Efficient Egocentric Action Recognition with Multimodal Data

Marco Calzavara
ETH Zurich

Ard Kastrati
ETH Zurich

Matteo Macchini
Magic Leap

Dushan Vasilevski
Magic Leap

Roger Wattenhofer
ETH Zurich

Abstract

The increasing availability of wearable XR devices opens new perspectives for Egocentric Action Recognition (EAR) systems, which can provide deeper human understanding and situation awareness. However, deploying real-time algorithms on these devices can be challenging due to the inherent trade-offs between portability, battery life, and computational resources. In this work, we systematically analyze the impact of sampling frequency across different input modalities—RGB video and 3D hand pose—on egocentric action recognition performance and CPU usage. By exploring a range of configurations, we provide a comprehensive characterization of the trade-offs between accuracy and computational efficiency. Our findings reveal that reducing the sampling rate of RGB frames, when complemented with higher-frequency 3D hand pose input, can preserve high accuracy while significantly lowering CPU demands. Notably, we observe up to a 3× reduction in CPU usage with minimal to no loss in recognition performance. This highlights the potential of multimodal input strategies as a viable approach to achieving efficient, real-time EAR on XR devices.

1. Introduction

Since the advent of head-mounted devices, researchers and industry have shown growing interest in leveraging video captured by wearable cameras to extract valuable information. Among these devices, AR glasses such as Magic Leap 2[1] stand out by integrating digital content into the physical world. The data they capture can unlock numerous applications, from task assistance to personalized recommendations and contextual awareness.

Egocentric Action Recognition (EAR) is a research field focused on identifying human actions from first-person data. It presents unique challenges, including variable viewpoints, frequent occlusions, and the computational constraints of wearable devices. On AR glasses, efficiency is especially critical, not only to preserve battery life but also because these devices typically lack dedicated hardware acceleration and must rely on limited CPU resources. This project tackles these challenges by targeting efficient and accurate real-time action recognition for AR glasses.

EAR methods are typically categorized as either uni-modal or multi-modal. Several studies rely solely on RGB data for EAR. For example, [9] adopts a Bag-of-Objects approach, while [6] uses frame-level features derived from object detections. Other methods extract features directly from RGB frames. Among these, [15] combines 2D and 3D CNNs with a ConvLSTM [12], also testing frame differences instead of raw frames. Although RGB-based systems perform well, some studies explore whether hand pose alone is sufficient. For instance, [1] uses a Spatio-Temporal Graph Convolutional Network on hand skeletons, which is inspired by [18]. These results highlight a strong link between hand motion and egocentric actions.

The main drawback of uni-modal systems is their inability to capture diverse cues, such as object identity or hand-object interactions. Multi-modal models address this limitation by combining multiple inputs. A popular example is the two-stream architecture from [13], which processes RGB and optical flow separately. In EAR, gaze-based attention is used in [8], while [16] shows improved performance by adding an optical flow branch to [15]. Beyond RGB and flow, other modalities have also been explored: [14] combines RGB and IMU, while [11] uses RGB with 3D hand poses, both avoiding the high computational cost of optical flow. Three-stream models extend this further: [3] combines RGB, optical flow, and object features, showing better results than any single stream. However, as noted in [17], more modalities do not always lead to better performance.

Building upon previous work, our approach investigates a targeted two-stream architecture that utilizes RGB frames and 3D hand pose keypoints to predict the current action, aiming to optimize both accuracy and resource efficiency. This paper makes the following contributions:

