Human Action Recognition and Custom Dataset for Bus Passenger Safety

Dong Gyu Lee, Dong Seog Han

Kyungpook National University, Electronic and Electrical Engineering,
Deagu, Republic of Korea
doe58@knu.ac.kr, dshan@knu.ac.kr

Although many buses are equipped with onboard CCTV to assist drivers in monitoring the cabin, identifying abnormal passenger conditions in real time remains challenging. This work aims to enable early detection of emergency situations by recognizing passenger behaviors from video captured by cameras installed inside special-purpose vehicles. In practice, the interior of bus-like vehicles is highly complex and cluttered, which makes robust behavior understanding difficult. Another major limitation is the lack of publicly available datasets that represent diverse special-vehicle interiors. To address these issues, we recreated multiple bus environments and collected in-vehicle data to build a dedicated dataset. Using the collected dataset, we benchmarked several deep learning models to explore suitable approaches for passenger behavior recognition, and our proposed method achieved higher recognition accuracy than the competing baselines.

DOI
https://doi.org/
10.18690/um.feri.4.2026.10

ISBN
978-961-299-111-1

Keywords:
human action recognition,
bus passenger safety,
deep learning,
custom dataset,
pose estimation

University of Maribor Press

DOI
https://doi.org/
10.18690/um.feri.4.2026.10

ISBN
978-961-299-111-1

Prepoznavanje človeških aktivnosti in razvoj podatkovne zbirke za varnost potnikov na avtobusu

Dong Gyu Lee, Dong Seog Han
Nacionalna univerza Kyungpook, elektrotehnika in inženiring,
Deagu, Južna Koreja
doe58@knu.ac.kr, dshan@knu.ac.kr

Ključne besede:
prepoznavanje človeških aktivnosti,
varnost potnikov na avtobusu,
globoko učenje,
namenska podatkovna zbirka,
ocena položaja telesa

Kljub temu da so številni avtobusi opremljeni z video-nadzornimi sistemi za pomoč voznikom pri spremljanju dogajanja v vozilu, pravočasna prepoznava nepredvidenih dogodkov, ki ogrožajo varnost potnikov, še vedno predstavlja velik izziv. Cilj tega dela je zgodnje odkrivanje nevarnih situacij s prepoznavanjem vedenja potnikov na podlagi videoposnetkov iz kamer, nameščenih v vozilih. Zanesljivo prepoznavanje človeških aktivnosti otežuje kompleksna notranjost vozil, kjer oprema in drugi potniki povzročajo številne okluzije. Dodatno omejitev predstavlja pomanjkanje javno dostopnih podatkovnih zbirk, ki bi zajemale raznolike pogoje v vozilih. Za reševanje teh izzivov smo zasnovali več različnih avtobusnih okolij in zbrali posnetke iz notranjosti vozil, s katerimi smo oblikovali novo podatkovno zbirko. Na podlagi zbrane zbirke smo izvedli primerjavo več pristopov globokega učenja za prepoznavo človeških aktivnosti, pri čemer naša metoda dosega najvišjo točnost v primerjavi z obstoječimi metodami.

Univerzitetna založba
Univerze v Mariboru

1 Introduction

In large vehicles such as buses, passengers may be injured when they collide with surrounding structures or lose balance during sudden acceleration or abrupt braking, as well as when they move inside the cabin while the vehicle is in motion. The severity of these incidents ranges from minor bruises to serious falls, and the risk can be higher for elderly passengers or those with medical conditions who may experience unexpected physical distress. Although many buses are equipped with onboard CCTV and safety fixtures such as handrails and grab handles to mitigate these risks, ensuring passenger safety remains a persistent challenge.

This study focuses on understanding passenger behavior from bus-mounted CCTV footage. Specifically, we aim to detect passengers and infer their current actions so that potentially hazardous or abnormal situations can be communicated to the driver in a timely manner. To this end, we evaluate multiple deep learning-based approaches for passenger behavior recognition and compare their performance to identify the most suitable method.

