenges of simultaneously attaching sensors to multiple individuals. This process can be time-consuming, intrusive, and uncomfortable for

the participants, limiting the usability and acceptance of such methods in real-world scenarios. Additionally, the maintenance and calibration of multiple sensor attachments for each person can be cumbersome, making it less practical for large-scale applications. Therefore, the limitations extend beyond cost and hardware complexity to encompass practicality, user comfort, and scalability. Furthermore, the approaches proposed for multiperson or group activity recognition are mainly focused on surveillance (Ullah et al., 2021) or sports activities (Gavrilyuk et al., 2020) and not physical exercises. Moreover, these state-of-the-art methods are based on heavyweight CNN models to achieve good accuracy, regardless of the limitations of these models. Such limitations include the need for significant computational resources, higher memory and storage requirements, slower inference times, and the potential for overfitting on scarce datasets.

To the best of our knowledge, there are currently no existing methods for handling multiperson activity recognition, which is lightweight and can be used for conducting group exercises of elderly individuals using social robots in real time. Dealing with multiperson activity recognition in changing environments where the position of persons in video frames keeps changing is a challenging task. Furthermore, activity recognition based on ground robot vision is more challenging than that based on stationary CCTV cameras due to various factors, such as occlusions, limited view, and unstructured or diverse environments. Limited publicly available datasets comprising routine upper and lower body exercises required for maintaining physical health in older people remain a significant problem. To address these knowledge gaps, our contributions in the present paper are as follows:

- We proposed a lightweight vision-based multiperson activity recognition system that can be easily integrated into a robot or other interactive technologies in real time for group exercise training of elderly individuals.
- A principal step for multiperson activity recognition is to localize the persons throughout the video stream by detecting and tracking each individual. The object detection models trained on data comprised of objects from general categories are not feasible in this task. Therefore, we have fine-tuned a lightweight CNN model and trained it on new data for person detection to achieve better results in a dynamic environment. In the following step, we used an object tracker, which can perform tracking at a very high speed, to track the persons in the video stream.
- LSTM is well known for learning sequential data. We presented an efficient method for activity recognition based on a frame skip strategy, feature extraction, and a Multilayer Long Short-Term Memory (MLSTM) network. The use of multiple LSTM layers in MLSTM enables the network to learn abstract and more complex representations of sequential data.
- The lack of publicly available datasets comprising routine exercise activities required for elderly individuals is a significant problem. Therefore, we curated a new dataset called the Routine Exercise Dataset (RED) in this work. This dataset consists of multiple upper-body and lower-body physical exercises specifically designed for the physical well-being of older adults. The dataset includes 19 different exercises performed 40 times by each of the 19 participants.
- We conducted an ablation study to evaluate the performance of the proposed activity recognition method with existing methods on various datasets ($n = 16$), including the RED dataset without noise, four noise-corrupted RED datasets, RED dataset combined with a publicly available dataset, and 10 publicly available activity recognition datasets. The proposed method proved generalizable to exercise and daily activities with high accuracy and minimal loss.
- We developed a Robot Operating System (ROS)-based (Stanford Artificial Intelligence Laboratory, 2018) application to deploy our proposed system on a humanoid robot and to test it in real-life scenarios of group exercises.

The remaining content of the paper is organized as follows: Section 2 presents an overview of the literature related to the present work. The implementation environment and the data acquisition process are presented in Sections 3 and 4, respectively. The methodology for the multiperson activity recognition system and our architecture of the ROS-based application for testing it in a real-life environment are described in Section 5. Section 6 details the experimentations and results, Section 7 presents the general discussion about the contributions of this research and some limitations of existing works, and finally, we offer some conclusions and potential directions for future research in Section 8.

2. Related work

Human-machine interaction requires developing systems that can identify and replicate specific actions executed by humans. Several methodologies allow robots to recognize human gestures (Ding & Chang, 2015; Shah et al., 2021), detect vocal directives (Ding & Shi, 2017), or both, facilitating natural interaction with human beings. Accurate and rapid detection of executed actions remains an obstacle to recognizing and acting on human activity. Human Activity Recognition (HAR) approaches have been proposed for identifying actions that can contribute to fall aversion, as falls pose a severe risk for elderly individuals (Flores-Barranco et al., 2015) or for aiding individuals in physical rehabilitation. This section reviews work related to activity recognition, group activity recognition, and publicly available datasets for exercise recognition.

2.1. Activity recognition

In the past few years, researchers have introduced numerous activity recognition techniques using hand-crafted and deep neural network-based methods (Jamil et al., 2022c; Sarkar et al., 2022). The integration of data from various sensors has enabled the precise identification of numerous human activities, as demonstrated by Gil-Martín et al. (2020) and Li et al. (2018). The processing of data from various sensors requires additional computational procedures for synchronization and filtration of pertinent information. Pose estimation models have been widely employed in HAR to localize the 25 skeletal joints of the human body. Utilizing skeletal data for storing examples results in a relatively small dataset in terms of size, despite the large set of examples that can be stored. According to previous studies (Neili et al., 2017), using skeleton joints has yielded favorable outcomes in activity recognition. At the same time, innovative techniques for activity recognition have also been proven to be successful on color and depth images obtained through the Kinect sensor, yielding favorable outcomes. These methods encompass naïve Bayes (Yang & Tian, 2014), dynamic Bayesian mixture models (Faria et al., 2014), and random forest (Krüger & Nguyen, 2015) in proximity networks. Moreover, Song et al. (2022) constructed various baselines and Lee et al. (2021) tested various data-driven machine learning-based methods for skeleton-based action recognition.

Several recent studies have suggested alternative methods for activity classification utilizing Convolutional Neural Networks (CNNs) (Huynh-The et al., 2019; Wan et al., 2020). In those studies, CNNs have been employed against different modalities of data from multiple sources, such as wearable sensors, smartphones, and RGB or RGBD cameras. Within this particular framework, the benefits of CNNs are twofold, as they are capable of functioning concurrently as both a feature extractor and a classifier, as noted by Tomas and Biswas (2017) and Caetano et al. (2019). A study by Neili et al. (2017) leveraged the characteristics of CNNs to propose a method for recognizing human poses. This study involved estimating joint positions and utilizing a Support Vector Machine (SVM) to achieve a pose classification outcome. Tomas and Biswas (2017) developed a deep learning-based architecture that combines a CNN with Stacked Auto-Encoders (SAE) for activity recognition. The CNN component is

responsible for learning the representations of motion, while the SAE component models the motions of the skeletal joints. Afterward, the composite scores of the class derived from all networks are aggregated to yield an ultimate score emanating from the selected frames. Kim and Reiter (2017) employed a three-dimensional (3D) CNN architecture to interpret the temporal features of human actions. Moreover, Andrade-Ambriz et al. (2022) used 3D convolutions in combination with an LSTM convolutional layer on RGB images for classifying daily activities and achieved good accuracy. Another study (Mim et al., 2023) used Gated Recurrent Units (GRU) along with an attention mechanism on wearable sensor data and extracted temporal features, followed by spatial feature extraction using an inception network (Ioffe & Szegedy, 2015), and finally classified the activity. In another study (Islam & Iqbal, 2020), the researchers proposed an algorithm based on a multimodal self-attention mechanism for daily life activity recognition and achieved good accuracy on multiple publicly available datasets. Kumie et al. (2023) also utilized a dual-attention network and achieved good accuracy on the datasets, which comprise daily life activities. Another multimodal HAR method (Yadav et al., 2022b) proposed for daily life activity recognition and fall detection also achieved good results in terms of accuracy. Various HAR methods have also been proposed for yoga action recognition. A two-stream network (Yadav et al., 2022a) based on skeletal and RGB data achieved good accuracy on a yoga action recognition dataset with real-time performance. Some of the recent works related to human activity recognition with their demerits are presented in Table 1.

2.2. Group activity recognition

The methods mentioned above commonly employ wearable sensors, smartphone sensors, or an RGB-D camera to capture human movements through the fusion of images and joint positioning, aiming only for single-person activity recognition. Recently, researchers have started paying more attention to group activity recognition (Gavrilyuk et al., 2020). The increased focus on group or multiperson activity recognition is due to the availability of publicly accessible datasets such as the Collective dataset (Choi et al., 2009) and Volleyball dataset (Ibrahim et al., 2016). Methods specifically designed for group or multiperson activity recognition extract information about the actions of each individual from video streams. This extracted information is then utilized to recognize individual actions and the overall group activity. Using wearable or smartphone sensors for multiperson activity recognition can be computationally or economically expensive, making it an impractical approach. Therefore, multiperson or group activity recognition using data captured through RGB cameras can be a viable solution and has been primarily explored in past research. In the initial stages, approaches depended on hand-crafted features of each person extracted from the video and subsequently analyzed by employing probabilistic graphical models (Choi & Savarese, 2013; Lan et al., 2012, 2011). The advancement of deep learning-based methods has led to a gradual improvement in the efficacy of group activity recognition. Particular Recurrent Neural Network (RNN)-based models have proven effective in their methodologies. In a method proposed by Ibrahim et al. (2016), LSTM was employed to create a model that captures the action dynamics of each person and integrates the data to forecast group activity. Deng et al. (2016) incorporated graphical models into RNNs to analyze the relations between different entities for group activity recognition. Shu et al. (2017) developed a Multilayered LSTM (MLSTM) architecture, which aimed to optimize both the accuracy of the predictions and the confidence level of the model's outputs. In another study, Bagautdinov et al. (2017) utilized an RNN to ensure the temporal consistency of box proposals, enabling the joint detection of each person on video, prediction of their actions, and identification of group activity in video footage. In a study conducted by Wang et al. (2017a), an LSTM layer-based model was employed for analysis of the dynamics of individual persons and within-group and between-group

Table 1
Recent works related to the human activity recognition.

Ref.	Main contribution	Limitations
Ronald et al. (2021)	Proposed a model based on Inception-ResNet for activity recognition and achieved an accuracy of 95.09% on the UCI-HAR dataset.	A large number of parameters resulted in a decline in accuracy.
Andrade-Ambriz et al. (2022)	Proposed an architecture based on temporal convolutional neural network using 3D CNN for human activity recognition. The model was best evaluated on the KARD and CAD-60 datasets with the precision of 100%.	The model was not evaluated on large-scale datasets, which may affect its ability to perform well on new data. Moreover, it can be challenging to adapt the proposed architecture for multiperson activity recognition.
Yadav et al. (2022a)	Proposed a dual-stream network comprised of time-distributed CNNs and LSTM using skeletal and RGB data for real-time yoga action recognition. Overall, a good investigation of using a dual-stream network for real-time yoga action recognition is presented.	The model was trained and evaluated on a small dataset comprised of 1,206 videos, which may limit its ability to generalize to the new data. The proposed network is prone to a decrease in real-time performance if there are more than two people in the scene.
Shojaedini and Beirami (2020)	Presented a deep learning structure to minimize the phenomenon of the accuracy situation as well as enhance the optimization ability of LSTM-CNN. The proposed method achieved accuracy of 82.38% and 96.32% on two sets of the WISDM dataset.	The proposed method experienced a slight decrease in accuracy against non-challenging activities.
Lee et al. (2021)	Proposed a method to derive optical flow information using skeletal data for human activity recognition. This method achieved high accuracy using simple machine learning algorithms.	Extracting skeletal data from RGB videos and deriving optical flow information later on can be computationally expensive. The proposed method was not tested against large-scale datasets.
Song et al. (2022)	Proposed a graph convolutional network for action recognition based on skeletal data. The proposed method achieved an accuracy of 92.1% on NTU RGB+D 60 dataset.	The presented model can be computationally expensive in training and deployment. Furthermore, it should be tested on more challenging datasets for better understanding.
Mim et al. (2023)	Proposed a hybrid model named GRU-INC for human activity recognition that is based on the Gated Recurrent Unit (GRU) for extracting temporal features and the Inception module for extracting spatial features. It achieved an F1-score of 96.27% on UCI-HAR dataset.	The proposed model was unable to perform well against complex activities.
Gao et al. (2021)	Presented an attention module-based residual network for blending the temporal attention module and channel information. The proposed method achieved an accuracy of 98.85% on the WISDM dataset.	A huge number of parameters increased the computational complexity of the model.
Proposed method	Presented a lightweight multiperson activity recognition framework. The proposed method was evaluated on multiple datasets comprised of various activities and outperformed state-of-the-art methods in terms of accuracy and computational complexity.	Our method utilizes 2D skeletal data, which may not be optimal for exercise quality assessment as compared to 3D skeletal data.

interactions. In a method proposed by Xu et al. (2015), a 2D CNN-based network and graph convolutional networks were used to construct a persons' relation graph that effectively captures the interplay between persons' appearance and position relations. Similar to Xu et al. (2015), another approach (Gavrilyuk et al., 2020) also employed person-level representations. However, this approach differed in utilizing the self-attention mechanism, which enabled it to selectively emphasize persons and group relations without requiring explicit graph construction. They used pose data, optical flow representations, and RGB images as input to their method. The existing approaches for multiperson activity recognition mainly focus on sports activities. To the best of our knowledge, real-time multiperson activity recognition to support group exercise among elderly persons has not been investigated in past research. In the present work, we propose a new framework for multiperson activity recognition to conduct group exercises through robotic coaches. We deployed the proposed framework on a social robot for testing in a real-world environment.

