FootSense: A Foot-Centric Dataset and Framework for Energy-Efficient HAR and Terrain Recognition

Dominique Nshimyimana[1,2] [0009−0009−8580−1248], Vitor Fortes Rey[1] [0000−0002−8976−5960], Bo Zhou[1,2], and Paul Lukowicz[1,2]

[1] German Research Center for Artificial Intelligence (DFKI)
[2] University of Kaiserslautern-Landau (RPTU), Germany

Abstract. Human Activity Recognition (HAR) using wearable sensors has traditionally relied on inertial measurement units (IMUs), which capture body motion efficiently but provide limited information about the surrounding environment. In this work, we introduce *FootSense*, a foot-centric multimodal sensing dataset that combines a downward-facing shoe-mounted camera with three body-worn IMUs to jointly recognize locomotion activities and terrain types in unconstrained real-world environments. The foot-mounted viewpoint naturally emphasizes ground-level context while reducing the capture of faces and surrounding social interactions compared with conventional wearable cameras. Furthermore, the foot is a promising location for biomechanical energy harvesting, offering long-term potential for self-powered wearable vision systems. The dataset contains synchronized multimodal recordings from eight participants performing nine locomotion activities across eight terrain types during continuous everyday movements. As a baseline, we present a lightweight multimodal recognition framework that fuses visual and inertial streams for activity and terrain classification. Experimental results demonstrate that foot-mounted visual information substantially improves terrain recognition while complementing inertial sensing for activity recognition. These findings suggest that foot-centric multimodal sensing is a promising direction for terrain-aware HAR and provides a foundation for future investigations of energy-efficient wearable sensing systems. We will make the dataset available on acceptance [3].

Keywords: human activity recognition · wearable sensors · terrain classification

1 Introduction

Human Activity Recognition (HAR) has become a fundamental research area in ubiquitous computing, enabling applications in health monitoring, rehabilitation, sports analytics, occupational safety, and context-aware interaction. Wearable sensing systems based on inertial measurement units (IMUs) have become the

[3] See this link for video sample https://anonmp4.help/v/UvKEoiQdlNX92ZL

dominant approach because they are lightweight, energy efficient, inexpensive, and comparatively privacy preserving while providing reliable measurements of human motion [1, 2].

Despite their success, IMU-based systems primarily capture body motion and provide only limited information about the surrounding environment. However, environmental context is particularly important for locomotion, where identical movements may occur on substantially different surfaces such as asphalt, grass, gravel, or stairs. Terrain awareness has long been recognized in robotics for adaptive locomotion [10], yet remains largely unexplored in wearable HAR datasets and benchmarks.

One of the natural way to capture this missing context is through vision. Conventional wearable cameras, however, often record faces, bystanders, and surrounding scenes, raising privacy concerns in everyday deployments [16]. In contrast, a downward-facing camera mounted on the shoe naturally focuses on the walking surface while reducing the likelihood of capturing faces and surrounding social interactions, as illustrated in Figure 1. This viewpoint therefore complements inertial sensing by providing localized environmental information while inherently limiting the observation of sensitive visual content as shown in the figure 2. Furthermore, the foot is a promising location for biomechanical energy harvesting, offering the potential for future self-powered wearable vision systems [19, 18, 17].

Fig. 1. The foot-mounted camera captures localized ground context for terrain-aware activity recognition while naturally reducing the capture of surrounding people and scenes.

Motivated by these observations, we introduce **FootSense**, a foot-centric multimodal sensing dataset consisting of synchronized recordings from a downward-facing shoe-mounted camera and three body-worn IMUs (wrist, pocket, and ankle). The dataset contains recordings from eight participants performing nine locomotion activities across eight terrain types in unconstrained real-world environments. The body-worn IMUs capture complementary body dynamics, while

the foot-mounted camera provides localized environmental context, enabling joint reasoning over human motion and terrain characteristics. In addition, we present a lightweight multimodal baseline that fuses visual and inertial information for activity and terrain recognition.

Our contributions are threefold:

- We introduce a publicly available foot-centric multimodal dataset comprising synchronized shoe-mounted video and body-worn IMUs with activity and terrain annotations collected in natural environments.
- We demonstrate that downward-facing foot-mounted vision provides complementary environmental information for terrain-aware activity recognition while offering a naturally privacy-reduced sensing perspective.
- We establish baseline results for multimodal activity and terrain recognition and discuss the implications of foot-mounted sensing for future energy-efficient wearable systems.

