

Video Behavior Recognition Running on Edge Devices to Realize a Patient Life Log System for Large-Scale Hospitals [†]

Madoka Inoue ^{*}, Shiomi Kino, Daiki Kobayashi and Kenichi Ogawa

Aiphone Co., Ltd., Nagoya 456-8666, Japan; shiomi_kino@aiphone.co.jp (S.K.); daiki_kobayashi@aiphone.co.jp (D.K.); kenichi_ogawa@aiphone.co.jp (K.O.)

^{*} Correspondence: madoka_inoue@aiphone.co.jp

[†] Presented at 8th International Conference on Knowledge Innovation and Invention 2025 (ICKII 2025), Fukuoka, Japan, 22–24 August 2025.

Abstract

Understanding the activities of hospitalized patients is important for hospital administrators in terms of preventing accidents and improving the efficiency of nursing care. To solve this problem, we have developed a technology to detect 10 types of patient activities from images. Since this image recognition technology operates on edge devices, it simultaneously understands the activities of several hundred beds of patients in a large hospital without being limited by network bandwidth due to video transmission, as is the case with server-based AI. In the experiment, 10 different behaviors of residents of an elderly care facility were detected, and logs of the residents' behaviors were collected. Analysis and utilization of the logs will be considered in future research.

Keywords: action recognition; life-log; image processing; elderly care

1. Introduction

The problem of falls and tumbles by elderly people living in hospitals and elderly care facilities has been the subject of various studies to date, including methods using pressure sensors [1], air spring mattresses and focal loss convolutional neural networks [2], wireless channel state information [3], three accelerometer sensors attached to the trunk of a person [4], and a method to analyze information from a thermopile sensor array (TPA) using a convolutional neural network–recurrent neural network (CNN-RNN) [5]. Conversely, research on fall prevention through image analysis has advanced significantly in recent years. One approach involves estimating a patient's condition from a single image by integrating bed position and posture estimation results to identify various postures, such as “getting out of bed,” “lying down,” “getting up,” and “sitting on the edge of the bed” [6]. Additionally, a method has been proposed for condition estimation using video footage, in which bed position data and time-series posture estimations are analyzed via long short-term memory networks to detect the precise moment a patient begins to rise from the bed [7].

Various hospitals are introducing technologies that combine cameras installed in hospital rooms with image analysis. However, the introduction of products and systems equipped with image recognition technology into hospitals poses a major challenge in terms of data transmission volume. As an example, we considered the introduction of an image analysis system in a hospital with 100 beds. Assuming that the image size of the camera is w (horizontal), h (vertical) (in pixels), the number of image channels is c , the



Academic Editors: Teen-Hang Meen, Chun-Yen Chang and Cheng-Fu Yang

Published: 3 February 2026

Citation: Inoue, M.; Kino, S.; Kobayashi, D.; Ogawa, K. Video Behavior Recognition Running on Edge Devices to Realize a Patient Life Log System for Large-Scale Hospitals. *Eng. Proc.* **2025**, *120*, 39. <https://doi.org/10.3390/engproc2025120039>

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

color tone of each channel is $n[\text{bit}]$, and the number of images taken per second is f , the data volume $a[\text{bps}]$ per second can be expressed by Equation (1).

$$a = w \times h \times c \times n \times f \quad (1)$$

As an operational example, if the camera image resolution is high definition (1280×780 pixels), the number of channels is 3, the color gradation is 8 bits, and the number of frames shot per second is 10 frames per second, the amount of data received from the camera per second is 239 megabytes per second (Mbps). If the image processing unit is located on a server, 239 Mbps of data transmission is required per camera, which means that a 100-bed hospital needs to secure a communication bandwidth of approximately 24 gigabytes per second.

One way to reduce the communication bandwidth is to compress the transmitted images, but there is concern that the recognition accuracy will be degraded due to the deterioration of image quality. To avoid these problems, we developed a method in which images are processed at the edge and only the processed results are sent to the server. The developed method is appropriate for detection on edge devices, but also detects 10 types of patient behaviors. The 10 detection behaviors and the definitions of the behaviors are listed in Table 1.

Table 1. The 10 types of detection actions we propose and their definitions.

Behavior	Definition of Behavior
Sit up	Change from lying down on the bed to sitting up on the bed.
Bed-exit	Change from sitting on the bed to standing on the floor.
Get into the bed	Change from standing on the floor to sitting on the bed
Sit on the edge of the bed	Change from sitting in the middle of the bed to sitting on the edge of the bed. Change from lying on the edge of the bed to sitting on the edge of the bed.
Lay down	Change from sitting position on bed to lying position on bed.
Slip off the bed	Change from lying on the edge of the bed to lying on the floor. Change from sitting on the edge of the bed to sitting on the floor.
Climb up the bed rail	Change from lying on the bed to lying on the railing.
Stand up in the bed	Change from a sitting position on the bed to a standing position on the bed.
Leave the room	The number of people detected standing on the floor changes from 1 to 0.
Enter the room	The number of people detected standing on the floor changes from 0 to 1.