2.3. Exercise datasets

Limited efforts have been made to curate datasets comprised of whole-body exercises required for maintaining the physical health of elderly individuals on a daily basis. Most existing datasets of rehabilitation exercises have utilized sensors affixed to the human body and are limited to specific health issues. Ebert et al. (2017) collected a dataset by utilizing five sensor devices that were affixed to the ankle, wrist, and chest regions. The purpose was to capture six distinct exercises that were executed by a cohort of 27 athletes. Additionally, data were annotated with a qualitative rating system ranging from one to five. The TRSP dataset (Dolatabadi et al., 2017) comprises 3D human pose approximations of stroke patients as well as healthy individuals who executed a series of movements utilizing a robot for stroke rehabilitation. The recorded data were annotated using a four-label system per-frame basis, including the following categories: no compensation,

shoulder elevation, leaning forward, and trunk rotation. Participants who had survived a stroke engaged in two distinct exercise modalities, reach-side-to-side and reach-forward-backward, and performed them bilaterally, using both their left and right hands. This study involved the participation of healthy individuals who executed the prescribed movements and emulated the typical compensatory actions exhibited by patients who have suffered a stroke.

Wearable sensors can pose inconveniences for patients due to factors such as their size and the need for specialized facilities to perform required movements. This can make the utilization of wearable sensors impractical or challenging for patients, affecting their comfort and willingness to engage with the technology. Some approaches utilize image sensors, such as cameras that detect color or depth, to monitor human movements. Many available image-centric datasets for activity recognition rely on using depth cameras, with the Kinect sensor (Lun & Zhao, 2015) being a popular choice. These depth cameras provide valuable information about the spatial characteristics of human movements, enabling more accurate and detailed analysis of activities. As proposed by Parisi et al. (2015), a dataset was acquired at the Kinesiology Institute of the University of Hamburg utilizing a Kinect sensor. In total, 17 athletes participated in the collection process of this dataset and performed three types of exercise activities. The dataset known as the University of Idaho-Physical Rehabilitation Movement Data (UI-PRMD) (Vakanski et al., 2018) comprises ten prevalent exercises for physical rehabilitation executed by ten healthy individuals. Each participant executed ten accurate and ten inaccurate (suboptimal) repetitions of the exercises. The motion data were captured using two different sensor technologies: a Kinect sensor and a Vicon optical tracking system. The KIMORE dataset proposed by Parisi et al. (2015) is recent. This comprises audiovisual recordings of 78 individuals engaged in rehabilitation exercises, consisting of 44 healthy controls and 34 patients. The data that have been gathered comprise joint positions, along with RGB and depth videos. The dataset is limited in scope because it includes only five gestures related to physical exercises for

Table 2
Implementation environment for proposed multi-person activity recognition system and robot application.

Name	Tools and technologies	Description
System components	Central Processing Unit (CPU)	Intel(R) Core(TM) i9-9900K @ 3.60 GHz
	Operating System (OS)	Ubuntu 16.04
	Random Access Memory (RAM)	64.0 GB DDR4 @ 2800 MHz
	Graphics Processing Unit (GPU)	GeForce RTX 2080 Ti
	Integrated Development Environment (IDE)	Visual Studio Code
Core libraries	Programming Language	Python 3.8
	OpenCV	Computer vision library (V 4.4.0)
	Keras	Open-source high-level neural networks API written in Python
	Matplotlib	Python library used for visualization of data
	Pandas	Python library for data preparation
	TensorFlow	Open-source deep learning framework
	Robot Operating System (ROS)	Support writing robot software
NAOqi SDK	Support writing software for Pepper Social Robot	

lower back pain. The dataset named IntelliRehabDS (IRDS) introduced by Miron et al. (2021) involved the observation of 15 actual patients and 14 healthy individuals who were instructed to perform nine distinct gestures. Compared to the datasets discussed above, this one is not confined to particular health issues. However, the number of samples included is limited. Moreover, there is a publicly available dataset named the Yoga Asana Recognition (YAR) database (Reyes-Ortiz et al., 2012) for yoga action recognition. It has 1206 samples collected using an RGB camera. As compared to the exercise datasets, there are more publicly available datasets for normal daily life activity recognition. A detailed description of various publicly available exercise and daily life activity-based datasets is presented in Table 6.

3. Implementation environment

In this section, we present tools and technologies used to collect the RED dataset and develop the proposed multiperson activity recognition system. The overall summary of these tools and technologies is presented in Table 2. The OpenCV library was initially used in data acquisition. Later, the OpenCV library and deep learning models incorporated by Keras and TensorFlow were executed over GPU for data preprocessing and feature extraction. Once the features were generated, we cleaned data to handle the outliers, missing values, and noisy data using Pandas and stored the features along with the class label in a .csv file. Afterward, the stored features were loaded and used to train the MLSTM architecture developed using Keras and TensorFlow. Finally, the robot application presented in Section 5.4 was developed using ROS and NAOqi to deploy the proposed multiperson activity recognition system through the Pepper robot.

4. Data acquisition

In this section, we describe our new curated dataset, comprised of various whole-body exercises needed for elderly individuals in daily life. The data collection process involved utilizing an Intel RealSense camera for capturing the RGB images at 30 frames per second from a single view. Based on the target application, the camera orientation in terms of height and angle was set in such a way that it captures visual information similar to the ground robot's vision. Our dataset comprises 19 exercises in total. These were selected based on the recommendations from physiotherapists working at a long-term care facility located in Norway. The exercises are not tailored to address particular medical conditions but rather comprise basic movements that physiotherapists incorporate in assessments of movement, rehabilitation regimens, or routine exercises for elderly clients. Based on the activities in our dataset, we named it the Routine Exercise Dataset (RED). The dataset was collected in the Social Robots Lab at NTNU, and the total number of participants involved in the data collection process was 19, aged between 26 and 48. It was not possible to collect such a large dataset with the help of elderly individuals because data

collection involved multiple repetitions of each exercise, which makes it a strenuous process. Each individual performed 40 repetitions for each exercise. All participants were healthy and could participate in a strenuous process to collect a large dataset. A total of 14,440 exercise samples were collected over multiple days. The length of each sample ranges from 3 s for the shortest exercise to 6 s for the longest exercise. The research study was approved by the Data Protection Official, who works with the Norwegian Centre for Research Data (ref.: 508625). We followed the protocols of 'General Data Protection Regulation (GDPR)' by European Union to keep the data secure and will ensure its anonymity before making it publicly available. The names of the 19 exercises included in our dataset are arm circle, chair stand, elbow flexion left, elbow flexion right, hip marching left, hip marching right, neck flexion front, neck flexion left, neck flexion right, shoulder abduction left, shoulder abduction right, shoulder flexion left, shoulder flexion right, shoulder front elevation, side leg raise left, side leg raise right, simple grapevine, upper body twist left and upper body twist right Fig. 5.

5. Methodology

This section discusses the core modules of the proposed method in detail. The framework of the proposed method, which is mainly divided into three core modules, is presented in Fig. 1, and the flowchart of the proposed algorithm is demonstrated in Fig. 2. In the first module, we extract the frames from the video and preprocess them to obtain key frames. The second phase extracts features of each actor in the video using pose estimation, person detection and tracking, and feature engineering. Finally, we feed features extracted against each actor in the video to our proposed Multilayer Long Short-Term Memory (MLSTM) network for activity recognition. The detailed algorithm designed to perform multiperson activity recognition is presented in Algorithm 1.

5.1. Preprocessing

In the preprocessing step, the frames are extracted from the video stream. Then, each frame is resized for uniformity of the input of the proposed framework. Each video frame is resized to $640 \times 480 \times 3$ to ensure that all frames are the same size, allowing consistent processing and analysis of the video frames.

5.2. Actor feature extraction

Actor (person) feature extraction refers to the process of extracting relevant and discriminative features representing each actor in a sequence of video frames. These features capture important information about the actors' poses, movements, or other relevant characteristics for activity recognition. This phase is divided into two branches: actor localization and pose estimation. Later, the final features are derived in the feature extraction phase using information generated by these two branches.

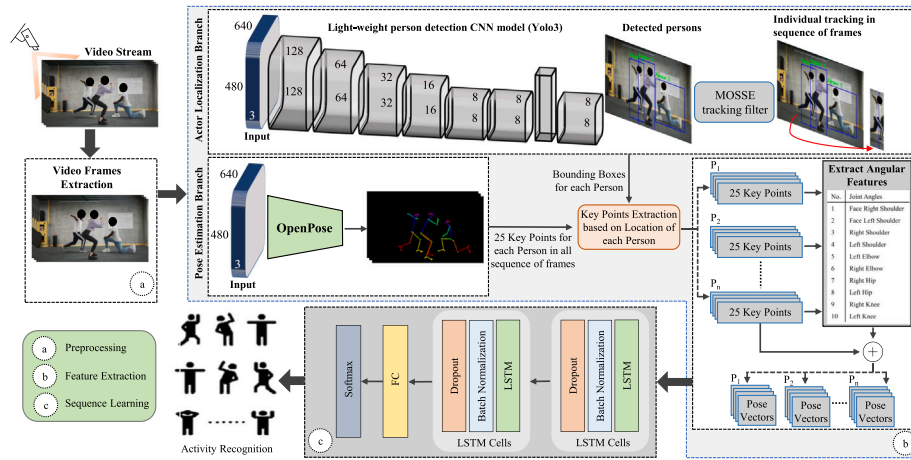


Fig. 1. Overview of the proposed framework.

After preprocessing, the input video frames are processed by two branches: pose estimation and person localization. The pose estimation branch generates pose key points (Skeletal joints), while the person localization branch tracks actors in the video frames and identifies bounding boxes around them. By matching the pose key points to each actor's location in the frames, angular features of each skeleton are obtained using 25 pose key points. Finally, these angular features are concatenated with the corresponding pose key points to obtain pose vectors, which are then fed as input to the Multi-layer LSTM (MLSTM) network for learning sequential patterns.

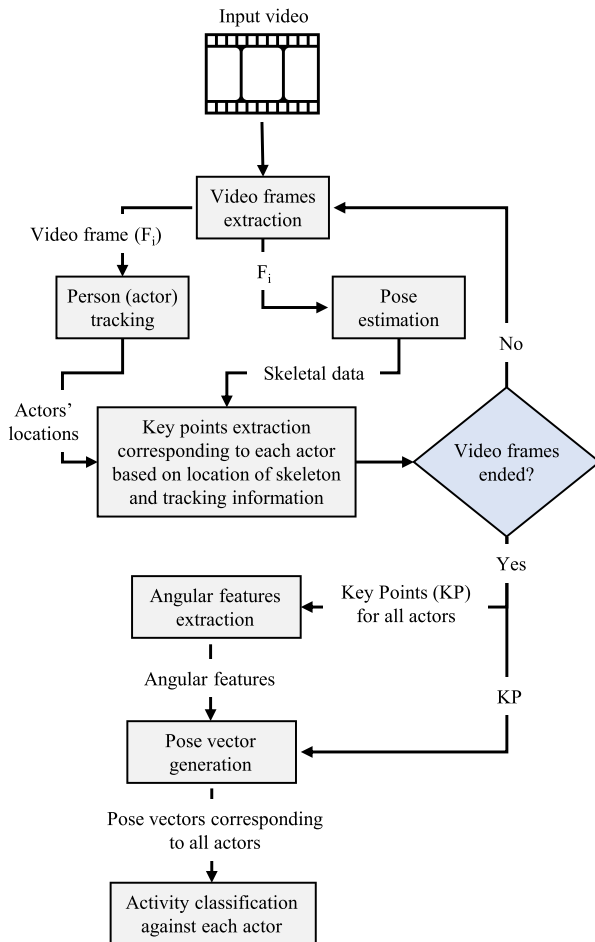


Fig. 2. Flowchart of the proposed algorithm.

5.2.1. Actor localization branch

The actor localization branch mainly includes two steps, i.e., actor detection and tracking. We fine-tuned a CNN-based model for actor detection and then applied a correlation-based tracker for actor tracking, as discussed below.

