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

MARIUS BOCK, University of Siegen, Germany
HILDE KUEHNE, University of Tuebingen, Germany
KRISTOF VAN LAERHOVEN, University of Siegen, Germany
MICHAEL MOELLER, University of Siegen, Germany

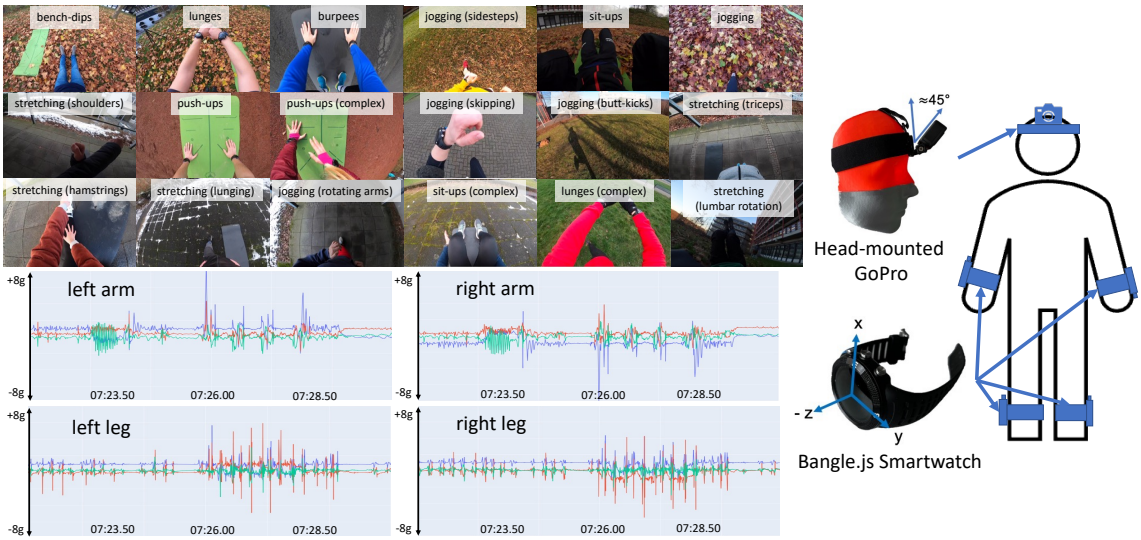

Fig. 1. Overview of the sensor setup and example data of the two types of wearable sensors employed in our WEAR dataset. 22 participants each performed 18 activities at 11 different outdoor locations while being equipped with four open-source smartwatches (one per limb) and a head-mounted camera.

Research has shown the complementarity of camera- and inertial-based data for modeling human activities, yet datasets with both egocentric video and inertial-based sensor data remain scarce. In this paper, we introduce WEAR, an outdoor sports dataset for both vision- and inertial-based human activity recognition (HAR). Data from 22 participants performing a total of 18 different workout activities was collected with synchronized inertial (acceleration) and camera (egocentric video) data recorded at 11 different outside locations. WEAR provides a challenging prediction scenario in changing outdoor environments using a sensor placement, in line with recent trends in real-world applications. Benchmark results show

Authors' addresses: Marius Bock, marius.bock@uni-siegen.de, Ubiquitous Computing, Computer Vision, University of Siegen, Siegen, Germany; Hilde Kuehne, h.kuehne@uni-tuebingen.de, Multimodal Learning, University of Tuebingen, Tuebingen, Germany; Kristof Van Laerhoven, kvl@eti.uni-siegen.de, Ubiquitous Computing, University of Siegen, Siegen, Germany; Michael Moeller, michael.moeller@uni-siegen.de, Computer Vision, University of Siegen, Siegen, Germany.

ACM 2474-9567/2024/12-ART175
https://doi.org/10.1145/3699776

that through our sensor placement, each modality interestingly offers complementary strengths and weaknesses in their prediction performance. Further, in light of the recent success of single-stage Temporal Action Localization (TAL) models, we demonstrate their versatility of not only being trained using visual data, but also using raw inertial data and being capable to fuse both modalities by means of simple concatenation. The dataset and code to reproduce experiments is publicly available via: mariusbock.github.io/wear/.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; • **Computing methodologies → Neural networks**.

Additional Key Words and Phrases: wearable activity recognition, inertial-based activity recognition, egocentric activity recognition, human activity recognition, temporal action localization, video activity recognition

ACM Reference Format:
Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. 2024. WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 175 (December 2024), 21 pages. https://doi.org/10.1145/3699776

1 INTRODUCTION & MOTIVATION

The physical activities that we perform in our daily lives have been identified as valuable information for a number of research fields and applications [3, 52, 69]. Research efforts have shown that physical activities can be detected using either wearable inertial sensors or camera-based approaches. The inertial sensors can continuously observe motion and gestures at particular body locations, whereas camera-based systems can typically observe the user's entire body, but can be hindered by (self-)occlusions. Inertial data manifests itself as multidimensional time series, but is unlike image data almost impossible to be interpreted by a human annotator after recording. Even though research has shown that both modalities are complementary to each other [62], available benchmark datasets that provide both camera and inertial-based sensor data remain scarce. We therefore introduce WEAR, an outdoor sports human activity recognition (HAR) dataset featuring workout activities performed by 22 participants while wearing inertial sensors on both wrists and ankles as well as a head-mounted camera capturing egocentric vision - see Figure 1. Unlike previous egocentric datasets [11, 22, 23], WEAR provides an inertial sensor placement which is in line with recent trends in real-world application scenarios, providing complementary information to the captured video stream. Further, amongst egocentric datasets which utilize limb-placed IMU sensors, WEAR is the first dataset being collected at different outdoor recording locations, with each location introducing different visual and surface conditions, yet not providing cues about the activity being performed. Our contributions in this paper are three-fold:

(1) We introduce a new inertial- and vision-based HAR dataset, called WEAR, consisting of data from 22 participants performing 18 different outdoor sports activities.
(2) Results of our performed benchmark analysis reveal that the application-driven sensor placement along with the challenging outdoor recording scenario makes WEAR an a suitable dataset to assess methods on how to combine strengths of both inertial- and vision-based approaches.
(3) We demonstrate that state-of-the-art TAL models from computer vision are particularly suited to not only process raw inertial data, but even successfully fuse multi-modal information significantly outperforming the best single-modality approach as well as beating the best possible (oracle) late fusion approach in terms of mAP.

2 RELATED WORK

2.1 Inertial- and Video-based HAR datasets

In Table 1 we show a curated list of datasets which provide both egocentric vision- (e.g. RGB, depth) and IMU-based (e.g. accelerometer, gyroscope, magnetometer) modalities in the context of HAR. We compare datasets regarding

Table 1. List of available egocentric vision datasets, which provide inertial data, compared with the WEAR dataset. We differentiate between number (#Class) and type of activity classes (S = Sports, D = Daily Living, C = Cooking), number of participants (#Participants), recording environment (outside or inside), limb placement of IMU sensors (LW = left wrist, RW = right wrist, LA = left ankle, RA = right ankle) and type of recorded videos (T = trimmed or U = untrimmed video sequences). *Subset which provides both IMU and activity annotations **Exact participant counts are not provided; labels summarize key-steps † Activity verb classes. ‡ Excludes calibration periods.

Dataset	General					Limb Placement				
	#Participants	#Class	#Hours	Type	Where	LW	RW	LA	RA	Video
MEAD [61]	2	20	<1	D,S	In,Out					T
DataEgo [53]	≈10	20	4	D,S	In,Out					U
Ego4D [22]	≈60	87†	160*	D,S,C	In,Out					T
Ego-Exo4D [23]	≤555**	689**	68*	D,S,C	In,Out					T
EPIC-Kitchens [11]	37	97†	100	C	In					U
UESTC-MMEA-CL [74]	10	32	30	D	In,Out					T
CMU-MMAC [12]	13	16†	2	C	In	✓	✓	✓	✓	U
ADL Dataset [16]	2	6	<1	D	In	✓	✓			U
ActionSense [13]	9	20	9‡	C	In	✓	✓	✓	✓	U
WEAR	**22**	**18**	**19**	**S**	**Out**	✓	✓	✓	✓	**U**

their recency, size, number of participants, number and type of activities performed, recording environment, limb placement position of IMU sensors and whether the dataset is provided on a clip-basis or a continuous stream. The three largest multi-modal corpuses currently available, the Ego4D [22], EPIC-Kitchens [11] and Ego-Exo4D [23] datasets only provide inertial data from the built-in IMU of the head-worn camera. Given the captured inertial data thus only decodes the head movement of participants, this position of an inertial sensor is likely to provide only limited complementary information to the egocentric vision modality. As evident by the rise in popularity of commercial head-mounted cameras and wrist-worn smartwatches for tracking sports, we thus decided to position the camera and IMU sensors used during collection of the WEAR dataset in line with the recent trends in real-world application scenarios. With the head and limbs being positions which do not limit participants in their freedom of movement, we deem said positions to further be most suited in capturing how participants interact with their environment and/ or objects. Further note that with the intended audience of the datasets residing mostly in the vision-based community, the CMU-MMAC [12], Ego4D, Ego-Exo4D and EPIC-Kitchens datasets are focusing on providing activity annotations following a vision-centric definition, e.g. *pick up object X*, or are close to a narration-style annotation, e.g. *adjust the stove heat*. Though these type of labels are differentiable from a vision-side, they are near impossible to distinguish using body-worn sensors. To nevertheless give an estimate of activities present in the dataset, we provide activity verb counts (Ego4D and EPIC-Kitchens) and high-level activity counts (Ego-Exo4D) in Table 1. Our chosen sensor placement thus makes the CMU-MMAC [12], ADL [16] and ActionSense [13] dataset most comparable to the WEAR dataset. Compared to these three datasets, the WEAR dataset is the largest in size and participant count. Furthermore, the three existing benchmark datasets were all recorded in a non-changing indoor (kitchen) environment, which, in case of the ActionSense and CMU-MMAC dataset, was artificially created in a laboratory. Unlike the outdoor setup of the WEAR dataset, the indoor setup thus limits each dataset's amount of variety captured in the visual data, as lighting conditions and surroundings remain the same throughout all participants. Lastly, as activities are mostly object-centric activities (e.g. cooking recipes) and action is mostly taking place in the POV of the user, prediction scenarios of the three datasets are more biased towards vision rather than inertial data.

