IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Smarter Aging: Developing A Foundational Elderly Activity Monitoring System with AI and GUI Interface

Ye Htet[1], Graduate Student Member, IEEE, Thi Thi Zin[2], Member, IEEE, Pyke Tin[2], Member, IEEE, Hiroki Tamura[2], Kazuhiro Kondo[3], Shinji Watanabe[3], and Etsuo Chosa[3]

[1]Interdisciplinary Graduate School of Agriculture and Engineering, University of Miyazaki, Miyazaki 889-2192, Japan
[2]Graduate School of Engineering, University of Miyazaki, Miyazaki 889-2192, Japan
[3]Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan

Corresponding author: Thi Thi Zin (e-mail: thithi@cc.miyazaki-u.ac.jp).

This work was partially supported by the JST SPRING under Grant JPMJSP2105. This study only involved human subjects. The protocols were approved by the Ethics Committee of the University of Miyazaki, Japan (protocol code O-0451, on January 28, 2019) and (protocol code O-1449 on November 20, 2023), with a waiver of written informed consent obtained from all participants.

ABSTRACT The global rise in the elderly population, which presents challenges to healthcare systems owing to labor shortages in caregiving facilities, necessitates innovative solutions for elderly care services. Smart aging technologies such as robotic companions and digital home gadgets, offer a solution to these challenges by improving the elderly's quality of life and assisting caregivers. However, limitations in data privacy, real-time processing, and reliability often hinder the effectiveness of the existing technologies. Among these, privacy concerns are a major barrier to ensuring user trust and ethical implementation. Therefore, this study proposes a more effective approach for smart aging through elderly activity monitoring that prioritizes data privacy. The proposed system utilizes stereo depth cameras to monitor the activities of the elderly. Data were collected from real-world environments with the participation of six elderly individuals from a care center and hospital. This system focuses on recognizing common daily actions of the elderly including sitting, standing, lying down, and seated in a wheelchair. Additionally, it recognizes transition states (in-between actions such as changing from sitting to standing) that are crucial for assessing balance issues. By integrating motion information with a deep-learning architecture, the system achieved a high accuracy of 99.42% in recognizing daily actions in real-time. This high accuracy was maintained even with minimal data from new environments through transfer learning, and the adaptability of this model ensured its potential for real-world applications. For intuitive interaction between the caregivers and the system, a user-friendly graphical interface (GUI) was also designed in the proposed approach.

INDEX TERMS deep learning architecture, elderly activity monitoring, GUI, motion information, real-time action recognition, smart aging, stereo depth cameras, transition state recognition

I. INTRODUCTION

The global population aged 65 and above is rapidly growing, placing significant strain on healthcare systems as the demand for services and nursing care increases [1, 2]. Declining mobility and health are common issues associated with aging that significantly affect the quality of life and independence of the elderly [3]. Therefore, it is necessary to develop effective solutions to promote mobility and independent living of the elderly without overburdening caregivers with excessive workloads.

One promising solution involves understanding the well-being of the elderly through the concept of 'smart aging' [4]. Smart aging can be defined as an innovative approach that enables the elderly population to live freely, securely, comfortably, healthily, and happily [2]. Although there are various ways to facilitate smart aging for the elderly, the utilization of modern technologies has increased in recent years, leveraging advanced software and hardware. Assisted living [5], [6] and healthcare monitoring [7], [8] are among the approaches aimed at

helping elderly individuals with independent living and smarter aging.

However, existing smart aging technologies often rely on physical sensors in the environment or require intrusive wearables, which can be inconvenient and limit the mobility of elderly individuals. In addition, privacy concerns can arise with certain monitoring methods to ensure user trust and ethical implementation. Hence, this study proposes a more effective approach to smart aging through indoor activity monitoring for the elderly by pushing the boundaries of the existing limitations.

The primary objective was to develop an activity monitoring system for elderly people in indoor settings using stereo depth cameras while keeping everything confidential. To achieve robust performance, Deep Learning (DL), a subset of Machine Learning (ML) and Artificial Intelligence (AI), recognized as a cornerstone of the Fourth Industrial Revolution (4IR or Industry 4.0, is employed. Notably, the DL approach is particularly well-suited to this application compared to ML or the Internet of Things (IoT). This is because traditional ML algorithms such as Support Vector Machine (SVM) and k-nearest neighbors (kNN) often require manual feature extraction from the data (depth information in this case). This can be a time consuming and domain-specific process that requires expert knowledge. In contrast, DL algorithms, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks excel at automatically learning these features directly from the data, making them less reliant on human intervention. Moreover, DL architectures, with their multi-layered structure, are adept at handling complex relationships within the data, leading to more accurate recognition of the intended actions. On the other hand, IoT generally focuses on connecting devices and sensors to collect and share data. Although depth cameras can be integrated into IoT systems, real-time action recognition often requires additional processing and analysis. By contrast, DL excels in this analysis by extracting meaningful data. Therefore, the proposed approach offers several advantages over other existing methods. It eliminates the need for wearable devices or sensors that may interfere with the elderly, allows for easy camera installation within the room, is cost-effective, achieves robust performance without manual feature extraction from the data, and preserves privacy by utilizing depth rather than color images.

To ensure the practicality of the system, data were collected from real-world environments, including a care center and hospital, with the participation of six elderly individuals (three from the care center and another three from the hospital). The focus was on recognizing seven common daily actions indoors: seated in a wheelchair, standing, sitting on the bed, lying on the bed, transition states between these actions, being outside the room, and receiving assistance.

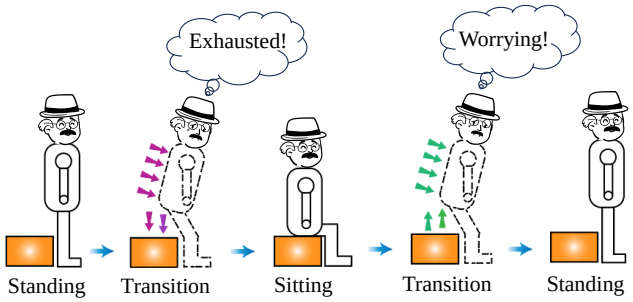


FIGURE 1. Illustration of transition states between sitting and standing actions.

Among them, transition states, which denote changes in body position and movement from one specific action to another (e.g., changing from sitting to standing), are crucial elements in the daily monitoring routines of the elderly. Fig. 1 illustrates the example transition states between sitting and standing actions. During transition states, elderly individuals may experience feelings of exhaustion due to the need for body balance or may be concerned about falling due to weakened physical conditions. Hence, recognizing transition states is important for health monitoring in the elderly. Such recognition addresses the challenges associated with impaired mobility and balance, thereby promoting the overall safety of elderly individuals.

Previously, action recognition models [9], [10], [11] were developed for elderly activity monitoring using various feature extraction methods and ML-based recognition approaches. While these models achieved reliable performance for specific actions, such as sitting, standing, and lying down, accurately recognizing transition states proved challenging. Transitional movements are often confused with specific actions because of their similar visual patterns, leading to lower recognition rates. To address this limitation, the proposed study refines the feature extraction and classification method for action recognition from prior studies to achieve a more effective distinction between transition states and specific actions. Motivated by this, accurate action classification is crucial for precisely recognizing the actions of the elderly. This study addresses this need by proposing an efficient classification method using DL algorithms.

The main findings of this study indicate that the system achieves significantly higher accuracy rates by integrating motion information into existing DL architectures. This novel architecture, named Motion-based Convolutional Recurrent Neural Networks (MotionCRNN), is described in detail in the methodology section. This model performs well, particularly in recognizing transition states, compared with the authors' prior studies, owing to the utilization of DL algorithms in action recognition. In addition to prioritizing the accurate recognition of transition states, the proposed method also

effectively recognizes other common daily actions and enables real-time processing. Furthermore, the adaptability of the proposed MotionCRNN model to transfer learning ensures its potential applicability to different environments.

In caregiving facilities, elderly residents typically rely on caregivers for monitoring. However, caregivers are unable to provide constant monitoring. The proposed system was developed to reduce the workload of caregivers while monitoring the well-being of the elderly. Therefore, prioritizing the interaction between the system and caregivers is more important than interaction with the elderly. To achieve this, a user-friendly Graphical User Interface (GUI) was designed in the proposed approach to assist caregivers and provide a convenient environment for seniors. The results obtained through the proposed DL model and GUI can be shared with the care staff, family members, and healthcare providers, enabling comprehensive monitoring and potentially leading to early interventions.

The main contributions of the paper are as follows:

1) Proposing an effective approach for real-time action recognition in indoor elderly activity monitoring by integrating stereo depth cameras and DL algorithms.
2) Enabling the proposed MotionCRNN to achieve promising recognition of transition states by integrating motion information into existing hybrid DL architectures.
3) Demonstrating MotionCRNN's capability for generalizability in recognizing elderly actions across different environments through transfer learning.
4) Illustrating the proposed system's reliability by developing and evaluating using real-world datasets.

The remainder of this paper is organized as follows: Section II reviews the literature related to the proposed system. Section III describes the methodology employed, and Section IV presents the experimental results. Section V provides a discussion, including limitations and future implications. Finally, Section VI concludes the paper.

II. RELATED WORKS

In this section, literature related to the proposed system is reviewed. Subsections A and B cover smart aging technologies and indoor elderly monitoring systems using various sensors and cameras. In Subsection C, different approaches to identifying transition states in elderly actions are discussed. Subsection D describes the relevant research on action recognition that employs DL algorithms.

A. SMART AGING TECHNOLOGIES

Smart aging technologies offer a range of innovative solutions to support elderly people in their daily lives and promote aging. These solutions encompass smart home products, gadgets, wearable devices, remote monitoring systems, and IoT-enabled healthcare applications [12], [13], [14]. These include features such as fall detection, electronic fences, temperature monitoring, and sleep monitoring. For example, a smart wearable device based on IoT has been designed to monitor physiological parameters in real time and provide remote access to the elderly's health status [15]. On the other hand, public entities deploy and operate smart mobility technologies to improve mobility and independence for older adults, while reducing operating costs [16]. Similarly, smart grid technology has been developed to provide useful information on the activities of daily living and monitor the short and long-term health of elderly individuals [17]. Owing to advancements in technology, AI has played a crucial role in developing smart aging systems to personalize healthcare for the elderly. For instance, AI tools such as ML and DL models are used to develop solutions that improve quality of life and autonomy and reduce caregiver burden [18], [19], [20], [21].

However, challenges arise in the implementation of personalized healthcare using smart aging technologies. Typical challenges include the potential disruption of existing care systems, technological literacy gaps, and privacy concerns due to constant monitoring [22], [23], [24]. Moreover, security vulnerabilities in IoT systems [25] and ethical considerations in AI must be addressed carefully. For instance, co-adaptation between technology and the elderly is crucial for user satisfaction and long-term adoption [26]. Therefore, a person-centered approach and sufficient governance are necessary to ensure generalizability, transparency, and effectiveness [24] in implementing smart aging technologies.

Overall, smart aging technologies offer promising solutions for aging and enhancing the well-being of the elderly. However, addressing security vulnerabilities, ethical considerations, and implementation challenges are crucial for successful adoption and impact. Motivated by this, this study addresses data privacy concerns in smart aging through indoor elderly activity monitoring using stereo depth cameras. This practical system, developed and evaluated for easy adoption in real-world environments, utilizes data collected from a care center and hospital with the participation of elderly individuals. As an ethical consideration, a waiver of written informed consent was obtained from all participants, and the data acquisition protocol received ethical approval for the experiment. Some related systems for indoor elderly activity monitoring are explained in the next subsection.

B. INDOOR ELDERLY MONITORING SYSTEMS

Elderly monitoring refers to an indoor system designed to process data related to the daily activities of the elderly, collected from sensors or cameras. It provides information concerning health conditions and behavioral status to aid in understanding the well-being of the elderly. A recent introduction to activity monitoring utilized wearable sensor

data and environment-independent fingerprints generated from wireless-fidelity channel state information using a hybrid DL model [27]. This system aimed to enhance the independence of the elderly and visually impaired individuals, achieving an accuracy of 99% in experiments conducted on two public datasets featuring various activities. However, sensor-based systems sometimes face challenges, including noisy data affecting the accuracy, unreliable readings owing to sensor placements, and the need for sophisticated data collection and processing. Additionally, they often require frequent charging, causing inconvenience for the elderly who may forget to use them.

In contrast, camera-based systems have gained popularity owing to several advantages [9]. They are non-invasive and comfortable, aligning well with the principles of smart aging to promote freedom and comfort among the elderly. Cameras offer a broader field of view, enabling monitoring of multiple activities using one device within a room or area. Importantly, they can serve multiple purposes beyond action recognition, including fall detection, medication monitoring, and remote communication. However, they also present challenges such as privacy concerns and limitations in environments with poor lighting or clutter.

On the other hand, depth cameras offer several distinct advantages over traditional RGB cameras. Whereas regular cameras capture 2D information, depth cameras provide 3D depth data, revealing the distance between the objects and camera sensor [28]. Thus, depth data offers privacy advantages because they capture distance information without facial details or identifiable features. Moreover, depth cameras perform well under low-light conditions, where regular cameras struggle, making them suitable for monitoring various indoor environments with limited lighting. However, depth cameras also have limitations such as limited sensing distance and low resolution [29]. Despite these limitations, the evolution and advancement of depth cameras in gaming, automotive, and medical fields have led to their increasing application in elderly care and smart homes.

Several studies have explored the use of depth cameras to monitor the elderly by analyzing their activity patterns [30], [31], [32], [33]. For example, a non-invasive sleep monitoring system was developed using a 3D depth camera (Microsoft Kinect II) [30] with the aim of long-term monitoring of sleep behaviors in seniors. Another study utilized depth-video-based methods for human activity recognition in indoor environments [31], and achieved efficient and robust results by experimenting with three publicly available depth datasets. In addition, a framework for fall detection that utilizes both accelerometer data and depth maps from a Kinect sensor was proposed [32], demonstrating a high performance in differentiating falls from other daily activities. The experiment was conducted on a public fall detection dataset and achieved a high performance. Furthermore, a solution was proposed that solely utilizes depth information from RGB plus depth (RGBD) cameras to monitor the elderly within indoor living spaces [33], enabling remote monitoring by family members and caregivers to understand their behavior and take appropriate action when needed.

Through a review of previous studies, it is evident that various categories are included for elderly monitoring purposes, such as sleep monitoring, fall detection, remote monitoring, and activity recognition. However, many of these systems rely on public datasets or performance datasets demonstrated by young people rather than testing actual elderly data. In addition, the camera view in most datasets is typically located in front of a person, which may be uncomfortable or impractical in real-world scenarios. In contrast, the system in this study collects real-world data on the elderly from care centers by using stereo depth cameras. The cameras were strategically placed above a person at a downward angle to minimize interference with the person's activities. This system aims for 24-hour monitoring and real-time action recognition processing for elderly residents in indoor settings. In the action recognition process, primitive actions such as sitting, standing, seated in the wheelchair, and lying down are identified and transition states that can pose potential risks, such as falling and abnormal activities are recognized. Different approaches to recognizing transition states for the elderly are explained in the following subsection.

C. TRANSITION-AWARE ACTION RECOGNITIONS

Transitions between actions are usually disregarded in action recognition because of their low incidence and short duration compared with other actions, affecting the performance of recognition systems if not properly addressed [34]. Hence, the real-time detection of transitions between actions is a valuable but somewhat untapped challenge, especially for continuous human daily action recognition [35], [36]. Although transition-aware action recognition faces several key challenges, it offers various benefits. The potential applications of transition-aware action recognition span online health monitoring, smart environments, and biomedical engineering, highlighting its relevance and impact in real-world scenarios.

Several studies have proposed effective systems for recognizing transition states in human action. For instance, real-time ML-based methods have been employed for automatic segmentation and recognition of continuous human daily action by integrating change point detection algorithms with smart home action recognition [37], [38]. Among them, an online change point detection strategy was introduced [36] that segmented continuous multivariate time-series smartphone sensor data and applied it to a transition-aware action recognition framework based on the hypothesis and verification principle. When paired with an ensemble classifier, the authors stated that their proposed strategy achieved recognition rates up to 99.8%. In another study, a transition-aware context network was proposed [39] to distinguish transition states. The network comprised two

components: a temporal context detector to extract long-term context information and a transition-aware classifier to classify actions and transition states. Utilizing spatiotemporal features, the network achieved a competitive performance and significantly outperformed state-of-the-art methods on the untrimmed UCF101 dataset. Moreover, CNN models were utilized to recognize transition actions, and the effectiveness of the approach was demonstrated through experiments with fuzzy logic [40].