By detecting ten distinct types of patient behaviors, it becomes possible to infer approximate daily activity patterns, such as wake-up and sleep times, as well as periods of physical activity. The collection and analysis of large-scale behavioral data are anticipated to yield novel insights, including correlations between activity trends and patient-specific attributes, as well as early indicators of potential incidents. Accordingly, this research not only seeks to prevent accidents through real-time behavior monitoring but also aims to establish a comprehensive data platform for analyzing risk factors and patient characteristics.

2. Methods

Recent research has used Transformer-based methods [8] for behavior detection, but incorporating a processor that executes the Transformer into an edge device requires a high-performance CPU, a graphics processing unit, or other computing devices, which leads to an increase in the unit cost of the device. Therefore, we propose an action detection method that reduces the overall system cost by combining image processing algorithms before using the Transformer, while still operating on edge devices. The method consists of five processes to detect human actions from input images. The first process is bed detection.

Detection of the bed position is necessary because most of the patient's actions, such as leaving and getting up, need to be interpreted relative to the bed. The second process is person detection. It identifies the presence or absence of a patient in a hospital room and provides a rough idea of the patient's location within the room. The third process is skeleton estimation, which takes the image area around the YOLO-X detection rectangle as input and extracts the skeleton of the person from the image area. The fourth process is the estimation of the person's posture and position.

We used a discriminator that simultaneously estimates posture, which outputs the posture of the detected person in three classes, "standing," "seated," and "supine," and position, which outputs the position of the detected person in two classes, "in bed" and "on the floor." For position and posture estimation, we use the visual attention (VA)-CNN proposed in Zhang et al.'s Skeleton-based Human Action Recognition [9]. In previous studies, VA-CNNs have been used for motion detection using coordinate information from multiple frames as input.

In this study, we used VA-CNNs as a function for estimating a person's posture at an arbitrary time. The reason for this is that when using a discriminator with a single frame as input, if the previous detector, such as person detection or posture estimation, fails to detect a person, the previous and next frames are used to estimate the person's posture at the current frame. The developed method inputs coordinate information for 20 frames. The fifth stage involves behavioral judgment, wherein changes in the patient's position and posture at a given time are analyzed to identify specific behaviors. A behavior is flagged when the detected posture change aligns with the criteria defined in Table 1. The overall detection process is illustrated in Figure 1.

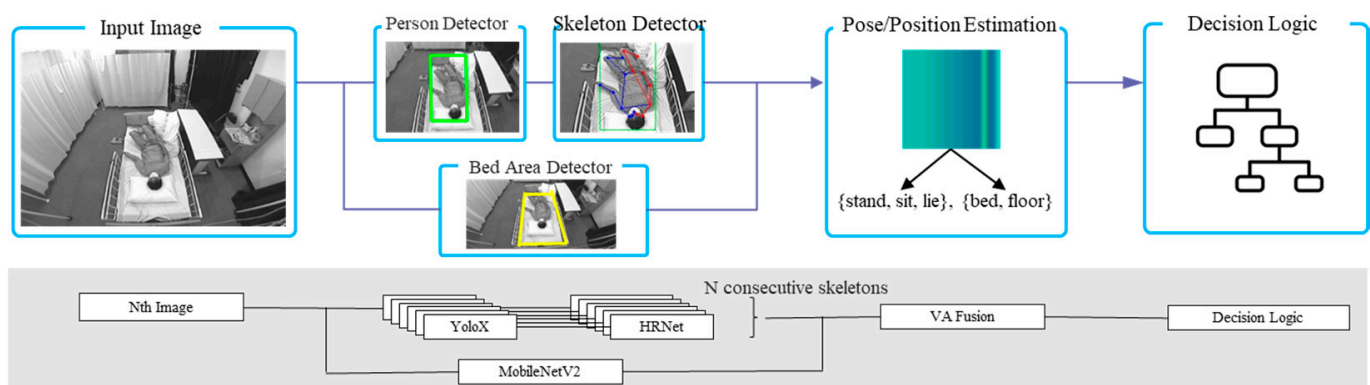


Figure 1. Developed detection process.

3. Results

The architecture employed for human detection is YOLOX-Tiny, comprising 5.06 million parameters. For skeletal estimation, the pose_hrnet_w32 model is utilized, containing 28.5 million parameters. Position and posture estimation is performed using MobileNet-V2, which consists of 3.504 million parameters. The experimental setup incorporates a hospital-grade camera. Specifically, the NLX-CA camera manufactured by Aiphone Co., Ltd. (Nagoya-city, Japan) is used, a device marketed for hospital security applications. This camera offers a wide horizontal field of view of 175 degrees and is equipped with near-infrared illumination, enabling continuous image capture under low-light conditions, including nighttime environments.

