WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition

Nishilkumar Balar Matr: 1638936

Savan Dihora Matr: 1648814

Neel Patel Matr: 1653726

Marius Bock marius.bock@uni-siegen.de

Abstract

Human Activity Recognition (HAR) using multimodal datasets, such as inertial and video data, with vision transformers remains an area of limited research. In this study, we leverage state-of-the-art vision transformer architectures, namely TriDet [1] and TemporalMaxer [2], to accomplish the objectives of HAR. Additionally, we explore the potential of using extracted features from the inertial dataset in combination with vision features as inputs. Our work achieves the highest average mean average precision (mAP) and F1 score by leveraging the combination of raw inertial features and extracted video features with the TriDet architecture. Furthermore, we attain the highest precision on the same dataset using the TemporalMaxer architecture, surpassing the previous work conducted by Bock et. al [3] in HAR on the WEAR dataset, a multimodal benchmark dataset. Code to reproduce our result is: `https://github.com/NishilBalar/wear/tree/raml_har`

1 Introduction

Detecting human physical activities using wearable devices, smartphones, and cameras has gained popularity due to its applications in various fields, such as healthcare and well-being, fitness and sports, security and surveillance, and transportation. Previous work in this area has primarily focused on using either inertial sensors or camera-based approaches. However, employing these modalities separately may not fully complement each other's drawbacks. To address these limitations, we utilize the WEAR dataset [3], which introduces a multimodal approach by leveraging both inertial and video data simultaneously for Human Activity Recognition (HAR).

The WEAR dataset contains 18 outdoor work-out activities performed by 18 participants at 10 different locations throughout the year. To capture the inertial data, 4 smartwatches are worn by each participant—2 on the wrists and 2 on the ankles. Additionally, video data is recorded using a head-mounted GoPro camera with a wide field-of-view, offering a resolution of 1080p and a frame rate of 60 Hz. The dataset also includes complex activities that involve multiple basic movements to perform an exercise, such as burpees, and null class activities where participants do not perform any activity. This multimodal dataset captures human movements with high accuracy, making it valuable for HAR research.

By using the WEAR dataset in conjunction with state-of-the-art models like TriDet [1] and TemporalMaxer [2], we aim to explore the application of vision-based transformer models for predicting human activities. Previous work by Bock et al. [3] demonstrated the success of using vision-based transformers, specifically Actionformer [4], by concatenating raw inertial features with video features extracted from two-stream I3D [5] models pretrained on Kinetics-400 [6]. They achieved the highest average mAP and F1 scores in a multimodal setting.

In this study, we also consider the window size of 1 second with 50 percent overlap, similar to previous work, to evaluate the performance of TriDet and TemporalMaxer architectures. The results of our experiments are presented in section 4.

2 Related Work

Considerable work has been conducted on inertial-based Human Activity Recognition (HAR) and vision-based HAR separately. However, only a few studies have explored the possibility of leveraging both modalities simultaneously. The inertial dataset offers the advantage of low confusion within background and workout classes but suffers from higher confusion among all classes for activity recognition. The well-known neural network architecture commonly used for inertial-based HAR is *DeepConvLSTM*, which combines convolutional and LSTM recurrent layers. Bock et al. has recently proposed a shallow DeepConvLSTM architecture that reduces computational requirements while achieving significant performance on the inertial sensor based HAR [7].

In contrast, vision-based HAR exhibits low confusion within workout classes but higher confusion for null classes. The focus of vision-based HAR includes action recognition, action detection, and anticipation. Action detection involves labeling activities in an untrimmed video and predicting their start and end times. On the other hand, in sensor-based HAR, activities are predicted within fixed-size sliding windows with specific temporal overlap. For temporal action detection (TAD), recent studies have utilized transformer architectures that do not rely on predefined anchor windows. These architectures employ an encoder, self-attention, and a decoder for TAD. In our work, we investigated two such transformer architectures, as mentioned earlier, to achieve competitive results on the WEAR dataset.