Another innovative approach [41] emphasizes the incorporation of realistic leg movements in valid human motion transitions. Referring to action-conditioned in-betweening learning, the approach focused on encouraging transition naturalness through leg movements. Experiments on three benchmark datasets demonstrated a high performance in terms of visual quality, prediction accuracy, and action faithfulness. However, to implement this approach, the legs of the person must be clearly detected. Furthermore, an algorithm based on Standard Deviation Trend Analysis (STD-TA) of sensor data was proposed [42] for recognizing transition states, defined as the transitional process between two different basic actions. The evaluation of the real data yielded an accuracy of over 80% with their model. A smartphone-based system [43] has also been developed for transition recognition. In previous research, transition-aware elderly action recognition systems were developed using a combination of the Hidden Markov Model (HMM) with ML algorithms and temporally dependent features extracted from body movements [10], [11]. However, both approaches face challenges such as confusion between transition states and specific actions, leading to unsatisfactory results.

The related studies mentioned above share a common approach of utilizing time-series or spatiotemporal features to identify transition states from other actions, although they employ different classifier models. Building on these concepts, the system in this study utilizes spatiotemporal features extracted from the body movements of the elderly to distinguish between transition states and primitive actions as well as among specific actions. Specifically, spatial features focus on body and object positioning, whereas temporal features capture the movement dynamics over time. To achieve this, the proposed system introduces a novel fusion of deep-learning algorithms, namely CNN and Recurrent Neural Networks (RNN), applied to depth data. These algorithms automatically extract spatiotemporal features from data, thereby enhancing the accuracy and effectiveness of the recognition process. The approach for extracting spatiotemporal features using DL algorithms and related works is described in the following subsection.

D. CRNN-BASED ACTION RECOGNITION MODELS

Research on spatiotemporal feature extraction and action recognition has explored traditional methods [44], [45], and DL techniques [46], [47], [48], [49], [50], [51]. DL models, such as CNNs and RNNs, are valuable because they can autonomously learn complex features and reduce reliance on manually crafted features. CNNs capture spatial features from video frames, whereas RNNs manage temporal dependencies by processing feature sequences over time. Integrating CNNs and RNNs for spatiotemporal feature extraction offers advantages in terms of accuracy and efficiency, as proven in existing literature. Therefore, the proposed system uses CNNs to encode spatial features and RNNs to decode temporal dependencies. These components were then fused and built into a single-model hybrid architecture for action recognition.

Several studies have investigated the application of CNNs and RNNs in action recognition. For example, one approach proposed recognizing human actions from videos using a combination of deep CNN and multi-layered RNN, specifically LSTM units [46]. CNNs extract features from individual video frames, whereas LSTMs processes the sequence of extracted features to capture temporal information. In their approach, different GoogLeNet architectures were used to extract various features from images. The extracted features were then converted into sequences and fed into multi-layered LSTMs. Finally, a softmax regression classifier categorizes the videos based on processed features. Notably, the network architecture utilizes both residual and inception blocks to handle convergence during the training process. Experiments showed that this approach, particularly the combination of multi-layered LSTMs with the Inception_Residual model, improved the evaluation performance.

Another study proposed a novel architecture using convolutional and recurrent networks for action recognition [47]. The approach incorporated separate layers to capture spatial and temporal information. In the first stage, that is, feature extraction, they utilized an improved p-non-local operation within a deep CNN. This operation is a simple and effective way to capture long-distance dependencies in video data. In the second stage, class prediction, they introduced a novel technique called fusion keyless attention. This was combined with a forward and backward bidirectional LSTM network to learn the sequential nature of the data, that is, how actions unfold over time. Their experiments on two datasets demonstrated that this model outperformed the traditional models.

In addition, researchers have explored various DL architectures for action recognition in videos [48]. They employed transfer learning from powerful pre-trained models to improve performance. The approach utilized two types of CNNs: one analyzing spatial information from RGB image frames and another capturing motion information through optical flow. Both leveraged pre-trained models for efficient feature extraction. In addition to separating the spatial and temporal feature extraction, their study also investigated combining them. They employed various CNN-RNN architectures, where CNNs (ResNet101, GoogleNet, and

VGG16) act as encoders to extract features and RNN variants (LSTM, Bi-directional LSTM, Gated Recurrent Unit (GRU), and Bi-directional GRU) act as decoders to handle the sequential nature of video data. The researchers proposed six additional aggregation networks after generating the individual models (one motion CNN model, three spatial CNN models, and twelve CNN-RNN fusion models). These networks used a technique called Average Fusion to combine the outputs from the spatial and temporal CNNs, as well as CNN-RNNs. This was aimed at further improving the overall action recognition performance.

Another approach utilized a Deep Bidirectional LSTM (DB-LSTM) network for action recognition in long videos [49]. The method combines a CNN for feature extraction and a DB-LSTM to handle the sequential nature of video data. The approach first extracts spatial features from every sixth frame of the video using a pre-trained CNN model (AlexNet) to reduce redundancy and complexity. A deep DB-LSTM network then processes the extracted features. By stacking multiple layers in both the forward and backward directions, the DB-LSTM learns long-term dependencies within the video sequence, making it suitable for analyzing longer videos. Experiments showed that this approach achieved state-of-the-art performance on the UCF-101, HMDB51, and YouTube action video datasets, outperforming other recent techniques.

Recent research has focused on overcoming the limitations of DL-based action recognition, particularly in terms of speed, scalability, and accuracy. One promising approach involves lightweight architecture and transformer neural networks. These techniques aim to address challenges such as high computational demands by offering reduced model size and faster processing. Additionally, transformer networks can effectively capture long-range dependencies within video data, potentially improving the recognition accuracy. For example, a recent study proposed Vision and Recurrent Transformer Neural Networks (ViT-ReT) for human action recognition in videos [50]. The framework combined a Vision Transformer (ViT) for efficient feature extraction and a Recurrent Transformer (ReT) to model the temporal information within a video sequence. Researchers compared ViT-ReT with traditional CNN and RNN-based approaches on several benchmark datasets. Their findings demonstrated that ViT-ReT achieved a significant speedup compared with the baseline method (ResNet50-LSTM) while maintaining comparable accuracy. Furthermore, ViT-ReT outperformed the state-of-the-art methods in terms of both accuracy and processing speed, making it suitable for resource-constrained and real-time activity recognition applications.

Previous approaches have developed a combination of CNN and RNN for action recognition by adjusting the model parameters and scaling architecture. However, many of these models extract spatial features from each single frame in a long sequence, potentially impacting real-time performance by processing the entire image. By contrast, the proposed system is simple and effective for real-time processing. First, the person in the image is segmented, and then spatial features are extracted from this segmented region rather than from the entire image, reducing the processing time. In addition, motion information is incorporated from two consecutive frames for the CNN to extract features, rather than a single frame, enhancing the model's capability. However, for the encoder-decoder aggregation of CNN and RNN, the same concept is built upon inspired by previous works, but with lightweight network architectures. Consequently, the proposed model is named MotionCRNN, and its details are explained in Section III.

In many cases, the results from DL models are not used directly; instead, they are refined using techniques, such as majority voting and reasoning, which are crucial for real-world applications. Several related studies have applied majority voting decisions and conditional reasoning to action recognition predictions. For example, a sliding window and majority voting skeleton-based approach was developed for online human action recognition using spatiotemporal graph convolutional neural networks [52]. The results demonstrated the high performance and efficiency of the majority-voting approach. Similarly, a model was developed to predict four different actions using majority voting for gameplay [53]. The findings indicated that majority voting yielded more accurate predictions with 92.59% accuracy, exceeding the peak accuracy value of individual pre-trained models. Subsequently, a model blending technique [54] was developed using majority voting in an ensemble of DenseNet-201 and ResNet-50 for melanoma classification. This method displayed satisfactory results, demonstrating the influence of majority voting decisions.

Regarding reasoning, one study [55] improved the performance of action recognition by modeling causal relationships based on preconditions and effects. The suggested cycle-reasoning model demonstrated improved action recognition performance through efficient reasoning about preconditions and effects. Additionally, an action reasoning framework [56] that uses prior knowledge was proposed to explain the semantic-level observations of video state changes. The experimental results indicated an improvement in recognition using this reasoning approach.

Related studies have demonstrated the effectiveness of majority voting and reasoning in refining the action recognition results. Motivated by this, majority voting is applied in the proposed approach to refine the prediction results and conditional reasoning is used to address the potential over-segmentation of transition states. The difference between the previous works and the proposed method is the utilization of sequential-based majority voting decisions and reasoning to reduce over-segmentation in transition states. This approach aims to enhance the accuracy and robustness of the system to effectively recognize the transition states.

III. METHODOLOGY

This section describes the details of the proposed process flow, including how data are obtained from the elder care center, the understanding of the recorded data, the visualization of the input images, and an overview of the system.

A. DATA ACQUISITION PROCESS

To develop the proposed system, experimental data were initially gathered from an elderly care center in Miyazaki City, Japan. The study involved three senior residents from that care center, and their comprehensive health profiles and recording criteria are presented in Table. I. All residents aged 65 years and older were diagnosed with cognitive decline or frailty. Depth image data were recorded inside separate rooms for a continuous period of 24 hours. This data acquisition protocol received ethical approval from the University of Miyazaki Ethics Committee (protocol code O-0451, on January 28, 2019), and a waiver of written informed consent was obtained from all participating individuals.

Fig. 2 illustrates the experimental environment and a representative sample. For data acquisition, stereo depth cameras were used to capture the daily activities of the elderly residents. The Mini PCs served as processing units for executing the recording procedure, and the resulting data were stored in external HDDs. To prevent accidental interactions with the cameras, they were strategically positioned above the curtain beside the bed and angled downward 45° toward the bed inside the room. The distance between the depth camera and bed was maintained at 2.5 meters, with the depth camera positioned 2.1 meters above the ground. During the recording, only depth data (distance information) were captured to preserve participants' privacy, with color images intentionally omitted. This data recording strategy lasted from a minimum of one day to a maximum of three days in each room.

B. UNDERSTANDING THE DATA

The initially recorded depth data, representing the distance values measured from the camera to the objects, were stored as raw floating-point data. To facilitate data retrieval, the floating-point values were saved in a Comma-Separated Value (CSV) file format, as illustrated in Fig. 3. Each frame was structured as an image with a resolution of 320×180 pixels, capturing the data at a frame rate of 5fps. For enhanced visualization and subsequent analysis, the retrieved depth images were colorized using the hue space colorization method, a process interpreted in a previous study [10]. Fig. 4 shows a colorized image of the sample data from each room. The specific details regarding the recorded data after removing the error frames are listed in Table. II.

During the recording period, elderly residents engaged in regular daily activities. Prominent and frequently occurring actions included "seated in the wheelchair," "standing," "sitting on the bed," and "lying on the bed," along with "transition states" changing from one action to another.

TABLE I
DATA ACQUISITION PROTOCOL FOR CARE CENTER

Participants	Three elderly residents (each allocated a separate room)
Selection Criteria	1) Residents aged 65 years and older, 2) Residents diagnosed with cognitive decline or frailty, 3) Residents in stable medical conditions, 4) Residents who have been fully informed about their participation in this study, 5) Residents who have voluntarily provided written consent after a comprehensive understanding of the study.
Collected Data Information	Recorded depth image data inside the room for a continuous period of 24 hours.

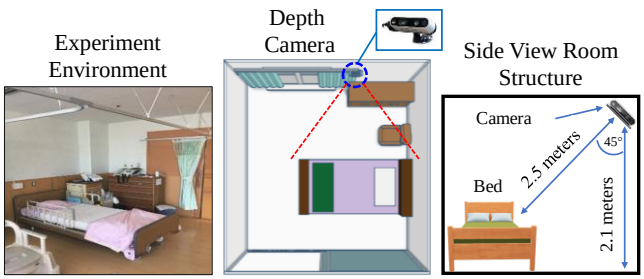


FIGURE 2. Illustration of the experimental environment (care center).

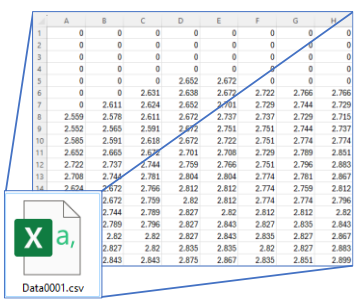


FIGURE 3. Example of depth data in CSV file format.

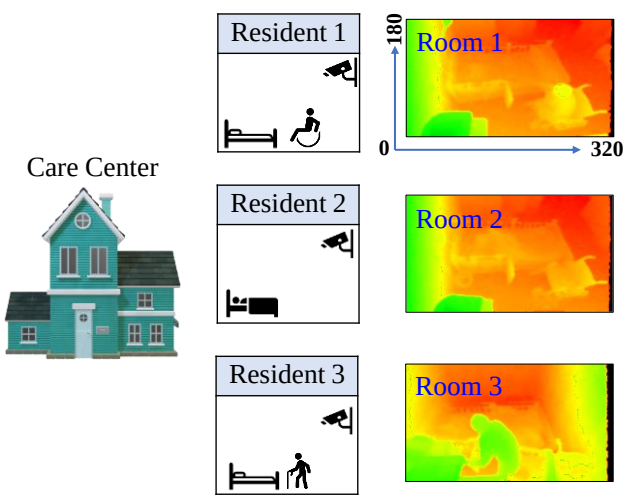


FIGURE 4. Sample colorized depth images from the care center.

TABLE II
RECORDED DATA FROM CARE CENTER

Room ID	Recorded Dates (yyyy/mm/dd_hh:mm) *Start Time*	*End Time*	Duration (hours)
1	2019/10/12_10:15	2019/10/13_04:14	18
1	2019/10/18_11:34	2019/10/24_13:21	146
2	2019/10/25_11:45	2019/10/28_06:53	67
3	2019/10/12_11:10	2019/10/13_05:10	18
3	2019/10/18_11:25	2019/10/22_20:33	105
3	2019/10/25_12:00	2019/10/28_07:22	68

Notably, there was typically only one elderly person inside the room, primarily present during bedtime or rest time, with the person being "outside" the room at other times. Another significant state acquired was when the elderly person was "receiving assistance," typically when a nurse or caregiver entered the room to provide aid. However, certain activities such as folding clothes and cleaning the cabinet near the bed were not considered significant components of their daily routines. Consequently, the importance of distinguishing among the four primitive actions (seated in the wheelchair, standing, sitting, and lying down) and three states (transition, outside, and receiving assistance) was emphasized for the elderly observed in this experiment.

C. ACTION VISUALIZATION

By utilizing the colorization approach, the visual representation of the depth images was enhanced, providing a clearer interpretation of each intended action and state, as illustrated in Fig. 5. In the figure, the term "outside" signifies the absence of any individual within the camera view. Other actions are represented when a person is inside the room. Firstly, "seated" indicates the person is in a wheelchair, positioned straight. Secondly, "standing" denotes that the person is in an upright posture. Thirdly, "sitting" represents the elderly individual resting on the bed. Then, "lying" signifies that the elderly person is lying down for rest in a straight posture. The term "transition" refers to an individual being in a state of change from one action to another, highlighting a period in which unexpected accidents may occur. Finally, "assistance" indicates that the elderly person is currently receiving support from a healthcare provider.

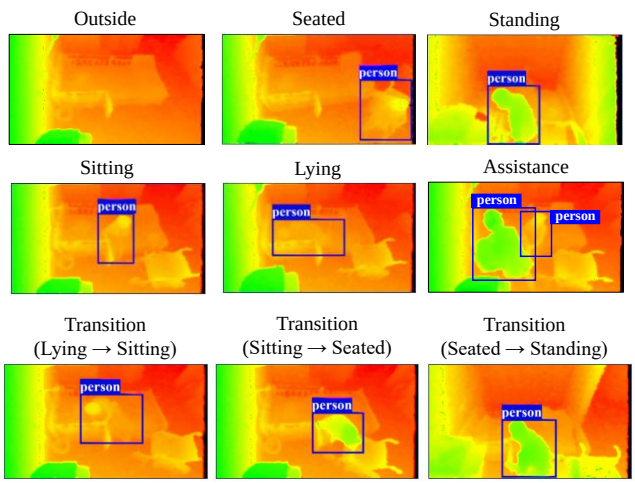


FIGURE 5. Visualization of actions and states.