[1] https://www.magicleap.com/magic-leap-2

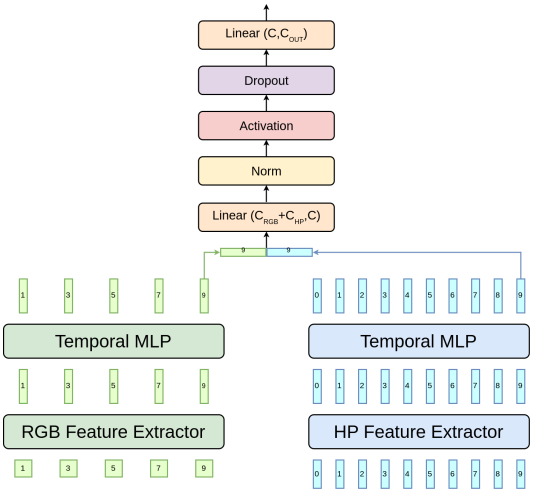
- We introduce a multi-modal EAR system that combines an RGB stream comprising a Vision Transformer (ViT) feature extractor with a hand pose stream.
- We present a systematic study of the interplay between sampling frequency, recognition accuracy, and computational cost across modalities. Our analysis reveals how adjusting the relative sampling rates of RGB and hand pose inputs can serve as a design lever to optimize EAR systems for resource-constrained devices.
- We demonstrate that downsampling RGB input — when complemented with higher-frequency hand pose signals — preserves accuracy while significantly improving efficiency. Specifically, we show that maintaining hand pose input at 30 Hz while reducing RGB input to 10 Hz achieves competitive accuracy with up to a 3× reduction in CPU usage compared to full-frame-rate RGB baselines. These results highlight the practical value of modality-aware sampling strategies for efficient on-device deployment of EAR systems.

2. Methods

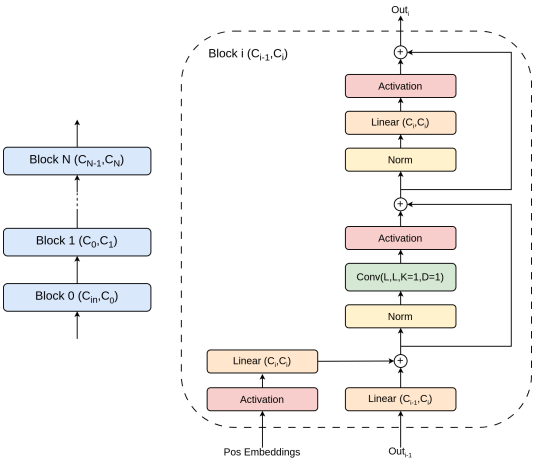
2.1. Model

Operating over sequences is crucial for accurate action recognition, as context from multiple consecutive time steps helps detect motion. In this work, we employ a multi-stream architecture consisting of modality-specific feature extractors and sequence models. The model, depicted in Figure 1a, processes the two modalities through dedicated modules: a LeViT [4] feature extractor and Temporal MLP for RGB features and a hand pose feature extractor paired with a Temporal MLP for hand pose features. The outputs from both streams, corresponding to the final time steps, are concatenated and processed through a series of layers for classification.

A key component of our approach is the Temporal MLP, which is responsible for modeling temporal dependencies in the feature representations. The architecture of the Temporal MLP, shown in Figure 1b, is inspired by [2] and is designed to efficiently capture long-range dependencies without relying on convolutional layers with large kernel sizes.



(a) Architecture of the Multimodal Temporal MLP. The model consists of two parallel processing streams: one for RGB features, utilizing a LeViT feature extractor and a Temporal MLP, and another for hand pose features, employing an MLP feature extractor paired with a Temporal MLP. The outputs from both streams at the final time step are concatenated and processed through additional layers for classification.



(b) Architecture of the Temporal MLP. This module captures long-range dependencies in temporal sequences without relying on convolutional layers with large kernel sizes. Inspired by [2], it employs MLP-based operations to model temporal relationships efficiently, making it well-suited for sequence-based tasks such as action recognition.

Figure 1. Architectures of the Multimodal Temporal MLP and Temporal MLP.

2.2. Experimental Conditions

Model training was performed using either two or four GPUs, depending on availability. The GPUs used for training included the Titan RTX (24 GB memory), GeForce RTX 3090 (24 GB memory), and Tesla V100 (32 GB memory). For CPU usage, a single thread of an AMD EPYC 7742 CPU (base clock: 2.25 GHz) was utilized.