Some multimodal datasets that bear resemblance to WEAR include [8] and [9]. Both datasets provide cooking activities in an artificial kitchen environment and adopt an object-centric perspective from the user's point of view. Other datasets, such as [10], lack a null class and inter-activity sequences. To address these limitations, we utilize the WEAR dataset, aiming for accurate predictions of natural human activities. In the context of multimodal-based HAR, Hu et al. [11] and Nakamura et al. [12] have conducted similar work involving preprocessing and feature extraction of inertial data. In contrast, Bock et al. proposed using raw inertial data without any preprocessing, with the aim of leveraging the strengths of both modalities to enhance the performance of vision transformers.

clip-L.	Model	Modalities	P	R	F1	Avg.mAP
1.0s	Shallow DeepConvLSTM	Inertial	74.98	79.66	76.03	24.20
1.0s	Attend-and-Discriminate	Inertial	70.52	84.24	75.13	23.80
1.0s	ActionFormer	Inertial	75.65	70.92	70.47	56.00
1.0s	ActionFormer	Camera	72.54	68.02	66.95	58.60
1.0s	ActionFormer	Inertial + Camera	**79.19**	76.28	76.12	**63.71**

Table 1: Results of human activity recognition approaches of Bock et al. [3] for a window size of 1 sec. on the WEAR dataset evaluated in terms of precision (P), recall (R), F1-score, and average mean average precision (mAP) for different temporal intersections over union (tIoU) thresholds. The results underline the complementary of the inertial and camera modalities. Best results are in bold.

Table 1 shows the results obtained by Bock et al. on the WEAR dataset. Their best results, marked in bold, were achieved with multimodal fusion using the Actionformer transformer architecture, particularly in terms of precision and average mAP. We further investigate the performance of other promising architectures, such as TriDet and TemporalMaxer, using a similar experimental setup. Our findings are discussed in Section 4.

3 Methodology

As mentioned earlier, we built upon the approach proposed by Bock et al. [3] and pursued two different methods to surpass the benchmark set by the authors on the WEAR dataset. Our approaches are as follows:

1. Application of state-of-the-art vision-based transformer architectures.
2. Use of extracted inertial features instead of raw features for multimodal HAR.

Additionally, in their study, Bock et al. [3] investigated the optimal window length and overlap for the task of Human Activity Recognition (HAR). They found that a window length of 1 second with a 50 percent overlap yielded the best performance across all three different HAR methods. Following their findings, we adopted the same windowing procedure for evaluating our two different approaches.

3.1 Application of State-of-the-art Vision Transformer

As discussed in the previous section, ActionFormer [4], a vision-based transformer architecture, demonstrated unexpectedly strong performance not only on vision-based modalities but also on inertial-based wearable activity recognition. Moreover, a simple concatenation of raw inertial data with two-stream Inflated 3D (I3D) feature embeddings of video clips, using a window length of 1 second for plain vision-based transformer architecture, yielded the highest average mean average precision (mAP) and close-to-best F1-scores. This observation has motivated the current project to explore state-of-the-art architectures that have shown the best results on the THUMOS-14 dataset [13].

In this study, we investigate two vision-based architectures, namely TriDet [1] and TemporalMaxer [2], to assess the applicability of vision transformers on inertial data and the fusion of both modalities. Shi et al. [1] proposed the TriDet temporal action localization model, an extension of Zhang et al.'s ActionFormer model [4], incorporating a trident-head to improve imprecise boundary predictions. Additionally, Tang et al. [2] recently introduced TemporalMaxer, which minimizes long-term temporal context modeling while maximizing information from the extracted video clip features, employing a basic, parameter-free, and local region operating max-pooling block that requires significantly fewer parameters and computational resources.