2.2 Inertial-based HAR

Compared to video-based modalities body-worn sensor systems bear a great potential in analyzing our daily activities with minimal intrusion, yielding various applications from the provision of medical support to supporting complex work processes [6]. Within the last decade deep learning based-methods have established themselves as the de facto standard in inertial-based HAR as they have shown to outperform classical machine learning algorithms [24, 25, 51]. One of the most well-known deep learning approaches for inertial-based HAR is the *DeepConvLSTM* which is a hybrid model combining both convolutional and recurrent layers [51]. By combining both types of layers the network is able to automatically extract discriminative features and model temporal dependencies. Following the success of the original DeepConvLSTM, researchers worked on extending the architecture [48, 72] or build up on the idea of combining convolutional and recurrent layers by proposing their own architectures [1, 73, 79, 84]. Within this publication we are reporting benchmark scores using the WEAR dataset inertial sensor-streams as input for two popular HAR models [1, 5]. Contrary to the belief that one needs to employ multiple recurrent layers when dealing with sequential data [34], Bock et al. [5] proposed an altered *shallow DeepConvLSTM* architecture which proved to outperform the original architecture by a significant margin. Differently, Abedin et al. [1] chose to build up on the idea of the DeepConvLSTM and introduced the *Attend-and-Discriminate* architecture which exploits interactions among different sensor modalities by introducing self-attention through a cross-channel interaction encoder and adding attention to the recurrent parts of the network.

2.3 Vision-based HAR

Predicting activities performed by humans based on visual-cues can broadly be categorized into three main application scenarios: action recognition, localization and anticipation. Action recognition systems [37, 45, 68] aim to assign a set of trimmed action segments an activity label. Contrarily, TAL systems [44, 76, 81] are tasked to identify start and end times of all activities in a untrimmed video by predicting a set of activity triplets *(start, end, activity label)*. Lastly, action anticipation systems [20, 56] aim to predict the label of a future activity having observed a segment preceding its occurrence. Though sensor-based HAR systems are employed using a sliding window approach and thus assign activity labels to a set of trimmed inertial-sequences, their ultimate goal is to identify a set of activities within a continuous timeline. We therefore deem vision-based TAL to be most comparable to inertial-based HAR and will focus on it in our benchmark analysis. Existing TAL methods can be divided into two categories: two- and single-stage approaches. Two-stage approaches [2, 21, 38, 40, 43, 54, 63, 65, 75, 80, 82, 83, 85] divide the process of TAL into two subtasks. First, during the action segment proposal generation, candidate video segments are generated which are then, classified with an activity label as well as refined regarding their temporal boundaries. Contrarily, single-stage approaches [9, 39, 41, 44, 46, 49, 58, 59, 76, 81] aim to localize actions in a single shot without using action proposals.

In light with the success of transformer architectures in natural language processing [14, 67] and computer vision [36, 45, 78], researchers have demonstrated their applicability for TAL [10, 42, 44, 59, 65, 81] breaking previously held benchmark scores of numerous popular datasets [11, 26, 33] without any additional training data by a significant margin. One of such architectures is the *ActionFormer* proposed by Zhang et al. [81], which is an end-to-end trainable transformer-based architecture, which unlike other single-stage approaches, does not rely on pre-defined anchor windows. The architecture combines multiscale feature representations with local self-attention and is trained through a classification and regression loss calculated by a light-weighted decoder. Building up on ActionFormer architecture, Shi et al. [58] proposed the *TriDet* model which suggest to replace the transformer layers of the ActionFormer with fully-convolutional, so-called SGP layers, as well as use a trident regression head which claims to improve imprecise boundary predictions via an estimated relative

probability distribution around the boundary. Given the rapid rise in popularity of single-stage TAL such as the ActionFormer, we decided said models to be a suited option to deliver a first benchmark for the WEAR dataset.

2.4 Multimodal (Inertial and RGB Video) HAR

With early works such that of Spriggs et al. [62] having shown the complementarity of inertial- and camera-based features, research has followed up by exploring different ways of combining the two modalities. One can categorize such methods broadly by the point in time at which the fusion of both modalities is performed. Late fusion approaches usually follow a two-stream architecture training both vision- and inertial-based modalities separately before merging together outputs of each stream through such as produced softmax probabilities e.g. via a weighted combination [70], attention mechanisms [19], pooling operations [30, 60], majority voting [17], a concurrent classifier [15, 28, 64, 71] or knowledge transfer between separate networks [55]. Early fusion approaches aim at jointly learning from both modalities by using feature embeddings calculated on one (or both) modalities to e.g. use the concatenation of both to train a concurrent network [8, 15, 16, 18, 27, 29, 31, 32, 47, 50, 61, 74, 77], enhance softmax probabilities used during late fusion [15, 16] or adding intermediate cross-view connections amongst the two modality streams [28]. With experiments showing that single-stage TAL models are able to produce competitive results on raw inertial data, this paper also tests the applicability of two state-of-the-art models, namely the ActionFormer and TriDet model, to fuse and combine cues of both modalities in an early-fusion style. Unlike other early fusion techniques, our approach is the first to directly use the raw inertial data by means of simple concatenation together with a vision-based feature embedding.

3 METHODOLOGY

3.1 Study Design & Scalable Pipeline

Participants of the WEAR dataset were recorded during separate recording sessions. Prior to their first session, participants were handed a recording plan which outlined the study protocol as well informed about any risks of harm, data collection, usage, anonymisation and publication, as well as how to revoke their data usage rights at any point in the future. The study design involving human participants was reviewed and approved by the University of Siegen (reference number: 03/2023 VVT). All participants were briefed and provided their written informed consent. Each participant was asked to perform all 18 workout activities detailed in the recording plan. The location and the time of day at which the sessions were performed, were not fixed and thus vary across participants (see Section B in the supplementary material). Participants were suggested to follow a two-session setup, i.e. 9 activities per session. Nevertheless, it was allowed to differ from this setup and split the 18 activities across as many (or as few) sessions as participants liked. This caused the amount of recording sessions to vary across participants, but also increased the amount of captured variability in weather conditions and recording locations. In order to avoid misunderstandings in the execution of the activities, the authors discussed all activities prior to each session and encouraged participants to ask questions during the session if something remained unclear. Participants were tasked to perform each activity for roughly 90 seconds. As activities varied in their intensity, it was not required to perform activities for 90 seconds straight and participants could include breaks as needed. Furthermore, to ensure that each participant was able to perform all workout activities properly, the recording plan detailed how activities could be altered in their execution, for instance so that they required less physical strength.

3.2 Experimental Protocol

In order to properly explain to participants the activities they needed to perform and give insights on the overall study design a recording plan (see Section D in the supplementary material) was provided to participants prior to their first session. The recording plan details all necessary materials and is written in such a way that the

can easily be reproduced by persons other than the authors. The plan further outlines the study protocol as well informs about any risks of harm, data collection, usage, anonymisation and publication, as well as how to revoke data usage rights at any point in the future. Besides a written description of each activity, the original document provides short video-clips of each activity, showing the correct execution of exercises. To avoid any misunderstandings, the participants further received a one-on-one session with the researchers being able to ask their questions about the plan and activities in it.

The recording plan provided with our dataset includes all necessary materials and is written in such a way that all activities and sessions can easily be reproduced by persons other than the authors. Besides the used sensors for video and acceleration recording, the exercises only require a yoga mat and a chair (or similar items). Sessions can be recorded at any location outside as long as the privacy of the participants and bystanders is ensured. The code repository of the dataset provides a detailed guide on how to work with the equipment and how to collect, postprocess and annotate newly collected data. We argue that this facilitates reproducibility, and with a minimal setup ensures that it is possible for others to extend our dataset at a later date.

3.3 Participant Information

We recorded data for 22 participants (13 male, 9 female) at 11 different locations and under varying weather conditions. The first 18 participants were recorded over a stretch of 5 months (October to February), totalling more than 15 hours. In an effort to provide a test set along with the WEAR dataset an additional 6 participants were recorded, totalling around 4 hours of additional data. To allow researchers to explore also personalised prediction approaches, the test dataset includes two participants which were already part of the first recordings, and which volunteered to complete the workout an additional time. Recordings of the test dataset took place during spring (new participants) as well as summer (re-recordings). Egocentric video data of the four new participants was recorded using a GoPro Hero 11 as opposed to a Hero 8 and two new participants were recorded at a previously unseen location. On average each participant contributed roughly 50 minutes of data. At the time of recording, the participants had a mean age of 27.59 years (standard deviation (SD): 4.69), a mean height of 175.2 cm (SD: 9.83), and a mean weight of 69.74 kg (SD: 11.23). In order to assess their sports level, participants filled in a post-session questionnaire. The questionnaire contained questions related to vital information (such as body height, weight and age), weekly workout frequency (min. 15 minutes duration) and experience in particular workout activities. The participants in the study worked out for a mean of 3.73 times per week (SD: 2.31), were familiar with a mean of 15 out of the 18 activities in advance (SD: 3.57), and regularly performed a mean of 5.71 of the recorded activities (SD: 4.14) as part of their private workouts. Participants reported for their personal workout schedules a wide-range of cardio- (running, hiking, cycling, dancing), strength- (weight lifting, freeletics, rowing, bouldering), team- (volleyball, basketball, (table-)tennis) and flexibility-focused (yoga, ballet) exercise types. Individual answers of each participant can be found in the supplementary material (see Section B).

