Supplemental Material: 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

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

A METHODOLOGICAL TRANSPARENCY APPENDIX & LICENSE

The source code that was used to conduct all experiments, postprocessing and evaluation steps mentioned in this paper is available via Github (mariusbock.github.io/wear/). The repository is written in such a way that other architectures (both inertial- and vision-based) can be added in the future.The repository provides Readme files which give details on the overall structure of the repository, how collect additional data and how to set up an Anaconda environment with the needed packages to run experiments. Experiments are defined via '.json'-format configuration files which allow for easy sharing of used hyperparameter settings. Experiments were conducted on a cluster-based system, with each run being assigned one Tesla V100 GPU, 10 GB of RAM and two AMD EPYC 7452 CPUs.

WEAR and all associated files are offered under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The dataset is hosted via the cloud-storage platform sciebo, which is a service hosted by hochschulcloud.nrw (https://hochschulcloud.nrw). It is a non-commercial cloud storage service for research, studying and teaching and is provided to participating institutions exclusively. With locations exclusively in North Rhine-Westphalia, sciebo is subject to the strict German directives on data protection and data security. The complete dataset can be downloaded via sciebo (https://bit.ly/wear_dataset). The dataset download is structured into the (1) '.json'-formatted annotations, (2) raw, synchronized inertial and vision data and (3) precomputed feature embeddings as mentioned in the main paper.

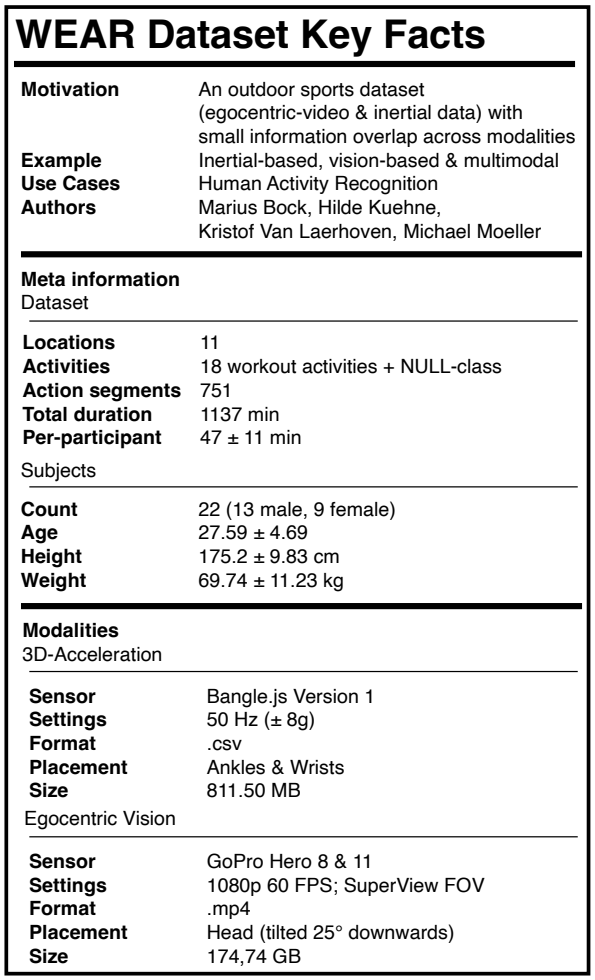
B DATASET OVERVIEW AND CONTENTS

The outdoor sports dataset WEAR features data of 22 participants performing each a total of 18 different workout activities with untrimmed inertial (acceleration) and camera (egocentric video) data recorded at 11 different outside locations. It provides a challenging prediction scenario marked by purposely introduced activity variations and an overall small information overlap across modalities. Figure 1 provides a dataset nutrition label inspired by Holland et al. [5] in a table-like manner.

B.1 Intended Uses and Ethical Considerations

Before participating in the study, participants were notified that by nature the data they provide can only be pseudonymised. This means that, though requiring a substantial amount of effort, the identity of a person can be

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.



WEAR Dataset Key Facts

Motivation	An outdoor sports dataset (egocentric-video & inertial data) with small information overlap across modalities
Example	Inertial-based, vision-based & multimodal
Use Cases	Human Activity Recognition
Authors	Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, Michael Moeller

Meta information

Dataset

Locations	11
Activities	18 workout activities + NULL-class
Action segments	751
Total duration	1137 min
Per-participant	47 ± 11 min

Subjects

Count	22 (13 male, 9 female)
Age	27.59 ± 4.69
Height	175.2 ± 9.83 cm
Weight	69.74 ± 11.23 kg

Modalities

3D-Acceleration

Sensor	Bangle.js Version 1
Settings	50 Hz (± 8g)
Format	.csv
Placement	Ankles & Wrists
Size	811.50 MB