2.3. Dataset

We conducted our experiments on the training and validation sets of the H2O dataset [7]. While the H2O dataset is designed for predicting the action performed over a video segment, it can also be used for frame-level prediction by assigning the segment's action label to each frame within the segment.

2.4. Preprocessing and Augmentation

RGB frames are cropped and normalized prior to feature extraction. For 3D hand keypoints, we apply a three-step normalization to ensure spatial consistency: (1) translating the wrists to the origin, (2) standardizing edge lengths be-

tween keypoints, and (3) rotating the hands to align specific vectors with canonical axes. Beyond preprocessing, we apply several data augmentation techniques. When both RGB frames and 3D hand pose keypoints are used, only shared augmentations are applied to preserve cross-modal consistency.

3. Results

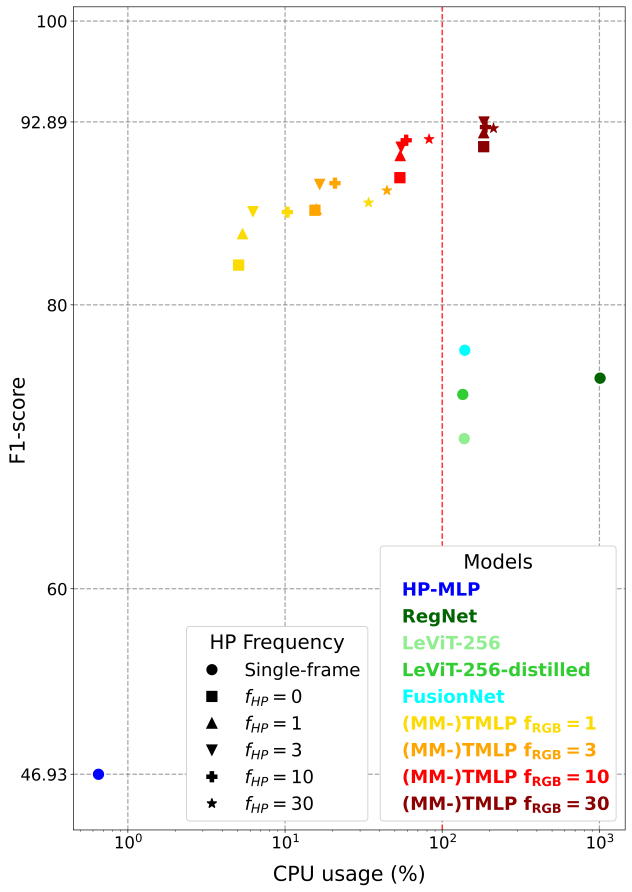


Figure 2. Relationship between F1-score and CPU usage (log scale) for different action prediction models. All the proposed sequence models outperform the single-frame models and highlight how reducing the RGB frequency significantly lowers the CPU usage with limited impact on performance, enabling adaptable design choices for different scenarios. CPU usage is estimated by running inference over a one-second input window.

We present a combined analysis of model performance and CPU usage. The results are displayed in Figure 2. Reported values refer to the macro F1-score. All sequence models were trained on 2-second sequences, corresponding to 60 time steps.

3.1. Single-frame models

We evaluated the performance of single-frame models. While RegNet [10] achieved a higher F1-score than LeViT-256, its higher CPU usage makes it less suitable for resource-limited environments.

Distillation [5] helped narrow the performance gap between LeViT-256 and RegNet-12GF while maintaining the same resource usage as the original LeViT-256. Therefore, we selected the LeViT-256-distilled model as the RGB feature extractor.

Additionally, we experimented with an MLP model referred to as HP-MLP. While effective for verb classification, its lower action prediction F1-score suggests that hand pose data lacks object-related cues. Here, verb prediction refers to identifying the general action type (e.g., "grab book" and "grab cappuccino" are both classified as "grab"). Despite its limitations, HP-MLP's low CPU usage makes it ideal for practical deployment.

Lastly, we explored single-frame fusion using the FusionNet model, which combines the outputs of the two feature extractors. FusionNet outperformed LeViT-256-distilled, highlighting the complementary nature of these modalities in the single-frame setting.

