engineering proceedings

MDPI

Proceeding Paper

Multi-Step Action Recognition for Long-Term Care Using Temporal Convolutional Network–Dynamic Time Warping–Finite State Machine and MediaPipe †

Feng-Jung Liu [1,*], Mei-Jou Lu [2] and Min Chao [3]

[1] Department of Digital Multimedia Design, Cheng Shiu University, Kaohsiung 833, Taiwan
[2] Department of Long-Term Care and Health Management, Cheng Shiu University, Kaohsiung 833, Taiwan; 0671@gcloud.csu.edu.tw
[3] Department of Civil Engineering and Geomatics, Cheng Shiu University, Kaohsiung 833, Taiwan; 0098@gcloud.csu.edu.tw
* Correspondence: k3013@gcloud.csu.edu.tw; Tel.: +886-7-7358800 (ext. 5506)
† Presented at the 7th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability 2025 (ECBIOS 2025), Kaohsiung, Taiwan, 23–25 October 2025.

Abstract

An intelligent multi-step action recognition system was designed for long-term caregiver training and assessment. Leveraging MediaPipe for precise and real-time human pose estimation, the system extracts detailed spatiotemporal body and hand keypoints. Temporal convolutional networks are employed to effectively capture temporal dependencies and complex features from sequential motion data. Dynamic time warping provides robust sequence alignment, allowing flexible comparison between performed actions and standard templates despite temporal variations in execution speed or style. A finite state machine imposes logical constraints by modeling expected action step sequences, enabling accurate detection of sequence anomalies or deviations. This hybrid architecture supports comprehensive evaluation and real-time feedback, facilitating improved caregiver skill acquisition, process adherence, and quality control within long-term care settings. The system aims to advance digital transformation in healthcare education by providing a scalable, precise, and adaptive training solution.

Keywords: TCN-DTW; FSM; MediaPipe; healthcare education; human activity recognition

1. Introduction

The growing demand for high-quality, long-term care, coupled with the increasing complexity of nursing procedures, highlights the need for effective and systematic caregiver training [1]. Traditional skill-training methods often rely on subjective instructor assessments, lack real-time feedback, and provide limited objective assessments or ways to track learners' skill progress, especially in multi-step tasks. These drawbacks can lead to inconsistent training quality, reduced safety, and difficulties in standardizing best practices. Recent advancements in AI, such as pose estimation and sequence modeling, offer innovative tools for objectively capturing, analyzing, and evaluating complex human movements [2]. For example, MediaPipe [3,4] tracks human body and hand movements in real-time and with high accuracy, while the time convolutional network (TCN) [5], combined with dynamic time warping (DTW) [6] and finite state machine (FSM), provides a robust framework for identifying and evaluating sequential nursing actions, accommodating execution variations and enforcing logical step sequences. Integrating these technologies is expected to

Check for updates

Academic Editors: Teen-Hang Meen, Cheng-Yi Chen and Cheng-Fu Yang

Published: 28 February 2026

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.

facilitate continuous, personalized, and accurate training, effectively addressing the major limitations of existing approaches. Therefore, we utilized these AI technologies to develop an intelligent, data-driven nursing skills education system as an auxiliary tool for future long-term care staff training and assessment, aiming to enhance adaptability in long-term care environments.

2. Related Works

Human activity recognition (HAR) and action sequence prediction are gaining attention in computer vision, ubiquitous computing, and smart healthcare. Traditional methods using handcrafted features from sensors or video often struggle with complex activities. Recent deep learning advances, especially spatial–temporal architectures, have greatly improved HAR robustness and generalizability.

One widely adopted framework is the convolutional neural network–long short-term memory (CNN-LSTM) hybrid model. CNNs are highly effective at extracting spatial representations from individual frames or skeletal keypoints, while LSTMs specialize in modeling temporal dependencies across sequences. This combination has been shown to achieve state-of-the-art performance on benchmark datasets such as the Mobile Health Dataset and the Human Activity Recognition Using Smartphones Dataset, often reporting accuracies above 98%. Compared with single-model baselines, CNN-LSTM architectures demonstrate superior ability to recognize repetitive movements and multi-step activities, which are particularly relevant in healthcare scenarios such as rehabilitation exercises, fall detection, and continuous patient monitoring [7–9].