Egocentric Vision

Sensor	GoPro Hero 8 & 11
Settings	1080p 60 FPS; SuperView FOV
Format	.mp4
Placement	Head (tilted 25° downwards)
Size	174,74 GB

Fig. 1. Dataset nutrition label of the WEAR dataset. The dataset nutrition label was originally proposed by Holland et al. [5]. Our adaptation is inspired by DelPreto et al. [4].

reconstructed. Although participants agreed to include their egocentric videos in a public dataset, it is essential to refrain from actively identifying the individuals featured in the WEAR dataset. If other researchers decide to contribute to the WEAR dataset by recording additional participants, societal and ethical implications should be considered. As with the participants part of the original release of the WEAR dataset, all participants must be briefed before their first recording, making them aware of all necessary information and implications that come with providing to the WEAR dataset. Recording locations should only be chosen if video recordings are allowed at said location and participants are given enough space to perform each activity safely. If the recording location involves pedestrians walking within close proximity, pedestrians should be notified that they are being recorded and, if applicable, captured faces should be blurred during postprocessing.

The WEAR dataset and associated code are made public for research purposes. With the accurate detection of physical activities that we perform in our daily lives having been identified as valuable information, the WEAR dataset focuses on one of the most popular application scenarios of wearable smartwatches and action cameras, i.e. self-tracking of workout activities. With the ease of reproducability we hope to make WEAR a collaborative, expanding dataset which researchers from different locations and backgrounds can contribute to. For example, as the current selection of participants is biased towards healthy, young people, we hope to overcome said limitation by including people from more diverse backgrounds and age groups in future iterations of the dataset.

Lastly, the authors took great care of avoiding any infringement of rights during the data collection process. Yet, in case of conflicts, they are of course committed to taking appropriate actions, such as promptly removing data associated with such concerns.

B.2 Class Distribution

Figure 2 gives an overview of the 18 activity classes featured in the WEAR dataset and provides number of coherent sequences as well as total duration per workout activity class.

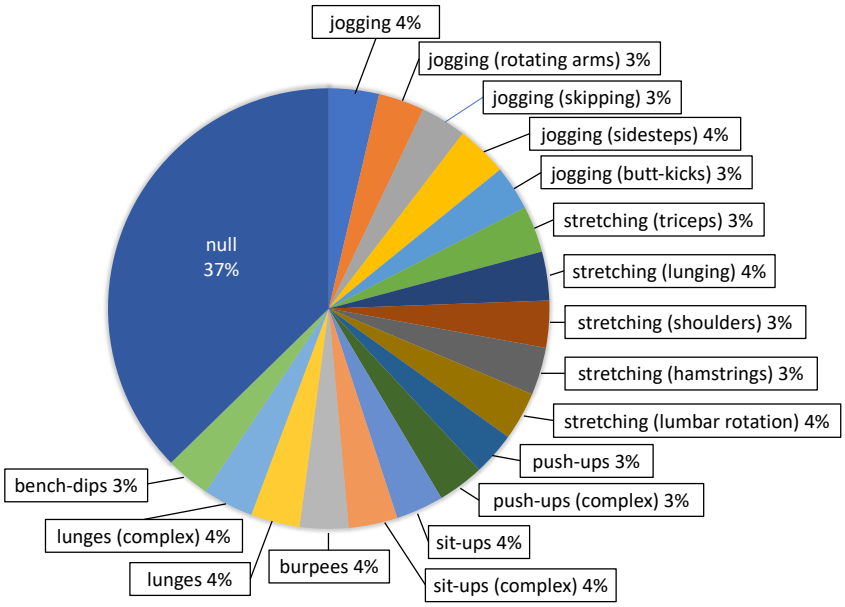


Fig. 2. Overview of the activity classes featured in the WEAR dataset. Percentages are measured relative to the total size of the dataset (around 1,200 minutes). The null class refers to samples not belonging to any of the classes of interest. A detailed description of each activity can be found in the recording plan attached at the end of the supplementary material.

B.3 Participant and Session Information

The location and the time of day at which the sessions were performed were not fixed and thus vary across subjects. As participants were allowed to split activities across more (or less) than two sessions, session counts vary across subjects. Table 1 provides information on all 11 recording locations that are part of the WEAR dataset. The table details general information such as surface conditions of the location as well as which direction the static camera seen in videos is facing. Table 2 provides supplementary information on all separate sessions contained in the dataset. For each session, we detail its overall length in minutes, the number of distinct activities performed by the participant, the location it was recorded at, the month and time of day it was recorded, as well as the overall weather conditions during the duration of the session.

After having completed all sessions, participants were asked to take part in a questionnaire which was used to gather vital information (gender, age, height and weight) as well as workout-specific questions, aiming towards assessing the overall fitness level and experience with the activities detailed in the study protocol. The workout-specific questions were:

(1) How many workouts (longer than 15 min) do you usually do per week?
(2) Which kind of workout do you usually do (cycling, team sport, gym, cardio, yoga etc.)?
(3) How many activities that are part of the workout plan did you know in advance?