Actor detection is one of the crucial steps in activity recognition, and different methods exist to perform this task. Their effectiveness and efficacy, however, are not up to the mark for activity recognition in the present application. One efficient existing object detection model is YOLOv3 (Redmon & Farhadi, 2018), which can detect actors as well. However, YOLOv3 is trained on datasets with different objects from the general categories and actors, which are irrelevant in the context of the present application. We have therefore fine-tuned the YOLOv3 model using two actor detection datasets, i.e., Caltech (Dollar et al., 2011) and SPID (Wang et al., 2017b). The effectiveness of training the YOLOv3 model on actor-specific data is better than the model trained on data of objects from the general category. As a result, it can detect actors in challenging video data with diverse poses, sizes, and lighting conditions. We extracted features using Darknet-53 as the backend for fine-tuning the YOLOv3 model in our approach. Small successive convolutional filters of 3×3 and 1×1 are present in Darknet-53, which assists in detecting actors of various scales, even for large distances. Logistic regression is used to detect the objects and confidence scores of their bounding boxes. Due to the better efficiency of Darknet-53, we used it as a backend model. It was also experimentally proven by Redmon and Farhadi (2018) that Darknet-53 possesses better efficiency than Darknet-19, ResNet-101, ResNet-152, etc., as shown in Table 3.

Darknet-53 performs better than the state-of-the-art techniques due to floating-point operations and its high speed. ResNet-101 and Darknet-53 perform similarly, but ResNet-101 is two times slower. Darknet-53 is 1.5 times faster than ResNet-101, and it performs better than ResNet-101. Our fine-tuned model achieved 32.56 Mean Average Precision (mAP) on the combined dataset, and it consumes only 22 ms in processing each frame (Redmon & Farhadi, 2018), making it suitable for our proposed multiperson activity recognition system in real-time. Further details about YOLOv3 can be read in work by Redmon and Farhadi (2018).

Tracking and analyzing the sequence of actions performed by actors in a video stream is one of the main processes in activity recognition. In the application environment, the activities are performed by each actor at a time, which needs to be tracked. Therefore, we track all actors in the sequence of video frames to capture the motion sequences for their localization and further processing after the detection of actors. Various deep learning-based architectures for actor tracking have been proposed in previous research, such as DeepSORT (Wojke et al., 2017). However, using deep learning models in the preprocessing step for

Algorithm 1: Multi-person Activity Recognition in Exercise Videos**Input:** Exercise Video V_{exer} **Preparation**

Load pretrained OpenPose model M_E
 Load fine-tuned person detection model M_Y
 Initialize MOSSE tracker M_T
 Load our trained MLSTM model M_L

Function Recognize_activity(V_{exer}):

Read frames (F) $\leftarrow (f_i, V_{\text{exer}})$
while ($i++ < \text{range}(\text{length}(F))$) **do**
 Pass frame f_i to M_E
for each actor a_j detected in f_i do
 | $KP_temp_list[i][j] \leftarrow$ Obtain pose key points (kp_j)
end
 Pass frame f_i to M_Y
for each a_j detected in f_i do
 | Compute bounding box b_j
 | Pass b_j to M_T for tracking a_j
 | $B \leftarrow$ Extract bounding box b_j and track ID t_j
end
for each b_j in B do
 | **for each kp_j in $KP_temp_list[i][j]$ do**
 | | $b_k \leftarrow$ Create bounding box using kp_j
 | | **if $b_j == b_k$ then**
 | | | $KP_list[i][j] \leftarrow$ Save kp_j corresponding to a_j
 | | | based on t_j
 | | **end**
 | **end**
end
for each kp_j in $KP_list[i][j]$ do
 | $af_j \leftarrow$ Compute angular features from kp_j
 | Pose vectors (Feature vectors) $PV[i][j] \leftarrow$ Obtain pose
 | vector by concatenating kp_j and af_j
end
end
for each a_j do
 | Forward pose vectors PV to M_L , and Predict activity class
 | $\leftarrow M_L(PV)$
end

return

Table 3
 Comparison of different backbone models for object detection.

Model	Ops Bn	FLOpBn/s	FPS	Top-1	Top-5
ResNet-101 (He et al., 2016)	19.7	1039	53	77.1	93.7
ResNet-152 (He et al., 2016)	29.4	1090	37	77.6	93.8
Darknet-19 (Redmon & Farhadi, 2017)	7.29	1246	171	74.1	91.8
Darknet-53 (Redmon & Farhadi, 2018)	18.7	1457	78	77.2	93.8

activity recognition reduces the overall efficiency of real-time application. Correlation-based object trackers are considered very fast (Shah et al., 2020). Therefore, we have used an object tracker named the MOSSE (Bolme et al., 2010) tracking filter for capturing actor sequences, which is ultrafast. It can process at a speed of more than 700 frames per second with high robustness toward different scales, abrupt movements, poses, and illumination changes. It can simultaneously track multiple actors and provides bounding boxes around them. We passed the locations of actors detected by our actor detection model in the first video frame to the MOSSE tracker for tracking in the remaining video frames.

5.2.2. Pose estimation branch

All actions performed by human beings involve the movement of skeletal joints. Hence, capturing the joint positions of each actor present

in the frames is essential. We used a pose estimation model M_E named OpenPose (Cao et al., 2017), which can be used for multiperson pose estimation in real time and is well suited to our framework. It takes an image frame F_i and outputs 25 pose Key Points (KP) (Skeletal joints) in the form of image coordinates for each actor A_j present in the image (Eqs. (1)–(3)).

$$KP = M_E(F_i) \quad (1)$$

$$KP^{(i,j)} = [kp_1^{(i,j)}, kp_2^{(i,j)}, \dots, kp_{25}^{(i,j)}] \quad (2)$$

$$\text{where, } i = 0, 1, 2, \dots, n; j = 0, 1, 2, \dots, n \quad (3)$$

The positions of actors can change over time while performing an activity. However, the OpenPose model does not provide tracking of the actors in the video. Therefore, we used the pose estimation branch and the actor localization branch for tracking and obtaining the skeletal joints against each actor in a video frame sequence, as shown in Fig. 1.

5.2.3. Feature extraction

As discussed in Section 5.2.2, the pose estimation branch outputs $KP^{(i,j)}$ against each actor j in frame i using the OpenPose model; on the other hand, the person localization branch (Section 5.2.1) tracks each actor in the video and finds bounding boxes with identification numbers against them. We obtain the KP corresponding to each actor in all frames by matching the bounding boxes generated by the person localization branch with the bounding boxes formed around the KP generated by the pose estimation branch. After obtaining the KP for each actor in all frames, we compute the angular features stated in Fig. 1 using those pose key points. As an example, the computation of the angle $\angle kp_1 kp_2 kp_3$ made at key point k_2 between vectors $\overline{kp_1 kp_2}$ and $\overline{kp_2 kp_3}$ keypoints k_1 , k_2 , and k_3 is determined as follows.

$$\angle kp_1 kp_2 kp_3 = \cos^{-1} \left(\frac{\mathbf{U} \cdot \mathbf{V}}{\|\mathbf{U}\| \|\mathbf{V}\|} \right) \quad (4)$$

Here, the vector \mathbf{U} represents the vector connecting kp_1 and kp_2 , i.e., $\overline{kp_1 kp_2}$, while the vector \mathbf{V} represents the vector connecting kp_2 and kp_3 , i.e., $\overline{kp_2 kp_3}$. Description of all parameters or symbols used in proposed framework are presented in Table 4. Finally, we obtain the pose vector (1D feature vector) by concatenating the KP with angular features. Each pose vector represents an actor in a single frame. The pose vectors obtained against each actor in F representing an activity sample are then fed into our proposed 1D MLSTM. We applied a frame skip strategy before feeding the pose vectors for sequential learning. The frame skip strategy is used to reduce the computational load by selecting a subset of intermediate frames from the input sequence of frames. Here, we have designed an adaptive frame skip strategy that skips the frames at an adaptive skip rate depending upon the length of the input sequence of pose vectors representing an activity sample to extract 20 intermediate pose vectors for sequential learning. This strategy allows us to extract a fixed sequence of pose vectors representing the intermediate movements of the actors. The time distribution of skeletal data (25 keypoints) in an activity sample for various activities is shown in Fig. 3. It shows which skeletal joints are primarily active in different physical exercises, such as right and left wrists and elbows in arm circle exercise (Fig. 3). The correlation between the skeletal joints and angular features formed by them plays an important role in achieving good model performance. Fig. 4 shows the strong correlations between the skeletal joints as well as the angular features related to the lower and upper body exercises. MLSTM processes these pose vectors to learn the hidden sequential patterns. A detailed explanation of RNN, its variants, and the proposed 1D MLSTM are presented in the next section.

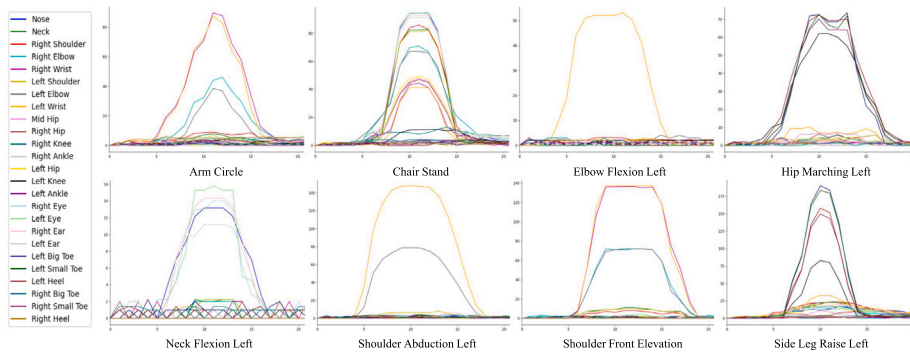


Fig. 3. Time distribution of the skeletal data in a sample of a few activities from the RED dataset.

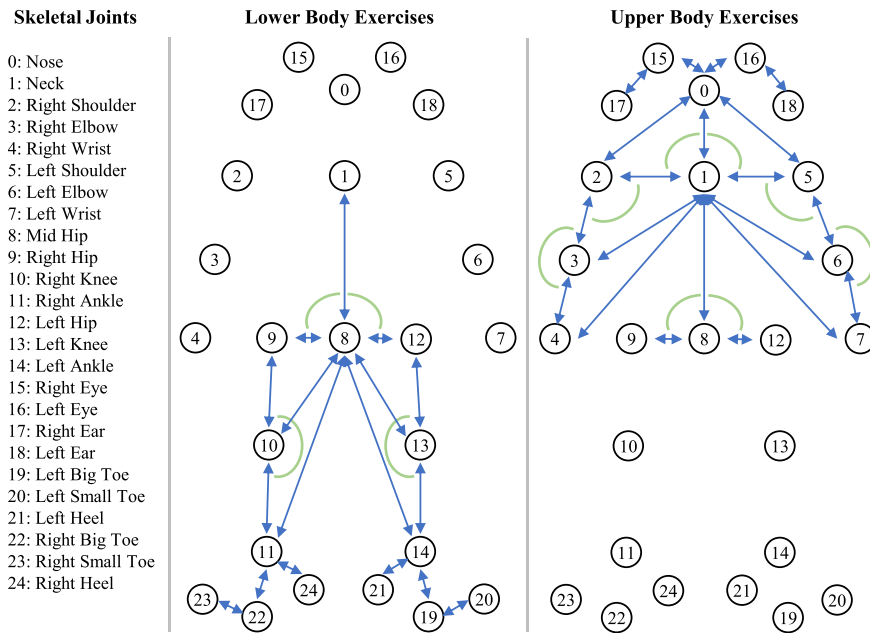


Fig. 4. Correlation between skeletal joints in different types of exercises.

Table 4

Description of mathematical symbols/parameters for different operations used in our proposed framework.

Symbols	Description
KP	Key points (skeletal joints).
V_i	Input video.
F_i	Frame i in V_i .
f_s	Number of the frames skipped during feature extraction.
x_t	Input to LSTM at time t .
f_i	Output of forget gate.
i_i	Output of input gate.
o_i	Output of output gate.
c_i	Output of current state of LSTM cell.
c_{i-1}	Previous state of LSTM cell.
w_f	Weights of forget gate of LSTM cell.
w_i	Weights of input gate of LSTM cell.
w_o	Weights of output gate of LSTM cell.
b_f	Biases of forget gate.
b_i	Biases of input gate.
b_o	Biases of output gate.
h_t	Final output of LSTM cell.

5.3. Sequence learning using multilayer LSTM

Various tasks, such as machine translation and activity recognition, can be expressed with sequences having variable lengths. The

need for a method that can learn hidden patterns in time-series data to solve problems related to sequential learning is crucial. A neural network, named RNN, has been introduced, which can extract and learn temporal features based on temporal relations from input time-distributed data and classify sequential data. RNNs predict the future output based on analysis of the hidden sequential patterns in both temporal and spatial dimensions by building connections between previous and current information. Many researchers have investigated various sequence learning problems using RNNs and achieved good results. However, despite being able to learn hidden patterns from sequential data efficiently, RNNs tend to forget earlier information while processing long-term sequences, which is called the vanishing gradient problem. A special variant of RNN, known as LSTM, can be used to solve this problem. Other methods, such as GRU and transformers, can also be used for sequence learning (EK et al., 2022). However, GRU suits simple tasks with fewer parameters, and transformers can be computationally complex. The LSTM network has proven efficient in tasks with long-range dependencies. Therefore, we have designed our proposed method using multilayer LSTM for better efficiency and less computational complexity to perform multiperson activity recognition.