3.4 Dataset Collection & Structure

The WEAR dataset provides participant-wise raw and processed acceleration and egocentric-video data (see Figure 1). We focus on 3D accelerometers especially as they cover a substantial amount of commercial fitness devices worn at the wrists and ankles. They furthermore are used in a large set of existing research and datasets focusing on wearable data for activity recognition, and they do not suffer from noise, drift, and other device-specific characteristics. 3D accelerometer data was collected at 50 Hz with a sensitivity of ± 8g using four open-source Bangle.js smartwatches running a custom, open-source firmware [66]. The watches were placed by the researchers in a fixed orientation on the left and right wrists and ankles of each participant. Egocentric video data was captured using either a GoPro Hero 8 or Hero 11 action camera, which was mounted using a head strap on each participant's head. The resulting '.mp4'-videos were recorded at 1080p resolution with 60 frames

per second and the camera being tilted downwards in a 45 degree angle. A second tripod-mounted camera was placed within the proximity of each participant to facilitate annotation recording the environment in which the workout was performed from a third-person-perspective. Using again a large FOV setting, the second camera was placed in a way such that as much area as possible was captured. To allow for even more freedom of movement, participants were allowed to move out of the FOV of the second camera, but were asked to start and end their activities within the camera's FOV. This allowed participants, especially during running exercises, to run straight distances and overall commence activities in a more natural way. For privacy reasons, the second camera's video and all audio captured are not part of the WEAR dataset.

The open-source firmware [66] running on each Bangle.js smartwatch stores the lossless, delta-compressed inertial data in separate files on the internal memory of each watch. During post-processing, said compressed files were extracted, uncompressed and concatenated to a single '.csv'-file per session. Being a common issue with accelerometers sampling at a high sampling rate, the Bangle.js smartwatch is not able to maintain an exact sampling rate of 50 Hz at all times throughout the experiment, with the true sampling rate being closer to 48 Hz with fluctuations ranging between ± 1 Hz. The firmware provides for each file a timestamp that was set by the on-board real-time clock, which allows correcting individual times of all delta-compressed samples. Therefore, in order to obtain the true sampling rate and correct the timestamps of the concatenated '.csv'-file, synchronisation jumps were performed by each participant at the start and end of each session. The synchronization jumps involved participants move in front of the tripod-mounted camera, stand still for approximately 10 seconds, jump three times along with raising the arms while jumping, and stand still for another 10 seconds. This allowed to map peaks in the inertial sensor streams to be mapped to points in the video stream and thus obtain a start and end point within both modality data streams. Lastly, assuming recorded inertial data records are equidistant, all records within the span of the start and end-point were evenly distributed across the experiment's duration and, as a final step, resampled to have a sampling rate of 50 Hz via linear interpolation. Similar to the inertial data, the video data recorded by the head-mounted GoPro was not recording at a true frame rate of 60 FPS, but slightly deviated from that (i.e. 59.94 FPS). We therefore also resampled the egocentric videos to be of a frame rate of 60 FPS.

In order to validate our synchronization process we made use of the similarities between sensor and audio data and converted each axis of the 3D accelerometer as well as their combined magnitude to four separate '.wav'-files. This approach is inspired by the works of Scholl et al. [57] and Bin Morshed et al. [4]. We calculated the magnitude as the summed norm of each individual inertial sensor channels, i.e. $\sqrt{x^2 + y^2 + z^2}$ with x, y and z being the x-, y- and z-axis of the 3D accelerometer data. Having converted the '.csv'-data to '.wav'-files allowed us to import both video data and inertial data into a standard video editing tool, in our case we used Final Cut Pro (see Figure 2). The user interface of Final Cut offers to see previews of sound files being in our case equivalent to a graph-like visualization of the acceleration data. This allowed us to verify our applied synchronization during annotation throughout the whole duration of each participant's session. On average, the combined magnitude proved to be most useful when verifying the correctness of our synchronization across time. Labels of the activities were added by a single expert annotator as subtitles in '.srt'-format. A final script then converted the exported '.srt'-file to '.csv'-format, filling gaps within the subtitles with a *NULL* label and appended this to the respective final inertial sensor data '.csv'-file.

4 BENCHMARKS AND BASELINE RESULTS

Though the WEAR dataset provides the possibility for a multitude of HAR use cases, this paper focuses on introducing one sample application scenario per data modality, namely: (1) inertial-based wearable activity recognition, (2) vision-based TAL, as well as, (3) a combined approach using both data modalities as input simultaneously. We chose to use said application scenarios because of their similarities with each other as they

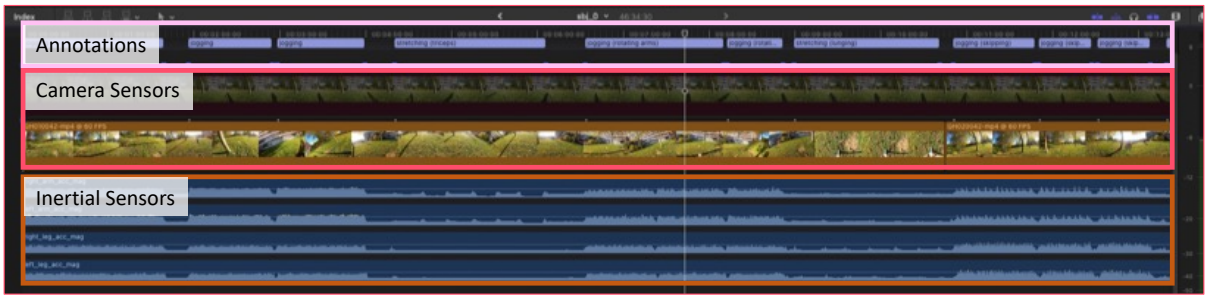

Fig. 2. Snapshot along with descriptions of the annotation process using Final Cut Pro. Importing the converted video and inertial data (as '.wav'-files) allowed for an easy validation of the synchronization process. Labels were added via subtitles, exported as '.srt'-files and converted such that they can be appended to the respective '.csv'-files.

both aim to detect a set of activities in an untrimmed sequence of data. Nevertheless, other HAR-specific (e.g. action anticipation and classification) and non-HAR application scenarios (e.g. hand detection, pose estimation or simultaneous localization and mapping (SLAM)) are applicable.

During each experiment we employ a Leave-One-Subject-Out (LOSO) cross-validation which has each participant being used as the validation set exactly one time while all other participants are used as trainig data. In order to minimize the risk of performance differences between experiments being the result of statistical variance, all experiments mentioned in this publication were repeated three times employing three different random seeds (1, 2 and 3). The final validation results of an experiment are then reported as the average across the three repeated runs. As mentioned in Chapter 3.1 a separate test dataset was collected which consists of 6 additional participants performing the same workout as all the other participants. The test dataset further includes additional sessions of two participants which volunteered to be recorded a second time. We use the test data to get an unbiased assessment of our chosen set of hyperparameters and applied postprocessing, which are determined during the single-modality and multimodal validation experiments.

Note that all models part of our benchmark analysis were applied without pretraining on existing benchmark datasets. Nevertheless, feature embeddings to train the vision models were extracted using a two-stream I3D feature extractor [7] which was pre-trained on Kinetics-400 [35]. With the standard error of evaluation metrics amongst runs being at maximum 2.5% and the majority of runs being below 1%, we only report average evaluation metrics in this paper. All mentioned experiments were conducted on a single NVIDIA Tesla V100 GPU and lasted no longer than 24 hours. Though sharing inherent similarities, vision-based action localization algorithms predict a collection of activity segments defined by a start and end time, while, contrarily, inertial-based HAR systems provide labels based on the pre-defined windowed segmentation. Given their difference in prediction output, different evaluation metrics are applied, with mean average precision (mAP) being most prominent metric in vision-based TAL and F1-score being the most prominent metric in inertial-based activity recognition. Therefore, to guarantee comparability amongst application scenarios and architectures, predictions of each algorithm are converted such that both vision- and inertial-based evaluation metrics can be calculated. More specifically, our reported benchmark evaluation metrics are (1) a record-based calculated recall, precision and F1-score, and (2) segment-based mean average precision (mAP) at different temporal intersection over union (tIoU) thresholds, commonly used to evaluate TAL datasets. To ensure readability we only provide a selection of visualized results. For a complete collection visualizations of all mentioned experiments, please see Chapter C.7 within the supplementary material.

4.1 Single-Stage Temporal Action Localization for Inertial Data

Though originally intended to be applied to video data, our work demonstrates that vision-based TAL models, such as the TriDet [58] and ActionFormer [81] models, prove to be applicable to inertial data as well as a means to fuse both modalities. Both TAL architectures take as input a collection of clip-wise feature embeddings, which are obtained by applying a sliding window approach on top of an input video. Using both classification and regression losses, the ActionFormer and TriDet model then try to localize activity segments, defined by an activity label, start, and end time, within a complete video. To obtain discriminative feature embeddings of each sliding video clip, the TAL community has resorted to using feature extraction methods such as a two-stream I3D feature extractor [7]. The feature extractors, usually pretrained on a larger vision corpus like the Kinetics-400 dataset [35], summarize the raw visual data into a one-dimensional feature embedding.

As both TAL and inertial-based architectures rely on a sliding window approach to preprocess data, our paper suggests a simple yet effective preprocessing technique such that raw inertial data can be used to train vision-based TAL models. Illustrated in Figure 3, we propose vectorizing the two-dimensional sliding window data commonly found in inertial-based architectures, creating one-dimensional raw inertial feature embeddings per sliding window that can be used to train TAL models such as ActionFormer and TriDet. Specifically, starting with the windowed inertial data of dimensions [*no. windows*, *window length*, *no. sensor axes*], we concatenate the individual sensor axes of each window, resulting in a one-dimensional feature vector of size [*window length*$\times$*no. sensor axes*]. In the case of the WEAR dataset, which provides 12 individual sensor axes per participant, we obtain (depending on the window length) feature vectors of size 300 (0.5 seconds), 600 (1 second), and 1200 (2 seconds) per video clip, i.e., sliding window. Even though our concatenation approach results in varying input dimensions, this change does not come at increased computational costs. More specifically, while the number of learnable parameters marginally increases (not more than 10%) with an increased input dimension, unlike other approaches, no additional embedding needs to be extracted from the inertial data, and raw data streams can be directly used. Furthermore, as demonstrated in the multimodal experiments in this paper, we show that by concatenating one-dimensional inertial data and I3D feature embeddings, the TAL architectures are capable of successfully fusing multi-modal information, significantly outperforming single-modality approaches.