Table 1. Description of the 11 locations featured in the WEAR dataset. For each location we provide information on surface conditions, overall surroundings and direction the static camera is facing.

Location ID	Description
1	Meadow in proximity to a larger building. Area is surrounded by trees with from November on-wards, fallen leaves laying on the ground. Static camera faces North-West.
2	Parking lot in proximity to building. Concrete surface. Static camera faces West.
3	Small square with concrete surface. Surrounded by bushes and buildings. Static camera faces West.
4	Meadow enclosed by bungalow-style living quarters. Static camera faces North-East
5	Covered walkway next to a building. Concrete surface. Walkway enclosed by building and bushes. Static camera faces North.
6	Football field with ash surface build behind a supermarket next a road and crop-fields. Long side of the football field is surrounded by bushes. Static camera faces mostly North-West.
7	Backyard in an urban-village with both concrete and grass surface. Terrace has a garden table and chairs standing around. Static camera faces mostly West.
8	Parking lot next to allotments in a city-area. Static camera faces mostly North-East.
9	Meadow next to a building. Static camera faces South.
10	City park in a metropolitan area behind a city mall. Park is surrounded by buildings, a playing ground, football and basketball fields. Static camera faces mostly North.
11	City park in a metropolitan area. Static camera faces mostly South.

(4) How many activities that are part of the workout plan do you perform regularly yourself as part of your own workouts?

Table 3 shows the answers to the questionnaire items for each participant. Note that, to protect the privacy of our study participants, we only asked for age, height and weight in ranges instead of exact values, and always provided the option to not answer the questions if preferred.

Table 2. Per-session meta-information. We provide the individual session count, duration of each session, number of activities performed during the session, location ID (LID) the session was performed at, approximate time of the year and day and weather conditions during recording time. More detailed information on each location can be found in Table 1 using the location ID. * Additional recordings performed by sbj_0 in summer. † Additional recordings performed by sbj_14 in summer.