2 Related Work

Wearable Human Activity Recognition. Wearable HAR has been extensively studied using inertial sensing, leading to numerous public datasets and benchmark models [3, 24]. However, many datasets are collected under scripted laboratory protocols or predefined outdoor routes [4, 5]. Recent multimodal datasets such as WEAR [7] and MM-Fit [8] incorporate additional sensing modalities but provide limited diversity in terrain and everyday environmental context.

Vision-based Wearable Sensing and Privacy. Vision-based HAR has achieved remarkable recognition performance through deep spatiotemporal models trained on large-scale video datasets [11, 20]. To mitigate the privacy concerns associated with continuous wearable vision, prior work has explored Privacy-by-Design principles [15, 14], privacy-preserving sensing pipelines, and abstract visual representations such as human pose estimation [13]. These approaches primarily reduce sensitive information after or during visual processing, whereas our work investigates whether the camera placement itself can naturally reduce the capture of sensitive content.

Foot-Centric Wearable Sensing. Foot-mounted wearable systems have been explored for gait analysis, locomotion monitoring, and terrain sensing by leveraging the unique interaction between the foot and the ground. Early work [26] demonstrated joint recognition of gait and ground type using wearable capacitive sensors mounted around the ankle. Later, CapSoles [27] proposed a smart insole that exploits capacitive ground coupling to recognize different floor materials during walking. More recently, LaserShoes [25] employed laser speckle imaging to accurately classify indoor and outdoor terrains using a lightweight footwear-mounted sensing system. These studies demonstrate the value of foot-centric sensing for terrain understanding and locomotion assistance, but primarily investigate specialized sensing modalities or focus on terrain recognition alone. In

contrast, existing work has largely focused on sensing systems rather than publicly available multimodal datasets. FootSense extends this line of research by providing a publicly available benchmark that combines downward-facing foot-mounted video with synchronized body-worn IMUs for joint activity and terrain recognition in unconstrained real-world environments.

3 Methodology

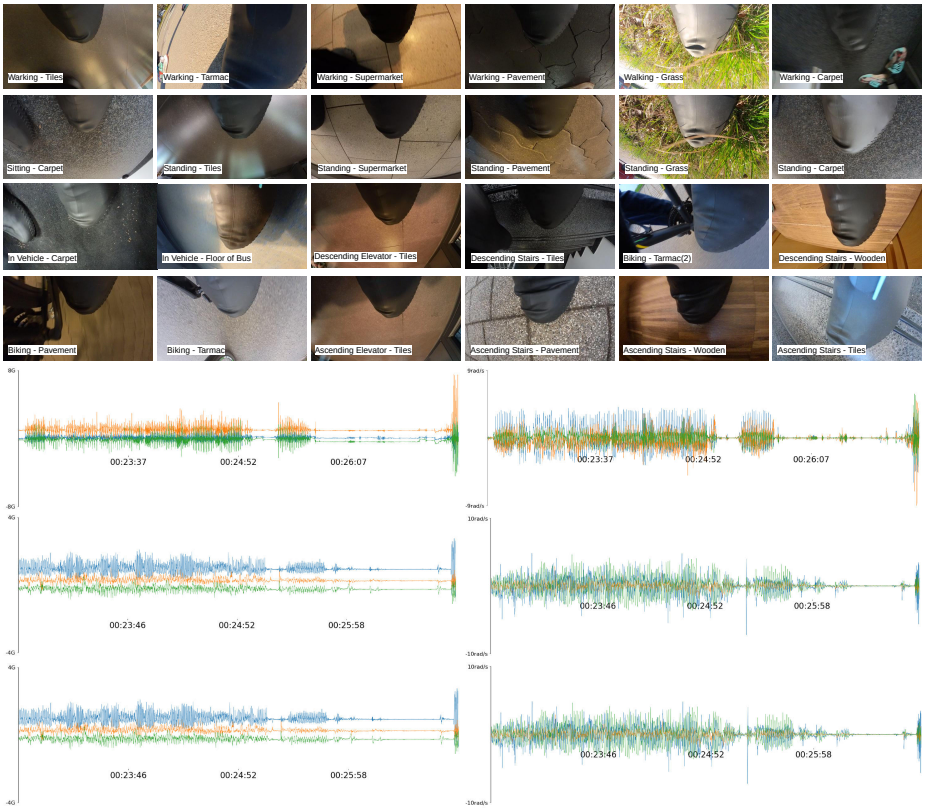

Fig. 2. Synchronized image sequences and target IMU waveforms across the distinct tracking activities performed at different types of terrain.