CNN-LSTM models are widely applied in healthcare for monitoring care activities, detecting abnormal movements (e.g., falls, bed exits), and supporting elderly or rehabilitation care in clinical and home settings. They flexibly handle heterogeneous inputs such as sensor data, video, and pose-estimation outputs (e.g., OpenPose, MediaPipe). To enhance performance and adaptability across populations, advanced variants incorporate attention mechanisms, bidirectional LSTMs, and data augmentation strategies [9–11].

In Taiwan and other East Asian regions, researchers have applied CNN-LSTM and pose-based deep learning systems to long-term care monitoring, including tasks such as violence detection in institutions and bed-exit/fall alerts for elderly residents. These implementations highlight the clinical value of such models in improving safety and enabling timely caregiver intervention [12–14].

Recent studies [5,6,15] have highlighted the potential of combining TCN, DTW, FSM, and MediaPipe pose estimation in healthcare applications. This integrated approach enables automated analysis of caregiver actions, clinical workflows, and patient behavior patterns. Looking ahead, the fusion of TCN's ability to capture long-term temporal dependencies, DTW's flexible sequence alignment, and FSM's logical step management with MediaPipe's real-time pose tracking is expected to greatly improve system efficiency, adaptability, and real-time responsiveness, making these techniques fundamental for next-generation intelligent healthcare monitoring and training systems.

2.1. MediaPipe Description

MediaPipe is an open-source machine learning framework developed by Google, designed for building pipelines to process time-series data such as video and audio. It is widely used in applications, including human pose estimation. MediaPipe is a cross-platform framework that supports a range of environments, including desktop and server systems, Android, iOS, and embedded devices like Raspberry Pi and Jetson Nano. This broad platform compatibility allows developers to build applications that perform consistently across various device ecosystems—desktop, mobile, web, and edge devices. Its

https://doi.org/10.3390/engproc2026129021

pose estimation module enables advanced capabilities such as posture analysis, movement tracking, and activity classification [3,4].

2.2. TCN-DTW Framework

TCNs are advanced deep learning models designed to capture long-range temporal dependencies in sequential data. Utilizing causal and dilated convolutions, TCNs efficiently process time series by ensuring no future information leakage and expanding the receptive field exponentially. This enables effective modeling of complex temporal patterns with high computational efficiency and parallelization. TCNs excel in large-scale sequence classification tasks where abundant labeled data is available, automatically extracting hierarchical features that represent temporal dynamics [5].

DTW, on the other hand, is a classical non-parametric alignment technique that computes an optimal match between two time series sequences, accommodating differences in length and speed by elastically warping the time axis. It serves as a robust similarity measure capable of aligning sequences that exhibit temporal distortions, making it particularly effective for anomaly detection and template matching in scenarios with limited training data. However, DTW's computational cost scales quadratically with sequence length and lacks the capacity to learn high-level features autonomously [6].

Integrating TCN with DTW unites deep feature learning and flexible temporal alignment. TCNs capture rich temporal patterns, while DTW handles variations in sequence speed and duration. This synergy improves robustness and accuracy in multi-step action recognition and anomaly detection, especially in healthcare settings with scarce and variable data. By combining learned representations with template matching, the approach enables efficient sequence classification adaptable to real-world temporal dynamics [15].

3. Proposed Methodology

The integration of TCN, DTW, FSM, and MediaPipe is well-suited for applications with limited data. TCNs offer efficient temporal feature extraction with fewer samples compared to complex deep learning models. DTW provides flexible, non-parametric sequence alignment without large training datasets. FSMs use rule-based state management for interpretable and logical process evaluation, minimizing reliance on extensive data. MediaPipe delivers real-time, precise skeletal keypoint extraction, ensuring reliable inputs even with small datasets. This method balances deep learning and rule-based methods, making it ideal for multi-step action recognition and assessment in clinical, educational, or smart caregiving environments where data is scarce and annotations costly. The proposed workflow is shown in Figure 1.

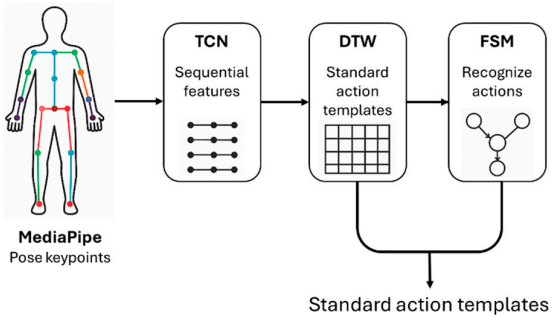


Figure 1. The system workflow.