3.2. RGB-only Sequence Models

We experimented with sequence models processing RGB frames (square markers in Figure 2). We observe a steady F1-score decline from the model with $f_{\text{RGB}} = 30$ Hz to the model with $f_{\text{RGB}} = 1$ Hz. However, CPU usage drops more significantly, roughly by a factor of three with each threefold reduction in frequency. These results indicate that RGB frequency can be adjusted to fit resource constraints with minimal performance loss.

3.3. Multimodal Sequence Models

We evaluated MM-TMLP models across various combinations of RGB and hand pose sampling frequencies. A comparison with RGB sequence models shows that, for a fixed f_{RGB}, incorporating the hand pose stream consistently improves performance, regardless of the hand pose sampling frequency f_{HP}.

For a fixed f_{HP}, the F1-score declines more sharply as f_{RGB} decreases, though the hand pose stream helps mitigate this drop. Additionally, the performance gap between the best (MM-TMLP with $f_{\text{RGB}} = 30$ Hz and $f_{\text{HP}} = 3$ Hz) and worst (MM-TMLP with $f_{\text{RGB}} = 1$ Hz and $f_{\text{HP}} = 1$ Hz) models is smaller than in RGB-only settings.

The CPU usage primarily depends on f_{RGB}, decreasing roughly threefold with each reduction in f_{RGB}, which is consistent with RGB-only models. Adding the hand pose modality increases CPU usage, particularly at higher f_{HP} values, as expected.

These trends highlight a key insight: configurations using 10 Hz RGB and 10–30 Hz hand pose frequencies achieve nearly the same F1-score as the full 30 Hz RGB and 30 Hz hand pose setup, while reducing CPU usage by approximately 3×. This demonstrates that significant efficiency gains can be achieved with minimal performance loss, enabling flexible deployment depending on available computational resources.

4. Conclusions

In this work, we systematically explored the tradeoffs between accuracy and CPU usage in egocentric action prediction for resource-constrained settings. Our results provide a complete characterization of these trade-offs. In particular, they highlight that lowering the sampling rate for modalities with heavier CPU usage such as RGB (e.g., decreasing f_{RGB} from 30 Hz to 10 Hz while keeping $f_{\text{HP}} = 30$ Hz) leads to only minor accuracy degradation while significantly improving CPU efficiency.

Future work includes enhancing both the efficiency and accuracy of EAR models. One promising direction is to optimize the RGB stream using lightweight architectures or quantization to further reduce computational cost. Another key objective is to develop a large, specialized dataset for egocentric single-frame action recognition, addressing the current data gap in this domain. Additionally, exploring new modalities—such as audio, gaze tracking, and head pose—may provide further opportunities for performance improvement.

References

[1] Pratyusha Das and Antonio Ortega. Symmetric sub-graph spatio-temporal graph convolution and its application in complex activity recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3215–3219. IEEE, 2021. 1

[2] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2023. 2

[3] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 6252–6261, 2019. 1

[4] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 2

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[6] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas PJJ Noldus, and Remco C Veltkamp. Object detection-based location and activity classification from egocentric videos: A systematic analysis. *Smart Assisted Living: Toward An Open Smart-Home Infrastructure*, pages 119–145, 2020. 1

[7] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 2

[8] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[9] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012. 1

[10] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 3

[11] Md Salman Shamil, Dibyadip Chatterjee, Fadime Sener, Shugao Ma, and Angela Yao. On the utility of 3d hand poses for action recognition. In *European Conference on Computer Vision*, pages 436–454. Springer, 2025. 1

[12] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 1

[13] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 1

[14] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–24. IEEE, 2009. 1

[15] Swathikiran Sudhakaran and Oswald Lanz. Convolutional long short-term memory networks for recognizing first person interactions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017. 1

[16] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. *arXiv preprint arXiv:1807.11794*, 2018. 1

[17] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), pages 20270–20281, 2023. 1

[18] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1