3.1 Dataset Description

The FootSense dataset combines synchronized inertial and visual data acquired from three body-worn IMUs and a shoe-mounted camera. Two Apple smartwatches were used to capture inertial measurements from the wrist (Apple Watch

Series 7) and ankle (Apple Watch Series 6), while an iPhone 12 placed in the dominant trouser pocket recorded the third IMU stream. All inertial sensors sampled tri-axial accelerometer and gyroscope signals at 50 Hz.

Visual data were acquired using an Insta360 GO 2 camera (52.9 mm × 23.6 mm × 20.7 mm) mounted on the upper shoe instep and oriented downward toward the walking surface. The camera recorded RGB video at a resolution of 1920 × 1080 pixels and 24 FPS, providing a localized view of the terrain while limiting the capture of surrounding people and scenes.

Eight participants ($\mu_{\text{age}} = 27.9$, SD= 3.54; height 167–188 cm; weight 53–86 kg) performed continuous recordings in unconstrained indoor and outdoor environments. Rather than following scripted laboratory routines, participants were instructed to reach nearby destinations such as supermarket, office or bus stations while naturally performing daily locomotion activities such as walking, cycling, stair traversal, elevator use, public transportation, and standing. This protocol captured realistic transitions between activities and terrain types that commonly occur in everyday mobility. The resulting dataset contains 3.8 hours of synchronized multimodal recordings covering nine activity classes and eight terrain categories.

Each sensing device stored its measurements locally throughout the recording session. To synchronize the modalities, a distinctive synchronization event (jumping) was performed multiple times at both the beginning and the end of every recording. These temporal markers were manually aligned offline for the video and each IMU, after which all inertial signals were resampled onto a common 50 Hz timeline to compensate for small clock drifts and differences in sampling frequency. After the video stream and synchronized IMUs were aligned, the data was ready for annotation and later model training.

The combination in Table 1 reflects the association between activities and their environmental contexts, where locomotion-related activities such as Walking and Standing span multiple surface types, while transition-related activities (e.g., Ascending/ Descending Stairs and Elevators) are observed in more constrained environmental settings.

Activity and terrain labels were obtained through frame-by-frame inspection by two independent annotators. Video frames were spatially downsampled to 480p to reduce computational cost while preserving sufficient visual detail for recognition, following common practice in wearable HAR datasets [7].

3.2 Baseline Recognition Framework

To establish baseline performance on the FootSense dataset, we implement a lightweight multimodal recognition framework that combines visual and inertial information (Figure 3). The framework jointly models visual and inertial information through complementary processing branches designed to capture both spatial-temporal motion patterns and fine-grained sensor dynamics.

The input to the system consists of a synchronized 2-second temporal window, where the video sequence is downsampled to 13 frames and paired with the corresponding IMU data ($100 \times C$ channels). The visual branch processes

Activity(%)	**Environment / Surface**
Sitting(3.6)	Carpet
Standing(18.0)	Carpet, Tiles, Market Tiles, Pavement, Grass
Walking(48.8)	Carpet, Tiles, Market Tiles, Pavement, Tarmac, Grass
Stairs Down(4.8)	Tiles, Wooden
Elevator Down(1.7)	Tarmac
Stairs Up(1.7)	Pavement, Tiles, Wooden
Elevator Up(9.7)	Tarmac
In Vehicle(9.7)	Bus Floor (Rubber/Vinyl), Carpet
Biking(5.7)	Tarmac, Pavement

Table 1. Recorded activities and associated environmental conditions used in the dataset. Floor types(%) are listed: Carpet (19.1), Tiles (21.9), Pavement (33.9), Tarmac (5.2), Grass (3.9), Market Tiles (9.5), Wooden (5.2), Bus Floor (1.2)