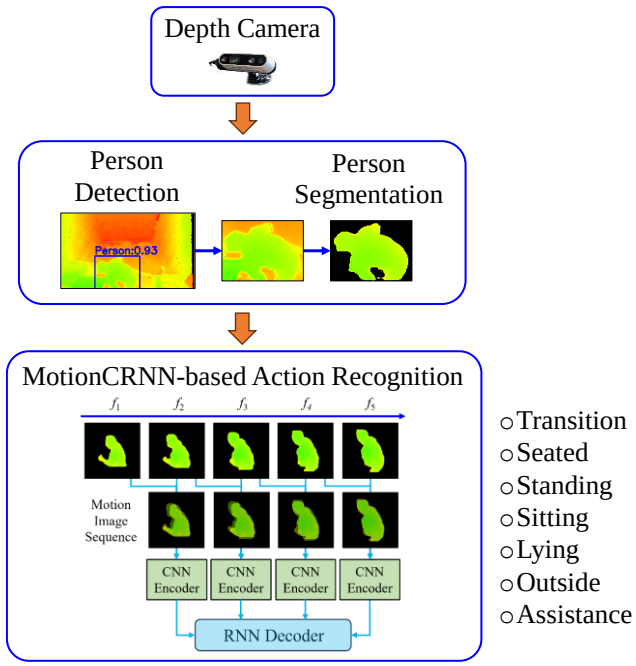


FIGURE 6. System overview.

D. SYSTEM OVERVIEW

The system overview is shown in Fig. 6, encompassing three key steps: depth-data processing, person detection and segmentation, and MotionCRNN-based action recognition. In general, the system uses depth image data from the camera as input and outputs information regarding the actions or states in which the targeted elderly person is engaged. The following subsections provide an in-depth explanation of each step of this process.

1) DEPTH-DATA PROCESSING

The data obtained from the depth camera were initially in the form of raw depth data, characterized by substantial noise and depth fluctuations. To address this, instead of utilizing raw data, depth-processing methods were employed to smooth the data and fill the missing pixel values. Building upon an approach similar to that used in a previous work [9], the method was refined for implementation in the proposed system. Depth processing in the proposed system involved the following sequential steps: (i) hole filling, (ii) depth-to-disparity conversion, (iii) application of bilateral spatial filtering, (iv) conversion of disparity back to depth, (v) thresholding within the depth range, and (vi) another round of hole filling applied to the resulting depth frame.

The hole-filling process employed the "filling-from-left" method, which was designed to address gaps in the depth image caused by black pixels, indicative of information loss. This method involved the leftward filling, starting from the leftmost pixel column. The choice to start from the left was influenced by the camera's reference point being the left camera, and shadow noise often appeared on the left background side. Subsequently, the resulting hole-filled depth image was converted into a disparity image using (1).

In this equation, f represents the focal length of the camera in pixels and b is the length of the baseline between the two imagers of the stereo depth camera in meters.

$$\text{Disparity} = \frac{f \times b}{\text{Depth}} \tag{1}$$

After converting the hole-filled depth image into a disparity image, spatial filtering was applied using bilateral filtering. The filtered disparity image was then transformed back into depth space using (2). Following this, depth thresholding was executed by limiting the minimum and maximum distances to 0.3 meters and 6 meters, respectively. Notably, this selected range accommodated the entire room, given that the distance between the camera and bed was 2.5 meters. To determine the depth process, the same hole-filling process was repeated.

$$\text{Depth} = \frac{f \times b}{\text{Disparity}} \tag{2}$$

Following depth processing, the resulting depth image was converted into a color image through depth image colorization in the hue color space, as explained in previous sections. As a general reminder, the hue color space was chosen for its ability to prevent extreme white or black colors, ensuring that the images maintain a balanced appearance without becoming excessively dark or washed out compared with simple representations such as grayscale. This colorization technique provides an additional advantage as it facilitated the utilization of colorized images as input to object detectors. By converting raw depth images into color images, a visual representation compatible with the commonly employed RGB-based object detection algorithms was created. These colorized images acted as a bridge between the depth domain and object detectors, encouraging a more efficient and effective analysis of the captured data.

2) PERSON DETECTION AND SEGMENTATION

The fusion of You Look Only Once version 5 (YOLOv5) and Segment Anything Model (SAM), as employed in a previous study [11], was integral to the system. The process involved using depth-colorized images as input to the YOLOv5 object detector, with the resulting person bounding box serving as prompts for SAM to extract person masks exclusively. In the proposed system, notable refinements were made to the YOLOv5 model. New training and validation images were organized, and the pre-processing of the input was adjusted. In this study, a more extensive dataset encompassing diverse images was used. The input image size for YOLOv5 was set to 320×320, maintaining the original image resolution (320×180) without resizing. This approach contributed to an improved person detection model, which subsequently influenced the SAM. The outcome of the person mask from SAM was then padded with zeros (black pixels) in both dimensions to achieve uniform bounding box sizes across all the images as illustrated in Fig. 7.

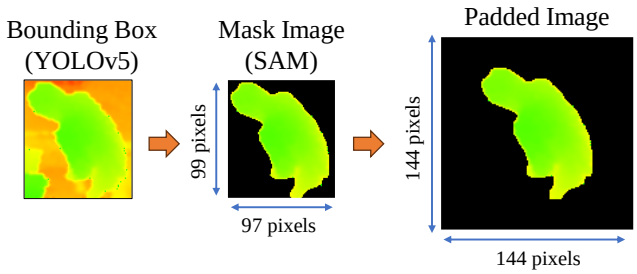


FIGURE 7. **Sample process for person detection and segmentation.**

3) ACTION RECOGNITION

The subsequent action recognition process, based on person segmentation outputs, is outlined in Fig. 8 to differentiate between various states and actions. The logic is described in the following conditions.

1) If no person was detected in the image, it was defined as a person being "outside" the room.
2) If more than one person was detected in the image, it was defined as an elderly individual who is "receiving assistance" from one or more health caregivers.
3) If only one person was detected in the image, the MotionCRNN-based action recognition was employed to distinguish between "seated in the wheelchair," "standing," "sitting," and "lying down" actions, as well as "transition states." It is important to note that all the transition states changing from one action to another were categorized as "transition" labels.

This approach ensured a clear interpretation of the scene, accounted for the presence or absence of individuals, and accurately characterized actions and states when a single person was detected.

When making action decisions, the focus was placed on the sequences of images rather than on individual frames. The emphasis on transition states is prominent because these states capture changes in body movements, whereas specific actions tend to exhibit more stable movements. To determine the optimal duration for decision-making, an in-depth analysis of ground-truth durations for transitions in three experimental rooms was conducted, revealing that the most frequent transition duration was 3 seconds. The histogram of the transition durations shown in Fig. 9 supports the selection of an optimal duration between 2 and 12 seconds. Considering the trade-off between delayed recognition for longer durations and potential inaccuracies for shorter durations, a duration range of 3-5 seconds was considered suitable. Given a processing rate of 1fps in the experiment, a duration of 5 seconds was selected, encompassing five frames in each sequence for robust action recognition. To process continuously throughout the long sequence, a sliding window method with a window size of 5 and 1-stride movement was applied in the experiment. The same sliding window method was employed for ground-truth labeling.

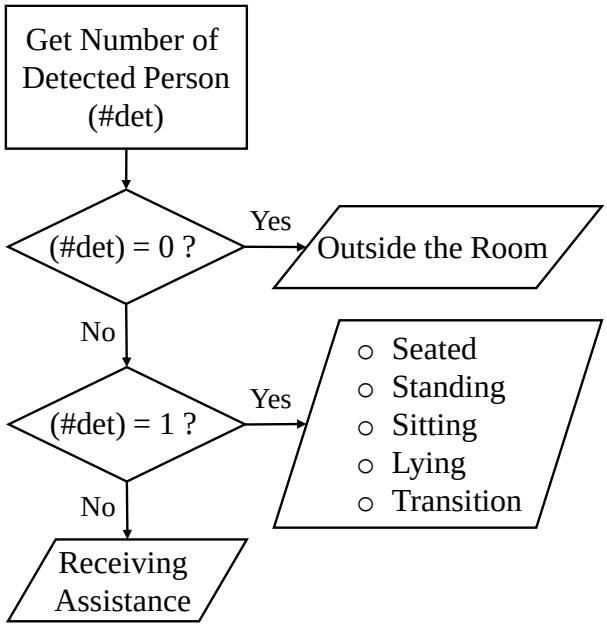


FIGURE 8. **Differentiating between different actions and states.**

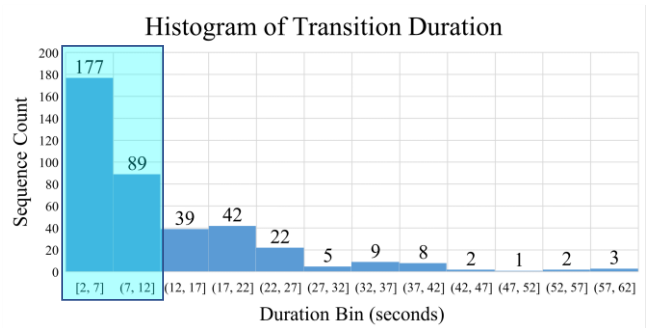


FIGURE 9. **Histogram of transition duration.**

The process flow of MotionCRNN-based action recognition, which is a notable contribution of the proposed method, is shown in Fig. 10. For this method, the integration of CNN and RNN was employed to extract spatiotemporal features. The process involved utilizing a CNN to extract spatial features and an RNN to extract temporal features. Here, CNN features were not extracted directly from the images; instead, they were derived from motion image sequences, in which the motion of two consecutive images was calculated from the normal image sequence. Therefore, the approach was denoted as MotionCRNN by extracting CNN features from the motion sequence and RNN for action prediction. Motion images were computed using segmented person-mask images, as shown in Fig. 11. To calculate the motion images, two consecutive images were first normalized by dividing them by 255. The normalized images were then summed channel-wise, and the resulting image was scaled up by multiplying it by 100.

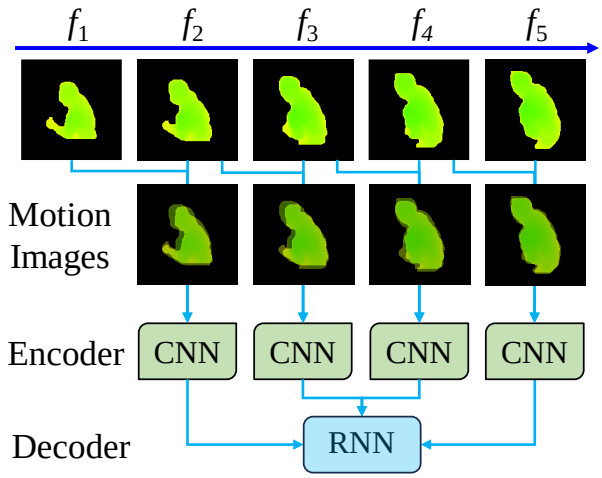


FIGURE 10. **Process of MotionCRNN-based action recognition.**

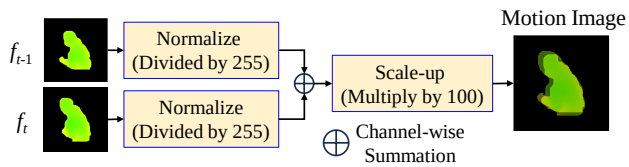


FIGURE 11. **Calculation of motion image.**

Subsequently, CNN transfer learning was applied to the EfficientNet architecture [57] (specifically EfficientNetB4) to extract spatial features. EfficientNet was chosen because of its efficiency and lightweight pre-trained weights. Regarding the RNN component, the Gated Recurrent Unit (GRU) [58] was chosen over LSTM because of its ability to handle time dependency more effectively than LSTM or basic RNN.

The specific model architectures of the CNN encoder and the RNN decoder are shown in Fig. 12. Transfer learning of the CNN encoder involved removing the last Fully Connected (FC) layer from EfficientNetB4 and adding two hidden FC layers, each followed by batch normalization and Rectified Linear Unit (ReLU) activation functions. Subsequently, a dropout layer was included to prevent overfitting, and another FC layer was added for feature embedding. The main layers and their respective parameters are presented in the top model plotting blocks as well as the full model architecture of the CNN encoder in the bottom 3D visualization of Fig. 12 (a). These CNN features were embedded in four consecutive motion images and used as inputs to the RNN. The proposed RNN comprised three unidirectional GRU layers, one FC layer followed by a dropout layer, and the final FC layer for classification as shown in Fig. 12 (b), with the respective parameters. This strategic integration of CNN and RNN into motion images allowed for the extraction of both spatial and temporal features, contributing to robust action prediction in the MotionCRNN framework.

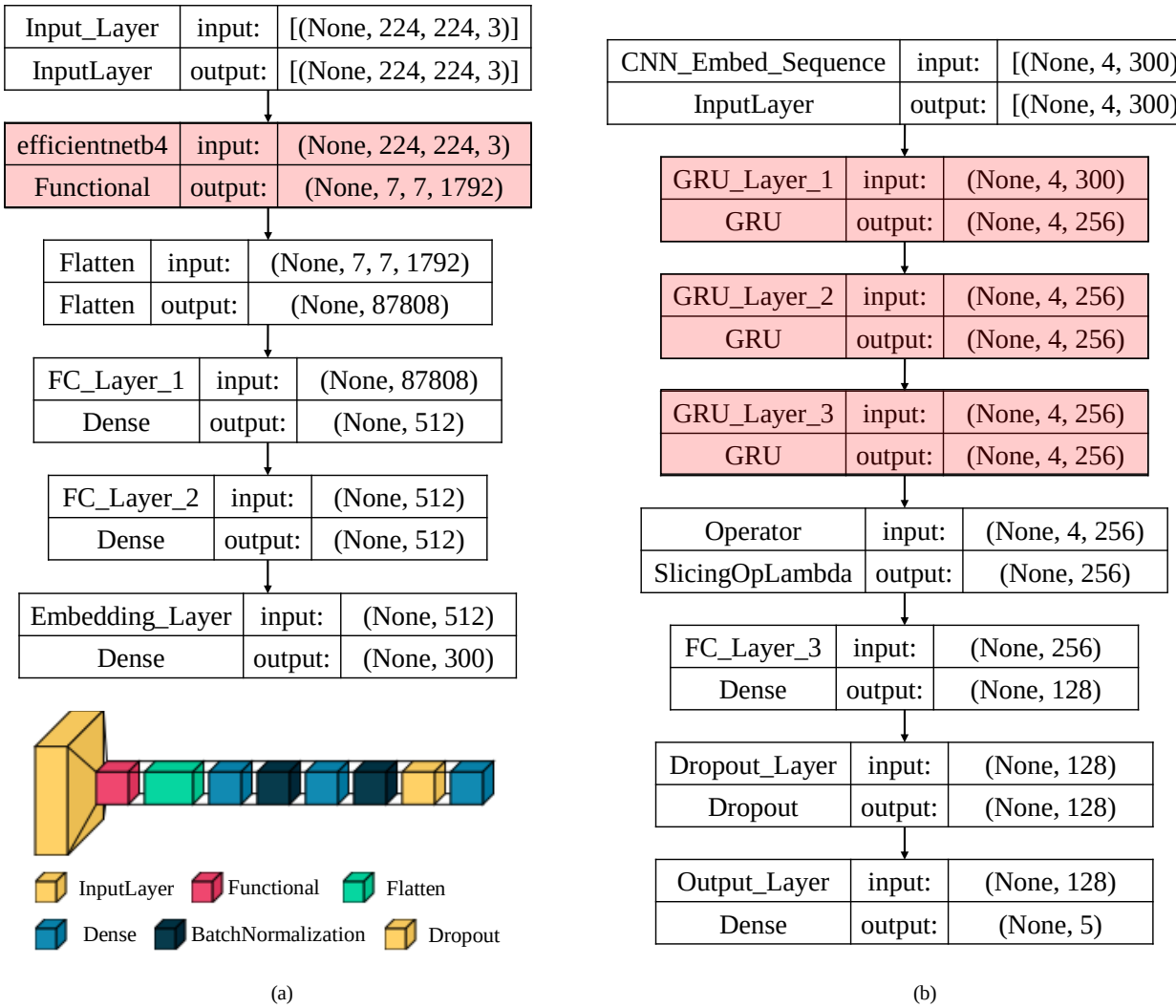


FIGURE 12. CRNN Architecture: (a) CNN Encoder, (b) RNN Decoder

4) EVALUATION METRICS

To determine whether the developed algorithm was reliable, different metrics were employed to assess the performance of the person detection and action recognition models.