3.1. Data Extraction with MediaPipe

Using MediaPipe, 33 key body landmarks are extracted in real time from video input, capturing the participant's movements. Each landmark is represented by its three-

dimensional coordinates (x, y, z). To ensure spatial consistency across different camera views and subjects, the raw landmark coordinates are preprocessed using relative positioning. Specifically, landmark coordinates are normalized relative to the pelvic center, and scaled by the participant's shoulder width to account for individual anatomical differences.

3.2. MediaPipe Data: Relative Positioning and Normalization

To normalize for spatial variability caused by camera position and participant placement, we adopt a relative positioning method based on both pelvic and shoulder landmarks. Specifically, we selected the pelvic center (MediaPipe landmark index 0) as the primary reference point for translational alignment. Additionally, the horizontal distance between the left and right shoulder points (landmark indices 11 and 12) is used to define a scale factor reflecting the participant's shoulder width. This scale is employed to normalize the size of the skeleton to reduce inter-subject body size variation. The relative and normalized coordinates $\left({x_i}', {y_i}', {z_i}'\right)$ for each landmark are computed as follows.

$$\left(x_i', y_i', z_i'\right) = \frac{\left(x_i - x_{pelvic}, y_i - y_{pelvic}, z_i - z_{pelvic}\right)}{d_{shoulders}} \quad (1)$$

where $d_{shoulders}$ is the Euclidean distance between left and right shoulder landmarks.

The processed relative and normalized landmark data serve as stable and standardized input features for downstream analysis. TCN extracts deep temporal features from these sequences, enabling precise classification, while DTW aligns variable-length sequences flexibly, facilitating robust assessment of healthcare-related movements.

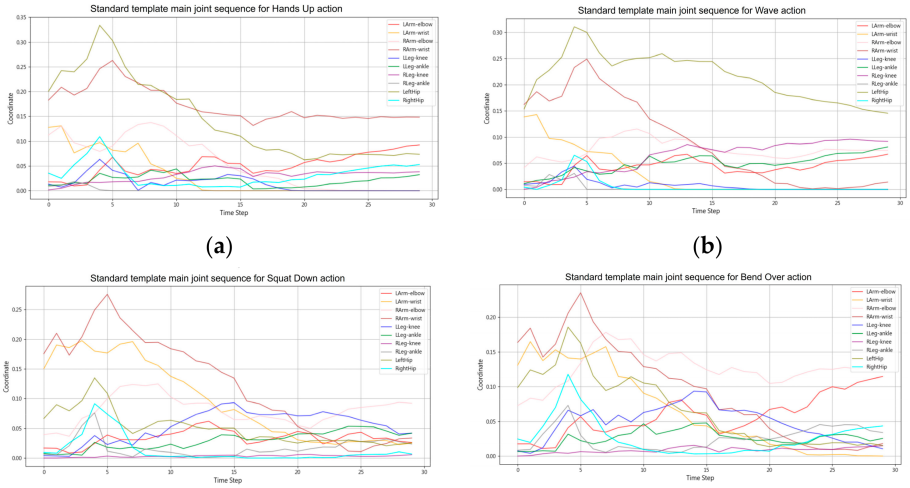
3.3. TCN+DTW Framework

The process begins with data preprocessing, where input sequences are extracted from MediaPipe skeleton-based data. To ensure consistency, the sequence length is normalized using padding or sliding windows, with a default length of 30 frames. These preprocessed sequences are then fed into TCN, which learns temporal-invariant representations and captures motion dynamics over time. The output of TCN is a feature embedding, represented either as a sequence of feature vectors or as a pooled representation.

Next, the extracted feature sequence is compared against stored standard templates using DTW. This method aligns sequences along the temporal dimension, compensating for differences in execution speed, and the alignment cost serves as a measure of similarity between the two actions. For each input sequence, DTW distances are computed against all action templates, and the action label corresponding to the minimum distance is selected. If the mean distance exceeds a predefined threshold, the action is classified as the rejected.

