Title: Robustness of Vision-Based Human Activity Recognition under Naturalistic Distribution Shifts
Keywords: robustness, out-of-distribution generalization, human activity recognition, multimodal fusion

Robustness to distribution shifts remains a challenge for vision-based recognition. While documented in video classification [1,2], robustness evaluations in human activity recognition (HAR) are often limited to controlled benchmarks or synthetic corruptions. Less is known about robustness to natural variations in environment, embodiment, and camera viewpoint, or its relation to task [3].

We study robustness using a multimodal dataset of daily activities, capturing unscripted recordings across multiple environments, body positions, and camera views [5]. The dataset provides natural shifts that approximate out-of-distribution conditions. We evaluate current vision models across three scenarios: cross-view, cross-body, and cross-environment recognition. Performance is analyzed at hierarchical annotation levels from coarse (L1) to fine-grained (L3), per recent video activity benchmarks [4].

Vision-based models show high accuracy in matched domains but degrade severely under distribution shifts. A ResNet baseline, for example, drops from 89% to 15% accuracy in cross-view evaluation (Figure 1). Depth and RGB modalities show different robustness at fine-grained levels, suggesting modality sensitivity is linked to activity granularity. Wearable sensing modalities (IMU, EMG, insole) show more stable performance across shifts (Figure 2). Multimodal fusion improves accuracy and reduces variability over single modalities.

Robustness failures in HAR are thus influenced by domain shifts, task granularity, and sensing modality. Although based on a single dataset, its multi-environment and multi-view structure presents challenges representative of real-world deployment. These results underscore the need for multimodal and hierarchical benchmarks when developing robust HAR methods.

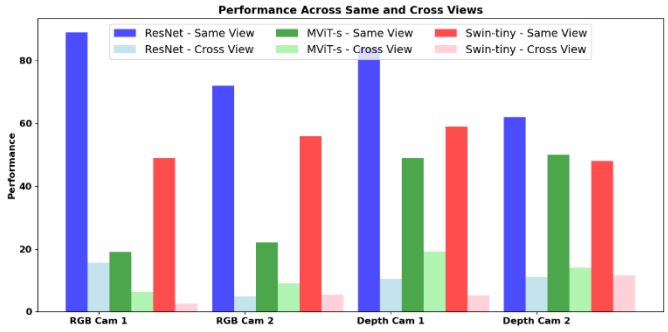

Figure 1: Accuracy of vision models across same-view and cross-view evaluation on RGB and depth cameras. Vision performance drops sharply under cross-view conditions.

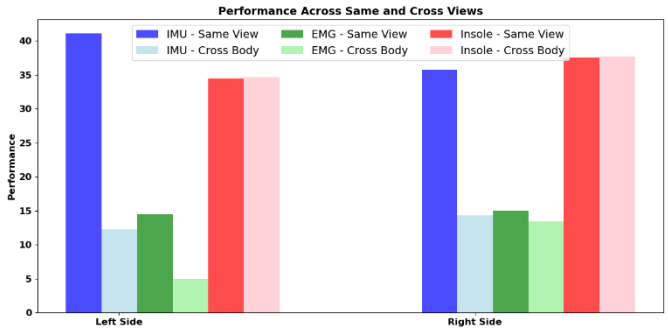

Figure 2: Accuracy of wearable modalities across same-side and cross-body evaluation. Wearables exhibit more stable performance compared to vision models under distribution shifts.

References:

[1] Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., & Steinhardt, J. (2022). PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16762–16771.

[2] Hara, K., Kataoka, H., & Satoh, Y. (2018). Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6546–6555.

[3] Wang, C., & Yan, J. (2023). A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition. *IEEE Access, 11*, 53880–53898. DOI:10.1109/ACCESS.2023.3282311.

[4] Xu, J., Zhao, G., Yin, S., Zhou, W., & Peng, Y. (2024). FineSports: A Multi-Person Hierarchical Sports Video Dataset for Fine-Grained Action Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21773–21782.

[5] Kaviani, G., Yarici, Y., Kim, S., Prabhushankar, M., AlRegib, G., Solh, M., & Patil, A. (2025). *Hierarchical and multimodal data for daily activity understanding*. arXiv preprint arXiv:2504.17696.