5.3.1. Multilayer LSTM network

The extension of RNN named LSTM (Hochreiter & Schmidhuber, 1997) is specifically designed to model and analyze (interpret) long-term sequences, which helps resolve the vanishing gradient problem

Table 5
Architecture of the proposed 1D MLSTM model.

Model layers	Layer attributes
LSTM	No. of units = 256 Activation function = tanh
Batch normalization	–
Dropout	Dropout rate = 0.05
LSTM	No. of units = 256 Activation function = tanh
Batch normalization	–
Dropout	Dropout rate = 0.05
Dense	No. of units = 64 Activation function = ReLU
Dense	No. of units = No. of classes Activation function = Softmax

faced in RNNs. Multiple cell units comprise the LSTM's internal structure. Each cell unit contains three gates, i.e., input, output, and forget gates, that control the information flow and process of sequential pattern recognition. The configuration of these gates is set in such a way that each gate receives the input from the previous step and transmits the computed output to the next gate. A sigmoid function controls each of these gates. For example, the portion of the information that needs to be updated is determined by the input gate $i_{h_{mt}}$, while the output gate o_t retains the information for the following sequence. When necessary, the forget gate f_t processes the information from the previous cell state before erasing it from memory and the input gate. The previous cell state c_{t-1} and the current input x_t are computed by using the \tanh activation function in the recurrent unit g , whereas h_t can be computed by multiplying the value of the output gate by the \tanh of the current cell state c_t . One can obtain the final output by passing h_t to the softmax classifier. Following are the mathematical equations of the operations executed by these gates:

$$i_t = \sigma(w_i * [h_{t-1}, x_t] + b_i) \quad (5)$$

$$o_t = \sigma(w_o * [h_{t-1}, x_t] + b_o) \quad (6)$$

$$f_t = \sigma(w_f * [h_{t-1}, x_t] + b_f) \quad (7)$$

$$g = \tanh(w_g * (x_t + c_{t-1}) + b_g) \quad (8)$$

$$c_t = ((c_{t-1} * f_t) + (g * i_t)) \quad (9)$$

$$h_t = (\tanh(c_t) * o_t) \quad (10)$$

$$output = softmax(h_t) \quad (11)$$

Multiple layers can be stacked to improve the performance of a deep neural network. Similarly, the capability of an LSTM to learn complex and hidden sequential patterns can be improved by increasing the number of network layers. Adding more layers to the network allows for capturing higher-level representations and abstract features from the input sequence. The choice of the number of layers in an LSTM network depends on the complexity of the problem, the size of available training data, and the computational resources. Thus, the architecture of our network also contains multiple layers. We adopt a 1D MLSTM architecture to verify the accuracy of our proposed network. Our 1D MLSTM architecture comprises six layers: two layers of LSTM, two layers of batch normalization, and two dropout layers. Hierarchical processing is applied to the input data fed to the MLSTM, which has several layers. The hidden state of the previous layer is fed as input to each layer in the network and then passes its output to the next layer. The computational process of the memory cell of the MLSTM is similar

to the standard LSTM. The output of the last dropout layer then goes through a sequence of dense layers, as given in Table 5, followed by a softmax activation function for classification. The proposed 1D MLSTM architecture is achieved after conducting numerous experiments, which involve modifying the layers, adjusting the number of units in each layer, and experimenting with different dropout rates.

5.4. Social robot application and experiments

As an interactive technology, social robots are more effective than 2D screens in improving cognitive function and reducing depression and loneliness among elderly individuals (Lim, 2023). However, 2D screens can also provide similar functionality more economically. Therefore, we have designed our application to efficiently deploy and function on any interactive technology. This section presents the architecture of the Robot Operating System (ROS)-based (Stanford Artificial Intelligence Laboratory, 2018) application developed in the present study and describes the experiments conducted in a real-world environment. We developed this application to deploy it on a humanoid social robot named 'Pepper' developed by SoftBank Robotics (SoftBank Robotics, 2023). ROS is specifically designed for developing robot software, providing advantages such as reusability and modularity. Hence, we designed our ROS-based application to easily integrate it with any digital interactive device, such as a robot or 2D screen, with minimal changes. Our application is more robust, modular, and adaptable compared to the basic applications developed in previous studies (Andrade-Ambriz et al., 2022). The ROS architecture of the proposed application is presented in Fig. 5. In Fig. 5, the entities shown in green color represent the ROS nodes, and the incoming arrow to a ROS node shows that it is subscribed by another ROS node and the outgoing arrow shows that it is subscribing to another ROS node. The ROS architecture consists of four modules presented from top left to bottom right in Fig. 5: the module in the upper left corner represents the activity recognition process, followed by the module in the upper right corner representing data management, and the modules at the bottom right and bottom left corners representing verbal and nonverbal interactions, respectively.

The experimental environment was created in the Social Robots Lab at NTNU, which included an RGB camera in front of the participants. The camera and the Pepper robot were connected to a computer that runs the proposed multiperson activity recognition system presented in Fig. 1. During the experiment, the robot plays the role of an artificial exercise coach to demonstrate the exercises and monitor the participants while they perform those exercises. Our multiperson activity recognition system then processes the video to recognize the activity performed by participants. Later, the robot responds to the situation using a dialog box based on the activity performed by the participants. Fig. 6 illustrates the experiment conducted in a real-world environment with multiple participants using the Pepper robot. Fig. 6(a) shows the interaction of participants with the robot, and intermediate frames from the video stream of participants performing a physical exercise are shown in Fig. 6(c). Moreover, Fig. 6(b) depicts the RViz visualization platform, which allows real-time analysis of the robot's state.

6. Experimental results

We performed a series of experiments on our proposed MLSTM sequence learning method using three datasets, including our dataset named the Routine Exercise Dataset (RED) (Section 4) without noise, four noise-corrupted RED datasets, and 11 publicly available benchmark datasets. Among all datasets, eight consist of daily life activities, and four comprise physical exercise activities. We also combined a physical exercise activity dataset with a daily life activity dataset to test the performance and generalizability of our proposed method. The motivation behind conducting experiments on datasets of different categories was to test the generalizability of our proposed method on

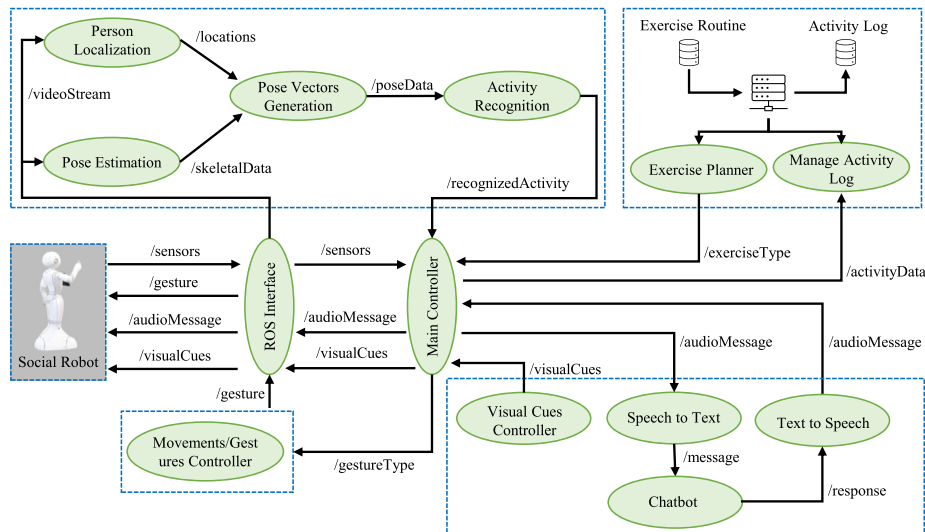
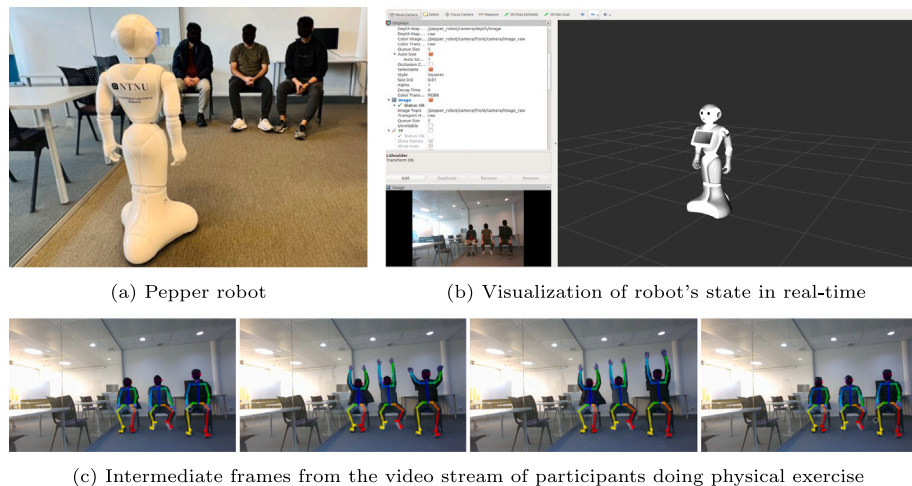


Fig. 5. ROS architecture for the proposed multiperson activity recognition system.



(c) Intermediate frames from the video stream of participants doing physical exercise

Fig. 6. Illustration of the experiments conducted in real-life environment with three participants.

various activities and find the potential of the proposed system in multiple applications in addition to supporting group exercise through intelligent virtual coaches. A detailed description of the benchmark datasets included in our experiments is provided in Section 6.1. Overall results showing the performance achieved by our proposed method and comparison with the related works are presented in Section 6.3.

6.1. Datasets used in our experiments

Extensive experiments were conducted on 13 datasets without noise and four noise-corrupted datasets. The RED dataset is described in Section 4, and a description of the remaining datasets is given below. The summary of datasets used in the experimental evaluation of the proposed activity recognition method is presented in Table 6.

6.1.1. Kinect activity recognition dataset (KARD)

The KARD dataset (Gaglio et al., 2014) is a benchmark dataset for human activity recognition. It contains 18 classes: horizontal arm wave, two-hand wave, high arm wave, high kick, draw tic, draw x, side kick, front kick, grab a hat, fold, throw paper, hand clap, hold an umbrella, phone call, drink, walk, sit, and stand. Ten different subjects performed each activity three times. This dataset provides RGB videos ($640 \times 480 \times 3$), the respective depth maps, and the joint positions.

6.1.2. IntelliRehabDS (IRDS)

The IRDS dataset (Miron et al., 2021) is based on activities involving physical rehabilitation movements. It is a benchmark dataset comprising nine classes: elbow flexion left, elbow flexion right, shoulder flexion left, shoulder flexion right, shoulder abduction left, shoulder abduction right, shoulder forward elevation, side tap left, and side tap right. A total of 29 subjects participated in the collection of this dataset, and all subjects performed each activity a variable number of times. The repetitions of activities performed by each subject were categorized as correct or incorrect. However, we included only those samples ($n=2047$) from all activities that were performed correctly.

6.1.3. Cornell activity datasets: CAD-60

CAD-60 (Sung et al., 2012) is a video dataset containing daily activities recorded using an RGB-D camera. It has 12 activities, including brushing teeth, rinsing mouth, putting in contact lenses, drinking water, talking on the phone, opening a pill container, cooking–stirring, cooking–chopping, talking on the couch, writing on a whiteboard, relaxing on the couch, and working on a computer. Four subjects participated in the data collection, resulting in 60 samples.

6.1.4. MSR daily activity 3D

Wang et al. (2012) prepared this dataset based on daily life activities. It has 16 activities, including eating, drinking, reading a book,

Table 6
Summary of the datasets used in experimental evaluation.