Participant	Session	Duration	# Activities	Month	Time-of-day	LID	Weather conditions
sbj_0	1	16:33:30	7	mid-Oct.	morning	1	sunny, ≈10°C
sbj_0	2	11:55:00	6	mid-Oct.	afternoon	1	partly-cloudy, ≈10°C
sbj_0	3	18:06:00	7	late-Oct.	afternoon	1	partly-cloudy, ≈20°C
sbj_1	1	20:20:00	9	late-Oct.	afternoon	1	sunny, ≈15°C
sbj_1	2	25:58:00	9	early-Nov.	afternoon	1	sunny, ≈10°C
sbj_2	1	32:24:00	9	early-Nov.	morning	1	sunny, ≈10°C
sbj_2	2	25:08:00	9	mid-Jan.	afternoon	2	cloudy, after rain, ≈0°C
sbj_2	3	01:52:00	1	mid-Feb.	afternoon	3	sunny, ≈5°C
sbj_3	1	33:34:00	10	mid-Nov.	afternoon	4	sunny, ≈5°C
sbj_3	2	25:52:00	6	mid-Nov.	afternoon	4	partly-cloudy, ≈10°C
sbj_3	3	06:24:00	2	mid-Nov.	afternoon	4	sunny, ≈10°C
sbj_3	4	03:41:00	2	late-Jan.	afternoon	5	cloudy, snowy, ≈-5°C
sbj_4	1	24:07:30	9	mid-Nov.	midday	1	foggy, cloudy, windy, ≈5°C
sbj_4	2	29:04:00	9	late-Nov.	afternoon	1	partly-cloudy, ≈5°C
sbj_5	1	19:48:30	9	mid-Nov.	afternoon	1	sunny, ≈10°C
sbj_5	2	16:02:00	9	end-Nov.	afternoon	1	cloudy, ≈5°C
sbj_6	1	23:52:00	10	end-Nov.	afternoon	1	foggy, ≈5°C
sbj_6	2	17:51:30	8	end-Jan.	morning	5	cloudy, snowy, ≈-5°C
sbj_7	1	22:48:00	9	late-Dec.	morning	6	partly-sunny, ≈10°C
sbj_7	2	24:45:00	9	late-Dec.	midday	6	partly-sunny, ≈10°C
sbj_8	1	20:00:00	9	late-Dec.	midday	6	partly-cloudy, ≈10°C
sbj_8	2	21:35:00	9	late-Jan.	afternoon	7	cloudy, ≈0°C
sbj_9	1	18:50:00	9	early-Jan.	afternoon	8	cloudy, ≈10°C
sbj_9	2	17:16:00	9	early-Jan.	afternoon	8	cloudy, ≈10°C
sbj_10	1	21:42:00	9	mid-Jan.	afternoon	5	rainy, windy, ≈5°C
sbj_10	2	21:04:00	9	early-Feb.	afternoon	5	rainy, windy, ≈5°C
sbj_10	3	23:39:00	9	mid-Feb.	afternoon	9, 3	sunny, cloudy, windy, ≈5°C
sbj_11	1	17:41:00	9	mid-Jan.	morning	5	cloudy, rainy, ≈5°C
sbj_11	2	19:21:00	9	mid-Jan.	midday	5	cloudy, rainy, ≈5°C
sbj_12	1	27:08:00	9	mid-Jan.	afternoon	5	cloudy, windy, ≈0°C
sbj_12	2	27:22:00	9	late-Feb.	afternoon	5	partly-sunny, windy, ≈0°C
sbj_13	1	30:08:00	9	mid-Jan.	afternoon	5, 3	sunny, ≈0°C
sbj_13	2	36:10:00	9	mid-Jan.	afternoon	5, 3	sunny, ≈0°C
sbj_14	1	22:18:00	9	mid-Jan.	afternoon	5, 3	sunny, ≈-5°C
sbj_14	2	31:03:00	9	mid-Jan.	afternoon	5, 3	cloudy, ≈-5°C
sbj_15	1	23:17:00	9	late-Jan.	afternoon	5, 3	cloudy, ≈0°C
sbj_15	2	20:06:00	9	late-Jan.	afternoon	5, 3	cloudy, ≈0°C
sbj_16	1	26:34:00	9	early-Feb.	midday	10	partly-sunny, ≈10°C
sbj_16	2	31:56:00	9	early-Feb.	midday	10	partly-sunny, ≈10°C
sbj_17	1	23:16:00	9	early-Feb.	afternoon	1	sunny, ≈0°C
sbj_17	2	28:15:00	9	early-Feb.	afternoon	3	sunny, ≈0°C
sbj_18*	1	19:15:00	9	mid-Aug.	afternoon	1	sunny, ≈25°C
sbj_18*	2	18:06:00	9	mid-Aug.	afternoon	1	sunny, ≈30°C
sbj_19†	1	25:48:00	9	mid-Aug.	afternoon	1	sunny, ≈25°C
sbj_19†	2	30:07:00	9	mid-Aug.	afternoon	1	sunny, ≈30°C
sbj_20	1	19:45:00	9	early-Mar.	afternoon	3	cloudy, ≈5°C
sbj_20	2	24:07:00	9	early-Apr.	afternoon	3	cloudy, ≈10°C
sbj_21	1	18:13:00	9	early-Mar.	midday	9	sunny, ≈15°C
sbj_21	2	16:56:00	9	early-Mar.	midday	9	sunny, ≈15°C
sbj_22	1	16:25:00	9	late-Mar.	midday	11	sunny, ≈10°C
sbj_22	2	15:23:00	9	late-Mar.	midday	11	sunny, ≈10°C
sbj_23	1	22:42:00	9	late-Mar.	midday	11	sunny, ≈10°C
sbj_23	2	19:54:00	9	late-Mar.	midday	11	sunny, ≈10°C

Table 3. Per subject answers to the questionnaire handed to participants after having completed all sessions. The questionnaire collected vital information (gender (G), left- or righthanded (L/R), age, height and weight) as well as workout-specific questions, i.e. frequency and type of private workouts and number of activities, part of the WEAR dataset, which were known in advance and regularly conducted in private workouts. Note that sbj_18 and sbj_19 are excluded from this list as they are identical to sbj_0 and sbj_14 respectively.

Subject	G	L/R	Age	Height	Weight	Private Workouts		Activities	
						Frequency	Type	Known	Regularly
sbj_0	M	R	≥40	180-189	70-79	5	Cycling	5	0
sbj_1	M	R	25-29	170-179	60-69	3	Hiking	11	0
sbj_2	M	R	25-29	180-189	80-89	5	Gym, Cardio	18	9
sbj_3	M	R	35-39	170-179	70-79	4-5	Gym, Basketball, Cardio	18	9
sbj_4	M	R	25-29	180-189	60-69	0	Table-tennis	18	0
sbj_5	F	R	30-34	160 -169	N/A	2-3	Freeletics	16	9
sbj_6	F	R	25-29	150-159	50-59	1	Gym	9	0
sbj_7	M	R	30-34	180-189	80-89	5	Gym, Cardio	18	5
sbj_8	F	R	25-29	170-179	60-69	2-3	Volleyball, Yoga	15	7
sbj_9	F	R	25-29	150-159	50-59	7	Gym, Bicycling, Cardio, Ballet	18	7
sbj_10	F	R	20-24	160-169	50-59	5	Gym, Dancing, Yoga	15	7
sbj_11	F	R	25-29	160-169	50-59	3	Volleyball, Cardio, Yoga	18	11
sbj_12	F	R	20-24	170-179	60-69	4	Gym	17	8
sbj_13	M	R	20-24	≥190	90-99	2	Gym, Cardio	16	8
sbj_14	M	R	30-34	170-179	80-89	0	N/A	11	2
sbj_15	F	L	25-29	180-189	60-69	8	Rowing, Gym, Cycling, Cardio	18	9
sbj_16	M	R	20-24	180-189	60-69	2-3	Gym	15	3
sbj_17	M	R	25-29	180-189	70-79	4	Badminton, Bouldering, Hiking	15	5
sbj_20	M	R	25-29	170-179	70-79	N/A	N/A	12	N/A
sbj_21	M	R	20-24	170-179	70-79	2-3	Bouldering, Cardio	12	2
sbj_22	M	R	25-29	170-179	70-79	4	Badminton, Team Sport, Hiking	18	15
sbj_23	F	R	25-29	170-179	60-69	3-4	Tennis, Gym	17	4