By using our approach, the TAL architectures process the inertial data in a different format compared to, for example, the shallow DeepConvLSTM. However, their performed feature extraction is comparable. Specifically, both architectures shift convolutional filters separately across each sensor axis, with the only difference being that the TAL models do not pad the start and end of each sensor axis, allowing for a slight overlap amongst axes at their point of concatenation. Nevertheless, we argue that our approach draws inspiration from visual feature embeddings such as the I3D features, which have proven effective in training architectures, yet, by design, will also cause an overlap when transitioning between flow and RGB features.

4.2 Single-modality Experiments

Similar to Zhang et al. [81] and Shi et al. [58], we opted to train the vision-based benchmark models using two-stream I3D feature embeddings with three different clip lengths (0.5, 1, and 2 seconds), employing a 50% overlap between clips. Besides extending the number of epochs to 100, we adopted the same training strategy that yielded the best results on the EPIC-Kitchens dataset [11], as reported by both architectures. In contrast to inertial-based approaches, TAL models are not trained to predict an explicitly modeled NULL-class. With both models set to predict up to 2000 action segments per video, each timestamp ended up being classified by an action segment, resulting in a prediction performance for the NULL-class close to or at 0% accuracy. To address this, we eliminated low-scoring segments by increasing the scoring threshold of both models to 0.1, significantly enhancing the accuracy of the NULL-class while only marginally affecting prediction performance of all other activity classes.

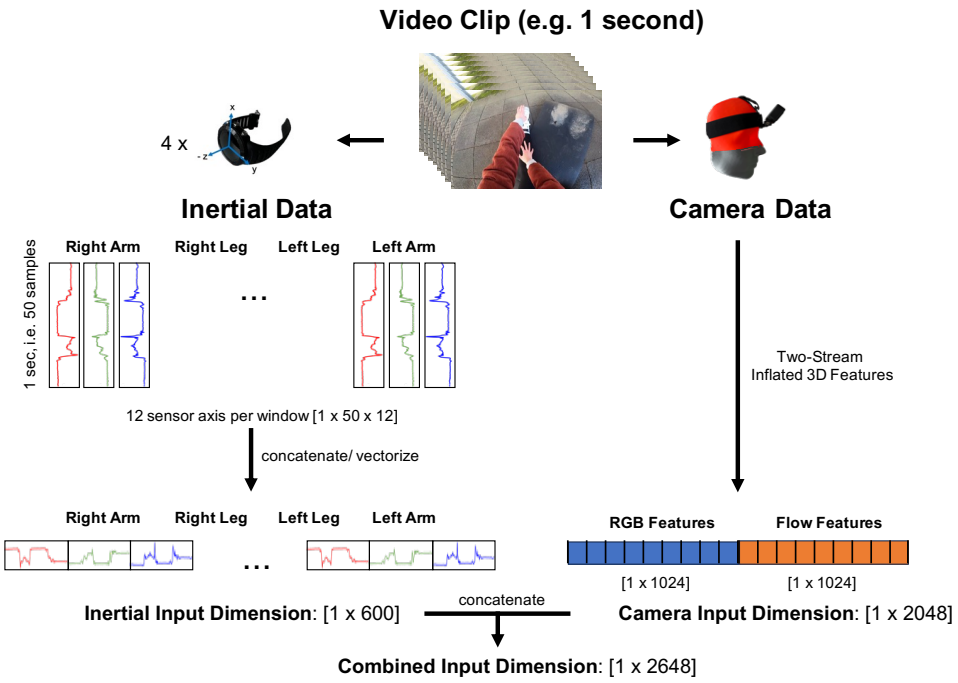

Fig. 3. Visualization of the applied preprocessing on inertial and camera data in order to make to create a feature embedding which can be used to train the TriDet and ActionFormer network.

As our inertial-based benchmark algorithms of choice, we employed the shallow DeepConvLSTM proposed by Bock et al. [5] and the Attend-and-Discriminate model proposed by Abedin et al. [1]. Throughout all experiments, we followed the training strategy suggested by Bock et al. [5], which has shown to yield reliable results on various inertial-based HAR datasets, only extending the number of epochs to match those of the vision-based experiments. To compensate for longer training times, we applied a step-wise learning rate schedule. Additionally, incorporating architecture changes recommended by Bock et al. [5], we modified the Attend-and-Discriminate model to utilize a one-layered instead of a two-layered recurrent module and adjusted the convolutional kernel size based on the sliding window and sampling rate of the WEAR dataset. Given that inertial-based architectures provide predictions on a per-window basis, frequent intermediate, short-lasting activity switches occur along the time axis, resulting in small coherent segments and ultimately lower mAP scores compared to the vision-based models presented in this paper. To mitigate these switches, predictions made by the inertial-based architectures were smoothed using a majority-vote filter of 10 seconds. For detailed ablation results on the design choices and chosen hyperparameters see Section C of the supplementary material.

Examining the results presented in Table 2, one can observe that for the TAL models, a clip length of 1 second yielded the best predictive performance, while for the inertial-based architectures, a larger window size of 2 seconds produced higher classification results. Analyzing per-class results, it is evident that when using only visual data as input, the models struggle to differentiate between different running styles, activities outside the field of view of the participant (e.g., triceps stretches), as well as normal and complex sit-ups. In contrast, since inertial sensors are not affected by occlusions and accelerometer data provides an encoded representation

of limb orientation due to the Earth's gravitational force, algorithms trained on inertial data do not exhibit these weaknesses and are particularly reliable during activities where limb orientation is the main discriminator (e.g., stretches). However, as indicated by a larger NULL-class accuracy compared to the inertial-based models, vision-based approaches are better at differentiating activities from breaks, which is a desirable trait of HAR systems given the often substantial size of the NULL-class [6].

Interestingly, using the same hyperparameters as those used during vision-based experiments, a plain ActionFormer and TriDet network not only produce competitive classification results based on inertial input data but also show less confusion among activity classes compared to inertial-based architectures. Comparing per-class results between the TAL models and classic inertial-based architectures, it is evident that the confusion in the TAL models mostly lies within the activity categories themselves (jogging, stretching, and strength), whereas inertial-based models exhibit a higher degree of overall confusion between the activities and the NULL-class. Given that inertial-based models like the DeepConvLSTM are trained to classify sliding windows individually, we assume that this overall larger confusion and intermediate activity switches are caused by the inertial models' inability to leverage context information beyond the window they are tasked to classify. Consequently, due to the intermediate activity switches, calculated mAP scores of the inertial architectures are significantly lower than those of the camera-based approaches. Nonetheless, inertial-based models are on average able to predict all workout activities more consistently and produce the highest classification metrics across all experiments.

4.3 Multimodal (Inertial and Egocentric Video) Experiments

As demonstrated in previous chapters, both the vision and inertial modalities present various strengths and weaknesses. Therefore, in an effort to leverage the strengths of both modalities, we evaluated a combined, multimodal training approach using the TriDet and ActionFormer model. To fuse the two-stream I3D feature embeddings with the inertial data early on, we vectorized the inertial data, as described in the previous section, such that each sliding window is represented by a one-dimensional inertial embedding vector. We then concatenated said vector with the I3D feature embedding of the corresponding video clip, resulting in an early fused representation of the a sliding window consisting of both visual and inertial information. Through simple concatenation of both modalities, both architectures achieved the highest average mAP and close-to-best F1-scores across all experiments (see Table 2). Comparing confusion matrices of all three approaches (see Figure 4) reveals that both vision models, applied in a plain fashion, are able to successfully combine inertial and vision data and leveraging the previously mentioned strengths of each modality. To assess how our early-fusion approach compares to voting-based late-fusion approaches such as proposed by Ijaz et al. [28], we implemented an *Oracle*-based late fusion, which creates perfectly late fused predictions of different models. The predictions are merged by comparing each of them with the ground truth data and only keeping, if predicted by one of the networks, the correct prediction. Interestingly, the first *Oracle*-late-fusion *O-LF(I, C)*, which late fuses predictions of the best inertial and best vision model, produces lower mAP scores than that of the best TAL model being trained on both modalities simultaneously. Furthermore, late-fusing the best inertial, vision and early-fusion approach (*O(I, C, I + C)*), increases classification and mAP scores of *O(I, C)* by as much as 10%, suggesting the early-fusion-based approach is capable of learning to differentiate activities both single-modality models failed to classify correctly (see e.g. misclassified stretching exercises in Figure 5). Nevertheless, classification results of the *Oracle*-based late fusion significantly outperform both single- and combined-modality approaches, indicating that the data set is far from being saturated.

4.4 Test Set Results

To verify chosen hyperparameters and suitability of the applied postprocessing as determined in the LOSO cross-validation, we collected an additional test set of six additional participants. The participants constitute of

Table 2. LOSO validation results of human activity recognition approaches based on body-worn IMU (Inertial), vision (Camera) and combined (Inertial + Camera) features for different clip lengths (CL) on the first 18 participants of the WEAR dataset evaluated in terms of precision (P), recall (R), F1-score and mean average precision (mAP) for different temporal intersection over union (tIoU) thresholds. The results underline the complementarity of the inertial and camera modalities. *O-LF()* corresponds to the *Oracle*-based late fusion, which creates perfectly late fused predictions of different models.

Best results per modality are in **bold**.