Dataset	Activity type	Total classes	Total samples
KARD (Gaglio et al., 2014)	Common Daily Life	18	540
IRDS (Miron et al., 2021)	Physical Exercise	9	2047
CAD-60 (Wang et al., 2012)	Common Daily Life	12	60
MSR daily activity 3D (Wang et al., 2012)	Common Daily Life	16	320
YAR (Yadav et al., 2022a)	Yoga Exercise	20	1206
UP-fall detection (Martínez-Villaseñor et al., 2019)	Common Daily Life	11	1122
UTKinect-Action3D (Xia et al., 2012)	Common Daily Life	10	200
UTD-MHAD (Chen et al., 2015)	Common Daily Life	27	861
UESTC RGB-D (Ji et al., 2019)	Physical Exercise	40	25,600
NTU RGB+D (Shahroudy et al., 2016)	Common Daily Life	40	37,920
RED	Physical Exercise	19	14,440
RED + KARD	Physical Exercise and Common Daily Life	37	14,980
Laplacian noise-corrupted RED (Mean (μ) = 0.5, Variance (σ^2) = 0.5)	Physical Exercise	19	14,440
Laplacian noise-corrupted RED ($\mu = 0$, $\sigma^2 = 1$)	Physical Exercise	19	14,440
Heteroscedastic noise-corrupted RED ($\mu = 0.5$, $\sigma^2 = 0.5$)	Physical Exercise	19	14,440
Heteroscedastic noise-corrupted RED ($\mu = 0$, $\sigma^2 = 1$)	Physical Exercise	19	14,440

writing on paper, calling on a cellphone, using a laptop, cheering up, using a vacuum cleaner, sitting still, playing a game, tossing a paper, walking, laying down on a sofa, playing the guitar, sitting down or standing up. A total of ten subjects participated in the data collection, resulting in a total of 320 samples. This dataset provides depth, vision, and skeletal information.

6.1.5. Yoga asana recognition (YAR) database

YAR is a benchmark dataset (Yadav et al., 2022a) comprising 20 yoga exercises, also known as asanas. This dataset was collected through the participation of 16 subjects. The participants performed each yoga exercise a variable number of times. Overall, the YAR dataset has 1206 samples, with a total video duration of more than six hours. This dataset was collected using an RGB camera with a frame rate of 30 fps and a resolution of 1280×720 .

6.1.6. UP-fall detection dataset

The UP-fall Detection dataset is a publicly available dataset proposed by Martínez-Villaseñor et al. (2019). It is a large-scale dataset focused on activities of daily living collected from 17 healthy participants. It consists of 11 daily life activities divided into six basic activities (standing, walking, sitting, jumping, picking up an object, laying down) and five falls (falling forward using knees, falling forward using hands, falling backward, falling, falling sideward, sitting in an empty chair). The UP-fall detection dataset is recorded using wearable, ambient, and vision sensors. We used only vision data collected through two RGB cameras from the front and lateral views in our experiments. Every participant performed each activity three times, resulting in 1122 samples and a size of 277 GB.

6.1.7. UTKinect-Action3D dataset

The UTKinect-Action3D dataset (Xia et al., 2012) consists of actions from daily life. It has ten classes: sit down, walk, stand up, carry, pick up, throw, pull, push, clap hands, and wave hands. This dataset was collected with the help of 10 participants who performed each action twice, resulting in 200 samples. The dataset was recorded using a Microsoft Kinect sensor with information on visual, depth, and skeleton joint locations.

6.1.8. UTD multimodal human action dataset (UTD-MHAD)

The UTD-MHAD dataset (Chen et al., 2015) is publicly available and consists of 27 daily life actions. It was collected using a Kinect camera and an inertial sensor. This dataset has 861 samples collected with the help of eight subjects. It provides visual, skeletal joint, depth, and inertial data information. We included only visual data in our experiments.

6.1.9. UESTC RGB-D varying-view action database

The UESTC RGB-D dataset (Ji et al., 2019) is based on activities involving 40 types of aerobic exercises. A total of 118 subjects participated in the data collection process for this dataset. The participants performed each exercise three times in a fixed direction. The dataset was collected using RGB-D cameras from eight views, resulting in 25,600 samples. The UESTC RGB-D dataset consists of visual, depth, and skeletal joint information, but we used only visual information in our experiments.

6.1.10. NTU RGB+D dataset

A large-scale dataset for 3D human activity analysis known as the NTU RGB+D dataset (Shahroudy et al., 2016) is a large-scale benchmark dataset. This dataset provides visual, depth, skeletal joint, and infrared data. It has 60 classes and a total of 56,880 samples. It is mainly divided into three categories, i.e., daily actions ($n=40$), medical conditions ($n=9$), and mutual actions or two-person interactions ($n=11$). In our experiments, we have included the data of daily actions, which has 40 classes and 37,920 total samples.

6.1.11. Noise corrupted RED dataset

To test the tolerance of our proposed method against noise-corrupted data, we added Laplacian and heteroscedastic noise to our RED dataset. We evaluated the performance of our model against two different values of means and variances (Table 6) for both types of noises. We randomly chose skeletal joints and samples from the overall dataset to add noise to it instead of adding noise to all 25 skeletal joints. Four noise-corrupted datasets were generated by adding noise to the RED dataset.

6.2. Implementation details

First, we processed all datasets using the preprocessing and feature extraction modules discussed in Section 5. Then, in accordance with generic machine learning protocols, we divided each shuffled dataset into training (90%) and testing (10%) sets. The data were stratified to ensure that the class distribution was maintained in the training and testing sets. We validated the outcome of our proposed 1D MLSTM architecture through K-fold cross-validation. We used stratified K-Fold to ensure that the class distribution was maintained in the training and validation sets. The tools and technologies used in the study are presented in Table 2. Initially, we employed the Adam optimizer and 'Categorical Cross Entropy' loss function with a default learning rate of 0.001, which resulted in overfitting. To address this, we performed vast experiments with different learning rates and achieved optimal results with a learning rate of 0.00003 and a batch size of 32. We also used a scheduling strategy called 'ReduceLRonPlateau', which adjusts

Table 7
Comparative analysis of the proposed method with existing state-of-the-art methods.

	Dataset	Precision (%)	Recall (%)	F-1 Score (%)	Accuracy \pm S.D (%)
Andrade-Ambriz et al. (2022)	KARD	100	100	100	–
Andrade-Ambriz et al. (2022)	CAD-60	100	100	100	–
Andrade-Ambriz et al. (2022)	MSR daily activity 3D	–	–	–	95.6
Khan et al. (2022)	KARD	84	87	85	86.8 \pm 0.73
Khan et al. (2022)	IRDS	97.6	98.4	98	96.14 \pm 0.64
Khan et al. (2022)	RED	87.9	89.1	88	91.1 \pm 0.81
Yadav et al. (2022a)	YAR	96.74	97.05	96.52	96.31
Yadav et al. (2022b)	UP-fall detection	96.90	96.70	96.60	96.60
Lee et al. (2021)	UTKinect-Action3D	–	–	–	97.00
Islam and Iqbal (2020)	UTD-MHAD	–	–	–	95.12
Kumie et al. (2023)	UESTC RGB-D	–	–	–	97.70
Song et al. (2022)	NTU RGB+D	–	–	–	95.70
Proposed method	KARD	100	100	100	99.27 \pm 0.53
	IRDS	100	100	100	99.61 \pm 0.64
	CAD-60	100	100	100	99.77 \pm 0.33
	MSR daily activity 3D	100	100	100	99.90 \pm 0.12
	YAR	98.65	99.01	99.01	98.79 \pm 0.21
	UP-fall detection	98.80	98.90	98.84	98.90 \pm 0.11
	UTKinect-Action3D	100	100	100	99.16 \pm 0.39
	UTD-MHAD	98.65	98.90	98.77	98.42 \pm 0.28
	UESTC RGB-D	98.60	98.69	98.64	98.90 \pm 0.33
	NTU RGB+D	97.60	98.21	97.95	97.75 \pm 0.37
	RED	99.3	99	99	98.88 \pm 0.66
	RED + KARD	99	99	99	99.16 \pm 0.48
	Noise-corrupted RED*	98.56	98.60	98.57	98.24 \pm 0.47
	Noise-corrupted RED**	98.05	98.20	98.12	98.15 \pm 0.69
Noise-corrupted RED***	97.98	98.10	98.03	98.11 \pm 0.53	
Noise-corrupted RED****	97.61	97.79	97.69	97.88 \pm 0.38	

* Laplacian noise ($\mu = 0.5, \sigma^2 = 0.5$).

** Laplacian noise ($\mu = 0, \sigma^2 = 1$).

*** Heteroscedastic noise ($\mu = 0.5, \sigma^2 = 0.5$).

**** Heteroscedastic noise ($\mu = 0, \sigma^2 = 1$).

the learning rate based on training loss. Furthermore, we examined our method over 50 epochs and gradually increased the number of epochs to 350.

6.3. Experimental evaluation and comparison with related works

Once the design of the proposed *MLSTM* with optimal parameters was completed, we trained it using datasets discussed in Section 6.1. This section presents the validation results of the proposed 1D *MLSTM* architecture and its comparison with related works. We used four evaluation metrics for performance evaluation, including precision, recall, F1-score, and accuracy. The results obtained using our proposed method and state-of-the-art approaches presented in related works on various datasets are presented in Table 7. It appears in Table 7 that the proposed method outperformed other existing methods with accuracies of 99.27%, 99.77%, 99.90%, 98.90%, 99.16%, 98.42%, and 97.75% on common daily life activity datasets, including KARD, CAD-60, MSR daily activity 3D, UP-fall detection, UTKinect-Action3D, UTD-MHAD, and NTU RGB+D, respectively. Moreover, the proposed method also performed better than the existing methods on exercise datasets including IRDS, YAR, UESTC RGB-D, and RED with accuracies of 99.61%, 98.79%, 98.90%, and 98.88%, respectively. These results prove the generalizability of the proposed method on different types of activities. An architecture proposed by Khan et al. (2022) achieved good accuracy on the exercise dataset but could not achieve the same performance on other datasets. The 3D CNN-based architecture proposed by Andrade-Ambriz et al. (2022) also achieved maximum precision and recall; however, it cannot be adapted for multiperson activity recognition.

Furthermore, the confusion matrices for various datasets are shown in Fig. 7. Fig. 7(a) demonstrates that our proposed architecture was able to classify almost all classes of the KARD dataset correctly. Moreover, all classes in the IRDS dataset were also accurately classified by the proposed architecture, as shown in Fig. 7(b). Fig. 7(c) shows a slight decrease in accuracy due to a high level of similarity between the

classes ‘Neck Flexion Front’ and ‘Neck Flexion Right’, resulting in an accuracy of 98.88% for the RED dataset. We also combined the RED and KARD datasets, resulting in 37 classes in total, and observed 99.16% accuracy. It can be observed from the training and validation accuracy and loss shown in Fig. 8 that our method achieved high accuracy with a minimal loss on various datasets. Overall, our method outperformed the state-of-the-art 3D CNN-based method (Andrade-Ambriz et al., 2022), CNN hybrid model (Khan et al., 2022; Yadav et al., 2022a, 2022b), and attention-based (Islam & Iqbal, 2020; Kumie et al., 2023) approaches for HAR in terms of precision, recall, accuracy, and computational complexity. Fig. 9 presents the analysis of K-fold cross-validation based on the average accuracy of the proposed 1D *MLSTM* architecture on various datasets. We have also performed sensitivity analysis of the proposed method with RED dataset to critically analyze different parameters’ effects on model’s performance in terms of accuracy. The sensitivity analysis is shown in Fig. 10. Furthermore, we added noise to the RED dataset (Section 6.1.11) to test the performance of the proposed method on noise-corrupted datasets. Our proposed method shows a minimal decrease in accuracy on noise-corrupted datasets compared to the performance on datasets without noise (Table 7).