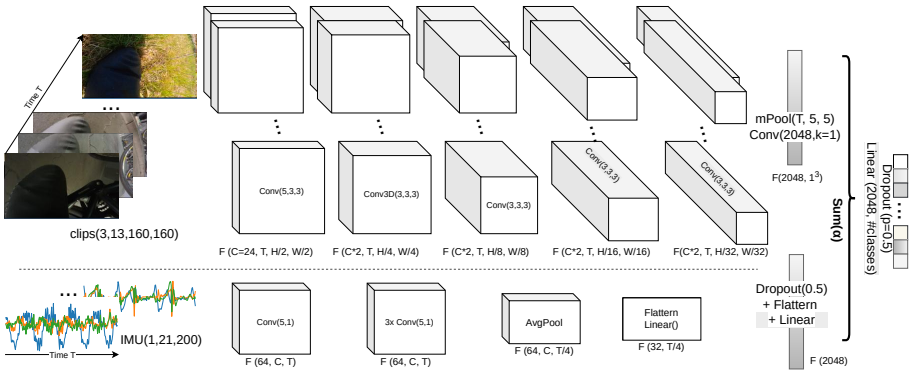

Fig. 3. Our baseline Structural Block Diagram.

short RGB clips consisting of 13 consecutive frames using a pretrained X3D encoder [20]. The resulting spatio-temporal representation is combined with motion features extracted from synchronized IMU sequences by a lightweight one-dimensional convolutional encoder. In parallel, the inertial branch, denoted as IMU-Encoder, processes multichannel IMU sequences using four consecutive convolutional layers with ReLU activations. The network applies convolutions over the temporal dimension while preserving sensor-channel structure, followed by average pooling for temporal downsampling. The extracted features are flattened and projected through two fully connected layers to produce a compact motion embedding, after which dropout regularization is applied. The visual and inertial embeddings are fused through a learnable weighted combination before being passed to a fully connected classification layer for activity or terrain prediction.

4 Evaluation and Results

Models were optimized using Cross-Entropy loss via an Adam optimizer (LR = 5×10^{-5}) for 100 epochs. Continuous sequences were segmented using a sliding window of 2 seconds [23] with overlap of 50%. Splitting configurations followed a fixed subject-independent breakdown (5 subjects for training, 3 for testing). To prevent data leakage, raw streams were normalized using exclusively training data means and standard deviations. The pipeline and models were implemented in PyTorch 1.8 running on a dedicated NVIDIA RTX 3090 Ti GPU.

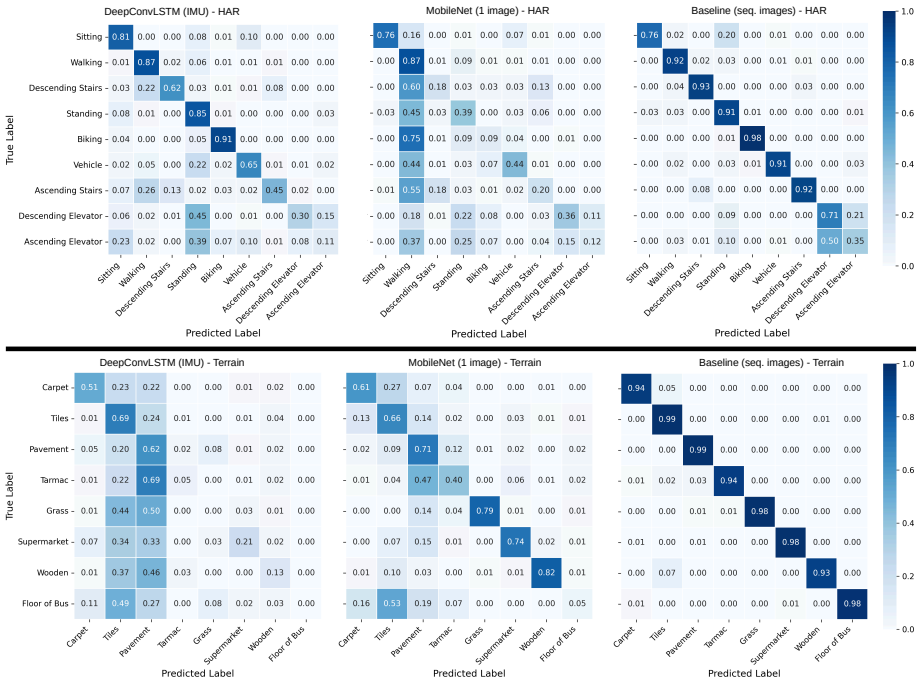

Fig. 4. Confusion Matrices for predicting HAR (top) and terrain (bottom) from a single modality. DC-LSTM takes as input only IMU data while MobileNet takes a single image.