For person detection, the focus was on the detection model's ability to correctly identify people (elderly in this context) and distinguish them from the background in colorized images. The two key metrics employed were "precision" and "recall" [59]. "Precision" measures the accuracy of positive detections whereas "recall" assesses the model's ability to capture all relevant detections. Additionally, the localization accuracy of the detection model was evaluated using "mAP@50" and "mAP@50-95" metrics where mAP represents the mean Average Precision [59]. These metrics assess how precisely the model locates people within the images. They are derived from the Intersection over Union (IoU), which measures the overlap between a predicted bounding box (the model's estimate of person location) and a ground-truth bounding box (actual location). "mAP@50" was calculated at an IoU threshold of 0.5, indicating a 50% overlap between predicted and actual bounding boxes. "mAP@50-95", calculated across varying IoU thresholds (0.5 to 0.95), indicates consistent accuracy even with stricter overlap requirements. The description, focus, and indication of higher values for each evaluation metric for person detection are presented in Table. III.

For action recognition, this study employed multi-class evaluation metrics to assess the performance of the model. These metrics included "accuracy," "precision," "recall," and "F1-score." "Accuracy" represents the overall proportion of actions correctly classified by the model. "Precision" focuses on the model's ability to identify specific actions accurately. "Recall", on the other hand, measures the model's ability to detect all instances of a specific action. Finally, the "F1-score" acts as a harmonic means of precision and recall, providing a balanced view of both the metrics. The description, focus, and indication of higher values for each evaluation metric in the case of transition state recognition are listed in Table. IV.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3405954

TABLE III
EVALUATION METRICS FOR PERSON DETECTION

Metric	Description	Focus	Higher Value Indicates
Precision	Accuracy of positive detections	Correctly identified elderly people	Fewer false positives (background clutter identified as elderly)
Recall	Completeness of detections	Capturing all actual elderly people	Fewer missed detections (actual elderly people not identified)
mAP@50	Localization accuracy	Precise location of elderly people	Better overlap between predicted and ground-truth bounding boxes (IoU ≥ 0.5)
mAP@50-95	Consistent localization accuracy	Accurate elderly location across varying overlap thresholds	Consistent performance even with stricter overlap requirements (0.5 ≤ IoU ≤ 0.95)

TABLE IV
EVALUATION METRICS FOR ACTION RECOGNITION (FOR "TRANSITION STATE" CLASS)

Metric	Description	Focus	Higher Value Indicates
Accuracy	Overall classification performance	Correctly classified actions	Higher proportion of all actions correctly classified
Precision	Specific class identification accuracy	Correct "transition state" classification	Less confusion with other actions for "transition state"
Recall	Completeness of specific class detection	Capturing all "transition states"	Fewer missed actual "transition state" instances
F1-score	Balanced view of precision and recall	Overall "transition state" classification performance	Good performance with minimal confusion between "transition state" and other classes

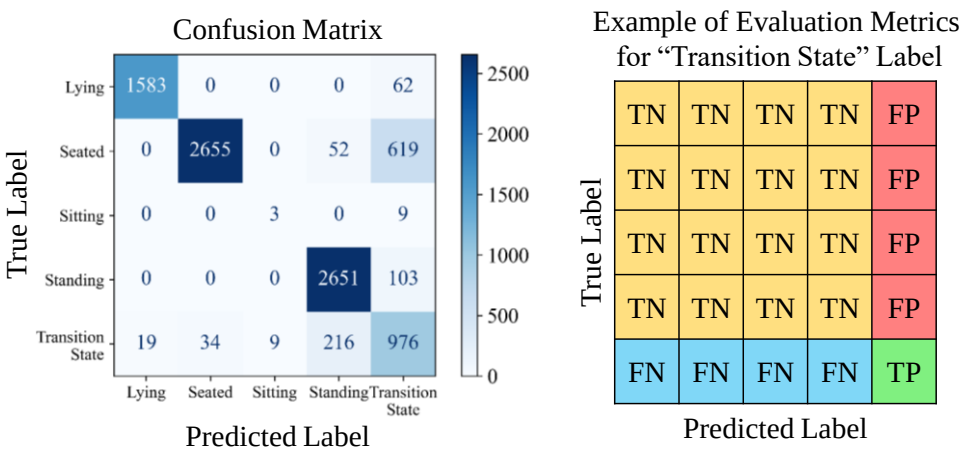


FIGURE 13. Sample evaluation for "transition state" class.

Fig. 13 shows an example of calculating these metrics for the "transition state" label in which True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are marked in the right matrix according to the left confusion matrix. The resulting percentages for each metric are described in (3)-(10).

Overall, achieving higher values of all these metrics across both person detection and action recognition tasks signifies a well-performing algorithm.

$$TP = 976 \tag{3}$$

$$FP = 103 + 9 + 619 + 62 = 793 \tag{4}$$

$$TN = 2651 + 3 + 52 + 2655 + 1583 = 6944 \tag{5}$$

$$FN = 216 + 9 + 34 + 19 = 278 \tag{6}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = 0.8809 = 88.09\% \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} = 0.5517 = 55.17\% \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.7783 = 77.83\% \tag{9}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 0.6470 = 64.70\% \tag{10}$$

IV. EXPERIMENTAL RESULT ANALYSIS

In this section, the dataset preparation process, evaluation performance for each process on different datasets, analysis of the results, and refining process are described. Furthermore, various comparisons are performed, and the GUI is explained.

A. DATASET PREPARATION

In the proposed system, trainable DL algorithms played a pivotal role in the implementation, encompassing the fusion of YOLOv5-SAM for person detection and segmentation, as well as EfficientNet and GRU for MotionCRNN-based action recognition. Diverse datasets were carefully prepared to facilitate the training of these algorithms. For the experiment, data from all three rooms of the care center were used, emphasizing the inclusion of a varied dataset. Dataset preparation involved the standard practice of splitting data into training, validation, and testing datasets. Importantly, the data used for each dataset did not overlap, thereby ensuring that the data used for training were distinct from those included in the validation and testing datasets. Specific datasets for each stage of the process are described in detail in the following sections.

B. TRAINING AND VALIDATION DATASETS

In the context of person detection using YOLOv5, 120,000 images were chosen from three rooms of the care center, each annotated with ground-truth bounding boxes. Subsequently, 70% of these images (84,000 images) were allocated to the training dataset, and the remaining 30% (36,000 images) were allocated to the validation dataset. For action recognition, multiple sequences of five consecutive images each were selected to train the MotionCRNN. These short sequences were collected from all the three rooms to ensure a balanced dataset. In total, 13,600 sequences were chosen, with 70% (9,520 sequences) designated for training, and the remaining 30% (4,080 sequences) assigned to the validation dataset. These datasets were selected and organized to include diverse situations and ensure robust training of the respective algorithms. The following subsections detail the specific training parameters, procedures, and performance evaluations for each stage of the proposed system.

1) PERSON DETECTION PERFORMANCE

The training parameters for the person detection process are presented in Table. V, employing the pre-trained weight yolov5l6 to train the custom dataset over 100 epochs with an IoU threshold of 0.6. The input image size was set to 320×320 pixels, preserving its relationship to the original image size. Performance evaluations of the training and validation datasets are presented in Table. VI. The model demonstrated robust performance for both datasets. It achieved over 99% precision, recall, and mAP@50, and over 97% for mAP@50-95. The detected bounding boxes were then passed into the SAM for segmentation. Fig. 14 shows some of the selected images resulting from the fusion of YOLOv5-SAM.

TABLE V
YOLOv5 TRAINING PARAMETERS

Pre-trained Weight	yolov5l6.pt
Batch Size	32
Image Size	320×320
IoU Threhsold	0.6
Confidence Threhsold	0.7
Epochs	100

TABLE VI
PERFORMANCE EVALUATION ON TRAIN AND VALID DATASETS

Dataset	Images	Evaluation Metrics (%)			
		Precision	*Recall*	*mAP@50*	*mAP@ 50-95*
Train	84,000	99.9	100	99.5	97.6
Valid	36,000	99.8	99.9	99.5	97.0

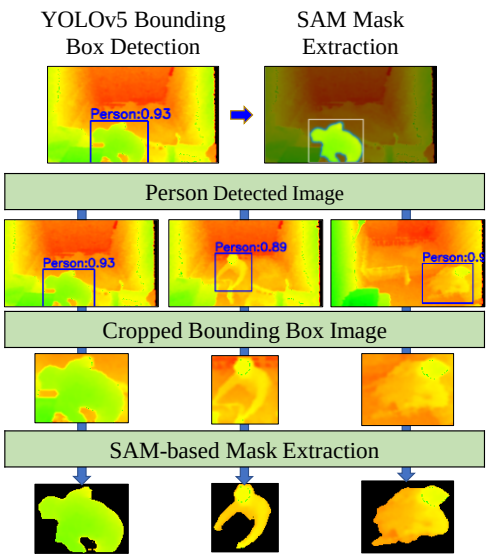


FIGURE 14. Resulting images from the fusion of YOLOv5-SAM.

2) ACTION RECOGNITION PERFORMANCE

The proposed system utilized the EfficientNetB4 pre-trained weight for CNN transfer learning and employed unidirectional GRUs for recurrent decision-making. In training the MotionCRNN, the cross-entropy loss function was calculated once every epoch for the CNN-RNN integration output, and the Adam optimizer was applied with the default learning rate (0.001). The model was trained for 20 epochs with a batch size of 64. The performance evaluation of the training and validation datasets are presented in Fig. 15. The confusion matrix for each dataset is shown in Fig. 16. Emphasizing the transition state class, it is evident that the training process performed well, achieving over 99% accuracy for all evaluation metrics on both the training and validation datasets. However, some false and missing predictions persisted as can be seen in the confusion matrices, indicating areas for potential improvement, despite the overall strong performance in training for recognizing transition states.

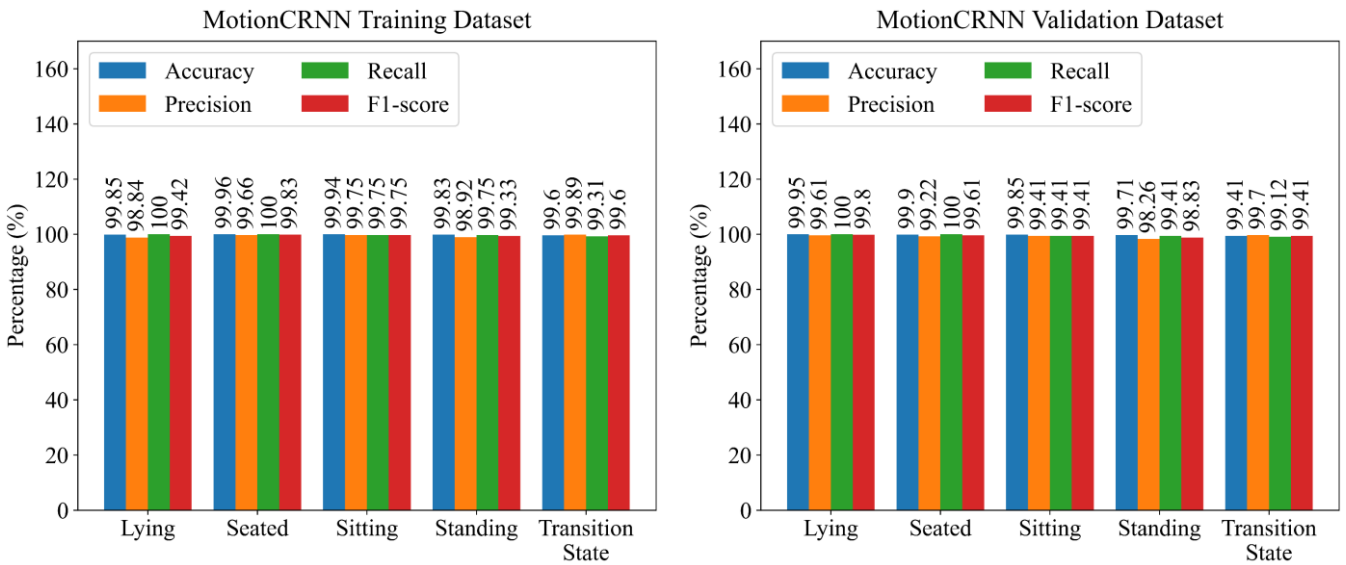


FIGURE 15. Action recognition performance evaluation for train and valid datasets.

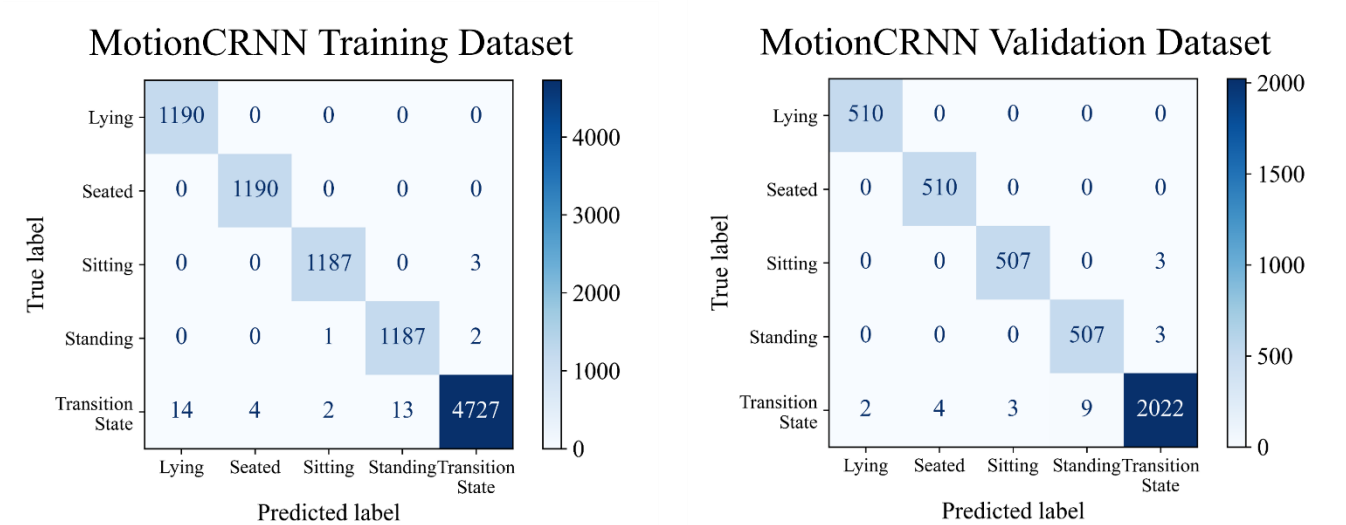


FIGURE 16. Confusion matrix of action recognition for train and valid datasets.

C. TESTING DATASET

The proposed algorithm was tested on all recorded datasets described in Table II. However, because of the time-consuming and intensive nature of the ground-truth labeling task, only one 15-hour duration (approximately) random long sequence from each room was selected for performance evaluation. The sequences comprised up to 54,800 frames at 1fps and were recorded during both the daytime and nighttime. Detailed information on the selected testing data is presented in Table. VII, where the actions included in each sequence are also described. Specifically, these testing data were selected to include the significant transition states observed in each room. For example, the Room 1 sequence highlighted transition states from seated in the wheelchair to standing and vice versa, whereas the Room 2 sequence covered a considerable number of transition states from sitting to lying down and vice versa. Finally, the Room 3 sequence mostly featured transition states from sitting to standing and vice versa, which are actions frequently performed by elderly residents. Reminding of the approach, a sliding window method was employed with a window size of 5 and a stride of 1 to process the long sequence.