	Model	CL	P	R	F1	mAP					
						0.3	0.4	0.5	0.6	0.7	Avg
Inertial	Shallow D.	0.5s	74.81	74.93	72.26	60.03	57.91	55.54	54.19	52.31	55.99
	A-and-D	0.5s	76.54	73.04	72.10	57.51	54.59	52.22	49.47	46.94	52.15
	ActionFormer	0.5s	63.09	78.17	66.80	81.12	77.47	69.08	53.98	38.56	64.04
	TriDet	0.5s	68.58	78.30	70.41	79.66	76.07	69.78	60.57	49.95	67.21
	Shallow D.	1s	77.71	77.41	75.44	63.37	61.35	58.91	57.02	55.05	59.14
	A-and-D	1s	79.97	75.92	74.67	61.72	58.77	56.29	53.83	51.59	56.44
	ActionFormer	1s	65.88	78.44	68.40	**81.91**	79.75	76.99	72.09	64.33	75.01
	TriDet	1s	68.18	**78.58**	70.36	81.89	**80.46**	**78.47**	**74.59**	**68.91**	**76.86**
	Shallow D.	2s	78.00	77.19	75.43	65.14	63.58	60.98	59.15	57.39	61.25
	A-and-D	2s	**80.96**	78.42	**77.10**	63.52	61.57	58.82	56.88	54.41	59.04
	ActionFormer	2s	61.56	76.63	64.57	80.14	77.24	74.41	70.24	63.18	73.04
	TriDet	2s	63.14	75.64	65.94	78.69	77.02	73.89	70.68	65.31	73.12
Camera	ActionFormer	0.5s	60.86	72.98	62.54	81.07	78.48	72.13	57.34	41.27	66.06
	TriDet	0.5s	64.69	72.89	64.84	80.46	76.95	72.62	64.46	55.52	70.00
	ActionFormer	1s	65.82	75.34	66.40	**85.19**	83.31	80.94	77.22	69.96	79.32
	TriDet	1s	**67.42**	75.21	**66.88**	84.88	**83.49**	**82.12**	**79.75**	**76.24**	**81.30**
	ActionFormer	2s	63.40	**76.25**	65.39	84.93	82.95	80.58	77.55	72.07	79.62
	TriDet	2s	65.53	75.36	65.91	82.94	82.00	80.22	78.13	75.20	79.70
Inertial + Camera	ActionFormer	0.5s	71.45	**83.77**	74.51	86.04	83.96	77.80	64.41	43.76	71.19
	TriDet	0.5s	**75.74**	82.79	**76.83**	84.45	82.07	77.12	68.81	57.16	73.92
	ActionFormer	1s	72.87	83.24	75.26	**86.82**	85.27	82.46	78.26	72.33	81.03
	TriDet	1s	73.39	82.00	75.09	86.42	**85.48**	**83.35**	**80.39**	**75.64**	**82.26**
	ActionFormer	2s	65.86	80.17	68.82	84.18	81.66	78.17	74.59	68.74	77.47
	TriDet	2s	69.19	81.09	71.98	84.66	82.79	80.66	78.11	74.15	80.07
	O-LF(I, C)	*0.5s*	*90.23*	*91.52*	*89.60*	*74.79*	*73.73*	*72.44*	*69.94*	*68.53*	*71.88*
	O-LF(I, C)	*1s*	*91.26*	*92.72*	*90.68*	*75.37*	*74.42*	*73.64*	*72.74*	*71.41*	*73.52*
	O-LF(I, C)	*2s*	*91.26*	*92.94*	*90.81*	*73.46*	*72.37*	*71.70*	*70.50*	*69.60*	*71.53*
	O-LF(I, C, I + C)	*0.5s*	*93.29*	*94.37*	*93.10*	*84.41*	*84.17*	*83.37*	*81.61*	*80.40*	*82.79*
	O-LF(I, C, I + C)	*1s*	*93.99*	*94.89*	*93.38*	*84.14*	*83.81*	*83.34*	*82.54*	*81.81*	*83.13*
	O-LF(I, C, I + C)	*2s*	*93.69*	*95.13*	*93.38*	*81.79*	*81.37*	*81.02*	*79.90*	*78.94*	*80.60*

four previously unseen participants as well as two participants, already present in the training data and which volunteered to be recorded a second time. Unlike recordings of the training data, the test data was recorded during spring and summer. Egocentric video data of the four new participants was recorded using a GoPro Hero 11 as opposed to a Hero 8 and two new participants were recorded at a previously unseen location. Figure 6 summarizes results of the benchmark algorithms applied on the test dataset using a 1 second sliding window with a 50% overlap. Training was performed identical to that described in previous chapters with the input data

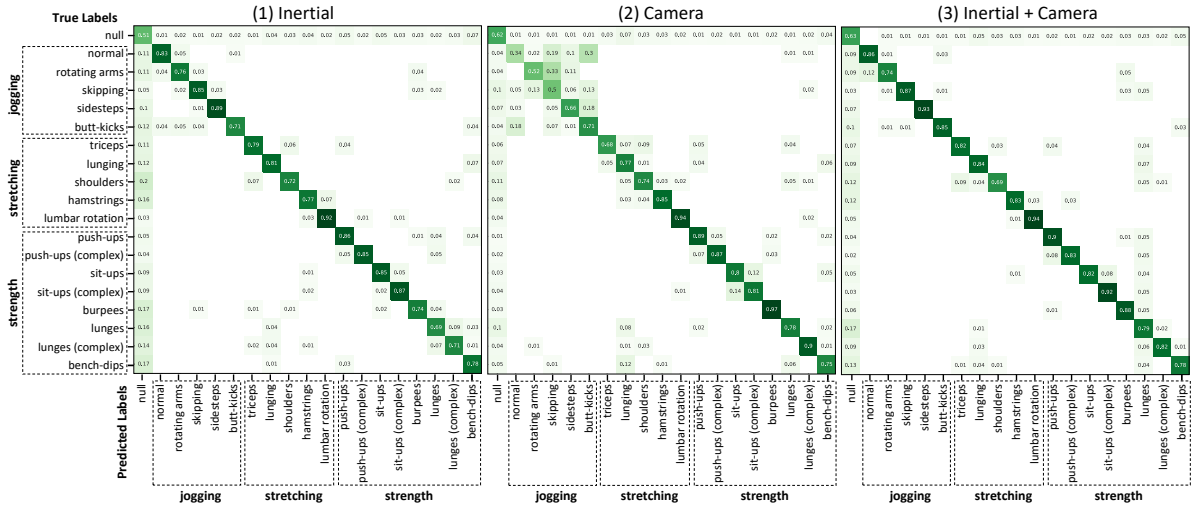

Fig. 4. Confusion matrices of the TriDet model [58] being applied using inertial, vision (camera) and both combined (inertial + camera) with a one second sliding window and 50% overlap. Compared to inertial-based architectures [1, 5] overall confusion (except for the NULL-class) is decreased. After combination strengths of each architecture are leveraged with e.g. jogging activities not getting confused anymore and overall confusion with the NULL-class decreases. Note that confusions which are 0 are omitted.

being the data of the 18 participants used for LOSO validation. Each experiment was repeated three times with different random seeds (i.e., 1, 2 or 3). One can see that observed trends are similar to those seen during LOSO cross-validation. Nevertheless, one can observe a increase in average mAP and classification metrics for both the inertial- and vision-based models, which does also translate to mAP and class increasing when combining camera and inertial data.

5 ABLATION EXPERIMENTS

5.1 Selection of Inertial Sensors

Although commercial fitness products worn on the ankle or foot (e.g., anklets, sock-embedded devices, or shoe-embedded units) have gained popularity in recent years, wrist-worn sensors like smartwatches remain the most popular body position for inertial sensors in the context of wearable activity recognition. The following experiments (see Figure 7) investigate how reported LOSO results in Chapter 4.2 and 4.3 are influenced by using only a subset of the inertial sensors, specifically by using only (1) acceleration recorded from the right wrist and (2) acceleration recorded from both the right wrist and right ankle. Results show that using only acceleration data obtained from the right wrist significantly decreases predictive performance across all algorithms and metrics. Moreover, the value of additionally measuring acceleration at the ankles of participants is clearly underlined, as results again significantly increase, mostly on par compared to using all four inertial sensor locations. Interestingly, unlike the inertial-based architectures, results of the vision-based TAL models improve when excluding data captured by the left wrist and left ankle inertial sensors, which could be due to the dataset being biased towards right-handed participants (see Table 3 in the supplementary material) and dominant hand movement being overall more consistent.

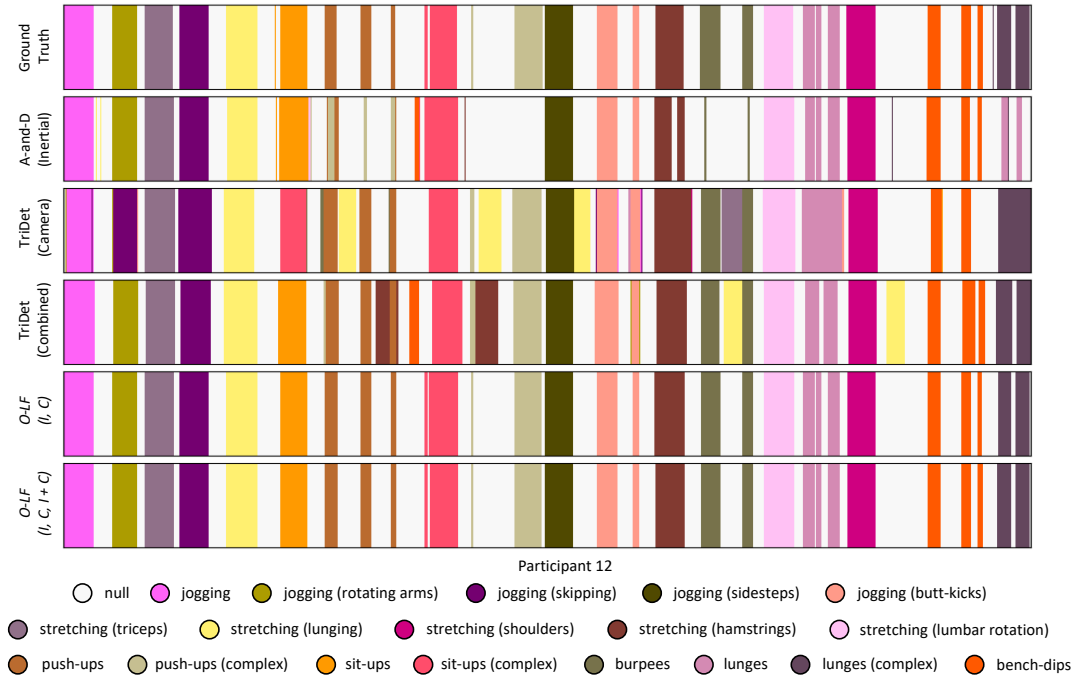

Fig. 5. Color-coded comparison of the ground truth data of a sample participant with the best inertial-based (A-and-D), camera-based (TriDet) and fusion-based model (TriDet) along with an oracle combination of the best fusion-based model (*O-LF(I, C)*) as well as an oracle combination the best camera, inertial and fusion-based-model (*O-LF(I, C, I + C)*) using a sliding window approach of 1.0 seconds with a 50% overlap. The visualisation underlines the similarities amongst the predictive streams of *O-LF(I, C)* and the fusion-approach as well advantages of learning from both modalities simultaneously.