We also evaluated the competence of the proposed method using the Receiver Operating Characteristic curve (ROC) and Area Under the Curve (AUC) values. The ROC calculates the contrast between the False Positive Rate (FPR) and the True Positive Rate (TPR) at different threshold values for classification decisions. It can be seen in Fig. 11 that the proposed method achieved the best AUC values and ROC curves out of the large-scale datasets, i.e., UESTC RGB-D, NTU RGB+D, and RED, used in our experiments. AUC values of 0.98, 0.94, and 0.99 were achieved for UESTC RGB-D, NTU RGB+D, and RED, respectively. Furthermore, we evaluated the performance of proposed method using precision–recall curve. As shown in Fig. 12, the proposed method achieved best precision–recall curves out of the large-scale datasets used in our experiments. We have also performed statistical tests to evaluate the performance of our model in terms of accuracy

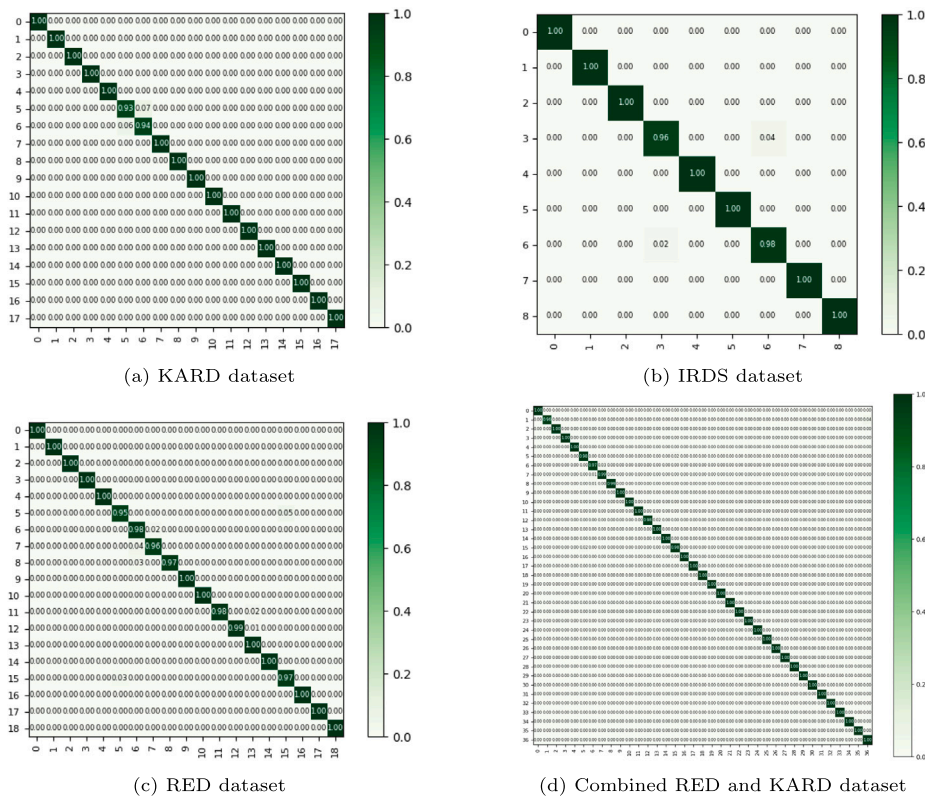


Fig. 7. Confusion matrices for all datasets, with the predicted labels displayed on the x-axis and the true labels on the y-axis.

Table 8

Number of parameters, training time and classification time of the proposed MLSTM architecture compared to state-of-the-art.

Method	Number of parameters (millions)	Dataset	Training time (m)	Classification time (milliseconds)
Kumie et al. (2023)	6.20	UESTC RGB-D	–	–
Song et al. (2022)	1.10	NTU RGB+D	–	–
Proposed method	0.85	KARD	1.12 ± 0.6	145 ± 10
		IRDS	2.19 ± 0.9	145 ± 10
		CAD-60	0.89 ± 0.3	145 ± 10
		MSR daily activity 3D	0.97 ± 0.2	145 ± 10
		YAR	1.83 ± 0.7	145 ± 10
		UP-fall detection	1.80 ± 0.5	145 ± 10
		UTKinect-Action3D	0.93 ± 0.3	145 ± 10
		UTD-MHAD	1.76 ± 0.47	145 ± 10
		UESTC RGB-D	23.19 ± 0.79	145 ± 10
		NTU RGB+D	29.80 ± 0.9	145 ± 10
		RED	14.50 ± 2.0	145 ± 10
RED + KARD	18.50 ± 2.1	145 ± 10		
Noise-corrupted RED	15.83 ± 1.64	145 ± 10		

on various datasets using K-fold cross-validation. Our null hypothesis states that various diverse datasets do not cause a significant difference in the performance of the proposed method. To test the hypothesis, we performed a Kruskal–Wallis test with a confidence interval of 95%. During the statistical tests, we found a p -value greater than 0.05 against all tests, which indicates that the various datasets did not cause a significant difference in the performance of the proposed method, and it performed equally well on each dataset.

6.4. Computational complexity of the proposed framework

A fast classification/execution time is one of the principal objectives of the proposed framework. The hardware components and core libraries used in the present work are described in Table 2. Table 8 presents the time required to train the proposed MLSTM architecture on various datasets. Variance in the processing time occurs due to

the variable size of the datasets. As Table 8 indicates, the proposed architecture takes 145 ms to classify the input sequence and has fewer parameters compared to state-of-the-art methods. Moreover, the time complexity analysis of the subsequent tasks involved in the proposed multiperson activity recognition system is visualized in Fig. 13.

6.5. Tracking results and discussion

In the proposed framework, accurate tracking plays a vital role because accurate tracking in real time yields precise activity recognition results. We have investigated different tracking methods and used the MOSSE tracker in the proposed framework for multiperson activity recognition due to its high accuracy and fast processing. A comparison to investigate tracking methods in terms of tracking visualization and processing rate in FPS is presented in Fig. 14. The state-of-the-art tracking methods, i.e., boosting (Grabner et al., 2008), TLD (Kalal et al.,

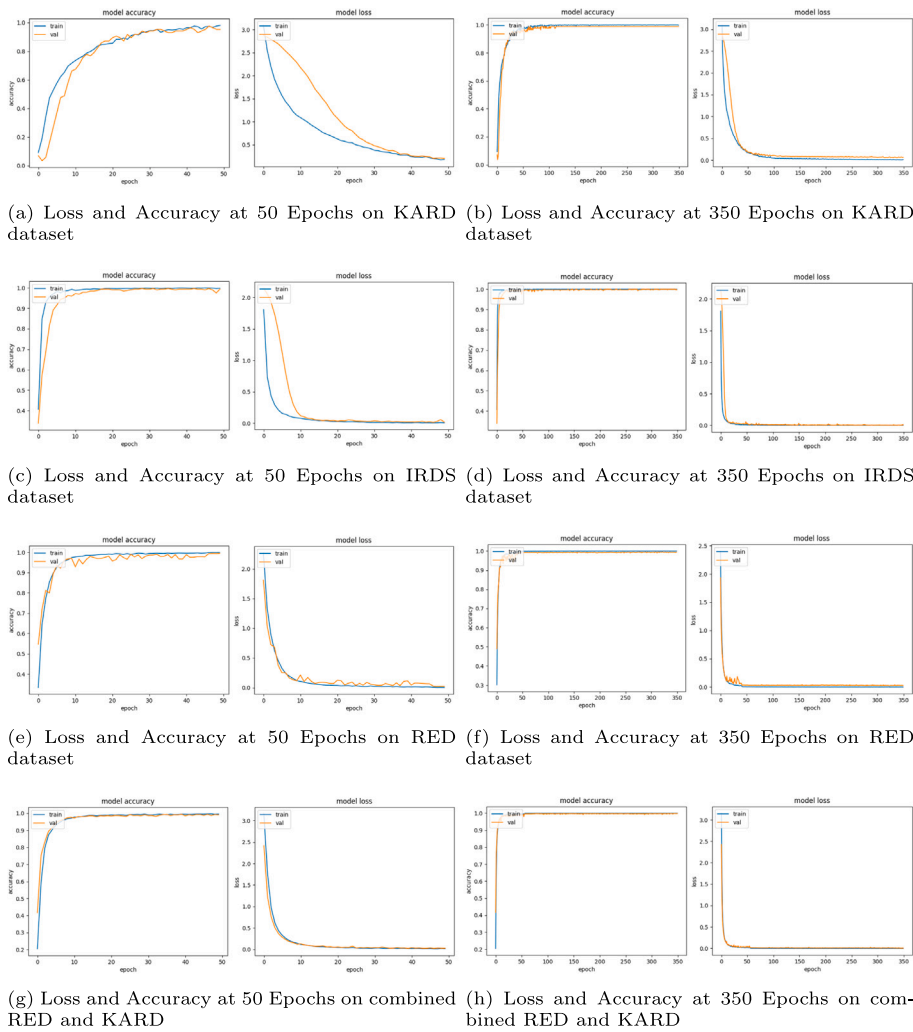


Fig. 8. Visualization of the training and validation loss and accuracy of the proposed *MLSTM* architecture on various datasets.

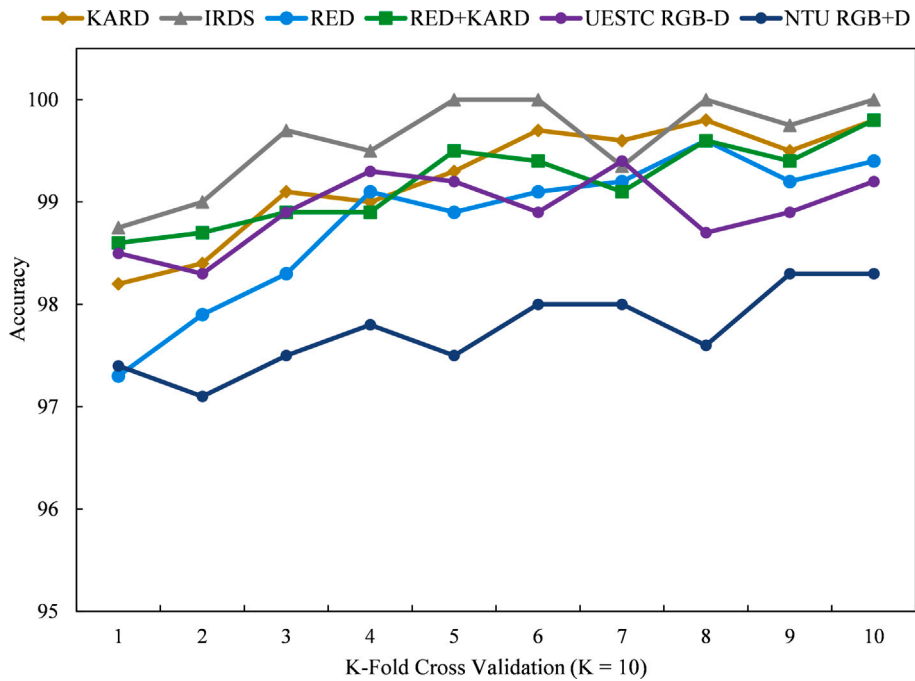


Fig. 9. Analysis of proposed 1D *MLSTM* architecture on various datasets using K-Fold cross-validation.

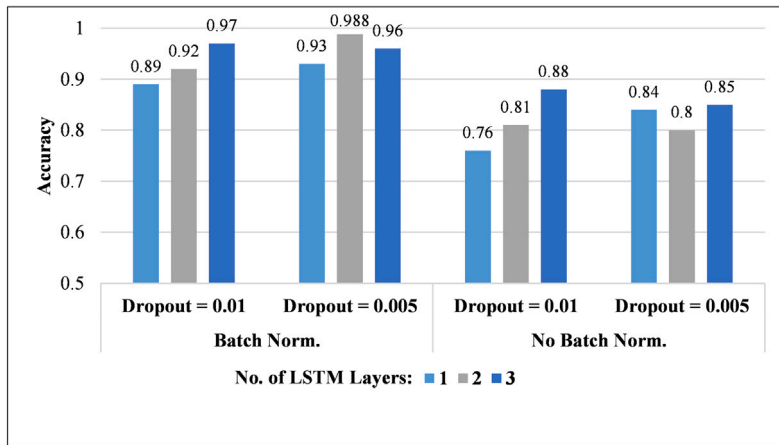


Fig. 10. The impact of varying LSTM layers, batch normalization, and dropout on the accuracy achieved by the proposed method.

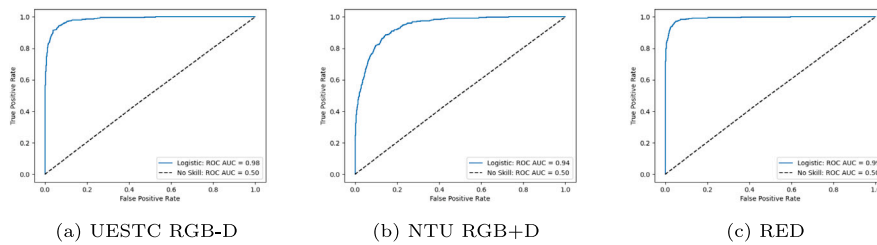


Fig. 11. ROC curves and AUC values achieved by the proposed method on UESTC RGB-D, NTU RGB+D, and RED datasets.

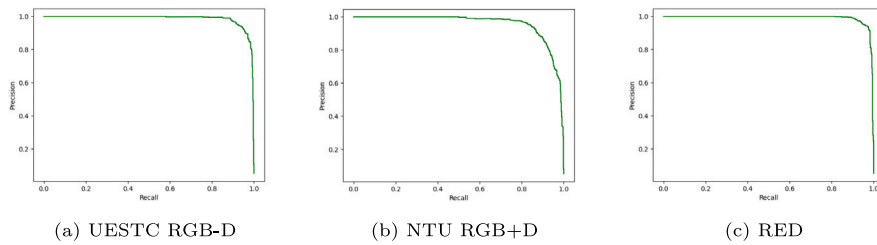


Fig. 12. Precision-Recall curve achieved by the proposed method on UESTC RGB-D, NTU RGB+D, and RED datasets.