1) PERSON DETECTION PERFORMANCE

The performance evaluation of the testing dataset for person detection, using a confidence threshold of 0.7 and an IoU threshold of 0.6 for YOLOv5, is shown in Table. VIII.

TABLE VII
TESTING DATASET INFORMATION

Room ID	Date and Time		Duration (hours)	Number of Frames	Included Action [a]
	Start Time	*End Time*			
1	2019/10/12 10:15:00	2019/10/13 00:39:00	14.4	51,840	*A, L, O, Se, St, Tr*
2	2019/10/25 11:50:00	2019/10/26 02:50:00	15	54,000	*A, L, O, Se, Si, Tr*
3	2019/10/12 11:10:00	2019/10/13 01:34:00	14.4	51,840	*A, L, O, Se, Si, Tr*

[a] *A*: Assistance, *L*: Lying, *O*: Outside, *Se*: Seated, *St*: Standing, *Si*: Sitting, *Tr*: Transition states

TABLE VIII
PERFORMANCE EVALUATION ON TESTING DATASET FOR YOLOV5

Room ID	Images (1fps)	Evaluation Metrics (%)			
		Precision	*Recall*	*mAP@50*	*mAP@50:95*
1	51,840	100.0	99.1	99.5	99.5
2	54,000	100.0	99.6	99.5	99.5
3	51,840	100.0	99.9	99.5	99.5

The results indicated that the proposed person detection model achieved over 99% accuracy for all metrics. It also achieved the highest precision rate which means that there was no false detection when the background was detected as a person. However, there were some missed detections in which a person could not be detected by the model (recall). Regarding person localization between the ground truth and prediction, both mAP@50 and mAP@50:95 had high rates. The sample results are shown in Fig. 17, where the resulting bounding boxes were used as prompts for person segmentation.

After analyzing the results of person detection, it was observed that the most challenging detection task occurred when a person was lying on the bed while being covered with a blanket. At that time, the distance information was not clearly visible or significant enough to distinguish the person from the bed. Despite utilizing numerous training annotations, some frames presented these types of missed detections. This also affected the recall rate.

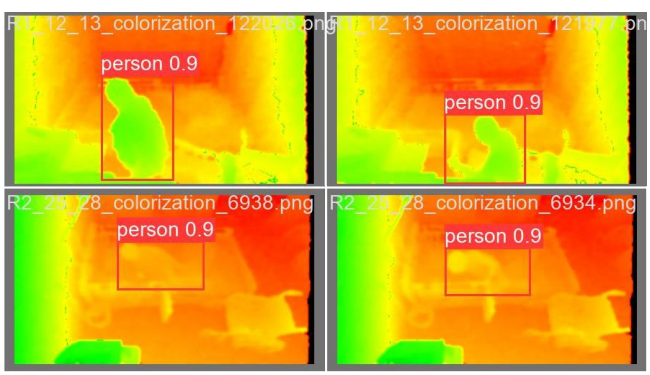


FIGURE 17. **Sample person detection testing results.**

To solve this problem, a condition is introduced at this stage, as shown in Fig. 18. If a person is not detected in the current frame, the algorithm checks the previous frame to determine whether a detection has occurred. If detection occurred in the previous frame, the previous bounding box coordinates were used in the current frame. Subsequently, a frame difference was performed within the defined bounding box by counting the number of pixels with intensity values. If the summation of this count was less than 30% of the defined bounding box area, it was determined that there was a small movement similar to the previous frame. In such cases, the algorithm replicated the previous box, cropped the image accordingly, and continued with the subsequent processes. Conversely, if the summation exceeded the 30% threshold, a large movement was determined, and no bounding box was assigned to the current frame.

2) ACTION RECOGNITION PERFORMANCE

After applying the person detection refinement to the bounding box recovery process, action recognition was performed using the MotionCRNN model. A visual representation of a sample 10-minute duration action recognition result from the Room 1 testing sequence is illustrated in Fig. 19, in which the top one is the scatter plot and the bottom two are the bar chart representations of ground truth and predictions in each time frame, respectively. By observing the visualization in this sample result, it can be seen that the person was in a transition state between standing and seated in the wheelchair frequently within the 10-minute duration. However, some of the results highlighted the occurrence of over-segmentation errors, particularly when the person was in a transition state, as represented by the red color in Fig. 19. This issue occurred primarily because decision making relied on the most probable action (Top-1 accuracy) for each prediction. To address this problem, a sequential-based majority voting decision and transition state reasoning were implemented.

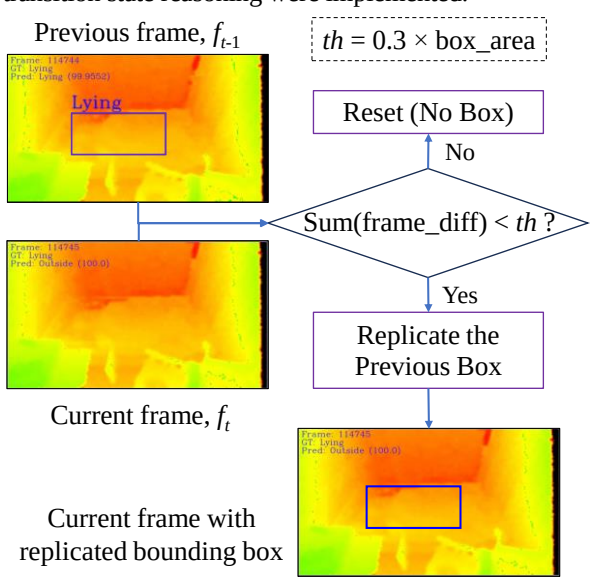


FIGURE 18. **Bounding box recovery process.**

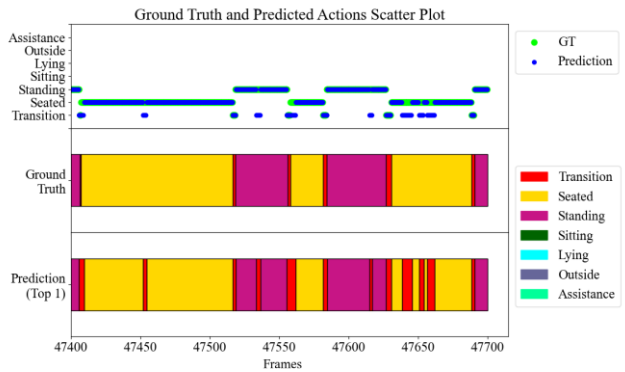
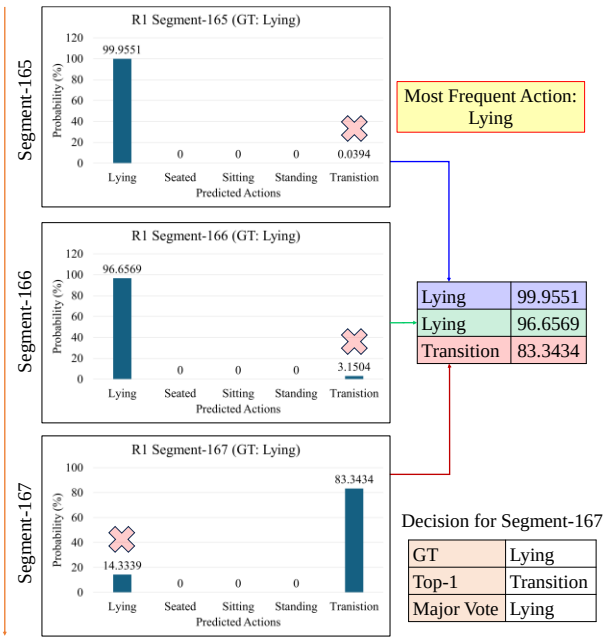


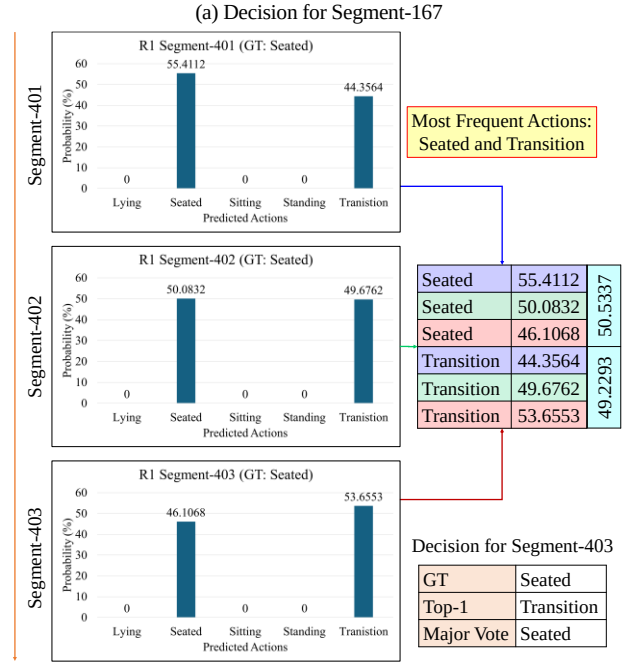
FIGURE 19. Visual representation of a sample action recognition result.

To implement the majority voting decision, Top-2 predicted labels were utilized, where "Top-2" refers to the two most probable predictions among the five prediction probabilities from the model. An illustration of the majority-voting decision is shown in Fig. 20. For instance, in Fig. 20 (a), to determine the predicted label for Segment-167 (bottom graph), the previous two segments (Segments 166 and 165) were considered, and the Top-2 predicted labels for each segment were checked. The small probability values were then removed using a threshold of 20 and the remaining probability values and labels were examined. In this example, two labels were identified as "lying" and one label as a "transition state." Hence, the most frequent action was determined as "lying" for Segment-167. It is evident that the ground-truth label was "lying," and the majority voting decision also indicated "lying," which achieved a better result than the Top-1 label, which was a "transition state." However, there were conditions in which the thresholded values resulted in the same number of labels, as shown in Fig. 20 (b). In such cases, the average probability values for each label were obtained and the decision was determined as the label with the highest average probability value. For this example, the "seated" label has the highest probability of 50.53%. Hence, even though the Top-1 label was a "transition state," majority voting correctly identified the action as "seated."

In this experiment, a transition state was generally defined as a state that changes from one action to another. However, even after applying majority voting, there were instances of false predictions as transition states throughout the long sequence. This problem occurred because of the model's lack of reasoning capabilities. Applying reasoning to the predictions of DL models is crucial for real-world effectiveness. Hence, to enhance the recognition results, a reasoning step was introduced that specified that a transition state should not occur between the same specific actions. In cases in which this condition occurred, the system refined the results after a certain period (1 hour in this experiment) by replacing the predicted transition states with specific actions before or after the transition state. This approach aimed to improve the accuracy and reliability of the recognition results by integrating conditional reasoning into prediction process.



(a) Decision for Segment-167



(b) Decision for Segment-403

FIGURE 20. Illustration of majority voting decisions.

A comparison between the Top-1 and the final refined results is shown in Fig. 21. Upon checking the visualization, it is evident that the two refined approaches (sequential-based majority voting and transition state reasoning) smoothened the action recognition results and reduced the over-segmentation errors among the predicted actions. By examining the resulting visualization, users can make decisions regarding the actions of their intended residents regarding health monitoring. An example of decision-making for the results in Fig. 21 could be: "Resident A is observed standing for a while, then transitioning to being seated in the wheelchair within 10 minutes. During this period, there are frequent transitions between seated and standing positions."

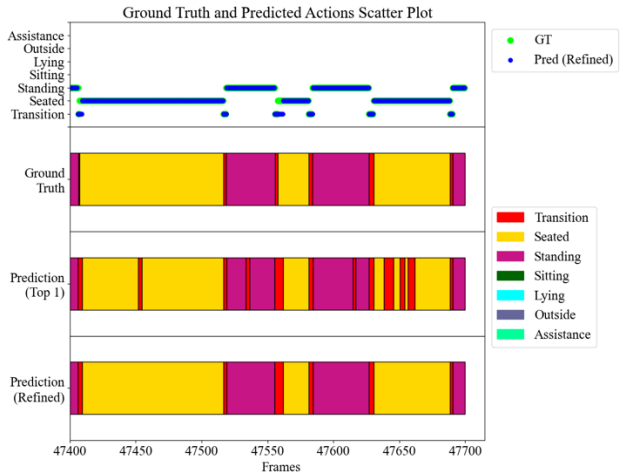


FIGURE 21. Visualization of action recognition with two refinements.

Such insights allow caregivers or observers to understand residents' activities over time, facilitating informed decision-making and appropriate interventions, as needed.

The action recognition performance on the evaluation metrics and confusion matrix after the refinements are shown in Fig. 22. In all testing sequences, the "outside" state was included, which is easy to identify even after person detection. Therefore, the evaluation was performed after excluding the "outside" state; however, it was included in the confusion matrix. In all three testing sequences, although the performance was promising, there was still some confusion between the actions. According to the result analysis, some false recognition cases were identified, primarily attributed to occlusion, low-quality segmented person masks, and misalignment of the transition states. First, occlusion occurs when an elderly individual inside the room is covered by a nurse or caregiver in front of the camera view. In these instances, the system detected only the nurse or caregiver and predicted their actions instead of the intended elderly resident. Because of this occlusion, false predictions between the "assistance" state and other actions occurred because the person detection model detected only one person instead of two persons. Second, most of the confusion between specific actions (seated in the wheelchair, sitting, standing, and lying down) occurred because of the inaccuracy of the extracted person masks segmented from the person segmentation process. In the third case, misalignment of transition states occurred when the occurrence of transition states in the ground-truth labeling did not align with those in the predictions. This misalignment resulted in the predicted transition states at different times relative to the ground truth. Sometimes, the transition states in the ground truth occurred prior to the predictions, and vice versa. Therefore, these challenges resulted in lower evaluation performance in terms of precision and recall. Moreover, by emphasizing transition state recognition, the results after refinements and excluding the "outside" state are shown in Fig. 23. Although there are still some areas for improvement, the experimental results are promising, highlighting the key contribution of this study. MotionCRNN with result refinement achieved an average accuracy of 99.19% and an average F1-score of 83.39%, demonstrating its effectiveness in differentiating transition states from other specific actions.

D. COMPARISON OF EXPERIMENTS

In this study, various tests based on different factors were conducted; thus, the results of these comparisons are detailed in the following subsections.

1) IMPACT OF REFINEMENT ON ACTION RECOGNITION

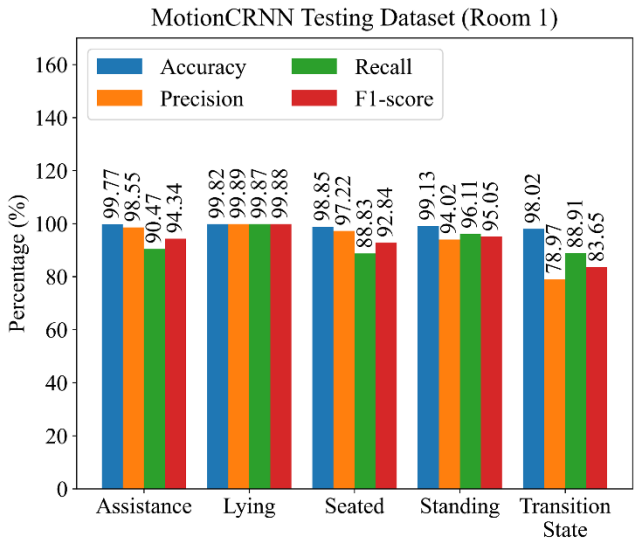
The overall results of action recognition, particularly for transition states, are compared in Fig. 24, with a primary focus on the impact of refinements, including bounding box recovery, majority voting, and reasoning for transition states. The results indicate that while the recall rate for all three rooms decreased slightly, improvements in precision and F1-score rates were evident, especially in Rooms 2 and 3, where they increased by up to 66% compared with the results without refinements. Among the refinement processes, conditional reasoning for transition states had the most significant impact on increased recognition rates.