5.2 Second execution of workout sessions

As mentioned in Chapter 3.1, as part of the test dataset provided along the WEAR dataset, we recorded all activities of two participants (sbj_0 and sbj_14) a second time in August. Both participants recording conditions significantly differed from their first recording, with temperatures being around 25 degrees Celsius with overall more sunny weather conditions. Further, as not all participants knew all activities beforehand (see Table 10 in the supplementary material), recording the same participants a second time would allow to analyse how a certain degree of familiarity with the recording setup can be seen in altered movements (e.g., via a smoother execution of activities) as well as participant-specific finetuning affects the overall recognition performance. Figure 8 compares validation results obtained on the first recording of sbj_0 and sbj_14 with their second execution of the workout plan in August. Unlike our prior experiments, each algorithm is trained using the validation data as reported in Chapter 4.2 and 4.3 except the participant which is to be analysed. All results are postprocessed as reported Chapter 4.2 and 4.3. While, one can see improved results regarding sbj_0, which only knew five of the workout activites prior to participating in the study, this trend does not apply to sbj_14. More specifcially, improvements and decline rates between the two recordings lie within the expected standard deviation across participants (between 15% to 20%). Though being a small sample size of only two participants, the results suggest that in order to guarantee a reliable detection of activities, each participant would need to be recorded multiple times under

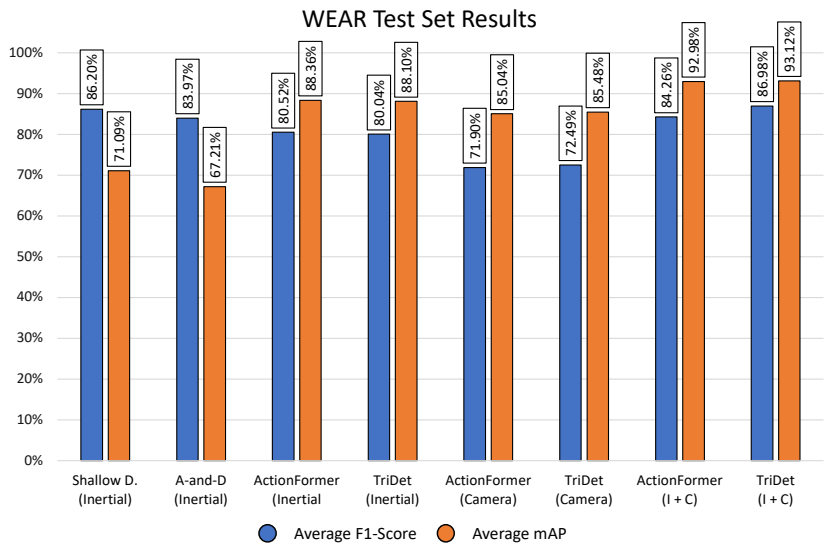

Fig. 6. Average F1-score and mAP on the WEAR test set. The test features 4 unseen participants as well 2 reoccuring ones, a unseen location, different weather conditions and a new camera sensor. One can see that observed trends are similar to those seen during LOSO cross-validation.

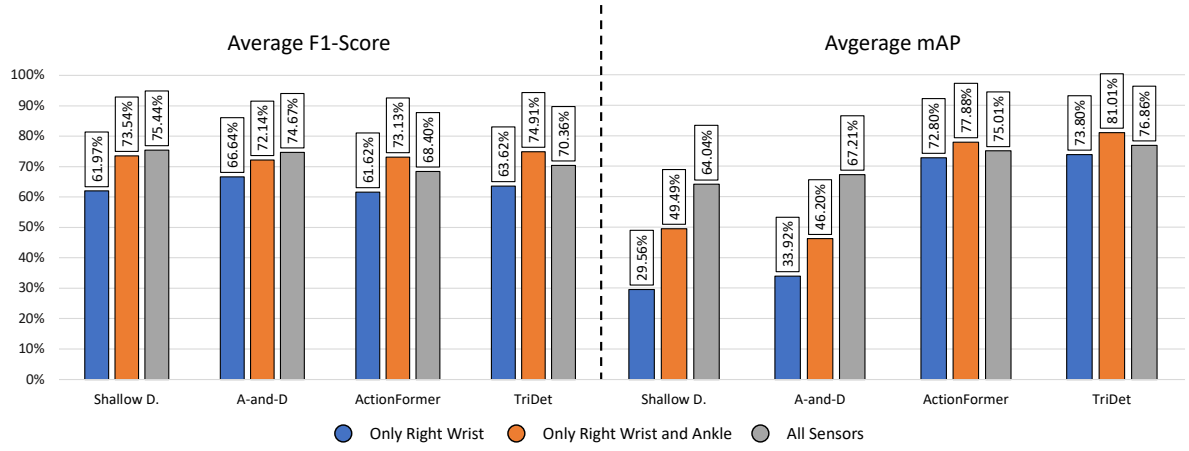

Fig. 7. Results using only inertial sensors placed on the right wrist (RW) and right wrist + ankle (RW + RA) compared with using all sensors for a clip length of 1 second on our WEAR dataset evaluated in terms of F1-score and mean average precision (mAP) averaged across different temporal intersection over union (tIoU) thresholds (0.3:0.7:0.1). Using only wrist-worn data can see a clear overall decrease across all evaluation metrics.

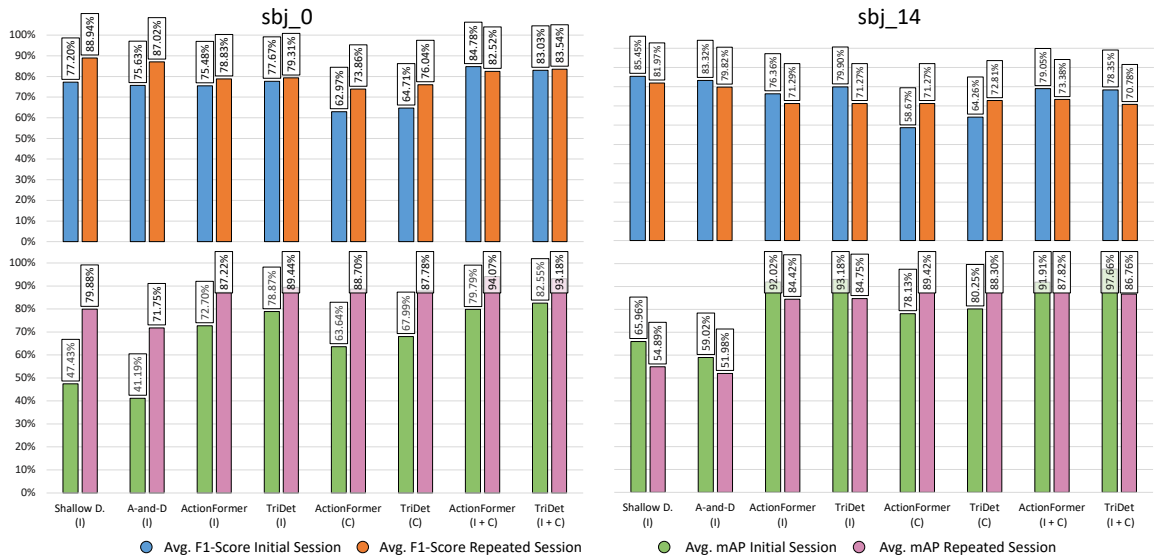

Fig. 8. Comparison of obtained results of repeated sessions for participants sbj_0 and sbj_14 for a clip length of 1 second on our WEAR dataset evaluated in terms of F1-score and mean average precision (mAP). The two participants were invited to perform the recording plan a second time. While one can see that improved results regarding sbj_0, suggesting potential learning effects of the correct execution of activities, this trend does not apply to sbj_14. Note that weather conditions (temperature and sunlight) significantly differ amongst the recordings – winter (first recording) compared to summer (2nd recording). These figures are, as in the earlier results, averaged across 3 runs using 3 different random seeds. Unlike our prior experiments, each algorithm is trained using the training data of all but the validation participants' recordings, ensuring the validation participants (sbj_0 and sbj_14) remain unseen during the training of each algorithm. All results are postprocessed as reported in the main paper.

different conditions. Nevertheless, in order to come up with reliable conclusions, future extensions of the WEAR dataset would need focus on re-recording more participants multiple times under varying conditions.

6 LIMITATIONS

Our dataset contributes a benchmark for human activity recognition classifiers, for the two leading wearable modalities of egocentric video and inertial data, using in particular a high variety of fitness exercises and outdoor scenes. With the current selection of participants, the WEAR dataset is biased towards young, healthy people. Given the ease of reproducibility, future extensions of the WEAR dataset could focus on featuring participants (1) of an older and/ or younger age, (2) with known physical impairments and (3) sessions recorded at new locations (outside of Germany) and at different times of the year (e.g. summer). As supplementary experiments already indicate, recording the same participants a second time would allow to analyse how a certain degree of familiarity with the recording setup can be seen in altered movements (e.g., via a smoother execution of activities) as well as give an intuition about robustness of learned approaches. Besides extending the amount of data recorded, further recordings could also involve other sensors such as higher-end commercial smartwatches to enable the study of increased sampling rates, the variability of the capturing devices, and the inclusion of additional modalities such as 3D gyroscopes, 3D magnetometers, or photoplethysmography (PPG) to obtain fitness-relevant information such as heart rate), as well as additional wearables, such as earables. Furthermore, as participants remain at the

same location during their workout, including recordings of the same participants at different locations would allow to assess classifiers based on their ability to learn general cues to predict the workout activities rather than overfitting on location-specific ones. Lastly, as a camera observing the participant is needed for annotation of the collected data, a future extension of the WEAR dataset could include releasing the 3rd-person video data of the workouts along with the egocentric data, enabling the dataset to be also used e.g. for pose estimation.