Human Activity Recognition Task As shown in Table 2, traditional standalone IMU models (DC-LSTM [29] and TinyHAR [28]) show stable base recognition accuracy but struggle with micro-activities. This can be seen in the top row of Figure 4, where we see degradation of performance in classes related to using the elevator, which can be easily confused with standing given the short window size. Regarding MobileNet, which takes as input a single image, it can only reliably predict walking and sitting. On the other hand, DC-LSTM has a better

overall performance, although bellow our multi-modal framework (see Figure 5). Our baseline processing video clip sequences substantially outperforms its single image baseline (MobileNet), achieving a performance of F_1-Macro = 0.815 when unified with inertial features, versus 0.782 when using only the image sequences. In this case adding IMU helps in reducing the confusion between sitting versus standing and "elevator up" and "elevator down". Figure 5 shows the classes that remain challenging, where we see some confusion between "standing" and elevator and Vehicle classes. Note that the Vehicle class includes sitting in the car or standing in the bus, so in our selected window size it is hard for the model to differentiate between "standing" in a room and standing in the bus. Other classes with higher confusion are ascending and descending the elevator, which are similar in the image and IMU domain, with changes only at the start and end of the activity in the IMU readings. Thus, for better performance, we would probably need bigger window sizes and a network with better temporal modeling than our baseline one, especially in the IMU branch.

Table 2. Performance (mean ± std) on activity classes.

Task: HAR	Method	F1Macro	Accuracy
IMUs	DC-LSTM	0.584 ± 0.027	0.782 ± 0.024
	TinyHAR	0.571 ± 0.070	0.759 ± 0.027
One image	MobileNet	0.389 ± 0.024	0.609 ± 0.012
IMUs + One image	Our Baseline	0.613 ± 0.020	0.796 ± 0.006
Image Seq. (Clip)	Our Baseline	0.782 ± 0.018	0.895 ± 0.011
IMUs + Clip	Our Baseline	0.815 ± 0.012	0.899 ± 0.011

Terrain/Flooring Recognition Task Table 3 shows the terrain recognition performance. Inertial classifiers struggle to differentiate textures purely from tactile gait rhythms ($\sim$ 46% accuracy). A single image (one frame from 2-seconds clip) outperforms IMUs significantly. In fact, IMU data only confuses the network if it only has access to a single image. This can be seen in the bottom row of Figure 4, where MobileNet, which takes as input a single image, outperforms substantially DC-LSTM, which can only predict, to some extent, the differences between carpet, tiles and pavement. On the other hand, sequences of images provide more accurate results and we even see a small improvement when IMU data is added, as we can see in the confusion matrix of Figure 6. Our multi-modal model has at least 92% accuracy for all classes, and can even differentiate between "Carpet" and "Floor of Bus", even though the latter constitutes only 1.2% of samples (see caption of Table 1). More importantly, including IMUs reduces the variance in performance from 4.9% to 0.3%, stabilizing predictions.