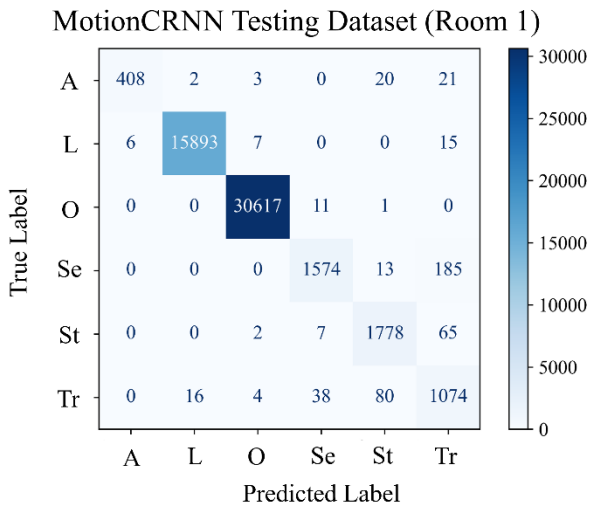
2) PROCESSING TIME COMPARISON

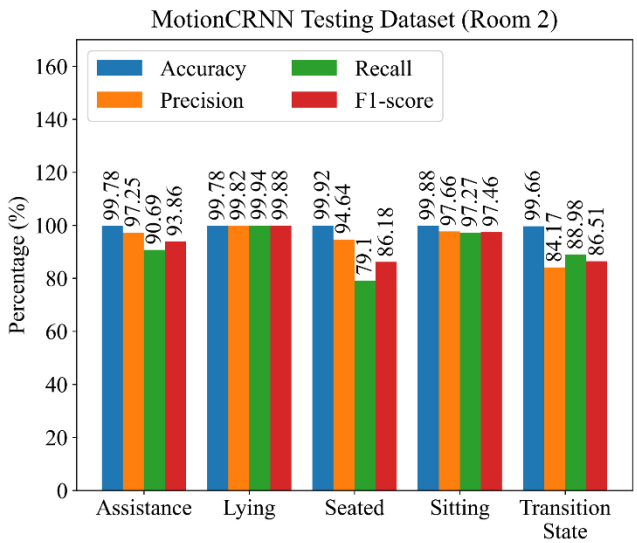
The testing process was performed using a 64-bit Intel (R) Core i9 PC with 64GB RAM and an NVIDIA GeForce RTX 4090 graphics card. Processing a 1-hour testing sequence at 1fps took approximately 0.5 hours on average. This translates to real-time processing capability, even with initial depth data processing and depth image colorization. Therefore, the proposed system effectively utilized the stereo depth camera and DL to achieve real-time action recognition with high accuracy in indoor elderly monitoring. The real-time capability with high accuracy is one of the contributions of this study. Notably, increasing the frame rate to 2.5fps resulted in a processing time of 1.5 hours without significant accuracy gains.

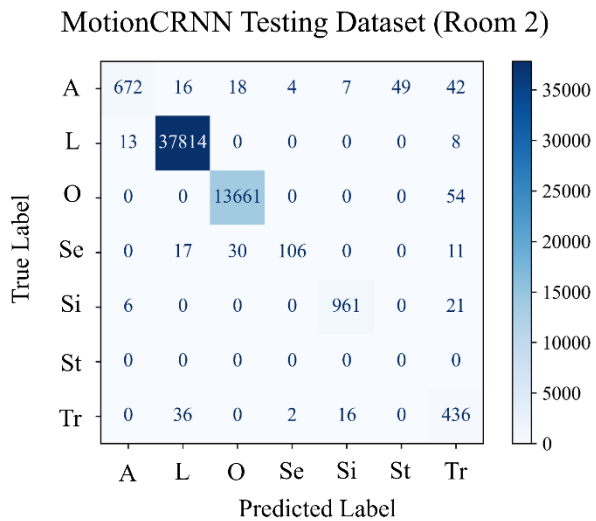
3) CNN BASE MODEL COMPARISON

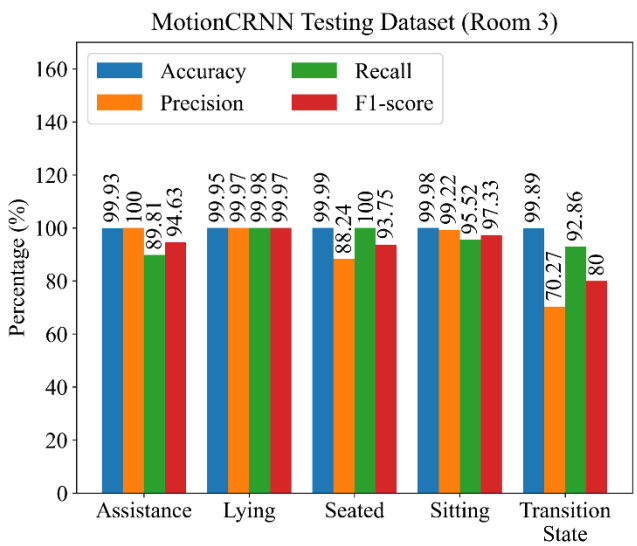
Furthermore, various EfficientNet architectures were tested to determine whether the model could be enhanced by changing its base model. Four model variants were used for comparison: one EfficientNetB4 and three EfficientNetV2 models [60] (V2L, V2M, and V2S), which were tested in three testing rooms. A comparison of the overall accuracy and processing time of each variant is presented in Table. IX, where the accuracy was calculated for all classes including "outside" and excluding "outside". The average processing time was calculated based on a 1-hour duration sequence. For example, the processing of the EfficientNetB4 model on the Room 1 testing dataset took 28.14 minutes on average for a 1-hour duration sequence. The results demonstrated that although the overall accuracy rates for all models showed no significant improvement, there was a significant difference in the average processing time duration.

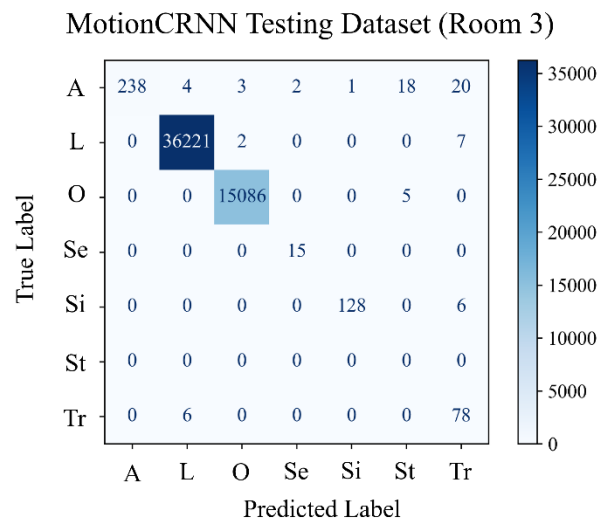












A: Assistance, L: Lying, O: Outside, Se: Seated, Si: Sitting, St: Standing, Tr: Transition State

FIGURE 22. **Action recognition performance after refinements for all testing sequences: (left) evaluation metrics, (right) confusion matrices.**

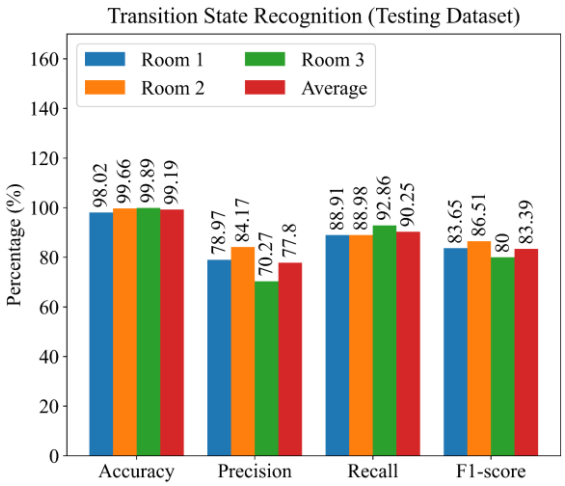


FIGURE 23. **Transition state recognition performance after refinements.**

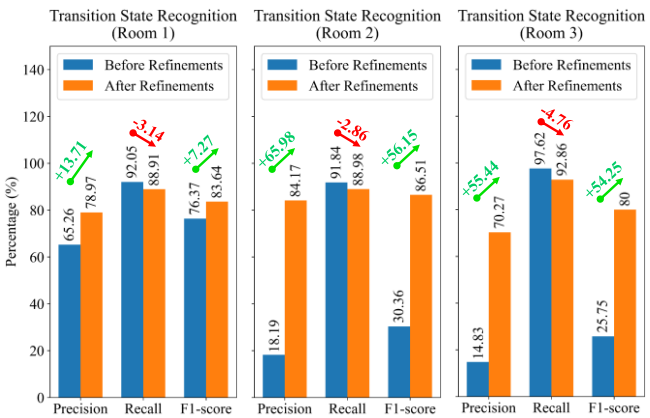


FIGURE 24. **Impact of refinements on transition state recognition.**

TABLE IX
OVERALL ACCURACY AND PROCESSING TIME COMPARISON (EFFICIENTNET VARIANTS)

EfficientNet Model	Room ID	Overall Accuracy (%)		Average Processing Time (Minutes)
		All Classes	*Excluding "Outside"*	
B4	1	99.04	97.79	28.14
	2	99.35	99.51	27.53
	3	99.86	99.88	28.01
V2L	1	98.67	96.85	37.05
	2	99.20	99.29	36.95
	3	99.90	99.88	37.27
V2M	1	98.59	96.73	33.64
	2	99.10	99.17	32.98
	3	99.90	99.88	29.94
V2S	1	98.84	97.32	28.78
	2	99.26	99.37	28.54
	3	99.89	99.87	28.81

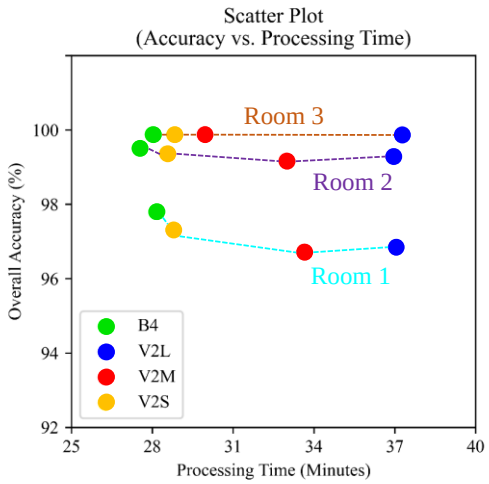


FIGURE 25. **Scatter plot for base model comparison.**

For clear visualization, a comparison scatter plot of the results from Table. IX is shown in Fig. 25, where the overall accuracy was calculated by excluding the "outside" class. The comparison demonstrated that the MotionCRNN with the EfficientNetB4 model achieved the best trade-off between overall accuracy and processing time.

4) SYSTEM COMPARISON

Finally, the reliability and effectiveness of the proposed system were compared with those of other related systems. Although the ultimate goal of recognizing the daily activities of elderly individuals remains consistent across these systems, various technologies have been implemented, employing different input data types and recognition models. In Table. X, a comparative analysis of the proposed system and recent studies that adopted distinct methodologies are presented. This comparison encompasses aspects such as input data types, usage of real-world data, awareness of transition state recognition, recognition model architecture, real-time processing capabilities, and privacy preservation considerations.

The comparison results indicate that the proposed method achieved an average accuracy of 99.42% for recognizing seven actions. This approach prioritized both privacy by utilizing depth data and real-world reliability through the use of real-world data, which is one of the contributions of this work. In addition, it captured the crucial transition states vital for elderly monitoring. The proposed system significantly surpassed the authors' prior works that used HMM and SVM with similar depth data and transition awareness [9], [10], [11]. While another sensor-based approach achieved transition-aware recognition [42], its accuracy was limited to 80%. Notably, while state-of-the-art hybrid DL recognition models [27], [50], [51] obtained high accuracy, they were not considered for the application with real-time processing, privacy concerns, or transition state recognition. However, it is remarkable that although most of the other systems used public datasets, this study used custom real data; thus, the accuracy results may vary according to the dataset scales.

TABLE X
(A) SYSTEM COMPARISON-1

Related Work (Year)	*Data Type*	*Real Data*	*Total Action*	*Transition Awareness*
[9] (2021)	Depth	✔	8	✔
[10] (2022)	Depth	✔	5	✔
[11] (2023)	Depth	✔	5	✔
[27] (2023)	Sensor	✘	6	✘
[42] (2020)	Sensor	✔	9	✔
[50] (2023)	RGB	✘	101	✘
[51] (2020)	RGBD	✘	27	✘
Ours	Depth	✔	7	✔

(B) SYSTEM COMPARISON-2

Related Work (Year)	*Recognition Model*	*Real-Time*	*Privacy-Preserving*	*Average Accuracy (%)*
[9] (2021)	SVM	✔	✔	93.73
[10] (2022)	HMM-SVM	✔	✔	84.08
[11] (2023)	SVM	✔	✔	88.26
[27] (2023)	CNN-LSTM	*N/A*	✘	99.00
[42] (2020)	STD-TA	*N/A*	✘	80.00
[50] (2023)	Vit-ReT	✔	✘	94.70
[51] (2020)	Deep CNN	*N/A*	✔	87.21
Ours	MotionCRNN	✔	✔	99.42

E. EXTENDED TESTING ON DIFFERENT DATASETS

To evaluate the generalizability, the system was tested on data from another hospital with significantly different camera positions and structures (compared to Fig. 2, see Table XI and Fig. 26). Three elderly patients from the hospital participated in the experiment. Depth image data from three separate rooms were recorded for a continuous period of 24 hours for three consecutive days. This data acquisition protocol also received ethical approval from the University of Miyazaki Ethics Committee (protocol code O-1449, on November 20, 2023).

The system utilized two trainable models: a YOLOv5 model for person detection and a MotionCRNN model for action recognition. The person detection model was first fine-tuned using 30,000 images from this extended dataset. While a smaller dataset of 600 new sequences was used for initial action recognition training, transfer learning enabled the effective evaluation of three 1-hour testing sequences (3,600 frames at 1fps each) from the new environment. The results presented in Table XII, ranging from 84.83% to 99.22% overall accuracy rates, demonstrate the potential of the system as a foundational model that can adapt to diverse settings with minimal additional data requirements. This highlights another key contribution of this study.

F. GRAPHICAL USER INTERFACE

A GUI specifically designed for end users, including family members and health caregivers, was developed to facilitate the real-time monitoring of elderly individuals and access detailed action information captured by the proposed action recognition system. The GUI consisted of two main windows, as shown in Fig. 27. The first window is the action detail window, where users can select the name or ID of the elderly resident they wish to monitor and input the desired start and end times to view the detailed information. The GUI then displays the recognized actions of the selected resident on a scatter plot, providing a second-by-second representation. Additionally, a bar chart summarizes the actions performed during the specified time frame. For a more comprehensive view of continuous actions, users can refer to a table that lists the specific durations of the consecutive actions. The GUI design ensures that end users and healthcare providers can easily access insightful information within a single window.

The second window in the GUI is a real-time monitoring window. Similar to the action detail window, users can input relevant information to either re-play or monitor the actions of the elderly residents in real time. This feature allows users to validate the accuracy of previously captured action details, thereby providing reassurance and confidence in the system performance. In summary, this GUI serves as a comprehensive tool for caregivers to monitor elderly in real time, access detailed action information, and interact with the analytics and recognition processes of system.