C SUPPLEMENTARY EXPERIMENTS AND FIGURES

C.1 Detailed Test Results

Table 4 provides additional results on the test set for different window lengths (0.5, 1 and 2 seconds). We chose to use hyperparameters and postprocessing in accordance with the LOSO cross-validation experiments. That is, predictions of the inertial-based architectures were smoothed using a majority filter of 10 seconds and predictions of the Temporal Action Localization (TAL) models were score thresholded using a threshold of 0.1 based on the prediction confidence of the model associated with each segment.

C.2 Ablation Study on Postprocessing

The following details ablation experiments conducted to demonstrate the effectiveness and validity of the applied postprocessing described in the experiments section of the main paper. Figure 3 illustrates the effect the majority vote filter has on the prediction stream of the inertial-based models. One can see that without applying a majority vote filter, inertial-based architectures produce a large amount of non-coherent segments. This is due to the fact that during training, inertial models are not explicitly trained to predict coherent segments, but rather predict a continuous stream of windowed data. The models therefore tend to show a lot of intermediate switches in-between activity labels which causes mAP scores of inertial-based architectures to be substantially lower than scores of vision-based models. We therefore make use of a majority vote filter to erase short activity-label switches. Table 5 and 6 shows experimental results of applying different-sized majority vote filters (5, 10, 15, 20 and 25 seconds) compared to applying no filter. Interestingly, results (see Table 5 and 6) not only demonstrate the effectiveness of the majority vote filter through a substantial increase in mAP scores, yet also show that said

Table 4. Test results of human activity recognition approaches based on body-worn IMU (Inertial), vision (Camera) and combined (I + C) features for different clip lengths (CL) on the test dataset mentioned in the main paper evaluated in terms of precision (P), recall (R), F1-score and mean average precision (mAP) for different temporal intersection over union (tIoU) thresholds. 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	88.47	86.02	85.55	75.86	74.24	73.06	70.78	68.10	72.41
	A-and-D	0.5s	85.42	79.24	79.67	65.36	63.51	59.64	56.73	56.01	60.25
	ActionFormer	0.5s	76.11	**87.24**	79.00	89.37	87.68	76.95	59.63	38.26	70.38
	TriDet	0.5s	81.48	86.63	82.00	90.26	88.02	80.21	68.33	58.38	77.04
	Shallow D.	1s	**90.10**	86.10	**86.20**	73.94	72.70	71.60	69.22	67.99	71.09
	A-and-D	1s	87.48	83.62	83.97	71.88	70.58	67.63	63.97	61.98	67.21
	ActionFormer	1s	78.69	86.72	80.52	**91.68**	**90.48**	**89.23**	87.23	83.17	**88.36**
	TriDet	1s	79.55	84.61	80.04	90.28	89.54	88.37	**87.49**	**84.83**	88.10
	Shallow D.	2s	88.06	84.73	84.38	69.31	67.73	65.61	63.55	60.85	65.41
	A-and-D	2s	88.67	85.36	85.43	72.81	70.47	67.39	64.90	62.59	67.63
	ActionFormer	2s	72.57	83.54	74.95	89.95	88.86	87.38	86.03	80.43	86.53
	TriDet	2s	76.49	83.24	77.25	89.85	89.17	87.97	85.37	80.77	86.63
Camera	ActionFormer	0.5s	65.58	76.04	67.32	83.72	82.85	78.84	58.43	37.52	68.27
	TriDet	0.5s	72.93	75.93	71.89	84.51	82.48	76.46	67.61	57.60	73.73
	ActionFormer	1s	70.70	78.17	71.90	86.09	86.09	85.98	**85.49**	81.56	85.04
	TriDet	1s	73.00	**78.64**	**72.49**	**86.65**	**86.42**	86.17	84.96	**83.21**	**85.48**
	ActionFormer	2s	68.97	75.79	68.39	86.12	85.71	85.21	84.00	79.11	84.03
	TriDet	2s	**73.41**	77.31	71.38	86.44	86.33	**86.23**	**85.49**	82.74	85.45
I + C	ActionFormer	0.5s	78.28	88.55	80.97	92.07	91.69	86.63	68.15	41.08	75.92
	TriDet	0.5s	**86.00**	89.12	86.23	92.18	90.15	83.11	75.24	64.59	81.06
	ActionFormer	1s	82.59	89.07	84.26	94.24	94.06	**93.80**	**93.29**	89.51	92.98
	TriDet	1s	85.72	**90.78**	**86.98**	**94.47**	**94.29**	93.77	92.69	**90.39**	**93.12**
	ActionFormer	2s	76.80	86.35	78.86	91.53	91.28	90.89	88.83	84.33	89.37
	TriDet	2s	80.94	87.19	82.19	93.43	92.81	92.01	91.34	89.63	91.84