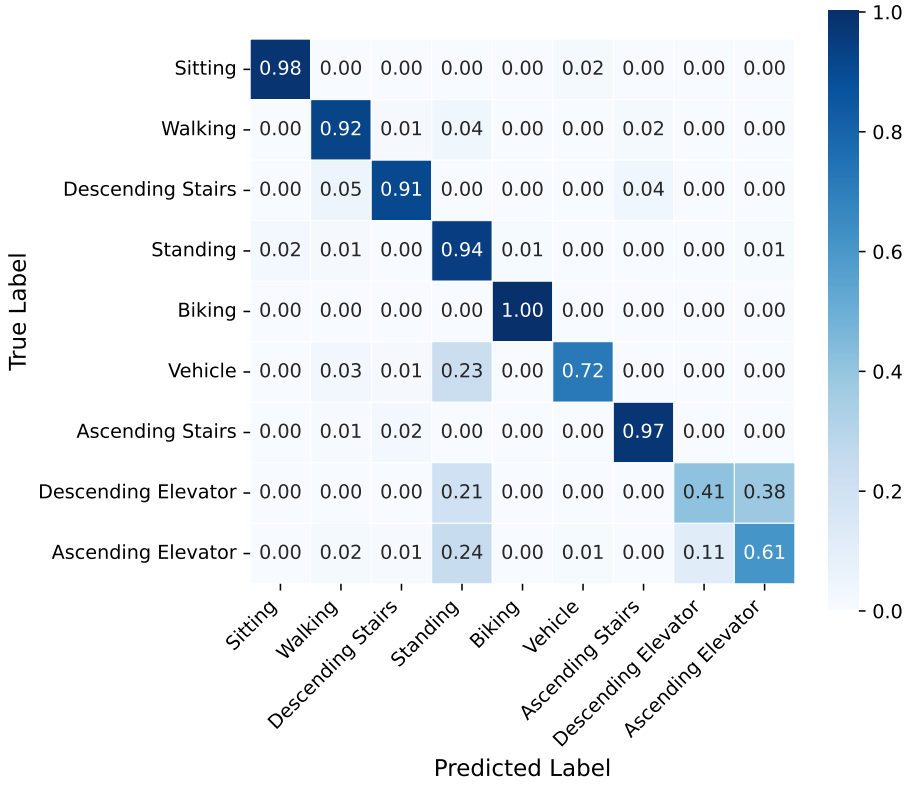

Fig. 5. Confusion matrix for activity predictions for 2-seconds window segments using IMU data and a sequence of 13 images. Results with our baseline network depicted in Figure 3.

5 Discussion & Energy Perspective

A key hindrance for wearable vision deployment is energy consumption. Standard wearable placements (e.g., wrist, chest) rarely see integration with dynamic kinetic harvesting structures due to high energy transport overheads across the body anatomy [24]. However, biomechanical energy harvesting at the shoe base represents the highest power output yield position on the human body during locomotion [19, 18, 17]. Co-locating our ultra-low-power vision sensor on the foot may allow the system to tap directly into localized power loops, avoiding body-spanning wiring networks and bringing self-powered egocentric vision closer to reality. This energy perspective serves as a design motivation for FootSense; a quantitative analysis of power consumption and harvesting feasibility is deliberately left to future work, for which the present dataset and baseline provide a foundation on the expected performance of both HAR and terrain recognition in this sensor placement.

Table 3. Performance on flooring/ground classification.

Task: Floor	Method	F1Macro	Accuracy
IMUs	DC-LSTM	0.269 ± 0.022	0.469 ± 0.016
	TinyHAR	0.280 ± 0.021	0.463 ± 0.013
One image	MobileNet	0.602 ± 0.030	0.663 ± 0.013
IMUs + One image	Our Baseline	0.514 ± 0.046	0.631 ± 0.036
Images Seq.(Clip)	Our Baseline	0.970 ± 0.049	0.975 ± 0.005
IMUs + Clip	Our Baseline	0.971 ± 0.003	0.977 ± 0.001