A key difficulty is that real bus interiors are visually cluttered and frequently involve occlusions caused by seats, poles, and other passengers. Moreover, publicly available datasets that capture passenger behaviors under such complex in-bus conditions are scarce, making model training and fair evaluation difficult. To address this limitation, we constructed a bus-like environment and collected image data of staged passenger actions, which we then used for training and benchmarking the proposed recognition models.

1.1 Related Work

Human Action Recognition (HAR) has been studied extensively due to its reliance on spatiotemporal cues and its broad applicability in surveillance, robotics, sports analytics, healthcare, and human–machine interaction. As deep learning architectures diversified, recent surveys have focused on organizing the field and clarifying the interplay between model families, sensing modalities, and evaluation protocols.

SMART-Vision provides a vision-centric survey and introduces a Venn-diagram taxonomy that explicitly captures hybridization among major paradigms such as two-stream models, 3D CNNs, graph-based methods, motion-centric designs, and transformers (AlShami et al., 2025). In addition to general RGB-based HAR, a dedicated survey of RGB-D action recognition summarizes algorithms and benchmarks that leverage depth information for improved robustness under viewpoint and illumination variations (Zhang and Wang, 2025). Few-shot action recognition has also been reviewed systematically, categorizing methods (e.g., generative and meta-learning based approaches) and emphasizing video-specific challenges such as temporal modeling and high-dimensional representations (Wanyan et al., 2024). Together, these survey works motivate a structured discussion of architectures and generalization settings in modern HAR.

Although video is the dominant input for HAR, action understanding from still images or sparse key frames remains important when temporal evidence is limited. Transfer-learning pipelines that fine-tune ImageNet-pretrained CNN backbones provide practical baselines for still-image action classification, demonstrating that strong spatial representations can partially compensate for missing motion cues (Kanangama and Thirukumaran, 2024). For sports scenarios, CBAM-enhanced ResNet variants combined with keypoint detection have been explored to emphasize discriminative spatial regions and pose-related patterns in video frames (Xu and Sun, 2024). Beyond classification, image-based action retrieval has been formulated as a distinct task: region-aware transformer models fuse person-centric and contextual features to retrieve semantically similar actions from static imagery (Wang et al., 2024). These studies highlight the value of region/context modeling when temporal information is weak or absent.

For video-based HAR, a key challenge is capturing temporal dynamics efficiently while maintaining discriminative capacity. Temporal Shift Module (TSM)-based backbones offer a lightweight mechanism for temporal interaction, and an ensemble-driven TSM system achieved strong performance in a competitive multi-modal recognition setting (Duong and Gomez-Krämer, 2024). Recent work also emphasizes fine-grained action details. FocusVideo introduces learnable local action queries to attend to action-relevant regions without explicit region annotations and couples global context with local refinement in transformer blocks, improving recognition for detail-sensitive classes (Wang et al., 2025).

Skeleton-based methods exploit human body structure via graph modeling to represent joints and their interactions. Multi-scale spatial graph convolution has been proposed to better capture both local and long-range joint dependencies by operating across multiple receptive fields (Fang and Wang, 2025). Beyond skeleton-only pipelines, hybrid fusion approaches aim to combine complementary cues from raw pixels and structured pose graphs. GCCIN adopts a dual-path architecture that fuses CNN-derived pixel features with ST-GCN-style skeleton features for multi-scale action recognition, seeking improved robustness under complex appearance variations (Huang and Zhang, 2025).