TABLE XI
DATA ACQUISITION PROTOCOL FOR HOSPITAL

Participants	Three elderly residents (three rooms)
Collected Data Information	Recorded depth image data for a continuous period of 24 hours for 3 days. - Room 4: 2024/01/05 to 2024/01/08 - Room 5: 2024/01/26 to 2024/01/29 - Room 6: 2024/01/31 to 2024/02/03

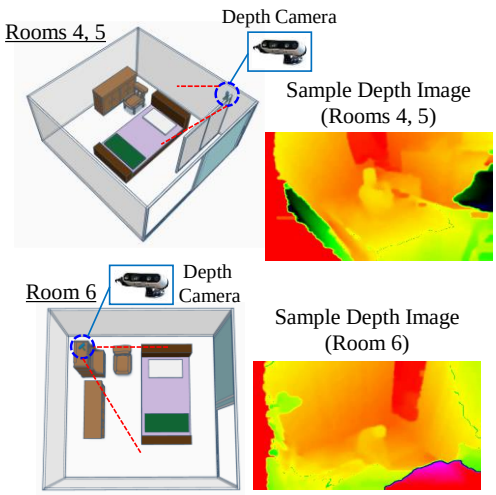


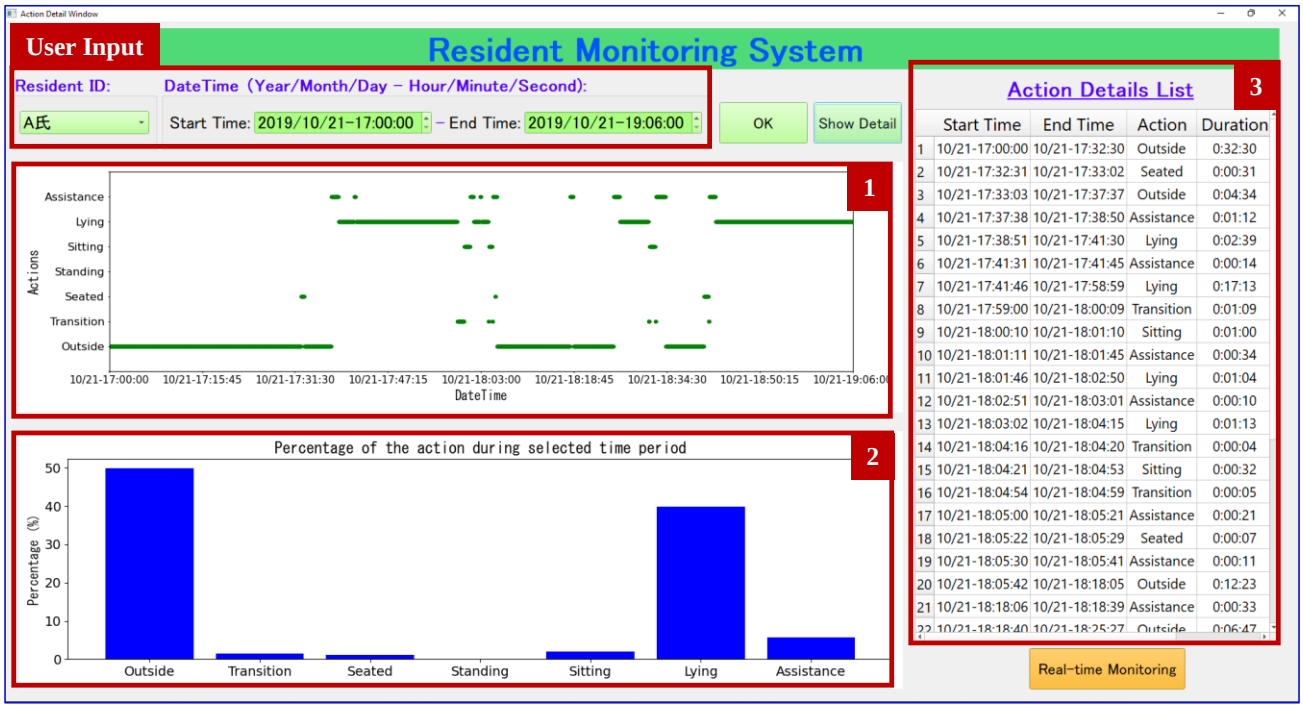
FIGURE 26. Illustration of environments (hospital).

TABLE XII
PERFORMANCE EVALUATION ON EXTENDED DATASET

Room ID	Date and Time *Start Time*	Date and Time *End Time*	Included Action [a]	Overall Accuracy (%)
4	2024/01/06 10:45:00	2024/01/06 11:45:00	*A, L, St, Tr*	99.22
5	2024/01/27 07:02:00	2024/01/27 08:02:00	*A, L, St, Si, Tr*	84.83
6	2024/01/31 19:56:00	2024/01/31 20:56:00	*A, L, O, Si, Tr*	94.89

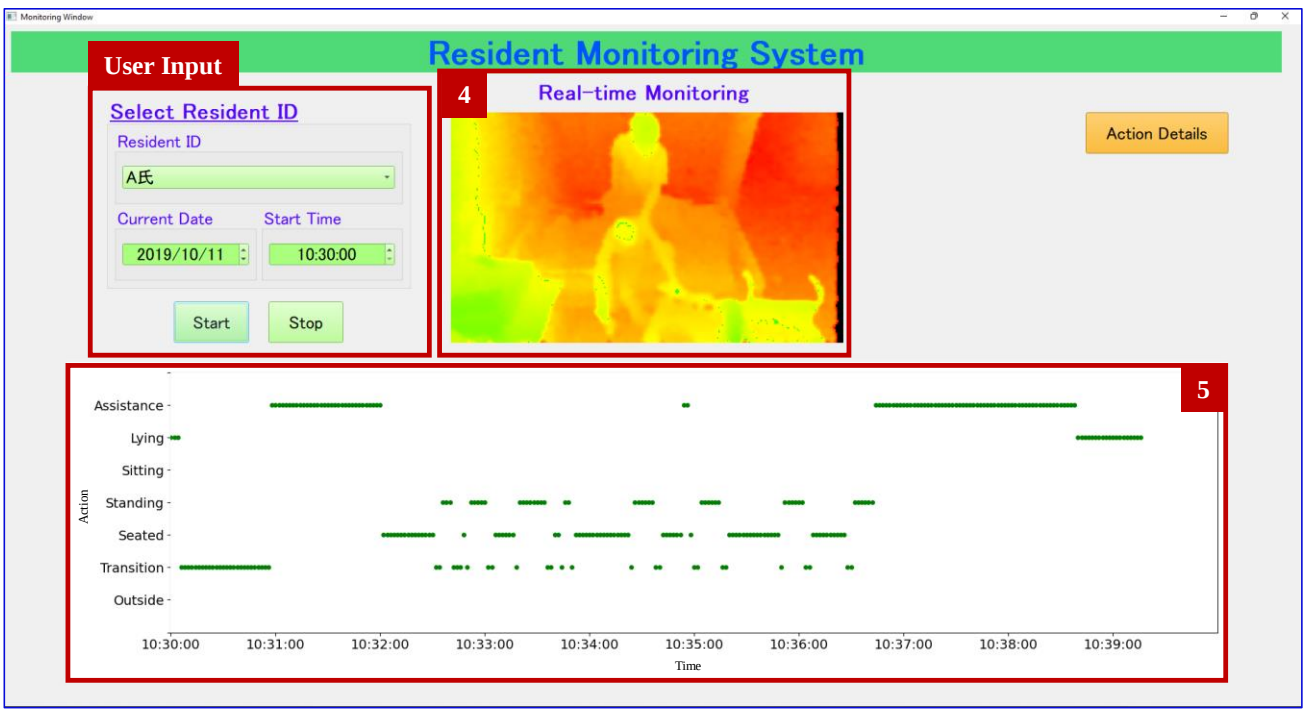
[a] *A*: Assistance, *L*: Lying, *O*: Outside, *St*: Standing, *Si*: Sitting, *Tr*: Transition states

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3405954



1 Action Details Plot 2 Action Summary Graph 3 Action Duration Table

(a) Action details window.



4 Real-time Video 5 Real-time Action Plot

(b) Real-time monitoring window

FIGURE 27. Graphical user interface.

V. DISCUSSION

This study developed a comprehensive elderly activity monitoring system that leverages real-world data from an elderly care center and hospital. This approach ensures data authenticity while prioritizing user privacy using depth images captured by stereo depth cameras. A resolution of 320×180 pixels was utilized for depth images, balancing image quality with efficient storage and real-time processing capabilities. While a higher resolution could offer better details, it is crucial to consider the trade-off with processing speed for optimal system performance.

The proposed system focuses on recognizing seven common daily actions, including transition states that can be used to predict potential risks. The results demonstrated improved performance compared to previous studies in recognizing these critical states. Person detection and segmentation were the fundamental components achieved by combining the YOLOv5 and SAM models. The analysis revealed a clear link between segmentation accuracy and action recognition accuracy, highlighting its importance. Although the current results are reliable, there is room for improvement, especially in person detection and segmentation. Upgrading to advanced YOLO versions or exploring alternative segmentation algorithms can enhance accuracy and processing speed.

A unique contribution of this experiment is the application of MotionCRNN to image sequences for action recognition. This approach incorporated motion information into a hybrid CNN-RNN architecture, which is valuable for identifying transition states that rely heavily on movement patterns. The system achieved a high accuracy of 99.42% in recognizing not only the transition states but also various specific actions in real-time. Through experimentation, the model architecture and parameters were optimized, further refining the results with bounding box recovery, sequential-based majority voting, and condition reasoning to enhance action recognition performance. It is acknowledged that exploring other advanced CNN base model architectures can potentially push the boundaries of accuracy even further.

Finally, a user-friendly GUI was designed to provide a platform for offline interaction between caregivers and the system, offering insights into the health trends and details of the activities of the elderly. In addition, various comparisons were performed to assess the reliability and effectiveness of the proposed system.

A. LIMITATIONS

The current method converts depth images into colorized images for compatibility with RGB-based object detectors. However, the optimal performance requires a camera-to-person distance to align with the training dataset. Significant variations in this distance may lead to false or missed detection. Leveraging 3D processing techniques that utilize distance information for detection can improve generalization and robustness across various environments. Eliminating the need for colorization and focusing on distance-based detection can yield more reliable results.

Another limitation is that the system is currently suitable for use in single-resident environments. In settings with multiple people, person tracking is necessary to enable accurate monitoring and address occlusion cases.

B. FUTURE IMPLICATIONS

The emphasis in implementing the elderly activity monitoring process has been on precisely recognizing daily actions including transition states. A potential future upgrade would involve integrating the system with cloud computing to automatically generate resident profiles. Another advancement is the utilization of Large Language Models (LLMs) to provide health summaries based on action recognition results.

Because one of the research goals was to develop a foundational model, utilizing a more diverse dataset during training could enhance its generalizability for real-world applications. Experiments using extended datasets demonstrated that the proposed action recognition model can be effectively applied to other datasets with varying camera positions and environmental conditions through transfer learning with minimal additional data. Future investigations could involve testing with camera streaming in actual environments such as hospitals, elderly care centers, and smart homes. In addition, modifying the model for deployment on devices with limited computational resources, such as Mini PCs and Raspberry Pi computers, could be explored.

VI. CONCLUSION

The proposed system offers several advantages. It demonstrated the effectiveness of using stereo depth cameras for indoor monitoring of the elderly, enabling 24-hour monitoring without additional lighting while preserving user privacy. The camera setup in this experiment was unobtrusive and did not interfere with residents' daily lives. Furthermore, real-time processing was successfully achieved at 1fps.

In conclusion, this system can aid elderly individuals to age safely, facilitating smarter living with the help of AI. Additionally, it can be deployed in smart care centers for remote monitoring and access to health details through a user-friendly GUI, promoting independent living, and assisting caregivers. Furthermore, the effective recognition of specific actions and transition states can provide valuable insights into the well-being of the elderly, aiding in the early detection of potential health issues related to mobility and balance. It is important to recognize that modern technology can benefit all generations. By educating and assisting the elderly in using smart devices and tools, they can be empowered to experience independent living and smarter aging, especially as the elderly population continues to grow.

ABBREVIATION TABLE IN ALPHABETICAL ORDER

Abbreviation	Full Term
AI	Artificial Intelligence
CSV	Comma-Separated Value
CNN	Convolutional Neural Network
CV	Computer Vision
DL	Deep Learning
FC	Fully Connected
FN	False Negative
FP	False Positive
GRU	Gated Recurrent Unit
GUI	Graphical User Interface
HMM	Hidden Markov Model
IoT	Internet of Things
IoU	Intersection over Union
kNN	k-Nearest Neighbors
LLM	Large Language Model
LSTM	Long Short-Term Memory
mAP	mean Average Precision
MotionCRNN	Motion-based Convolutional Recurrent Neural Network
ML	Machine Learning
RGBD	RGB plus Depth
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SAM	Segment Anything Model
STD-TA	Standard Deviation Trend Analysis
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
Vit-ReT	Vision and Recurrent Transformer Neural Network
YOLO	You Look Only Once

REFERENCES

[1] "Demography - elderly population - OECD data," theOECD. Accessed: Feb. 20, 2024. [Online]. Available: http://data.oecd.org/pop/elderly-population.htm

[2] I.-Y. Song, M. Song, T. Timakum, S.-R. Ryu, and H. Lee, "The landscape of smart aging: topics, applications, and agenda," *Data & Knowl. Eng.*, vol. 115, pp. 68–79, May 2018, doi: 10.1016/j.datak.2018.02.003.

[3] E. Freiberger, C. C. Sieber, and R. Kob, "Mobility in older community-dwelling persons: a narrative review," *Front. Physiol*, vol. 11, p. 881, Sep. 2020, doi: 10.3389/fphys.2020.00881.

[4] L. W. Keeler and M. J. Bernstein, "The future of aging in smart environments: four scenarios of the United States in 2050," *Futures*, vol. 133, p. 102830, Oct. 2021, doi: 10.1016/j.futures.2021.102830.

[5] M. P. De Freitas, V. A. Piai, R. H. Farias, A. M. R. Fernandes, A. G. De Moraes Rossetto, and V. R. Q. Leithardt, "Artificial intelligence of things applied to assistive technology: a systematic literature review," *Sensors*, vol. 22, no. 21, p. 8531, Nov. 2022, doi: 10.3390/s22218531.

[6] M. E. N. Gomes, D. Macêdo, C. Zanchettin, P. S. G. de-Mattos-Neto, and A. Oliveira, "Multi-human fall detection and localization in videos," *Comput. Vis. Image Und.*, vol. 220, p. 103442, Jul. 2022, doi: 10.1016/j.cviu.2022.103442.

[7] E. Teixeira et al., "Wearable devices for physical activity and healthcare monitoring in elderly people: a critical review," *Geriatrics*, vol. 6, no. 2, p. 38, Apr. 2021, doi: 10.3390/geriatrics6020038.

[8] M. Buzzelli, A. Albé, and G. Ciocca, "A vision-based system for monitoring elderly people at home," *Appl. Sci.*, vol. 10, no. 1, Art. no. 1, Jan. 2020, doi: 10.3390/app10010374.

[9] T. T. Zin et al., "Real-time action recognition system for elderly people using stereo depth camera," *Sensors*, vol. 21, no. 17, Art. no. 17, Jan. 2021, doi: 10.3390/s21175895.

[10] Y. Htet, T. T. Zin, P. Tin, H. Tamura, K. Kondo, and E. Chosa, "HMM-based action recognition system for elderly healthcare by colorizing depth map," *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, Art. no. 19, Jan. 2022, doi: 10.3390/ijerph191912055.

[11] Y. Htet, T. T. Zin, H. Tamura, K. Kondo, and E. Chosa, "Temporal-dependent features based inter-action transition state recognition for eldercare system," in *Proc. 13th Int. Conf. Consum. Electron. - Berlin (ICCE-Berlin)*, Berlin, Germany: IEEE, Sep. 2023, pp. 106–111. doi: 10.1109/ICCE-Berlin58801.2023.10375682.

[12] P. Jayashree, S. Shrinidhi, V. Aishwarya, and A. Sravanthi, "Smart assistive technologies for aging society: requirements, response and reality," in *Proc. 8th Int. Conf. Adv. Comput. (ICoAC)*, IEEE, pp. 111-116, Jan. 2017. doi: 10.1109/icoac.2017.7951755.

[13] C. M. M. Mansoor, S. K. Chettri, and H. M. M. Naleer, "A remote health monitoring system for the elderly based on emerging technologies," in *Proc. Int. Conf. Emerg. Global Trends in Eng. and Technol.*, Singapore, pp. 513-524, Apr. 2022. doi: 10.1007/978-981-99-4362-3_47.

[14] A. H. Sapci, and H. A. Sapci, "Innovative assisted living tools, remote monitoring technologies, artificial intelligence-driven solutions, and robotic systems for aging societies: systematic review," *JMIR Aging*, vol. 2, no. 2, p. e15429, 2019. doi: 10.2196/15429.