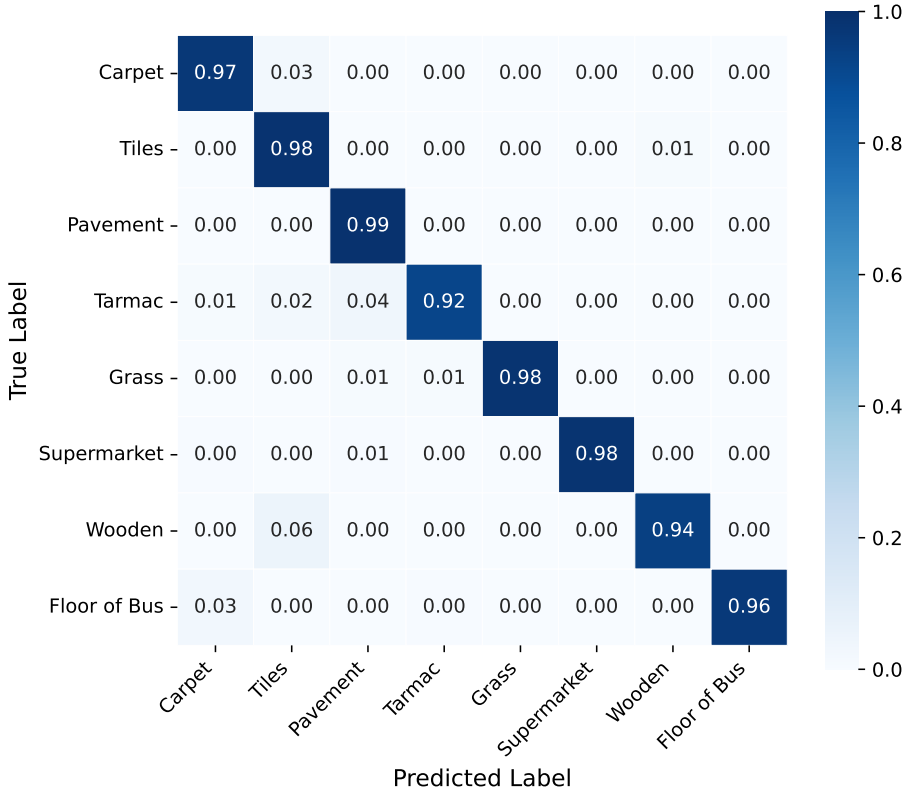

Fig. 6. Confusion matrix for flooring type predictions for 0.5 window segments using IMU data and a sequence of 13 images. Results with our baseline network depicted in Figure 3.

6 Limitations and Future Work

While our downward orientation reduces privacy, extreme angular foot movements can occasionally capture off-ground context info. To remedy this, we intend to implement a mechanical stabilization gimbal to ensure a consistent ground vector. Furthermore, future investigations will focus on the cross-modal alignment between unconstrained body cameras and structural foot systems to explore complementary environmental modelling. Future work includes power consumption analysis for the different modalities as well as direct energy harvesting implementation. We also plan to experiment with other low power architectures including spiking networks.

7 Conclusion

This paper presented a privacy-aware, foot-centric HAR framework and dataset motivated by energy harvesting potential and privacy constraints. We include HAR and Terrain classification as tasks and show that our proposed baseline model achieves high accuracy on human activity tracking while establishing robust context identification on challenging terrain variants. We hope our multimodal dataset and benchmark can support further research in low power, realistic and privacy-preserving HAR and context awareness in general.

References

1. M. Ciliberto, V. Fortes Rey, A. Calatroni, P. Lukowicz, and D. Roggen, "Opportunity++: A multimodal dataset for video-and wearable, object and ambient sensors-based human activity recognition," *Frontiers in Computer Science*, vol. 3, p. 792065, 2021.
2. M. Ciliberto, D. Roggen, and F. J. Ordóñez Morales, "Exploring human activity annotation using a privacy preserving 3D model," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*, Heidelberg, Germany, 2016, pp. 803–812. doi: 10.1145/2968219.2968290.
3. T. S. Qureshi, M. H. Shahid, A. A. Farhan, and S. Alamri, "A systematic literature review on human activity recognition using smart devices: advances, challenges, and future directions," *Artificial Intelligence Review*, vol. 58, no. 9, p. 276, 2025. doi: 10.1007/s10462-025-11275-x.
4. L. Pogrzeba, E. Muschter, S. Hanisch, V. Y. P. Wardhani, T. Strufe, F. H. P. Fitzek, and S.-C. Li, "A Full-Body IMU-Based Motion Dataset of Daily Tasks by Older and Younger Adults," *Scientific Data*, vol. 12, no. 1, p. 531, 2025. doi: 10.1038/s41597-025-04818-y.
5. M. Palermo, S. M. Cerqueira, J. André, A. Pereira, and C. P. Santos, "From raw measurements to human pose – a dataset with low-cost and high-end inertial-magnetic sensor data," *Scientific Data*, vol. 9, no. 1, p. 591, 2022. doi: 10.1038/s41597-022-01690-y.