The TCN-DTW framework offers robust temporal feature extraction, ensuring invariance to variations in execution speed and delivering reliable recognition even with limited training data. In Figure 2, the four sub-action visual templates trained using the TCN model are shown in this experiment. To enhance clarity, only the main skeletal features are presented in these images. To address the differences in movement speeds between caregivers, as driven by the design of streaming DTW [16,17], we employ frame buffer management and utilize computational power to implement online inference, overcoming the limitations of the TCN model in real-time sub-action alignment. This integration supports continuous, real-time comparison and matching of ongoing motion sequences, enabling precise and adaptive assessment in healthcare applications. By effectively managing the frame buffer, the system performs inference at 0.5 s intervals, thereby enhancing both real-time responsiveness and processing efficiency.

https://doi.org/10.3390/engproc2026129021



(**a**) (**b**)

(**c**) (**d**)

Figure 2. Visual standard templates of main joint sequence for 4 sub-actions: (**a**) Standard template for Hands Up action; (**b**) Standard template for Wave action; (**c**) Standard template for Squat Down action; (**d**) Standard template for Bend Over action.

3.4. FSM Controllor

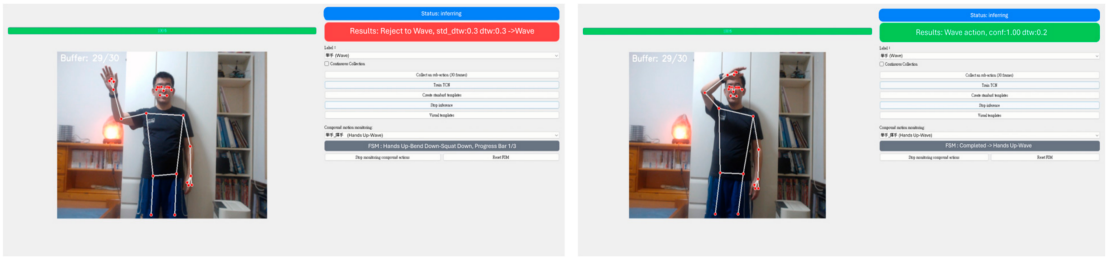
In this system, the FSM controller manages the flow of multi-step actions using predefined standard models derived from expert caregiver demonstrations. Temporal and spatial features form reference patterns, and sub-actions are marked complete only when the performed data satisfactorily match these models, enabling process evaluation, anomaly detection, and instructional feedback. To regulate the action sequence, FSM controls the workflow with states representing each sub-action step, as follows.

$$[S_0\text{: } Start \rightarrow S_1\text{: } Lift \rightarrow S_2\text{: } Turn \rightarrow S_3\text{: } Place \rightarrow End.] \quad (2)$$

Transitions between states occur only after the correct completion of the preceding sub-action, ensuring strict adherence to the standardized care procedure and enabling precise monitoring and quality assurance throughout the process.

3.5. Training and Optimization

Figure 3 shows the TCN-DTW-FSM motion recognition system prototype, highlighting its functions: collecting action data, training the TCN model, creating templates, starting inference, and monitoring compound motions. The model will use multi-step ground truth sub-action sequences for supervised learning, with data augmentation and regularization to enhance generalization and robustness.



(**a**) (**b**)

Figure 3. The TCN-DTW-FSM Motion Recognition System Executions: (**a**) Refusal to adjudicate if DTW distance exceeds the pretrained threshold; (**b**) Inference results: Confidence: 1.00 DTW: 0.3 (wave action).

4. Conclusions

This integrated TCN-DTW-FSM framework offers a scalable and interpretable solution for multi-step action recognition in healthcare training and monitoring, particularly focusing on long-term care personnel training. It supports both real-time training via live camera feeds and offline processing through annotation and extraction from prerecorded videos. By capturing fine-grained sequential motions with temporal dependencies, the system facilitates accurate skill assessment, error detection, and real-time feedback in clinical and remote care settings. To handle caregiver variability and ensure timely inference, we implemented online matching inference by efficient buffer management.

By combining deep learning with rule-based methods, the framework balances efficiency and robustness, making it ideal for multi-step action recognition in clinical, educational, and caregiving environments, especially where data scarcity and high annotation costs are challenges. Currently in the prototype stage, the system has not been trained or evaluated on real-world datasets. Future development will refine dataset preparation, learning configurations, data augmentation, training protocols, and evaluation metrics to improve model performance and system reliability, while also adding multi-person tracking for interactive or group scenarios.