Deployment-oriented HAR often faces challenging capture conditions and heterogeneous sensors. Drone-based action recognition pipelines address aerial viewpoints characterized by motion blur, dynamic backgrounds, and scale changes (Abbas and Jalal, 2024). For low-illumination settings, Dark Transformer studies cross-domain adaptation with video transformers to transfer knowledge from source domains to actions observed “in the dark” (Ulhaq, 2024). In distributed multi-camera environments with weak supervision, action selection learning has been proposed for multi-label multi-view recognition, learning to select action-relevant frames and to fuse wide-range views using only video-level tags (Nguyen et al., 2024). Finally, multimodal fusion of video and wearable IMU signals has been investigated in sports analytics via CNN–LSTM pipelines with adaptive weighting mechanisms, demonstrating that complementary sensing can improve recognition and enable quantitative motion assessment (Chen et al., 2025).
Overall, the above literature indicates that modern HAR systems are increasingly hybrid in architecture and modality, and that robustness and generalization are central considerations for real-world deployment.

2 Experiments

Due to the lack of publicly available passenger-safety datasets that reflect the characteristics of Korean bus interiors (city bus, express bus, emergency bus), we built a dedicated acquisition environment and collected our own data. Four cameras were mounted inside a test bus to capture passenger behaviors from multiple viewpoints. Five participants acted according to predefined scenarios designed to cover both a single-passenger setting and a crowded setting with multiple passengers, enabling us to model behavior under different levels of interaction and occlusion.

During recording, each participant repeatedly performed the target actions, and fixed time intervals were introduced between actions to ensure clear temporal separation for labeling.

To reflect realistic in-bus conditions, we intentionally included segments where the full body of a passenger is clearly visible as well as segments where the subject is partially occluded by seats, handrails, or other people. We defined five behavior classes commonly observed in buses: sitting, standing, holding (grabbing), walking, and falling. After data collection, we developed an in-house preprocessing tool to convert the raw videos into a format suitable for our recognition pipeline. For each camera, recordings were captured from diverse angles with durations ranging from approximately 3 to 5 minutes, and stored as 45-frame clips. In total, the dataset contains about 14 GB of video data.

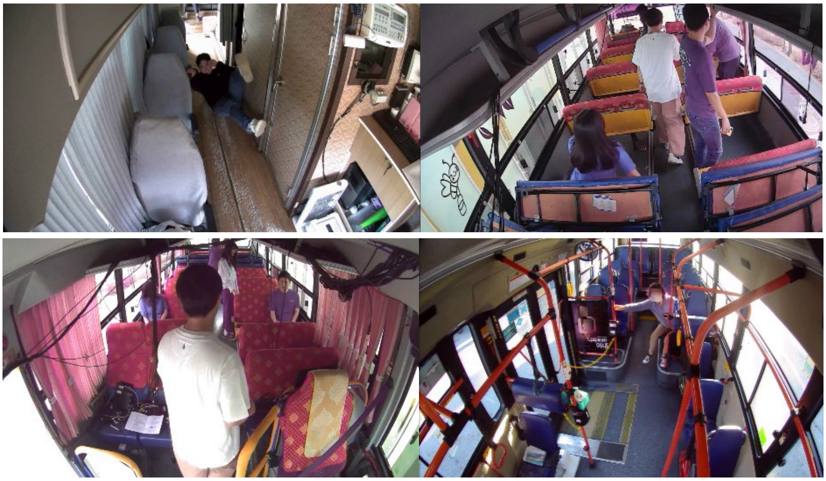

Figure 1: Dataset for various bus environments
Source: Own.

All experimental models were trained and tested using our collected custom dataset. For a fair comparison, we resized the input to 256×192 pixels and used approximately 16K training images in total. Performance was evaluated in terms of Average Precision (AP).

Among the baselines, DeepPose (Toshev and Szegedy, 2014) showed the weakest results, largely because it relies on a relatively simple CNN design; its accuracy dropped substantially when the target person was partially occluded by surrounding objects. HRNet (Wang and Sun K,2020), which preserves high-resolution representations through parallel high/low-resolution branches and repeated feature exchange, did not achieve the expected gains in our setting. We attribute this to the constrained input resolution, which likely limited HRNet's ability to exploit fine-grained spatial cues.