3.2 Extracted Inertial Features for Multimodal HAR

In this approach, our main focus is on multimodal activity recognition, where we employ the early fusion technique to combine inertial and video-based modalities. Building upon the work of Bock et al., who concatenated camera-based two-stream I3D features with raw inertial data and achieved the best results across all cases, we explore the use of extracted features from the raw inertial dataset. These features are obtained using a pre-trained DeepConvLSTM model [14], and we investigate whether extracted features can offer more informative representations compared to raw data for activity recognition.

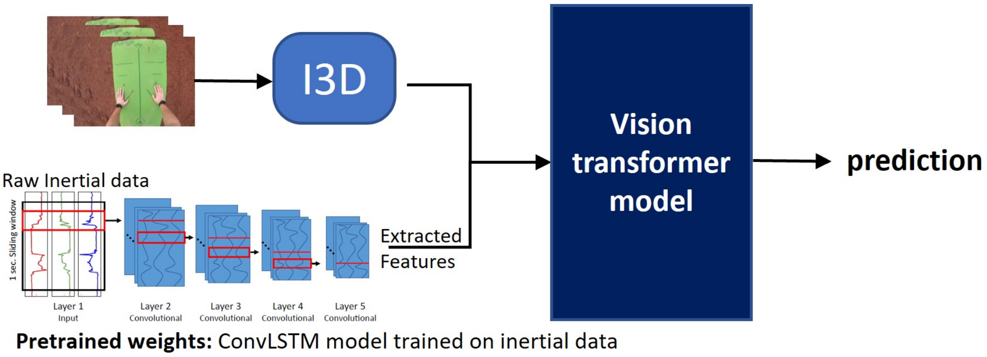


Figure 1: Concatenation of camera-based two-stream I3D feature embeddings with pretrained DeepConvLSTM model-extracted inertial features as input for vision-based transformer architecture in multimodal human activity recognition.

As depicted in Figure 1, we utilize the state-of-the-art model DeepConvLSTM [14] for inertial-based HAR. First, the model is trained on the raw inertial dataset, and its weights are saved for further processing. To extract meaningful features, we discard the LSTM layers of DeepConvLSTM and retain the convolution layers with their previously trained weights, effectively transforming the model into a feature extractor for raw inertial data. It's important to note that for all splits of cross-validation, different inertial features are extracted according to the learned weights specific to each split during the training of the DeepConvLSTM model. As in previous work, these combined features are fed to both the newly introduced vision transformers to compare the performance with respect to the raw inertial features-based multimodal HAR approach.

In the next section, we present detailed experimental results for the described approaches.

4 Experiments

The WEAR dataset includes three sets of pre-computed feature embeddings, which are computed with varying window sizes (0.5, 1, and 2 seconds). These feature embeddings consist of: (1) Feature embeddings from the two-stream I3D model[5], which were pre-trained on Kinetics-400[6]. These embeddings are derived from the raw video stream. (2) Vectorized inertial data, where the data from all axes of the inertial sensors are concatenated for each window to make the vectors of size [window length × no. of sensor axis]. (3) A fusion of both modalities, achieved by combining the feature embeddings from the two-stream I3D model and the vectorized raw inertial data.