The experimental data were collected in a laboratory designed to simulate a hospital room. As this study represents an early stage of research, volunteers were recruited to mimic patient behaviors, such as leaving the bed or sitting on its edge. The evaluation dataset comprises 120 video scenes depicting various movements. For instance,

among these scenes, the supine position appears 36 times, while slipping and falling occur 30 times. A detailed breakdown of behavior occurrences is presented in Table 2.

Table 2. Detection rate and false positive rate for target actions. The numbers in parentheses for the detection rate are the number of detected actions in the denominator and the number of detections in the numerator. The numbers in parentheses for the false positive rate are the number of scenes in the denominator and the number of false positives in the denominator.

Action	Detect Rate [%]	False Detect Rate [%]
Sit up	97.61 (41/42)	1.67 (2/120)
Bed-exit	100.0 (19/19)	3.33 (4/120)
Get into the bed	93.33 (14/15)	6.67 (8/120)
Sit on the edge of the bed	86.05 (37/43)	7.50 (9/120)
Lay down	100.0 (21/21)	2.50 (3/120)
Slip off the bed	93.33 (28/30)	0.83 (1/120)
Climb up the bed rail	100.0 (6/6)	2.50 (3/120)
Stand up in the bed	100.0 (6/6)	1.67 (2/120)
Leave the room	88.89 (24/27)	2.50 (3/120)
Enter the room	91.30 (21/23)	2.50 (3/120)

The training dataset varies by task. For human detection and skeletal estimation, the MS COCO 2017 dataset is used (training: 118,278 images; evaluation: 5000 images), supplemented with images of volunteers in the simulated environment (training: 58,131 images; evaluation: 15,605 images). For position and posture estimation, only the volunteer images from the simulated environment are used.

Detection performance is evaluated using the aforementioned architectures and training data. The system processes input at 5 frames per second (fps). During evaluation, a total of 23,726 images are analyzed, comprising the 15,605 training images and additional frames captured before and after each scene to account for system output latency. The proposed method exhibits a delay of 15 frames between the completion of a movement and the generation of the evaluation result. To ensure accurate performance assessment, a temporal margin is incorporated into the evaluation video. Consequently, each scene contains approximately 200 frames, corresponding to a duration of roughly 40 s.

The evaluation method involves flagging each scene with the frame in which the target behavior occurs. A detection is considered successful if the system identifies the correct behavior within 4 s of the flagged frame. If the detection occurs outside this window or the identified behavior does not match the flagged action, it is classified as undetected. Additionally, any detection without a corresponding flag within 4 s is regarded as a false positive. The detection rates for volunteer behaviors are analyzed and summarized.

4. Discussion

We evaluate the relevance of the ten behavioral categories targeted for detection in the experiment. Four specific actions—“sitting up,” “sitting on the edge of the bed,” “exiting the bed,” and “slipping off the bed”—were identified as essential for reducing the risk of indoor accidents among patients. Additionally, the action of “leaving the room” was included to address potential hazards that may occur outside the patient’s room, such as in hallways or restrooms.

Based on a preliminary survey conducted in hospitals and nursing homes, two additional behaviors—“climbing onto the bed rail” and “standing up in the bed”—were judged to pose significant safety risks and were therefore added to the detection targets. The survey involved unassisted elderly individuals and focused on the diversity of movements performed around their beds. Observations revealed that participants frequently engaged

in daily activities such as “getting up” and “exiting the bed,” as well as more hazardous actions like climbing onto bed rails and standing upright on the bed. Given the elevated risk associated with these behaviors, their inclusion in the detection framework was deemed necessary.

To monitor patient activity, we incorporated the behaviors “lying down,” “getting into bed,” and “entering the room.” These actions were selected as symmetrical counterparts to “getting up,” “exiting the bed,” and “leaving the room,” thereby enabling a more complete understanding of movement patterns and transitions. The video segments capturing “entering” and “leaving” the room were insufficient in duration, limiting the system’s ability to accurately assess these behaviors. The absence of footage immediately before and after these transitions hindered reliable classification. Furthermore, as illustrated in Figure 2, the postures associated with “sitting on the edge of the bed” and “slipping off the bed” were highly similar, leading to misclassification. This overlap in posture representation highlights a challenge that warrants further refinement in future iterations of the detection system.