increase does not come at the cost of a decreased F1-score, but rather an increase. Table 5 and 6 further show a majority vote filter of 10 seconds being most effective resulting in the highest F1-score.

TAL models are not trained on an explicitly modelled NULL-class. This means, that unlike inertial models such as the DeepConvLSTM, TAL models are only able to predict segments with activity labels other than the NULL-class. With both models being set to predict up to 2000 action segments per video, the unprocessed prediction results resulted in activity streams such as illustrated in Figure 4. One can see that almost all samples have been assigned an activity label, leaving only a few data to be predicted as NULL, ultimately resulting in a substantially lower NULL-class accuracy than compared to inertial-based models mentioned in this paper. We therefore increased the score-threshold of both the ActionFormer and TriDet model, eliminating low-scoring segments and replacing them with NULL (see Figure 4). This improved classification performance of the ActionFormer (see Table 7) and TriDet model (see Table 8) significantly across all experiments (i.e. using inertial, vision and a combined setup as input data), while only marginally decreasing mAP scores. Table 8 further shows 0.1 being the most effective identified threshold of our ablation study, resulting in the highest F1-score of the TAL models.

Table 5. Ablation experiments on the effect of different-sized majority vote (MV) filters (5, 10, 15, 20 and 25 seconds) on the raw prediction results (0 seconds) of the shallow DeepConvLSTM model. We report results on the three employed window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Best results per clip-length are in **bold**.

	MV filter	P	R	F1	mAP					
					0.3	0.4	0.5	0.6	0.7	Avg
0.5 window	0 sec	68.7	68.42	65.96	5.62	4.53	3.55	3.01	2.66	3.87
	5 sec	75.23	74.69	72.05	50.99	48.52	46.70	44.54	42.22	46.59
	10 sec	74.81	74.93	72.26	60.03	57.91	55.54	54.19	52.31	55.99
	15 sec	74.38	74.85	72.24	62.59	61.03	58.82	56.71	54.59	58.75
	20 sec	74.19	74.55	71.96	64.35	62.85	60.19	57.97	55.34	60.14
	25 sec	73.77	74.08	71.53	65.09	63.58	60.96	58.91	56.39	60.99
1.0 window	0 sec	74.00	72.11	70.36	12.16	10.50	9.16	7.58	6.89	9.26
	5 sec	77.84	77.08	75.04	55.53	53.27	51.33	48.98	46.28	51.08
	10 sec	77.71	77.41	75.44	63.37	61.35	58.91	57.02	55.05	59.14
	15 sec	77.59	77.32	75.31	66.49	64.55	62.51	60.70	58.18	62.49
	20 sec	77.34	76.99	75.05	67.90	66.14	64.37	61.86	59.26	63.91
	25 sec	76.82	76.55	74.66	69.02	67.53	65.38	62.50	59.67	64.82
2.0 window	0 sec	75.85	73.61	72.08	23.54	20.86	17.93	16.04	14.37	18.55
	5 sec	78.18	76.72	75.06	56.09	53.73	51.33	49.26	46.41	51.36
	10 sec	78.00	77.19	75.43	65.14	63.58	60.98	59.15	57.39	61.25
	15 sec	77.60	77.08	75.25	67.10	65.89	63.24	61.27	59.24	63.35
	20 sec	77.39	76.88	75.02	67.74	66.34	64.15	62.25	60.10	64.11
	25 sec	76.22	71.39	70.71	64.05	61.10	58.24	55.78	52.66	58.37

Table 6. Ablation experiments on the effect of different-sized majority vote (MV) filters (5, 10, 15, 20 and 25 seconds) on the raw prediction results (0 seconds) of the improved Attend-and-Discriminate model. We report results on the three employed window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Best results per clip length are in **bold**.