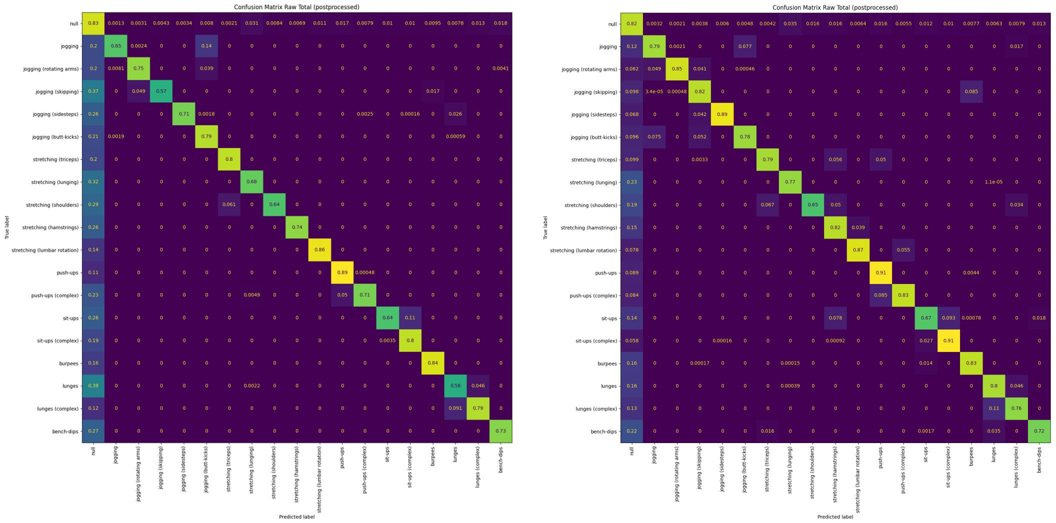
This work focuses on introducing extracted inertial features and application of these data modalities, namely: (1) raw-inertial features, (2) vision-based features(I3D), (3) combined raw-inertial and vision-based features(I3D), and (4) Extracted-inertial and vision-based features(I3D), to two vision-based architectures. In every experiment, a three-fold validation approach is utilized, where the data is split into three sets. For each experiment, 12 subjects are used for training, and 6 subjects are set aside for validation. The validation process ensures that each subject is part of the validation set exactly once, and the final evaluation metrics are calculated as the average across these three splits. To reduce the potential impact of statistical variance on performance differences between experiments, the evaluation metrics are averaged over three runs. Each run applies a different random seed, which helps ensure a more reliable and robust assessment of the model's performance. We chose to use the same training strategy and number of epochs as explained in [3]. Examining the results presented in [3] reveals that among the vision-based models, the optimal predictive performance was achieved when using a clip length of 1 second. In order to break previous benchmark results we used only a 1sec window length with 50% overlap. The evaluation metrics we presented in our benchmark results are the same as [3] which include two types: (1) Record-wise calculated recall, precision, and F1-score: These metrics are computed on a per-sample basis, allowing us to assess the performance of our method for individual records. (2) Mean Average Precision (mAP) at different temporal Intersection over Union (tIoU) thresholds: We used mAP, a widely used evaluation measure for temporal action localization datasets, to assess the accuracy of our approach under various tIoU thresholds[0.3, 0.4, 0.5, 0.6, 0.7].

4.1 Results

Through a straightforward concatenation of raw-inertial and vision-based features, both architectures achieve the highest average mean Average Precision (mAP) and the best F1 scores across all experiments, as shown in Table 2. By examining the results of all four approaches, presented in table 2, it becomes evident that both vision transformers, when applied in a basic manner, effectively integrate inertial and vision data. This integration allows them to leverage the unique strengths of each modality, resulting in improved performance for the tasks.

clip-L.	Model	Modalities	P	R	F1	Avg.mAP
1.0s	TriDet	Raw-inertial	82.78	76.88	77.40	72.45
1.0s	TriDet	Camera	70.14	65.74	67.25	68.81
1.0s	TriDet	Raw-inertial + Camera	81.44	**80.95**	**80.67**	**81.98**
1.0s	TriDet	Extracted-inertial + Camera	77.97	78.76	77.96	78.93
1.0s	TemporalMaxer	Raw-inertial	81.31	67.21	72.83	71.24
1.0s	TemporalMaxer	Camera	72.08	63.88	66.17	65.00
1.0s	TemporalMaxer	Raw-inertial + Camera	**83.07**	73.56	77.43	74.58
1.0s	TemporalMaxer	Extracted-inertial + Camera	80.29	72.74	76.33	73.37

Table 2: Results of our different approaches based on raw features and extracted features of WEAR for clip length of 1 second(with 50% overlap) trained on TriDet[1] and TemporalMaxer[2] evaluated in terms of Precision(P), Recall(R), F1-score and mean average precision(mAP) for different temporal intersection over unions thresholds(tIoU= 0.3,0.4,0.5,0.6,0.7). The best results among our experiments are in **bold**. The results which break the previous benchmark are in red.