[15] X. Zhou, X. Yi, Y. Liu, S. Xu, Z. Liu, and Z. Yan, "Design of intelligent wearable device based on embedded system," in *Proc. 9th Int. Forum Elect. Eng. Automat. (IFEEA)*, IEEE, pp. 202-205, Nov. 2022. doi: 10.1109/ifeea57288.2022.10037788.

[16] X. Chen, "Smart technologies and aging society," in *Smart Cities and Smart Commun.: Empowering Citizens through Intell. Technol.*, Singapore: Springer Nature Singapore, pp. 131-146, 2022. doi: 10.1007/978-981-19-1146-0_7.

[17] T. Thomas, C. Cashen, and S. Russ, "Leveraging smart grid technology for home health care," in *Proc. Int. Conf. Consum. Electron.*, pp. 274-275, Jan. 2013. doi: 10.1109/icce.2013.6486892.

[18] S. Iqbal, "Artificial intelligence tools and applications for elderly healthcare-review," in *Proc. 9th Int. Conf. Comput. Artif. Intell.*, pp. 394-397, Mar. 2023. doi: 10.1145/3594315.3594347.

[19] C. H. Lee, C. Wang, X. Fan, F. Li, and C. H. Chen, "Artificial intelligence-enabled digital transformation in elderly healthcare field: scoping review," *Adv. Eng. Inform.*, vol. 55, p.101874, 2023. doi: 10.1016/j.aei.2023.101874.

[20] S. Salomé, and E. Monfort, "The digital revolution and ageism: the ethical challenges of artificial intelligence for older people," *NPG Neurologie-Psychiatrie-Gériatrie*, 2023. doi: 10.1016/j.npg.2023.09.004.

[21] M. Koc, "Artificial intelligence in geriatrics," *Turkish J. Geriatrics*, vol. 26, no. 4, 2023. doi: 10.29400/tjgeri.2023.362.

[22] M. T. Harris, K. A. Blocker, and W. A. Rogers, "Older adults and smart technology: facilitators and barriers to use," *Front. Comput. Sci.*, vol. 4, p. 835927, 2022. doi: 10.3389/fcomp.2022.835927.

[23] G. Rubeis, "The disruptive power of artificial intelligence. ethical aspects of gerontechnology in elderly care," *Arch. Gerontology and Geriatrics*, vol. 91, p. 104186, 2020. doi: 10.1016/j.archger.2020.104186.

[24] T. Shiwani, S. Relton, R. Evans, A. Kale, A. Heaven, A. Clegg, and O. Todd, "New horizons in artificial intelligence in the healthcare of older people," *Age and Ageing*, vol. 52, no. 12, 2023. doi: 10.1093/ageing/afad219.

[25] Y. Yamout, T. S. Yeasar, S. Iqbal, and M. Zulkernine, "Beyond smart homes: an in-depth analysis of smart aging care system security," *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1-35, 2024. doi: 10.1145/3610225.

[26] K. M. Kokorelias, A. Grigorovich, M. T. Harris, U. Rehman, L. Ritchie, A. Levy, K. Denecke, and J. McMurray, "Coadaptation between smart technologies and older adults over time: protocol for a scoping review," *JMIR Res. Protocols*, vol.12, no. 1, 2023. doi: 10.2196/51129.

[27] K. Deepa, N. Bacanin, S. S. Askar, and M. Abouhawwash, "Elderly and visually impaired indoor activity monitoring based on wi-fi and deep hybrid convolutional neural network," *Sci. Rep.*, vol. 13, no. 1, p. 22470, Dec. 2023, doi: 10.1038/s41598-023-48860-5.

[28] M. S. Momin, A. Sufian, D. Barman, P. Dutta, M. Dong, and M. Leo, "In-home older adults' activity pattern monitoring using depth sensors: a review," *Sensors*, vol. 22, no. 23, p. 9067, Nov. 2022, doi: 10.3390/s22239067.

[29] A. Kadambi, A. Bhandari, and R. Raskar, "3D depth cameras in vision: benefits and limitations of the hardware: with an emphasis on the first-and second-generation kinect models," *Comput. Vis. Mach. Learn. RGB-D Sensors*, 2014, pp. 3–26. doi: 10.1007/978-3-319-08651-4_1.

[30] J. Park, J. Kim, J. Park, D. Adams, and C. Branstrom, "Development of an unobtrusive sleep monitoring system using a depth sensor," *J. Sleep Disorders: Treatment and Care*, Aug. 2020, Accessed: Feb. 20, 2024. [Online]. Available: https://www.scitechnol.com/abstract/development-of-an-unobtrusive-sleep-monitoring-system-using-a-depth-sensor-11454.html

[31] A. Jalal, S. Kamal, and D. Kim, "A depth video-based human detection and activity recognition using multi-features and embedded hidden markov models for health care monitoring systems," *Int. J. Interactive Multimedia Artif. Intell.*, vol. 4, no. 4, p. 54, 2017, doi: 10.9781/ijimai.2017.447.

[32] R. Jansi and R. Amutha, "Detection of fall for the elderly in an indoor environment using a tri-axial accelerometer and Kinect depth data," *Multidim. Syst. Sign. Process*, vol. 31, no. 4, pp. 1207–1225, Oct. 2020, doi: 10.1007/s11045-020-00705-4.

[33] C. J. Debono, M. Sacco, and J. Ellul, "Monitoring indoor living spaces using depth information," in *Proc. 10th Int. Conf. Consum. Electron. (ICCE-Berlin)*, Nov. 2020, pp. 1–5. doi: 10.1109/ICCE-Berlin50680.2020.9352158.

[34] J. H. Li, L. Tian, H. Wang, Y. An, K. Wang, and L. Yu, "Segmentation and recognition of basic and transitional activities for continuous physical human activity," *IEEE Access*, vol. 7, pp. 42565-42576, 2019. doi: 10.1109/access.2019.2905575.

[35] S. Aminikhanghahi, and D. J. Cook, "Using change point detection to automate daily activity segmentation," in *Proc. Int. Conf. Pervasive Comput. Commun. Workshops*, IEEE, pp. 262-267, Mar. 2017. doi: 10.1109/percomw.2017.7917569.

[36] D. Thakur and S. Biswas, "Online change point detection in application with transition-aware activity recognition," *IEEE Trans. Human-Machine Syst.*, vol. 52, no. 6, pp. 1176–1185, Dec. 2022, doi: 10.1109/THMS.2022.3185533.

[37] S. Irfan, N. Anjum, N. Masood, A. S. Khattak, and N. Ramzan, "A novel hybrid deep learning model for human activity recognition based on transitional activities," *Sensors*, vol. 21, no. 24, p. 8227, 2021. doi: 10.3390/s21248227.

[38] S. Aminikhanghahi, and D. J. Cook, "Enhancing activity recognition using cpd-based activity segmentation," *Pervasive and Mobile Comput.*, vol. 53, pp. 75-89, 2019. doi: 10.1016/j.pmcj.2019.01.004.

[39] L. Song, S. Zhang, G. Yu, and H. Sun, "TACNet: transition-aware context network for spatio-temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11987–11995. Accessed: Feb. 18, 2024. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Song_TACNet_Transition-Aware_Context_Network_for_Spatio-Temporal_Action_Detection_CVPR_2019_paper.html

[40] J. Kang, J. Kim, S. Lee, and M. Sohn, "Transition activity recognition using fuzzy logic and overlapped sliding window-based convolutional neural networks," *J. Supercomputing*, vol. 76, no. 10, pp. 8003-8020, 2020. doi: 10.1007/s11227-018-2470-y.

[41] C. Gu, C. Zhang, and S. Kuriyama, "Orientation-aware leg movement learning for action-driven human motion prediction." *arXiv*, Feb. 05, 2024. doi: 10.48550/arXiv.2310.14907.

[42] J. Shi, D. Zuo, and Z. Zhang, "Transition activity recognition system based on standard deviation trend analysis," *Sensors*, vol. 20, no. 11, Art. no. 11, Jan. 2020, doi: 10.3390/s20113117.

[43] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, Jan. 2016, doi: 10.1016/j.neucom.2015.07.085.

[44] N. Manouchehri and N. Bouguila, "Human activity recognition with an hmm-based generative model," *Sensors*, vol. 23, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/s23031390.

[45] L. Wang, Y. Zhou, R. Li, and L. Ding, "A fusion of a deep neural network and a hidden Markov model to recognize the multiclass abnormal behavior of elderly people," *Knowl. Syst.*, vol. 252, p. 109351, Sep. 2022, doi: 10.1016/j.knosys.2022.109351.

[46] C. Zhao, J. G. Han, and X. Xu, "CNN and RNN based neural networks for action recognition," *J. Phys.: Conf. Ser.*, vol. 1087, no. 6, p. 062013, Sep. 2018, doi: 10.1088/1742-6596/1087/6/062013.

[47] H. Zhao and X. Jin, "Human action recognition based on improved fusion attention cnn and rnn," in *Proc. 5th Int. Conf. Comput. Intell. Appl. (ICCIA)*, Jun. 2020, pp. 108–112. doi: 10.1109/ICCIA49625.2020.00028.

[48] S. Chopra, L. Zhang, and M. Jiang, "Human action recognition using multi-stream fusion and hybrid deep neural networks," in *Proc. IEEE Int. Conf. Syst., Man, and Cybern. (SMC)*, Oct. 2023, pp. 4852–4858.

[49] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018, doi: 10.1109/ACCESS.2017.2778011.

[50] J. Wensel, H. Ullah, and A. Munir, "Vit-Ret: vision and recurrent transformer neural networks for human activity recognition in videos," *IEEE Access*, 2023. doi: 10.1109/access.2023.3293813.

[51] A. S. Rajput, B. Raman, and J. Imran, "Privacy-preserving human action recognition as a remote cloud service using RGB-D sensors and deep CNN," *Expert Syst. Appl.*, vol. 152, p. 113349, Aug. 2020, doi: 10.1016/j.eswa.2020.113349.

[52] M. Dallel, V. Havard, Y. Dupuis, and D. Baudry, "A sliding window based approach with majority voting for online human action recognition using spatial temporal graph convolutional neural networks," in *Proc. 7th Int. Conf. Mach. Learn. Technol. (ICMLT). New York, NY, USA: Assoc. Comput. Mach.*, Jun. 2022, pp. 155–163. doi: 10.1145/3529399.3529425.

[53] T. S. Apon, A. Islam, and MD. G. Rabiul Alam, "Action recognition using transfer learning and majority voting for csgo," in *Proc. 13th Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*, Oct. 2021, pp. 235–240. doi: 10.1109/ICTS52701.2021.9608407.

[54] K. Safdar, S. Akbar, and A. Shoukat, "A majority voting based ensemble approach of deep learning classifiers for automated melanoma detection," in *Proc. Int. Conf. Innov. Comput. (ICIC)*, Nov. 2021, pp. 1–6. doi: 10.1109/ICIC53490.2021.9692915.

[55] H. Yoo, H. Li, Q. Ke, L. Liu, and R. Zhang, "Precondition and effect reasoning for action recognition," *Comput. Vis. Image Und.*, vol. 232, p. 103691, Jul. 2023, doi: 10.1016/j.cviu.2023.103691.

[56] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli, "Explainable video action reasoning via prior knowledge and state transitions," in *Proc. 27th ACM Int. Conf. Multimedia, in MM '19. New York, NY, USA: Assoc. Comput. Mach.*, Oct. 2019, pp. 521–529. doi: 10.1145/3343031.3351040.

[57] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks." *arXiv*, Sep. 11, 2020. doi: 10.48550/arXiv.1905.11946.

[58] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv*, Dec. 11, 2014. doi: 10.48550/arXiv.1412.3555.

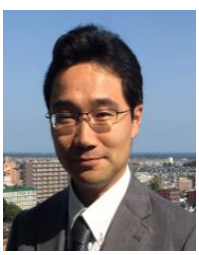
[59] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *Proc. Int. Conf. Syst., Sig. Image Processing (IWSSIP)*, Jul. 2020, pp. 237–242. doi: 10.1109/IWSSIP48289.2020.9145130.

[60] M. Tan and Q. V. Le, "EfficientNetV2: smaller models and faster training," *arXiv*, Jun. 23, 2021. doi: 10.48550/arXiv.2104.00298.

Ye Htet (Graduate Student Member, IEEE) received the B.E. degree and M.E. degree in Electronic Engineering from the University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar, in 2017 and 2020, respectively. Then, he worked as a researcher at the Graduate School of Engineering, University of Miyazaki, Miyazaki, Japan for two years. He is currently a Ph.D. student at the Interdisciplinary Graduate School of Agriculture and Engineering, University of Miyazaki, Japan. His research interests include computer vision, artificial intelligence, deep learning, and human behavior understanding.

Thi Thi Zin (Member, IEEE) received the B.Sc. degree (with honor) in Mathematics in 1995 from Yangon University, Myanmar, and the M.I.Sc. degree in Computational Mathematics in 1999 from the University of Computer Studies, Yangon, Myanmar. She received her master's and Ph.D. degrees in Information Engineering from Osaka City University, Osaka, Japan, in 2004 and 2007, respectively. From 2007 to 2009, she was a Post-Doctoral Research Fellow of Japan Society for the Promotion of Science (JSPS). She is currently a Professor at the Graduate School of Engineering, University of Miyazaki, Miyazaki, Japan. Her research interests include human behavior understanding, intelligent transportation systems, cow behavior analysis, health care monitoring systems, and image recognition.

Pyke Tin (Member, IEEE) received the B.Sc. degree (Hons.) in mathematics from the University of Mandalay, Myanmar, in 1965, the M.Sc. degree in computational mathematics from the University of Rangoon, Myanmar, in 1970, and the Ph.D. degree in stochastic processes and their applications from Monash University, Australia, in 1976. He was a Rector of the University of Computer Studies, Yangon, and a professor of computational mathematics. He is currently a distinguished Professor at the Graduate School of Engineering and a visiting Professor with the International Relations Center in University of Miyazaki, Miyazaki, Japan. His research interests include image search engines, queuing systems, computer vision, stochastic processes, and their applications to image processing.



Hiroki Tamura received B. E. and M. E. degrees from the University of Miyazaki in 1998 and 2000, respectively. From 2000 to 2001, he was an Engineer in Asahi Kasei Corporation, Japan. In 2001, he joined the University of Toyama, Toyama, Japan, where he was a Technical Official in the Department of Intellectual Information Systems. In 2006, he joined the University of Miyazaki, Miyazaki, Japan, where he was an Assistant Professor in the Department of Electrical and Electronic Engineering. Since 2015, he has been a Professor in the Department of Environmental Robotics. His main research interests include Neural Networks and Optimization Problems. In recent years, he has become interested in Biomedical Signal Processing using Soft Computing.

Kazuhiro Kondo graduated from Miyazaki Medical College in 1983, obtained a doctor's qualification, and worked as a surgeon of digestive organs. He received his Ph.D. in Medicine from the University of Miyazaki in 1994. He worked as a Director of Miyazaki Municipal Facilities for Geriatric Health Services, "Sazanka-En," from April 2015 to March 2020. He is currently an Honorable Professor of the Community Medical Center at the University of Miyazaki Hospital and a former director of Miyazaki Tano Municipal Hospital. His research interests include community medicine and gastroenterology.

Shinji Watanabe graduated from Miyazaki Medical College, Japan, in 1992, and obtained a doctor's qualification. Since 2016, he has been working as an orthopaedic surgeon at Miyazaki City TANO Hospital, mainly treating joint diseases of the lower extremities. Since April 2023, he has been the director of Miyazaki City TANO Hospital. His research interests include joint surgery, lower limb surgery, osteoporosis, and pediatric orthopedics.

Etsuo Chosa received the M.D. degree from Oita Medical University, Japan, in 1984, and a Ph.D. degree from Miyazaki Medical University, Japan. He is currently a professor and chairman in the Department of Orthopaedic Surgery, Faculty of Medicine, University of Miyazaki, Japan. Currently, he is the director of the University of Miyazaki Hospital, Japan. He has conducted research on a wide variety of diseases, including osteoarthritis, trauma, rheumatoid arthritis, osteoporosis, and tumors. He has over 200 peer-reviewed publications in international journals and has provided numerous national and international presentations.