We also evaluated SimCC (Li Yang et al, 2022), which estimates joints via coordinate classification rather than heatmap regression. Since SimCC requires an additional detector, we adopted an HRNet-based detector in our implementation. Representing joint locations using bin-style coordinates yielded higher performance than vanilla HRNet, presumably because it mitigates quantization errorscommonly introduced by heatmap discretization.

Table 1: Performance comparision table

	Input size	AP	AP^{50}	AP^{75}
DeepPose	256 x 192	0.499	0.664	0.523
HRNet	256 x 192	0.540	0.677	0.631
SimCC	256 x 192	0.671	0.799	0.736
YOLO-Pose	256 x 192	0.704	0.849	0.748
RTMPose	256 x 192	0.721	0.866	0.795
PS-Pose(ours)	256 x 192	0.725	0.877	0.802

YOLO-Pose (Maji and Nagori,2022), slightly outperformed SimCC. This improvement appears to stem from differences in both detector quality and the training strategy described in the original method. Moreover, by leveraging CSP-Darknetand PANetwith multi-head predictions, YOLO-Pose remained relatively robust under occlusion-heavy scenes.

RTMPose (Jiang et al., 2025), incorporating more recent design choices, achieved the strongest overall performance in our experiments and was particularly suitable for crowded bus scenarios with frequent occlusions. Interestingly, YOLO-Pose produced better AP50scores, which we interpret as an effect of detector disparity-our RTMPose configuration used YOLOv3 as the detector. Despite this, RTMPose provides a practical balance by combining advantages of coordinate classification (computational efficiency) with strong pose estimation accuracy, making it scalable for real-world deployment.

Finally, we considered PS-Pose, an enhanced variant built upon RTMPose. In our study, we integrated an additional module designed to exploit surrounding contextual cues in crowded environments, where interactions with nearby objects and passengers are common. This context-aware design improved discrimination between visually similar behaviors, leading to a measurable gain in recognition accuracy under multi-passenger and occlusion-dominant conditions.

3 Conclusion

Using bus-mounted cameras, this study collected scenario-driven images under diverse conditions-including single-passenger and crowded-cabin cases-to build a dataset tailored to complex in-bus environments. We then investigated effective approaches for recognizing passenger behaviors by benchmarking a skeleton-based keypoint method against a spatial-information fusion model and comparing their performance.

References

Abbas, Y., & Jalal, A. (2024). Drone-based human action recognition for surveillance: a multi-feature approach. In 2024 International Conference on Engineering & Computing Technologies (ICECT)(pp. 1–6). doi:10.1109/ICECT61618.2024.10581378

AlShami, A. K., Rabinowitz, R., Lam, K., Shleibik, Y., Mersha, M., Boult, T., & Kalita, J. (2025). Smart-vision: survey of modern action recognition techniques in vision. Multimedia tools and applications, 84(27), 32705–32776. doi:10.1007/s11042-024-20484-5

Benavent-Lledo, M., Mulero-Pérez, D., Ortiz-Perez, D., Garcia-Rodriguez, J., & Argyros, A. (2024). Enhancing action recognition by leveraging the hierarchical structure of actions and textual context. arXiv preprint arXiv:2410.21275. https://arxiv.org/abs/2410.21275

Chen, Y., Chen, Y., Lou, K., Chen, Y., Li, H., & Jiang, Z. (2025). Multimodal Action Recognition and Quantitative Analysis Model for Pickleball Based on CNN-LSTM with Adaptive Weight Fusion. In 2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)(pp. 855-866). doi:10.1109/ICICML67980.2025.11333691

Duong, A. K., & Gomez-Krämer, P. (2024). Action recognition using temporal shift module and ensemble learning. In International Conference on Pattern Recognition(pp. 302–313). doi:10.1007/978-3-031-88217-3_23

Fang, Y., & Wang, B. (2025). Multi-scale Spatial Graph Convolutional Network for Skeleton-based Action Recognition. In 2025 2nd International Conference on Digital Image Processing and Computer Applications (DIPCA)(pp. 247-250). doi:10.1109/DIPCA65051.2025.11042736