Author Contributions: Conceptualization, F.-J.L. and M.C.; methodology, F.-J.L.; software, F.-J.L.; validation, F.-J.L., M.-J.L., and M.C.; formal analysis, F.-J.L.; investigation, F.-J.L.; resources, M.C.; data curation, M. J. L.; writing—original draft preparation, F.-J.L.; writing—review and editing, F.-J.L.; visualization, F.-J.L.; supervision, F.-J.L.; project administration, F.-J.L.; funding acquisition, M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Glauberman, G.; Ito-Fujita, A.; Katz, S.; Callahan, J. Artificial Intelligence in Nursing Education: Opportunities and Challenges. *Hawaii J. Health Soc. Welf.* **2023**, *82*, 302–305. [PubMed]
2. Ramirez-Alcocer, U.M.; Tello-Leal, E.; Romero, G.; Macías-Hernández, B.A. A Deep Learning Approach for Predictive Healthcare Process Monitoring. *Information* **2023**, *14*, 508. [CrossRef]
3. LearnOpenCV. MediaPipe—The Ultimate Guide to Video Processing. Available online: https://learnopencv.com/introduction-to-mediapipe/ (accessed on 21 August 2025).
4. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M.; Lee, J.; et al. MediaPipe: A Framework for Perceiving and Processing Reality. In Proceedings of the 3rd Workshop on Computer Vision for AR/VR at CVPR, Long Beach, CA, USA, 17 June 2019. Available online: https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf (accessed on 20 August 2025).
5. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal Convolutional Networks for Action Segmentation and Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1003–1012. [CrossRef]
6. Shulman, Y. Dynamic Time Warp Convolutional Networks. *arXiv* **2019**, arXiv:1911.01944. [CrossRef]
7. Chen, J.; Wang, J.; Yuan, Q.; Yang, Z. CNN-LSTM Model for Recognizing Video-Recorded Actions Performed in a Traditional Chinese Exercise. *IEEE J. Transl. Eng. Health Med.* **2023**, *11*, 351–359. [CrossRef] [PubMed]
8. Tan, X.-N. Human Activity Recognition Based on CNN and LSTM. *J. Comput.* **2023**, *34*, 221–235. [CrossRef]

9. Khatun, M.A.; Yousuf, M.A.; Ahmed, S.; Uddin, M.Z.; Alyami, S.A.; Al-Ashhab, S.; Akhdar, H.F.; Khan, A.; Azad, A.; Moni, M.A. Deep CNN-LSTM with Self-Attention Model for Human Activity Recognition Using Wearable Sensor. *IEEE J. Transl. Eng. Health Med.* **2022**, *10*, 2700316. [CrossRef] [PubMed]
10. Hassan, N.; Miah, A.S.M.; Shin, J. A Deep Bidirectional LSTM Model Enhanced by Transfer-Learning-Based Feature Extraction for Dynamic Human Activity Recognition. *Appl. Sci.* **2024**, *14*, 603. [CrossRef]
11. Alshammari, S.A.; Albalawi, N.S. Enhancing Healthcare Monitoring: A Deep Learning Approach to Human Activity Recognition Using Wearable Sensors. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 18843–18848. [CrossRef]
12. Cheng, Y.-C. Research on the Application of Deep Learning Technology in Long-Term Care Action Recognition. Master's Thesis, National Chin-Yi University of Technology, Taichung, Taiwan, 2022. Available online: https://hdl.handle.net/11296/76u4xc (accessed on 20 August 2025).
13. Chang, W.-J. *A Bed-Exit and Bedside Fall Warning System Based on Deep Learning Edge Computing Techniques*; Project Report; Southern Taiwan University of Science and Technology: Tainan, Taiwan, 2020. Available online: https://www.futuretech.org.tw/futuretech/index.php?action=product_detail&prod_no=P0008700005608&web_lang=en-us (accessed on 18 August 2025).
14. Industrial Technology Research Institute (ITRI). Smart Monitoring and Assisted Care System. Available online: https://ictjournal.itri.org.tw/xmdoc/cont?xsmsid=0M236556839091904564&sid=0O087625319827470907 (accessed on 18 August 2025).
15. Moor, M.; Horn, M.; Rieck, B.A.; Roqueiro, D.; Borgwardt, K.M. Temporal Convolutional Networks and Dynamic Time Warping Can Drastically Improve the Early Prediction of Sepsis. *arXiv* **2019**, arXiv:1902.01659.
16. Cai, X.; Xu, T.; Yi, J.; Huang, J.; Rajasekaran, S. DTWNet: A Dynamic Time Warping Network. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 14 December 2019; pp. 11640–11650.
17. Sakurai, Y.; Faloutsos, C.; Yamamuro, M. Stream Monitoring under the Time Warping Distance. In Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE), Istanbul, Turkey, 15–20 April 2007; pp. 1046–1055. [CrossRef]

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.