7 DISCUSSION & CONCLUSION

In this paper, we introduced a benchmark dataset for both inertial- and vision-based Human Actvity Recognition (HAR), to explore the learning of HAR across these modalities. The dataset comprises data from 22 participants performing each 18 different sports activities with the two common types of wearable sensors delivering inertial (3D acceleration) and camera (egocentric video) data. Unlike previous egocentric datasets, the application-driven sensor setup of WEAR provides a challenging prediction scenario across both modalities marked by changing outdoor environments along with a small information overlap between the inertial and vision data, putting forward the necessity of exploring techniques to combine both modalities.

Benchmark results obtained using each modality separately show that each modality interestingly offers complementary strengths and weaknesses in their prediction performance. In light of the recent successes of single-stage TAL models following the architecture design as proposed by Zhang et al. [81], we demonstrate their versatility by applying them in a plain fashion without any pretraining using preextracted I3D features and raw inertial data as input. Using the inherent similarities of approaches of both communities, we suggest a simple, yet effective preprocessing technique of inertial data, which vectorizes sliding window such that they can be used to train TAL models. Vision-based TAL such as the ActionFormer [81] have thus far neither been explored in inertial nor in the combination of inertial- and vision-based human activity recognition. Results show that the vision-based models are not only able to produce competitive results using inertial data, but also can function as an architecture to fuse both modalities by means of simple concatenation with vision data. In experiments that combined raw inertial with extracted vision-based feature embeddings, the plain, vision-based TAL models were able to produce the highest average mAP and close-to-best F1-scores. Lastly, to give an intuition about a possible upper bound for future fusion-approaches, we evaluated an oracle-merged late fusion of the best inertial- and vision-based model predictions.

With WEAR, we provide both communities (inertial- and vision-based HAR) a common, challenging benchmark dataset to assess the applicability of combined human activity recognition approaches.

ACKNOWLEDGMENTS

We gratefully acknowledge the DFG Project WASEDO (grant number 506589320) and the University of Siegen's OMNI cluster.

REFERENCES

[1] Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. 2021. Attend and Discriminate: Beyond the State-Of-The-Art for Human Activity Recognition Using Wearable Sensors. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–22. https://doi.org/10.1145/3448083

[2] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. 2020. Boundary Content Graph Neural Network for Temporal Action Proposal Generation. In *European Conference on Computer Vision*. https://doi.org/10.1007/978-3-030-58604-1_8

[3] Ling Bao and Stephen S. Intille. 2004. Activity Recognition From User-Annotated Acceleration Data. *Pervasive Computing* (2004), 158–175. https://doi.org/10.1007/978-3-540-24646-6_1

[4] Mehrab Bin Morshed, Harish Kashyap Haresamudram, Dheeraj Bandaru, Gregory D. Abowd, and Thomas Ploetz. 2022. A Personalized Approach for Developing a Snacking Detection System using Earbuds in a Semi-Naturalistic Setting. In *ACM International Symposium on Wearable Computers*. https://doi.org/10.1145/3544794.3558469

[5] Marius Bock, Alexander Hoelzemann, Michael Moeller, and Kristof Van Laerhoven. 2021. Improving Deep Learning for HAR With Shallow Lstms. In *ACM International Symposium on Wearable Computers*. https://doi.org/10.1145/3460421.3480419

[6] Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors. *Comput. Surveys* 46, 3 (2014), 1–33. https://doi.org/10.1145/2499621

[7] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2017.502

[8] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2016. Fusion of Depth, Skeleton, and Inertial Data for Human Action Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/ICASSP.2016.7472170

[9] Guo Chen, Yin-Dong Zheng, Limin Wang, and Tong Lu. 2022. DCAN: Improving Temporal Action Detection via Dual Context Aggregation. In *AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v36i1.19900

[10] Feng Cheng and Gedas Bertasius. 2022. TallFormer: Temporal Action Localization With a Long-Memory Transformer. In *European Conference on Computer Vision*. https://doi.org/10.1007/978-3-031-19830-4_29

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision* (2022). https://doi.org/10.1007/s11263-021-01531-2

[12] de la Torre, Fernando, Jessica K. Hodgins, Javier Montano, and Sergio Valcarcel. 2009. Detailed Human Data Acquisition of Kitchen Activities: The CMU-Multimodal Activity Database (CMU-MMAC). In *Conference on Human Factors in Computing Systems*. http://www.cs.cmu.edu/~ftorre/web_page/humansensing.cs.cmu.edu/projects/CMU-MMAC.html

[13] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. 2022. ActionSense: A Multimodal Dataset and Recording Framework for Human Activities Using Wearable Sensors in a Kitchen Environment. In *Neural Information Processing Systems Track on Datasets and Benchmarks*. https://action-sense.csail.mit.edu

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Confernce of the North American Chapter of the Association for Computational Linguistics*. https://arxiv.org/abs/1810.04805

[15] Alexander Diete and Heiner Stuckenschmidt. 2019. Fusing Object Information and Inertial Data for Activity Recognition. *MDPI Sensors* 19, 19 (2019). https://doi.org/10.3390/s19194119

[16] Alexander Diete, Timo Sztyler, and Heiner Stuckenschmidt. 2019. Vision and Acceleration Modalities: Partners for Recognizing Complex Activities. In *IEEE International Conference on Pervasive Computing and Communications Workshops*. https://doi.org/10.1109/PERCOMW.2019.8730690

[17] Alexander Diete, Timo Sztyler, Lydia Weiland, and Heiner Stuckenschmidt. 2018. Improving Motion-Based Activity Recognition With Ego-Centric Vision. In *IEEE International Conference on Pervasive Computing and Communications Workshops*. https://doi.org/10.1109/PERCOMW.2018.8480334

[18] Muhammad Ehatisham-Ul-Haq, Ali Javed, Muhammad Awais Azam, Hafiz M. A. Malik, Aun Irtaza, Ik Hyun Lee, and Muhammad Tariq Mahmood. 2019. Robust Human Activity Recognition Using Multimodal Feature-Level Fusion. *IEEE Access: Practical innovations, Open Solutions* 7 (2019), 60736–60751. https://doi.org/10.1109/ACCESS.2019.2913393

[19] Ziqi Gao, Yuntao Wang, Jianguo Chen, Junliang Xing, Shwetak Patel, Xin Liu, and Yuanchun Shi. 2023. MMTSA: Multi-Modal Temporal Segment Attention Network for Efficient Human Activity Recognition. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–26. https://doi.org/10.1145/3610872

[20] Rohit Girdhar and Kristen Grauman. 2021. Anticipative Video Transformer. In *IEEE/CVF International Conference on Computer Vision*. https://doi.org/10.1109/iccv48922.2021.01325

[21] Guoqiang Gong, Liangfeng Zheng, and Yadong Mu. 2020. Scale matters: Temporal scale aggregation network for precise action localization in untrimmed videos. In *IEEE International Conference on Multimedia and Expo*. https://doi.org/10.1109/ICME46284.2020.9102850

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, and Xingyu Liu. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR52688.2022.01842

[23] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, and et al. 2024. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[24] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition Using Wearables. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–28. https://doi.org/10.1145/3090076

[25] Nils Y. Hammerla, Shane Halloran, and Thomas Ploetz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables. In *Twenty-Fifth International Joint Conference on Artificial Intelligence*. https://dl.acm.org/doi/10.5555/3060832.3060835

[26] Fabian Caba Heilbron, Juan Carlos Niebles, Victor Escorcia, and Bernard Ghanem. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2015.7298698

[27] Menghao Hu, Mingxuan Luo, Menghua Huang, Wenhua Meng, Baochen Xiong, Xiaoshan Yang, and Jitao Sang. 2023. Towards a Multimodal Human Activity Dataset for Healthcare. *Multimedia Systems* 29, 1 (2023), 1–13. https://doi.org/10.1007/s00530-021-00875-6

[28] Momal Ijaz, Renato Diaz, and Chen Chen. 2022. Multimodal Transformer for Nursing Activity Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. https://doi.org/10.1109/cvprw56347.2022.00224

[29] Javed Imran and Balasubramanian Raman. 2020. Evaluating Fusion of RGB-D and Inertial Sensors for Multimodal Human Action Recognition. *Journal of Ambient Intelligence and Humanized Computing* 11, 1 (2020), 189–208. https://doi.org/10.1007/s12652-019-01239-9

[30] Javed Imran and Balasubramanian Raman. 2020. Multimodal Egocentric Activity Recognition Using Multi-Stream CNN. In *11th Indian Conference on Computer Vision, Graphics and Image Processing*. https://doi.org/10.1145/3293353.3293363

[31] Md Mofijul Islam and Tariq Iqbal. 2021. Multi-GAT: A Graphical Attention-Based Hierarchical Multimodal Representation Learning Approach for Human Activity Recognition. *IEEE Robotics and Automation Letters* 6, 2 (2021), 1729–1736. https://doi.org/10.1109/LRA.2021.3059624

[32] Md Mofijul Islam and Tariq Iqbal. 2022. MuMu: Cooperative Multitask Learning-Based Guided Multimodal Fusion. In *AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v36i1.19988

[33] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. 2014. THUMOS challenge: Action Recognition With a Large Number of Classes. http://crcv.ucf.edu/THUMOS14/

[34] Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and Understanding Recurrent Networks. *CoRR* abs/1506.02078 (2015). http://arxiv.org/abs/1506.02078

[35] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *CoRR* abs/1705.06950 (2017). http://arxiv.org/abs/1705.06950

[36] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Ninth International Conference on Learning Representations*. https://arxiv.org/abs/2010.11929

[37] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *IEEE/ CVF Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr52688.2022.00476

[38] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. 2020. Fast Learning of Temporal Action Proposal via Dense Boundary Generator. In *IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1609/aaai.v34i07.6815

[39] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. 2021. Learning Salient Boundary Feature for Anchor-Free Temporal Action Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr46437.2021.00333

[40] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In *IEEE/CVF International Conference on Computer Vision*. https://doi.org/10.1109/iccv.2019.00399

[41] Qinying Liu and Zilei Wang. 2020. Progressive Boundary Refinement Network for Temporal Action Detection. In *AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v34i07.6829