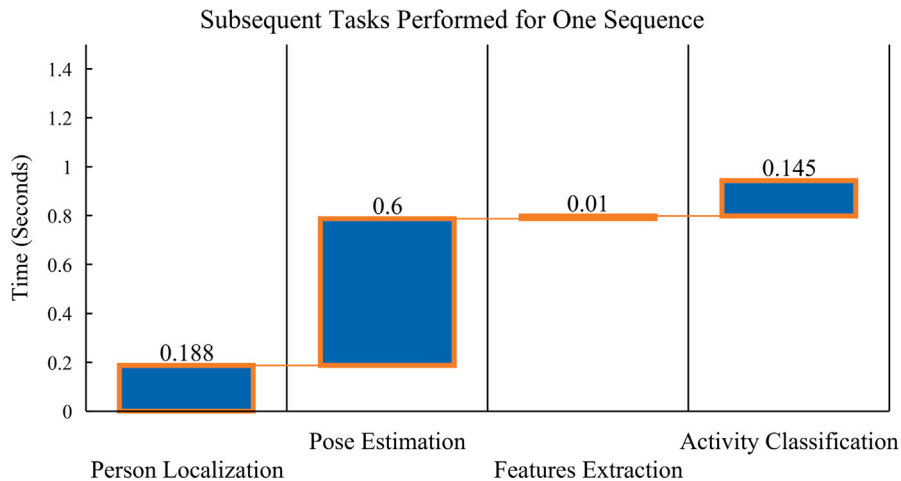


Fig. 13. Time required for the subsequent tasks for one video sequence analysis.

2011), KCF (Henriques et al., 2014), and CSRT (Lukezic et al., 2017), drift in the case of fast-moving objects and challenging situations. In addition, their processing speed is very slow. Fig. 14 shows that the

MOSSE tracker achieved better accuracy and a fast processing rate (FPS), making it most suitable for real-time activity recognition. It is vulnerable to drifting; however, most failures and drifting occur in

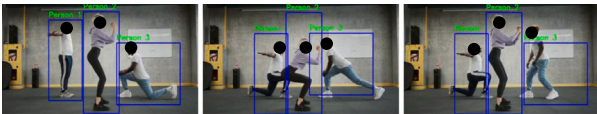
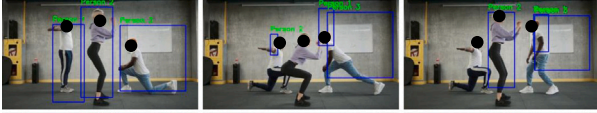
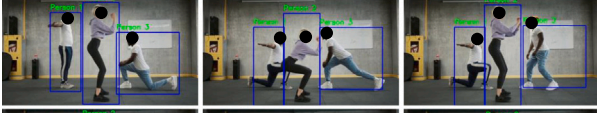
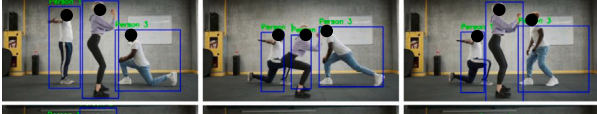
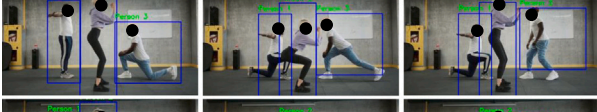
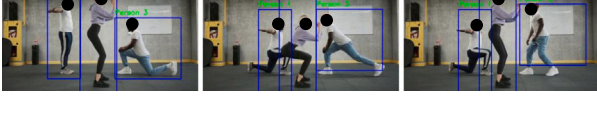
Method	Tracking visual results in intermediate frames	Processing Speed (FPS)
Boosting		15
TLD		22
KCF		44
CSRT		28
DeepSORT		28
MOSSE		275

Fig. 14. Comparison of the results achieved by different tracking algorithms.

the case of a large out-of-plane rotation of the object being tracked. The reason behind this problem is that the MOSSE tracker uses naive filtering (Bolme et al., 2010). Therefore, we used Peak-to-Sidelobe Ratio filtering (MOSSE PSR) rather than naive filters in the proposed framework. MOSSE PSR filtering can efficiently track failure or detect occlusions by measuring the correlation peak's strength to halt the online update and resume tracking the object when it reappears with the same appearance. The detection-based tracker known as DeepSORT (Wojke et al., 2017), which introduced deep learning in tracking, also achieved good accuracy. However, its processing speed is not feasible for real-time multiperson activity recognition.

7. Discussion

In this work, we proposed a vision-based, lightweight multiperson activity recognition framework for group exercise training of elderly individuals. Moreover, we curated a new dataset named Routine Exercise Dataset (RED), which consists of 19 daily physical exercise activities for elderly people. A total of 19 subjects participated in the data collection process held in the Social Robots Lab at NTNU. Besides testing the proposed method on datasets comprising exercise activities, we tested it on other datasets comprising various activities such as yoga actions and daily life activities. A detailed comparison of the proposed method with multiple RNN-based, CNN-based, transformer-based, and hybrid models was performed to evaluate its performance. Our method was able to outperform state-of-the-art methods on various datasets in terms of accuracy (Table 7) and computational complexity (Table 8). We also tested our proposed system by deploying it on a social robot and executing a group exercise session in a real-life environment. A limitation of the existing methods focused on skeleton-based activity recognition is the lack of testing on noise-corrupted datasets. There are significant chances of receiving noisy skeletal data in real-life environments. Therefore, we added various noises to our RED dataset (Section 6.1.11) for testing the performance of the proposed method against noise-corrupted datasets. Our method was able to perform well

on noise-corrupted datasets with a minimal rise in training time, which makes it robust to noisy skeletal data in real-life environments.

The proposed method generalizes well to different types of activities, making it suitable for various applications in eldercare. The target application for the proposed research is multiperson activity recognition for group exercise by elderly individuals. Due to the high generalizability of our developed framework, there can be various real-life applications of the proposed research, including exercise recognition, daily life activity recognition, and fall detection in multiperson environments. The proposed system can be deployed with different interactive technologies, such as 2D screens or social robots, to achieve various real-life applications for eldercare. Moreover, it can also be deployed to monitor distributed environments with various numbers of people at a time. Furthermore, the proposed system can also be utilized for activity recognition in lone or social virtual reality (metaverse)-based applications.

8. Conclusions and future work

Human Activity Recognition (HAR) is considered of prime importance for efficient human-machine interaction. It can play a vital role in various fields of life, including healthcare, and has been a popular research domain in the last few years. However, past research focusing on HAR applications in healthcare has mainly focused on single-person rather than multiperson or group activity recognition. Moreover, these methods mainly use data from wearable sensors, sensors placed on the body of users or sensors in mobile phones, making them cumbersome to use for multiperson activity recognition. On the other hand, the existing HAR methods using vision data employ heavyweight CNNs, which makes them less favorable for real-time applications. Maintaining good accuracy and efficiency in activity recognition systems reported in previous research is challenging. Another challenge is the limited number of publicly available datasets comprised of physical exercise activities.

To accurately and efficiently address these challenges, we proposed a real-time method for multiperson activity recognition with a primary focus on group exercise training of elderly individuals. The main steps

included in this framework are person detection and tracking, pose estimation, feature extraction, and using the proposed 1D MLSTM for activity classification. We fine-tuned and utilized a lightweight CNN model for person detection and an ultrafast object tracker for person tracking. Then, a pose estimation model is used to obtain skeletal data, and features are extracted for each person based on these data. Finally, the MLSTM architecture trained to learn the sequential patterns in a sequence of frames is used to classify the activity. Experiments on 16 datasets in total, including our newly curated Routine Exercise Dataset (RED) without noise presented in this paper, four noise-corrupted RED datasets, 10 benchmark activity recognition datasets, and one combined dataset, confirm the efficiency of the proposed framework in terms of accuracy, generalizability, and computational complexity. Moreover, testing the proposed system in a real-life scenario by integrating it into a social robot confirms its efficiency in real-time applications.

Currently, our present work has some limitations we wish to address in future research. In this work, we used 2D pose estimation that cannot be used to derive angular features in 3D space. In future research, we will investigate 3D pose estimation in a multiperson environment to obtain 3D angular features in real time for better analysis of exercise dynamics and quality assessment of exercise. Moreover, we will collect and contribute another exercise dataset that can assist in the quality assessment of exercise activities tailored to elderly individuals. Finally, we want to conduct user studies in long-term care facilities to analyze the behavior of elderly individuals with the proposed system and discuss its overall impact on society. The present framework can help healthcare professionals conduct group exercise training and recognize other daily life activities to monitor elderly individuals for better health and safety.

CRedit authorship contribution statement

Syed Hammad Hussain Shah: Conceptualization, Methodology, Algorithm development, Data curation, Software, Experiments, Writing – original draft, Writing – review & editing. **Anniken Susanne T. Karlsson:** Supervision, Analysis, Planning, Writing – review & editing. **Mads Solberg:** Supervision, Analysis, Planning, Writing – review & editing. **Ibrahim A. Hameed:** Supervision, Analysis, Planning, Methodology, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors are grateful to the Norwegian University of Science and Technology (NTNU) for supporting the project and all the cordial participants and professionals who made this study possible.

References

- Ahmad, S., Umirzakova, S., Jamil, F., & Whangbo, T. K. (2022). Internet-of-things-enabled serious games: A comprehensive survey. *Future Generation Computer Systems*, 136, 67–83.
- Andrade-Ambriz, Y. A., Ledesma, S., Ibarra-Manzano, M.-A., Oros-Flores, M. I., & Almanza-Ojeda, D.-L. (2022). Human activity recognition using temporal convolutional neural network architecture. *Expert Systems with Applications*, 191, Article 116287.

- Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., & Savarese, S. (2017). Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4315–4324).
- Blindheim, K., Solberg, M., Hameed, I. A., & Alnes, R. E. (2023). Promoting activity in long-term care facilities with the social robot pepper: a pilot study. *Informatics for Health and Social Care*, 48(2), 181–195.
- Bolme, D. S., Beveridge, J. R., Draper, B. A., & Lui, Y. M. (2010). Visual object tracking using adaptive correlation filters. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 2544–2550). IEEE.
- Caetano, C., Brémond, F., & Schwartz, W. R. (2019). Skeleton image representation for 3d action recognition based on tree structure and reference joints. In *2019 32nd SIBGRAPI conference on graphics, patterns and images* (pp. 16–23). IEEE.
- Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7291–7299).
- CareerStaff Unlimited (2023). Biggest challenges of long-term care. <https://www.careerstaff.com/healthcare-staffing-blog/biggest-challenges-of-long-term-care/>.
- Chen, C., Jafari, R., & Kehtarnavaz, N. (2015). 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* (pp. 168–172). IEEE.
- Choi, W., & Savarese, S. (2013). Understanding collective activities of people from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6), 1242–1257.
- Choi, W., Shahid, K., & Savarese, S. (2009). What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops* (pp. 1282–1289). IEEE.
- Cippitelli, E., Gasparrini, S., Gambi, E., & Spinsante, S. (2016). A human activity recognition system using skeleton data from rgbd sensors. *Computational Intelligence and Neuroscience*, 2016.
- Deng, Z., Vahdat, A., Hu, H., & Mori, G. (2016). Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4772–4781).
- Ding, J., & Chang, C.-W. (2015). An eigenspace-based method with a user adaptation scheme for human gesture recognition by using Kinect 3D data. *Applied Mathematical Modelling*, 39(19), 5769–5777.
- Ding, J., & Shi, J.-Y. (2017). Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots. *Computers & Electrical Engineering*, 62, 719–729.
- Dolatabadi, E., Zhi, Y. X., Ye, B., Coahran, M., Lupinacci, G., Mihailidis, A., Wang, R., & Taati, B. (2017). The toronto rehab stroke pose dataset to detect compensation during stroke rehabilitation therapy. In *Proceedings of the 11th EAI international conference on pervasive computing technologies for healthcare* (pp. 375–381).
- Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4), 743–761.
- Ebert, A., Beck, M. T., Mattausch, A., Belzner, L., & Linnhoff-Popien, C. (2017). Qualitative assessment of recurrent human motion. In *2017 25th European signal processing conference* (pp. 306–310). IEEE.
- EK, S., Portet, F., & Lalanda, P. (2022). Lightweight transformers for human activity recognition on mobile devices. arXiv preprint arXiv:2209.11750.
- Faria, D. R., Prenebida, C., & Nunes, U. (2014). A probabilistic approach for human everyday activities recognition using body motion from RGB-D images. In *The 23rd IEEE international symposium on robot and human interactive communication* (pp. 732–737). IEEE.
- Flores-Barranco, M. M., Ibarra-Mazano, M.-A., & Cheng, I. (2015). Accidental fall detection based on skeleton joint correlation and activity boundary. In *Advances in visual computing: 11th international symposium, ISVC 2015, Las Vegas, NV, USA, December 14-16, 2015, proceedings, part II 11* (pp. 489–498). Springer.
- Gaglio, S., Re, G. L., & Morana, M. (2014). Human activity recognition process using 3-D posture data. *IEEE Transactions on Human-Machine Systems*, 45(5), 586–597.
- Gao, W., Zhang, L., Teng, Q., He, J., & Wu, H. (2021). DanHAR: Dual attention network for multimodal human activity recognition using wearable sensors. *Applied Soft Computing*, 111, Article 107728.
- Gavrilyuk, K., Sanford, R., Javan, M., & Snoek, C. G. (2020). Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 839–848).
- Gil-Martín, M., San-Segundo, R., Fernandez-Martinez, F., & Ferreiros-López, J. (2020). Improving physical activity recognition using a new deep learning architecture and post-processing techniques. *Engineering Applications of Artificial Intelligence*, 92, Article 103679.
- Grabner, H., Leistner, C., & Bischof, H. (2008). Semi-supervised on-line boosting for robust tracking. In *Computer vision—ECCV 2008: 10th European conference on computer vision, Marseille, France, October 12-18, 2008, proceedings, part I 10* (pp. 234–247). Springer.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Henriques, J. F., Caseiro, R., Martins, P., & Batista, J. (2014). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3), 583–596.