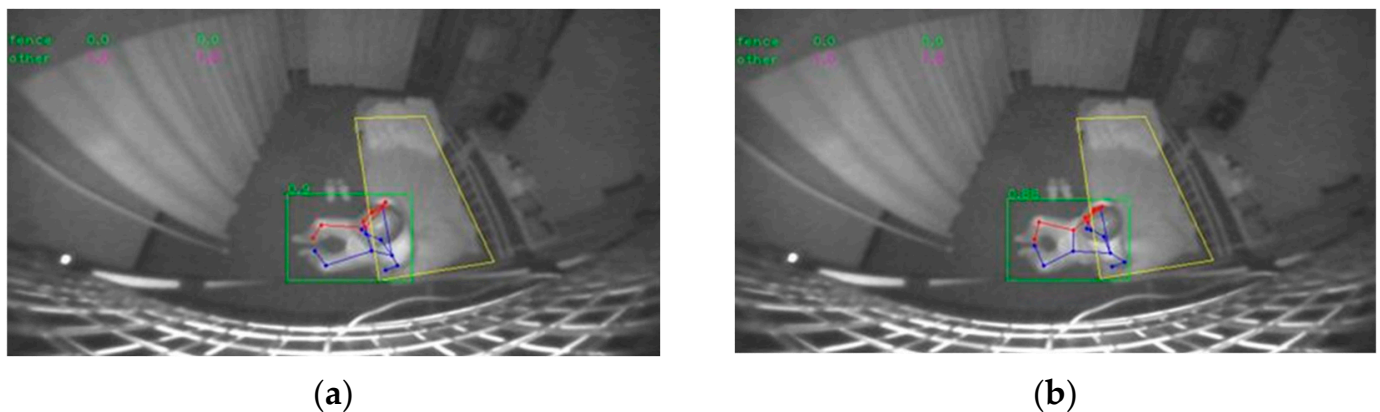


Figure 2. Examples of difficult posture judgments: (a) Sitting on the edge of the bed; (b) Sitting on the floor.

5. Conclusions

We developed a method for detecting 10 types of patient behavior and understanding patient activity trends. The method can be implemented on edge devices, considering the burden of video transmission when installed in a large-scale hospital. In the future, demonstration experiments will be conducted on the accuracy of detection in nursing homes and hospitals, and investigate the acquisition of patient activity trends and the usefulness of the acquired activity trends.

Author Contributions: Conceptualization, M.I.; methodology, M.I., K.O. and D.K.; software, K.O., D.K. and S.K.; validation, K.O., D.K. and S.K.; formal analysis, D.K.; investigation, K.O.; resources, M.I. and K.O.; data curation, K.O.; writing—original draft preparation, M.I.; writing—review and editing, K.O., D.K. and S.K.; visualization, M.I.; supervision, M.I.; project administration, M.I. and K.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the participation of professional actors who were fully informed of the research purpose and provided explicit consent for the use of their recorded data.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the participants to publish this paper.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The raw video data are not publicly available due to contractual restrictions and privacy considerations regarding the participants. Access to processed numerical data may be granted for research purposes upon reasonable request.

Conflicts of Interest: The authors are employees of Aiphone Co., Ltd. The study describes basic research that may support the future development of products by the company. The authors declare no other conflict of interest.

References

1. Lee, C.-N.; Yang, S.-C.; Li, C.-K.; Liu, M.-Z.; Kuo, P.-C. Alarm system for bed exit and prolonged bed rest. In Proceedings of the 2018 International Conference on Machine Learning and Cybernetics (ICMLC), Chengdu, China, 15–18 July 2018; Volume 2.
2. Meng, F.; Liu, T.; Meng, C.; Zhang, J.; Zhang, Y.; Guo, S. Method of bed exit intention based on the internal pressure features in array air spring mattress. *Sci. Rep.* **2024**, *14*, 27273. [[CrossRef](#)]
3. das Chagas, A.D.; Lovisolo, L.; Tcheou, M.P.; e Souza, J.B. A proposal for fall-risk detection of hospitalized patients from wireless channel state information using Internet of Things devices. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108628. [[CrossRef](#)]
4. Härmä, A.; Kate, W.T.; Espina, J. Bed exit prediction based on movement and posture data. In Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Valencia, Spain, 1–4 June 2014.
5. Morawski, I.; Lie, W.-N.; Aing, L.; Chiang, J.-C.; Chen, K.-T. Deep-learning technique for risk-based action prediction using extremely low-resolution thermopile sensor array. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 2852–28633. [[CrossRef](#)]
6. Chen, L.-B.; Chang, W.-J.; Yang, T.-C. BedEye: A Bed Exit and Bedside Fall Warning System Based on Skeleton Recognition Technology for Elderly Patients. *IEEE Access* **2025**, *13*, 60403–60423. [[CrossRef](#)]
7. Inoue, M.; Taguchi, R. Bed exit action detection based on patient posture with long short-term memory. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; IEEE: Piscataway, NJ, USA, 2020.
8. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
9. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.