Huang, W., & Zhang, J. (2025). Graph Convolution Combined with Image Pixel Information Networks for Multi-Scale Action Recognition. In 2025 5th International Conference on Artificial Intelligence, Automation and High Performance Computing (AIAHPC)(pp. 157-163). doi:10.1109/AIAHPC66801.2025.11290605

Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., & Chen, K. (2023). RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. arXiv preprint arXiv:2303.07399. doi:10.48550/arXiv.2303.07399.

Kanangama, K. A. N. C., & Thirukumaran, S. (2024). Enhancement of Image Based Human Action Recognition using Transfer Learning with Pre-Trained CNN Architectures. In 2024 9th International Conference on Information Technology Research (ICITR)(pp. 1-6). doi:10.1109/ICITR64794.2024.10857746

Lan, J., & Yuan, T. (2024). V-APT: Action Recognition Based on Image to Video Prompt-Tuning. In 2024 2nd International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII)(pp. 63-67). doi:10.1109/ICMIII62623.2024.00018

Li, S., Li, W., Ren, Q., & Lin, D. (2025). Few-Shot Human Action Recognition with CLIP-Based Semantic-Guided Network. In 2025 5th International Conference on Computer, Control and Robotics (ICCCR)(pp. 573-577). doi:10.1109/ICCCR65461.2025.11072638

Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W., & Xia, S.-T. (2022). SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation. Computer Vision – ECCV 2022. doi:10.48550/arXiv.2107.03332. (Available at arXiv:2107.03332).

Maji, D., Nagori, S., Mathew, M., & Poddar, D. (2022). YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. arXiv preprint arXiv:2204.06806. doi:10.48550/arXiv.2204.06806.

Nguyen, T. T., Kawanishi, Y., Komamizu, T., & Ide, I. (2024). Action Selection Learning for Multi-label Multi-view Action Recognition. In Proceedings of the 6th ACM International Conference on Multimedia in Asia(pp. 1–1). doi:10.1145/3696409.3700211

Toshev, A., & Szegedy, C. (2014). DeepPose: Human Pose Estimation via Deep Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/CVPR.2014.214. (Available at arXiv:1312.4659).

Ulhaq, A. (2024). Dark Transformer: A Video Transformer for Action Recognition in the Dark. arXiv preprint arXiv:2407.12805. https://arxiv.org/abs/2407.12805

Wang, H., Zhao, J., & Gui, J. (2024). Region-aware image-based human action retrieval with transformers. Computer Vision and Image Understanding, 249, 104202. doi:10.1016/j.cviu.2024.104202

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., et al.(2020). Deep High-Resolution Representation Learning for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).doi:10.1109/TPAMI.2020.2983686. (Available at arXiv:1908.07919).

Wang, M., Huang, Z., Kong, X., Shen, G., Dai, G., Wang, J., & Liu, Y. (2025). Action Detail Matters: Refining Video Recognition with Local Action Queries. In Proceedings of the Computer Vision and Pattern Recognition Conference(pp. 19132–19142). doi:10.1109/CVPR52734.2025.01782

Wanyan, Y., Yang, X., Dong, W., & Xu, C. (2024). A comprehensive review of few-shot action recognition. arXiv preprint arXiv:2407.14744. https://arxiv.org/abs/2407.14744

Xu, C., & Sun, L. (2024). Convolutional Block Attention Mechanism with ResNet18 for Sports Action Recognition based on Video Image Key Point Detection. In 2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC)(pp. 1-5). doi:10.1109/ICDSCNC62492.2024.10941504

Zhang, Y., & Wang, Y. (2025). A comprehensive survey on RGB-D-based human action recognition: algorithms, datasets, and popular applications. EURASIP Journal on Image and Video Processing, 2025(1), 15. doi:10.1186/s13640-025-00677-0