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Huynh-The, T., Hua, C.-H., & Kim, D.-S. (2019). Learning action images using deep convolutional neural networks for 3D action recognition. In *2019 IEEE sensors applications symposium* (pp. 1–6). IEEE.
- Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., & Mori, G. (2016). A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1971–1980).
- Ihianle, I. K., Nwajana, A. O., Ebeuwa, S. H., Otuka, R. I., Owa, K., & Orisatoki, M. O. (2020). A deep learning approach for human activities recognition from multimodal sensing devices. *IEEE Access*, 8, 179028–179038.
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456). pmlr.
- Islam, M. M., & Iqbal, T. (2020). Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm. In *2020 IEEE/RSJ international conference on intelligent robots and systems* (pp. 10285–10292). IEEE.
- Jaarsma, T., Klompstra, L., Ben Gal, T., Boyne, J., Vellone, E., Bäck, M., Dickstein, K., Fridlund, B., Hoes, A., & Piepoli, M. F. (2015). Increasing exercise capacity and quality of life of patients with heart failure through wii gaming: the rationale, design and methodology of the hf-wii study; a multicentre randomized controlled trial. *European Journal of Heart Failure*, 17, 743–748.
- Jamil, F., Ahmad, S., Whangbo, T. K., Muthanna, A., & Kim, D.-H. (2022a). Improving blockchain performance in clinical trials using intelligent optimal transaction traffic control mechanism in smart healthcare applications. *Computers & Industrial Engineering*, 170, Article 108327.
- Jamil, F., Ibrahim, M., Ullah, I., Kim, S., Kahng, H. K., & Kim, D.-H. (2022b). Optimal smart contract for autonomous greenhouse environment based on IoT blockchain network in agriculture. *Computers and Electronics in Agriculture*, 192, Article 106573.
- Jamil, H., Qayyum, F., Iqbal, N., Jamil, F., & Kim, D. H. (2022c). Optimal ensemble scheme for human activity recognition and floor detection based on AutoML and weighted soft voting using smartphone sensors. *IEEE Sensors Journal*, 23(3), 2878–2890.
- Ji, Y., Xu, F., Yang, Y., Shen, F., Shen, H. T., & Zheng, W.-S. (2019). A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition. arXiv preprint arXiv:1904.10681.
- Kalal, Z., Mikolajczyk, K., & Matas, J. (2011). Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7), 1409–1422.
- Khan, I. U., Afzal, S., & Lee, J. W. (2022). Human activity recognition via hybrid deep learning based model. *Sensors*, 22(1), 323.
- Kim, T. S., & Reiter, A. (2017). Interpretable 3d human action analysis with temporal convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition workshops* (pp. 1623–1631). IEEE.
- Krüger, J., & Nguyen, T. D. (2015). Automated vision-based live ergonomics analysis in assembly operations. *CIRP Annals*, 64(1), 9–12.
- Kumie, G. A., Habtie, M. A., Ayall, T. A., Zhou, C., Liu, H., Seid, A. M., & Erbad, A. (2023). Dual-attention network for view-invariant action recognition. *Complex & Intelligent Systems*, 1–17.
- Lan, T., Sigal, L., & Mori, G. (2012). Social roles in hierarchical models for human activity recognition. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 1354–1361). IEEE.
- Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., & Mori, G. (2011). Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8), 1549–1562.
- Lee, Y.-T., Pengying, T., Yayilgan, S. Y., & Elezaj, O. (2021). Data-driven machine learning approach for human action recognition using skeleton and optical flow. In *Intelligent technologies and applications: Third international conference* (pp. 163–175). Springer.
- Li, Q., Lin, W., & Li, J. (2018). Human activity recognition using dynamic representation and matching of skeleton feature sequences from RGB-D images. *Signal Processing: Image Communication*, 68, 265–272.
- Lim, J. (2023). Effects of a cognitive-based intervention program using social robot PIO on cognitive function, depression, loneliness, and quality of life of older adults living alone. *Frontiers in Public Health*, 11, 313.
- Lukezic, A., Vojir, T., Čehovin Zajc, L., Matas, J., & Kristan, M. (2017). Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6309–6318).
- Lun, R., & Zhao, W. (2015). A survey of applications and human motion recognition with microsoft kinect. *International Journal of Pattern Recognition and Artificial Intelligence*, 29(05), Article 1555008.
- Martínez-Villaseñor, L., Ponce, H., Brieua, J., Moya-Albor, E., Núñez-Martínez, J., & Peñafort-Asturiano, C. (2019). UP-fall detection dataset: A multimodal approach. *Sensors*, 19(9), 1988.
- Mekruksavanich, S., & Jitpattanakul, A. (2021). Lstm networks using smartphone data for sensor-based human activity recognition in smart homes. *Sensors*, 21(5), 1636.
- Mim, T. R., Amatullah, M., Afreen, S., Yousuf, M. A., Uddin, S., Alyami, S. A., Hasan, K. F., & Moni, M. A. (2023). GRU-INC: An inception-attention based approach using GRU for human activity recognition. *Expert Systems with Applications*, 216, Article 119419.
- Miron, A., Sadawi, N., Ismail, W., Hussain, H., & Grosan, C. (2021). IntelliRehabDS (IRDS)—A dataset of physical rehabilitation movements. *Data*, 6(5), 46.
- Neili, S., Gazzah, S., El Yacoubi, M. A., & Amara, N. E. B. (2017). Human posture recognition approach based on ConvNets and svm classifier. In *2017 international conference on advanced technologies for signal and image processing* (pp. 1–6). IEEE.
- Parisi, G. I., von Stosch, F., Magg, S., & Wermter, S. (2015). Learning human motion feedback with neural self-organization. In *2015 international joint conference on neural networks* (pp. 1–6). IEEE.
- Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., Zhao, H., Miao, X., Liu, R., & Fortino, G. (2022). Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. *Information Fusion*, 80, 241–265.
- Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263–7271).
- Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.
- Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., & Parra, X. (2012). Human activity recognition using smartphones. <http://dx.doi.org/10.24432/CS4S4K>, UCI Machine Learning Repository.
- Ronald, M., Poulouse, A., & Han, D. S. (2021). iSPInception: An inception-ResNet deep learning architecture for human activity recognition. *IEEE Access*, 9, 68985–69001.
- Sarkar, A., Banerjee, A., Singh, P. K., & Sarkar, R. (2022). 3D human action recognition: Through the eyes of researchers. *Expert Systems with Applications*, 193, Article 116424.
- Schrader, L., Vargas Toro, A., Konietzny, S., Rüping, S., Schäpers, B., Steinböck, M., Kreuer, C., Müller, F., Güttler, J., & Bock, T. (2020). Advanced sensing and human activity recognition in early intervention and rehabilitation of elderly people. *Journal of Population Ageing*, 13, 139–165.
- Shah, S. H. H., Hameed, I. A., Karlsen, A. S. T., & Solberg, M. (2022a). Towards a social vr-based exergame for elderly users: An exploratory study of acceptance, experiences and design principles. In *Virtual, augmented and mixed reality: Design and development: 14th international conference, VAMR 2022, Held as part of the 24th HCI international conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I* (pp. 495–504). Springer.
- Shah, S. H. H., Han, K., & Lee, J. W. (2020). Real-time application for generating multiple experiences from 360° panoramic video by tracking arbitrary objects and viewer's orientations. *Applied Sciences*, 10(7), 2248.
- Shah, S. H. H., Karlsen, A. S. T., Solberg, M., & Hameed, I. A. (2022b). A social VR-based collaborative exergame for rehabilitation: codesign, development and user study. *Virtual Reality*, 1–18.
- Shah, S. H. H., Steinnes, O.-M. H., Gustafsson, E. G., & Hameed, I. A. (2021). Multi-agent robot system to monitor and enforce physical distancing constraints in large areas to combat covid-19 and future pandemics. *Applied Sciences*, 11(16), 7200.
- Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1010–1019).
- Shojaedini, S. V., & Beirami, M. J. (2020). Mobile sensor based human activity recognition: distinguishing of challenging activities by applying long short-term memory deep learning modified by residual network concept. *Biomedical Engineering Letters*, 10, 419–430.
- Shu, T., Todorovic, S., & Zhu, S.-C. (2017). CERN: confidence-energy recurrent network for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5523–5531).
- SoftBank Robotics (2023). Pepper. Website, <https://us.softbankrobotics.com/pepper>.
- Song, Y.-F., Zhang, Z., Shan, C., & Wang, L. (2022). Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2), 1474–1488.
- Stanford Artificial Intelligence Laboratory (2018). Robotic operating system. <https://www.ros.org/>.
- Stumpf, J. F. (2010). *Motion capture system: US patent 20100304931A1*, URL: <https://patents.google.com/patent/US20100304931A1/en>.
- Sung, J., Ponce, C., Selman, B., & Saxena, A. (2012). Unstructured human activity detection from rgbd images. In *2012 IEEE international conference on robotics and automation* (pp. 842–849). IEEE.
- Taylor, W., Shah, S. A., Dashtipour, K., Zahid, A., Abbasi, Q. H., & Imran, M. A. (2020). An intelligent non-invasive real-time human activity recognition system for next-generation healthcare. *Sensors*, 20(9), 2653.
- Tomas, A., & Biswas, K. (2017). Human activity recognition using combined deep architectures. In *2017 IEEE 2nd international conference on signal and image processing* (pp. 41–45). IEEE.
- Ullah, A., Muhammad, K., Ding, W., Palade, V., Haq, I. U., & Baik, S. W. (2021). Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications. *Applied Soft Computing*, 103, Article 107102.
- Vakanski, A., Jun, H.-p., Paul, D., & Baker, R. (2018). A data set of human body movements for physical rehabilitation exercises. *Data*, 3(1), 2.
- García de Villa, S., Casillas-Pérez, D., Jimenez-Martin, A., & García-Domínguez, J. J. (2022). Simultaneous exercise recognition and evaluation in prescribed routines: Approach to virtual coaches. *Expert Systems with Applications*, 199, Article 116990.
- Wan, S., Qi, L., Xu, X., Tong, C., & Gu, Z. (2020). Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications*, 25, 743–755.

- Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 1290–1297). IEEE.
- Wang, M., Ni, B., & Yang, X. (2017a). Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3048–3056).
- Wang, D., Zhang, C., Cheng, H., Shang, Y., & Mei, L. (2017b). SPID: surveillance pedestrian image dataset and performance evaluation for pedestrian detection. In *Computer vision–ACCV 2016 workshops: ACCV 2016 international workshops* (pp. 463–477). Springer.
- Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing* (pp. 3645–3649). IEEE, <http://dx.doi.org/10.1109/ICIP.2017.8296962>.
- Xia, L., Chen, C.-C., & Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 20–27). IEEE.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048–2057). PMLR.
- Yadav, S. K., Agarwal, A., Kumar, A., Tiwari, K., Pandey, H. M., & Akbar, S. A. (2022a). YogNet: A two-stream network for realtime multiperson yoga action recognition and posture correction. *Knowledge-Based Systems*, 250, Article 109097.
- Yadav, S. K., Luthra, A., Tiwari, K., Pandey, H. M., & Akbar, S. A. (2022b). ARFDNet: An efficient activity recognition & fall detection system using latent feature pooling. *Knowledge-Based Systems*, 239, Article 107948.
- Yang, X., & Tian, Y. (2014). Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1), 2–11.
- Zaabar, B., Cheikhrouhou, O., Jamil, F., Ammi, M., & Abid, M. (2021). HealthBlock: A secure blockchain-based healthcare data management system. *Computer Networks*, 200, Article 108500.