6. T. M. Wiles, S. K. Kim, M. Mangalam, J. H. Sommerfeld, K. J. Brink, A. Grunkemeyer, M. K. Manifrenti, A. E. Charles, N. Shakerian, M. Haghighatnejad, *et al.*, "NONAN GaitPrint: An IMU gait database of healthy older adults," *Scientific Data*, vol. 12, no. 1, p. 143, 2025. doi: 10.1038/s41597-024-04359-w.
7. M. Bock, H. Kuehne, K. Van Laerhoven, and M. Moeller, "WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)*, vol. 8, no. 4, art. 175, 2024. doi: 10.1145/3699776.
8. D. Strömbäck, S. Huang, and V. Radu, "MM-Fit: Multimodal Deep Learning for Automatic Exercise Logging across Sensing Devices," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 4, art. 168, 2020. doi: 10.1145/3432701.
9. V. T. Lee, T. Alan, S. Kim, E. Ustun, A. Suleiman, A. Krisshna, T. Balbekov, A. Alaghi, and R. Newcombe, "Full System Architecture Modeling for Wearable Egocentric Contextual AI," *arXiv preprint arXiv:2512.16045*, 2025.
10. R. González, F. Rodríguez, and J. L. Guzmán, "Autonomous tracked robots in planar off-road conditions," *Modeling, Localization, and Motion Control*, Springer, Berlin, Germany, 2014.
11. J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the Kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.
12. N. Pervin, T. F. Sanam, and H. Imtiaz, "Privacy-preserving human activity recognition using principal component-based wavelet CNN," *Signal, Image and Video Processing*, vol. 18, no. 12, pp. 9141–9155, 2024.
13. C. Hinojosa, J. C. Niebles, and H. Arguello, "Learning privacy-preserving optics for human pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2573–2582.
14. M. Carvalho, O. Ennaffi, S. Chateau, and S. A. Bachir, "On the design of privacy-aware cameras: A study on deep neural networks," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 223–237.
15. A. Cavoukian, "Privacy by design: The definitive workshop. A foreword by Ann Cavoukian, Ph.D.," *Identity in the Information Society*, vol. 3, no. 2, pp. 247–251, 2010.
16. Y. Yang, P. Hu, J. Shen, H. Cheng, Z. An, and X. Liu, "Privacy-preserving human activity sensing: A survey," *High-Confidence Computing*, vol. 4, no. 1, p. 100204, 2024.
17. F. Qian, T.-B. Xu, and L. Zuo, "Design, optimization, modeling and testing of a piezoelectric footwear energy harvester," *Energy Conversion and Management*, vol. 171, pp. 1352–1364, 2018.
18. J. Zhao and Z. You, "A shoe-embedded piezoelectric energy harvester for wearable sensors," *Sensors*, vol. 14, no. 7, pp. 12497–12510, 2014.
19. G. Xu, Y. Yang, Y. Zhou, and J. Liu, "Wearable thermal energy harvester powered by human foot," *Frontiers in Energy*, vol. 7, no. 1, pp. 26–38, 2013. doi: 10.1007/s11708-012-0215-9.
20. C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 203–213.
21. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114, PMLR.
22. S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

23. M. Bock, A. Hölzemann, M. Moeller, and K. Van Laerhoven, "Improving Deep Learning for HAR with Shallow LSTMs," in *Proceedings of the 2021 ACM International Symposium on Wearable Computers (ISWC '21)*, Virtual, USA, 2021, pp. 7–12. doi: 10.1145/3460421.3480419.
24. C. Beach and A. J. Casson, "Inertial Kinetic Energy Harvesters for Wearables: The Benefits of Energy Harvesting at the Foot," *IEEE Access*, vol. 8, pp. 208136–208148, 2020. doi: 10.1109/ACCESS.2020.3037952.
25. Z. Yan, Y. Lin, G. Wang, Y. Cai, P. Cao, H. Mi, and Y. Zhang, "LaserShoes: Low-Cost Ground Surface Detection Using Laser Speckle Imaging," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg, Germany, 2023, Article 853, pp. 1–20. DOI: 10.1145/3544548.3581344.
26. J. Cheng, O. Amft, G. Bahle, and P. Lukowicz, "Designing Sensitive Wearable Capacitive Sensors for Activity Recognition," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3935–3947, 2013. doi: 10.1109/JSEN.2013.2259693.
27. D. J. C. Matthies, T. Roumen, A. Kuijper, and B. Urban, "CapSoles: Who Is Walking on What Kind of Floor?" in *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*, Vienna, Austria, 2017, Article 9, pp. 1–14. doi: 10.1145/3098279.3098545.
28. Y. Zhou, H. Zhao, Y. Huang, T. Riedel, M. Hefenbrock, and M. Beigl, "Tinyhar: A lightweight deep learning model designed for human activity recognition," in *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, 2022, pp. 89–93.
29. F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.