(a) TemporalMaxer

(b) TriDet

Figure 2: Confusion matrices of the TriDet[1] and TemporalMaxer[2] model being applied using combined raw-inertial and camera features with a one-second sliding window and 50% overlap, as they break the previous benchmarks.

Moreover, TriDet is able to give improved F1-score and Average mAP values than the previous best results shown in Table 1. Also, TemporalMaxer outperforms the previous benchmark results of Precision. However, our experiments have shown that TemporalMaxer[2] architecture is less computationally expensive than TriDet [1]. From the results, it is visible that extracting features from inertial data with our approach was not a successful attempt as simple raw inertial data combined with vision-based features gives the best results among all modalities. Figure 2 shows the confusion matrices of the approaches which actually break previous benchmarks. These confusion matrices show less confusion amongst the activity classes when compared with the previous best results. However, same as previous results shown in [3], the models are much more confused among different jogging activities compared to other classes and the confusion between the null class and all the other activities remained the highest.

5 Conclusion

In conclusion, The WEAR dataset presents a difficult prediction task involving two types of data (inertial and vision). This situation highlights the need to investigate methods for effectively merging and utilizing both types of data. During the experiments that integrated raw-inertial data with extracted vision-based feature embeddings, the two vision-based transformer models(TriDet and TemporalMaxer) demonstrated superior performance by achieving better average mAP (mean average precision), precision, and F1-scores than previous benchmark results. However, extracting inertial features using a pre-trained DeepConvLSTM model [14], and using extracted features combined with vision-based features did not improve the results.

5.1 Outlook

One potential avenue for future research involves extracting features from raw inertial data during the training process using the same feature extractor layers, such as Conv1d (1D Convolution), from the ConvLSTM architecture. These extracted features could then be concatenated with I3D video features and utilized in vision-based transformer models. This integration would require manipulation of the Backbone part of such architectures to accommodate the raw inertial dataset. By pursuing this approach, it is anticipated that the performance of the models could be enhanced compared to the existing approach, which relies solely on pretrained weights.

References

[1] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling, 2023.

[2] Tuan N. Tang, Kwonyoung Kim, and Kwanghoon Sohn. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization, 2023.

[3] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. Wear: An outdoor sports dataset for wearable and egocentric activity recognition, 2023.

[4] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer:localizing moments of actions with transformers, 2022.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Will Kay, Jo~ao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[7] Marius Bock, Alexander Hölzemann, Michael Moeller, and Kristof Van Laerhoven. Improving deep learning for HAR with shallow LSTMs. In *2021 International Symposium on Wearable Computers*. ACM, sep 2021. doi: 10.1145/3460421.3480419. URL `https://doi.org/10.1145%2F3460421.3480419`.

[8] Fernando de la Torre, Jessica K. Hodgins, Javier Montano, and Sergio Valcarcel. Detailed human data acquisition of kitchen activities: the cmu-multimodal activity database (cmu-mmac). *Conference on Human Factors in Computing Systems*, 2009.

[9] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2022.

[10] Sibo Song, Vijay Chandrasekhar, Ngai-Man Cheung, Sanath Narayan, Liyuan Li, and Joo-Hwee Lim. Activity recognition in egocentric life-logging videos. *Asian Conference on Computer Vision*, 2015.

[11] Menghao Hu, Mingxuan Luo, Menghua Huang, Wenhua Meng, Baochen Xiong, Xiaoshan Yang, and Jitao Sang. Towards a multimodal human activity dataset for healthcare, 2023.

[12] Katsuyuki Nakamura, Serena Yeunga, Alexandre Alahi, and Li Fei-Fei. Jointly learning energy expenditures and activities using egocentric multimodal signals. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[13] Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos "in the wild". *CoRR*, abs/1604.06182, 2016.

[14] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 2016. doi: 10.3390/s16010115. URL `https://www.mdpi.com/1424-8220/16/1/115`.