	MV filter	P	R	F1	mAP					
					0.3	0.4	0.5	0.6	0.7	Avg
0.5 window	0 sec	70.61	65.85	65.32	3.60	3.04	2.42	1.95	1.60	2.52
	5 sec	76.72	72.87	71.90	47.90	44.53	42.34	39.80	38.09	42.53
	10 sec	76.54	73.04	72.10	57.51	54.59	52.22	49.47	46.94	52.15
	15 sec	76.43	72.68	71.80	61.04	58.36	55.96	53.03	50.74	55.83
	20 sec	76.28	72.06	71.29	62.94	59.99	57.14	54.34	52.06	57.30
	25 sec	77.10	76.51	74.67	69.09	67.81	65.51	63.45	61.51	65.47
1.0 window	0 sec	73.54	69.83	68.95	9.18	7.66	5.97	5.26	4.56	6.53
	5 sec	80.04	75.77	74.60	51.35	48.40	45.63	43.09	40.22	45.74
	10 sec	79.97	75.92	74.67	61.72	58.77	56.29	53.83	51.59	56.44
	15 sec	79.01	75.68	74.41	63.54	61.64	59.26	57.27	54.63	59.27
	20 sec	78.33	75.18	73.96	65.15	62.76	60.87	58.42	55.62	60.56
	25 sec	77.90	74.61	73.51	65.88	64.21	61.72	58.75	56.70	61.45
2.0 window	0 sec	76.51	73.63	72.36	18.80	16.00	13.31	10.78	9.76	13.73
	5 sec	81.10	77.78	76.60	53.55	51.18	47.96	45.72	42.31	48.15
	10 sec	80.96	78.42	77.10	63.52	61.57	58.82	56.88	54.41	59.04
	15 sec	80.30	78.29	76.92	67.27	65.35	62.55	59.97	57.31	62.49
	20 sec	79.77	77.88	76.49	67.77	65.78	63.17	60.69	58.23	63.13
	25 sec	79.48	77.35	76.01	68.58	66.80	64.81	62.24	59.56	64.40

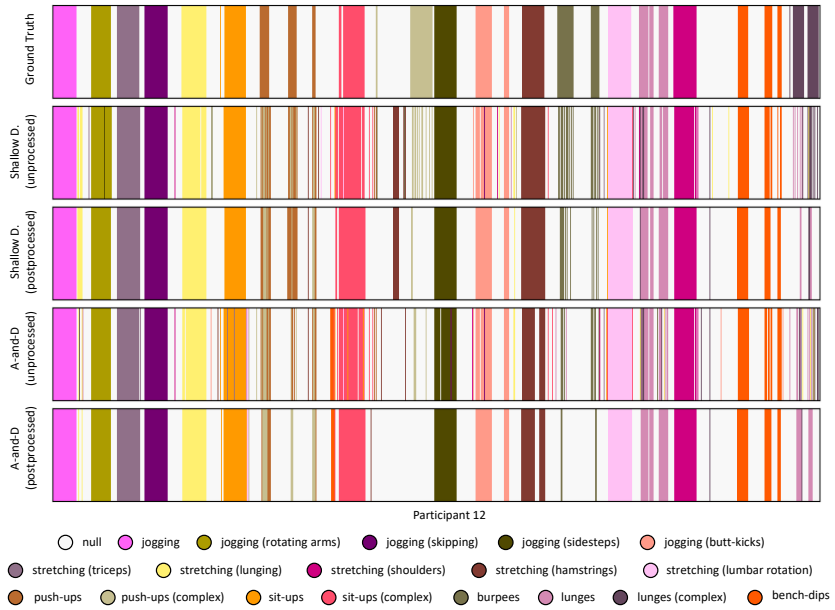


Fig. 3. Color-coded comparison of the ground truth data (top row) with the raw and postprocessed (10 sec majority vote filter) activity streams of the shallow DeepConvLSTM and Attend-and-Discriminate (A-and-D) model. The illustrated activity stream is of a sample subject having trained using inertial data which is windowed using a 1 second sliding window with 50% overlap.