[42] Xiaolong Liu, Song Bai, and Xiang Bai. 2022. An Empirical Study of End-To-End Temporal Action Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr52688.2022.01938

[43] Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, and Philip H. S. Torr. 2021. Multi-Shot Temporal Event Localization: A Benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr46437.2021.01241

[44] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. 2022. End-To-End Temporal Action Detection With Transformer. *IEEE Transactions on Image Processing* 31 (2022). https://doi.org/10.1109/TIP.2022.3195321

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *IEEE/CVF International Conference on Computer Vision*. https://doi.org/10.1109/iccv48922.2021.00986

[46] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. 2019. Gaussian Temporal Awareness Networks for Action Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2019.00043

[47] Yantao Lu and Senem Velipasalar. 2018. Human Activity Classification Incorporating Egocentric Video and Inertial Measurement Unit Data. In *IEEE Global Conference on Signal and Information Processing*. https://doi.org/10.1109/GlobalSIP.2018.8646367

[48] Vishvak S. Murahari and Thomas Plötz. 2018. On Attention Models for Human Activity Recognition. In *ACM International Symposium on Wearable Computers*. https://doi.org/10.1145/3267242.3267287

[49] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. 2022. Proposal-Free Temporal Action Detection via Global Segmentation Mask Learning. In *European Conferencee on Computer Vision*. https://doi.org/10.1007/978-3-031-20062-5_37

[50] Katsuyuki Nakamura, Serena Yeung, Alexandre Alahi, and Li Fei-Fei. 2017. Jointly Learning Energy Expenditures and Activities Using Egocentric Multimodal Signals. In *IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2017.721

[51] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *MDPI Sensors* 16, 1 (2016). https://doi.org/10.3390/s16010115

[52] Donald J. Patterson, Dieter Fox, Henry Kautz, and Matthai Philipose. 2005. Fine-Grained Activity Recognition by Aggregating Abstract Object Usage. In *Ninth IEEE International Symposium on Wearable Computers*. https://doi.org/10.1109/ISWC.2005.22

[53] Rafael Possas, Sheila Pinto Caceres, and Fabio Ramos. 2018. Egocentric Activity Recognition on a Budget. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2018.00625

[54] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. 2021. Temporal Context Aggregation Network for Temporal Action Proposal Refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr46437.2021.00055

[55] Valentin Radu and Maximilian Henne. 2019. Vision2Sensor: Knowledge Transfer Across Sensing Modalities for Human Activity Recognition. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–21. https://doi.org/10.1145/3351242

[56] Debaditya Roy and Basura Fernando. 2022. Predicting the Next Action by Modeling the Abstract Goal. *CoRR* abs/2209.05044 (2022). https://doi.org/10.48550/arxiv.2209.05044

[57] Philipp M. Scholl, Benjamin Völker, Bernd Becker, and Kristof Van Laerhoven. 2019. A Multi-Media Exchange Format for Time-Series Dataset Curation. In *Human Activity Sensing*, Nobuo Kawaguchi, Nobuhiko Nishio, Daniel Roggen, Sozo Inoue, Susanna Pirttikangas, and Kristof Van Laerhoven (Eds.). Springer, 111–119. https://doi.org/10.1007/978-3-030-13001-5_8

[58] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. 2023. TriDet: Temporal Action Detection With Relative Boundary Modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr52729.2023.01808

[59] Dingfeng Shi, Yujie Zhong, Qiong Cao, Jing Zhang, Lin Ma, Jia Li, and Dacheng Tao. 2022. ReAct: Temporal Action Detection With Relational Queries. In *European Conference on Computer Vision*. https://doi.org/10.1007/978-3-031-20080-9_7

[60] Sibo Song, Vijay Chandrasekhar, Bappaditya Mandal, Liyuan Li, Joo-Hwee Lim, Giduthuri Sateesh Babu, Phyo Phyo San, and Ngai-Man Cheung. 2016. Multimodal Multi-Stream Deep Learning for Egocentric Activity Recognition. In *IEEE conference on computer vision and pattern recognition workshops*. https://doi.org/10.1109/CVPRW.2016.54

[61] Sibo Song, Ngai-Man Cheung, Vijay Chandrasekhar, Bappaditya Mandal, and Jie Lin. 2016. Egocentric Activity Recognition With Multimodal Fisher Vector. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. https://doi.org/10.1109/icassp.2016.7472171

[62] Ekaterina H. Spriggs, Fernando De La Torre, and Martial Hebert. 2009. Temporal Segmentation and Activity Classification From First-Person Sensing. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. https://doi.org/10.1109/CVPRW.2009.5204354

[63] Deepak Sridhar, Niamul Quader, Srikanth Muralidharan, Yaoxin Li, Peng Dai, and Juwei Lu. 2021. Class Semantics-Based Attention for Action Detection. In *IEEE/ CVF International Conference on Computer Vision*. https://doi.org/10.1109/iccv48922.2021.01348

[64] David Strömbäck, Sangxia Huang, and Valentin Radu. 2020. MM-Fit: Multimodal Deep Learning for Automatic Exercise Logging across Sensing Devices. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22. https://doi.org/10.1145/3432701

[65] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. 2021. Relaxed Transformer Decoders for Direct Action Proposal Generation. In *IEEE/CVF International Conference on Computer Vision*. https://doi.org/10.1109/iccv48922.2021.01327

[66] Kristof Van Laerhoven, Alexander Hoelzemann, Iris Pahmeier, Andrea Teti, and Lars Gabrys. 2022. Validation of an Open-Source Ambulatory Assessment System in Support of Replicable Activity Studies. *German Journal of Exercise and Sport Research* 52, 2 (2022), 262–272. https://doi.org/10.1007/s12662-022-00813-2

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1706.03762

[68] Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. ActionCLIP: A New Paradigm for Video Action Recognition. *CoRR* abs/2109.08472 (2021). https://arxiv.org/abs/2109.08472

[69] Jamie A. Ward, Paul Lukowicz, Gerhard Tröster, and Thad E. Starner. 2006. Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2006), 1553–1567. https://doi.org/10.1109/TPAMI.2006.197

[70] Haoran Wei and Nasser Kehtarnavaz. 2020. Simultaneous Utilization of Inertial and Video Sensing for Action Detection and Recognition in Continuous Action Streams. *IEEE Sensors Journal* 20, 11 (2020), 6055–6063. https://doi.org/10.1109/JSEN.2020.2973361

[71] Tz-Ying Wu, Ting-An Chien, Cheng-Sheng Chan, Chan-Wei Hu, and Min Sun. 2017. Anticipating Daily Intention Using On-Wrist Motion Triggered Sensing. In *IEEE International Conference on Computer Vision*. https://doi.org/Âť

[72] Rui Xi, Mengshu Hou, Mingsheng Fu, Hong Qu, and Daibo Liu. 2018. Deep Dilated Convolution on Multimodality Time Series for Human Activity Recognition. In *IEEE International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN.2018.8489540

[73] Cheng Xu, Duo Chai, Jie He, Xiaotong Zhang, and Shihong Duan. 2019. InnoHAR: A Deep Neural Network for Complex Human Activity Recognition. *IEEE Access* 7 (2019). https://doi.org/10.1109/ACCESS.2018.2890675

[74] Linfeng Xu, Qingbo Wu, Lili Pan, Fanman Meng, Hongliang Li, Chiyuan He, Hanxin Wang, Shaoxu Cheng, and Yu Dai. 2023. Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning. *CoRR* abs/2301.10931 (2023). https://arxiv.org/abs/2301.10931

[75] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. 2020. G-TAD: Sub-Graph Localization for Temporal Action Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr42600.2020.01017

[76] Min Yang, Guo Chen, Yin-Dong Zheng, Tong Lu, and Limin Wang. 2023. BasicTAD: An Astounding RGB-Only Baseline for Temporal Action Detection. *Computer Vision and Image Understanding* 232 (2023). https://doi.org/10.1016/j.cviu.2023.103692

[77] Haibin Yu, Guoxiong Pan, Mian Pan, Chong Li, Wenyan Jia, Li Zhang, and Mingui Sun. 2019. A Hierarchical Deep Fusion Framework for Egocentric Activity Recognition Using a Wearable Hybrid Sensor System. *MDPI Sensors* 19, 3 (2019). https://doi.org/10.3390/s19030546

[78] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-To-Token ViT: Training Vision Transformers From Scratch on Imagenet. In *IEEE/CVF International Conference on Computer Vision*. https://doi.org/10.1109/iccv48922.2021.00060

[79] Yuta Yuki, Junto Nozaki, Kei Hiroi, Katsuhiko Kaji, and Nobuo Kawaguchi. 2018. Activity Recognition Using Dual-Convlstm Extracting Local and Global Features for Shl Recognition Challenge. In *ACM International Joint Conference and International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. https://doi.org/10.1145/3267305.3267533

[80] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2019. Graph Convolutional Networks for Temporal Action Localization. In *IEEE/CVF International Conference on Computer Vision*. https://doi.org/10.1109/iccv.2019.00719

[81] Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing Moments of Actions With Transformers. In *European Conference on Computer Vision*. https://doi.org/10.1007/978-3-031-19772-7_29

[82] Chen Zhao, Ali K. Thabet, and Bernard Ghanem. 2021. Video Self-Stitching Graph Network for Temporal Action Localization. In *IEEE/CVF International Conference on Computer Vision*. https://doi.org/10.1109/iccv48922.2021.01340

[83] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. 2020. Bottom-up Temporal Action Localization With Mutual Regularization. In *European Conference on Computer Vision*. https://doi.org/10.1007/978-3-030-58598-3_32

[84] Yexu Zhou, Haibin Zhao, Yiran Huang, Till Riedel, Michael Hefenbrock, and Michael Beigl. 2022. TinyHAR: A Lightweight Deep Learning Model Designed for Human Activity Recognition. In *ACM International Symposium on Wearable Computers*. https://doi.org/10.1145/3544794.3558467

[85] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. 2021. Enriching Local and Global Contexts for Temporal Action Localization. In *IEEE/CVF International Conference on Computer Vision*. https://doi.org/10.1109/iccv48922.2021.01326

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009