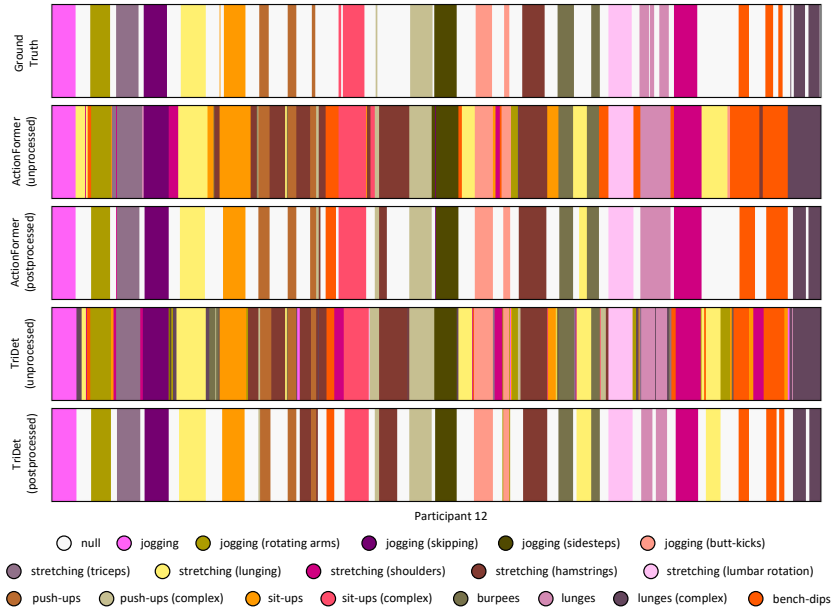


Fig. 4. Color-coded comparison of the ground truth data (top row) with the raw and score-thresholded (0.1) activity streams of the TriDet model. The illustrated activity stream is of sample subject having trained the model using both inertial and vision data which is windowed using a 1 second sliding window with 50% overlap.

Table 7. ActionFormer score thresholding results ablation experiments on the effect of different score thresholds (0.05, 0.1, 0.15, 0.2 and 0.25) on the raw prediction results (0.0 threshold). We report results using three clip length window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Best results per modality are in **bold**.

	Threshold	CL	P	R	F1	mAP					
						0.3	0.4	0.5	0.6	0.7	Avg
Inertial	0.0	0.5s	52.51	77.38	59.62	85.50	81.74	72.97	58.06	41.71	67.99
	0.05	0.5s	58.47	77.76	61.62	83.74	79.91	71.27	56.54	40.48	66.38
	0.1	0.5s	63.09	78.17	66.80	81.12	77.47	69.08	53.98	38.56	64.04
	0.15	0.5s	65.87	76.17	68.46	78.26	74.64	66.29	50.99	36.38	61.31
	0.2	0.5s	66.00	71.29	66.72	73.47	70.02	61.97	47.03	34.04	57.30
	0.25	0.5s	62.29	63.63	61.40	67.21	63.99	55.91	42.26	31.22	52.12
	0.0	1.0s	55.94	77.49	59.74	85.90	83.55	80.83	75.59	66.74	78.52
	0.05	1.0s	61.03	78.23	63.36	84.21	82.01	79.38	74.35	65.69	77.13
	0.1	1.0s	65.88	78.44	68.40	81.91	79.75	76.99	72.09	64.33	75.01
	0.15	1.0s	68.50	77.25	70.33	79.45	77.44	74.80	70.56	63.12	73.07
	0.2	1.0s	69.24	74.71	70.16	76.61	74.59	72.16	68.38	61.47	70.64
	0.25	1.0s	69.60	72.15	69.33	74.23	72.25	69.84	66.40	59.89	68.52
	0.0	2.0s	52.26	76.33	57.19	83.96	81.04	78.12	73.70	66.10	76.58
	0.05	2.0s	56.08	76.73	59.52	82.48	79.56	76.73	72.52	65.24	75.30
	0.1	2.0s	61.56	76.63	64.57	80.14	77.24	74.41	70.24	63.18	73.04
	0.15	2.0s	64.76	75.03	66.77	77.63	74.88	72.21	68.33	61.78	70.97
	0.2	2.0s	67.22	72.67	67.67	75.11	72.54	70.10	66.38	60.14	68.86
	0.25	2.0s	67.33	69.61	66.60	71.34	68.96	66.74	63.16	57.59	65.56
Camera	0.0	0.5s	48.38	72.36	54.31	85.47	82.48	76.21	61.63	45.21	70.20
	0.05	0.5s	57.02	73.50	59.01	84.92	81.93	75.69	60.82	44.10	69.49
	0.1	0.5s	60.86	72.98	62.54	81.07	78.48	72.13	57.34	41.27	66.06
	0.15	0.5s	61.74	70.15	62.56	75.94	73.45	67.82	53.54	38.37	61.83
	0.2	0.5s	60.12	65.24	60.16	69.51	67.27	61.75	48.47	35.46	56.49
	0.25	0.5s	55.92	58.25	54.97	60.86	58.95	54.24	43.59	32.19	49.97
	0.0	1.0s	52.62	74.21	55.41	87.52	85.55	83.00	78.97	71.43	81.30
	0.05	1.0s	61.39	75.98	63.20	87.06	85.02	82.42	78.45	70.99	80.79
	0.1	1.0s	65.82	75.34	66.40	85.19	83.31	80.94	77.22	69.96	79.32
	0.15	1.0s	67.98	73.95	67.37	82.34	80.55	78.50	74.94	68.11	76.89
	0.2	1.0s	67.65	71.81	66.57	78.58	77.22	75.30	71.95	65.29	73.67
	0.25	1.0s	66.56	69.73	65.43	74.34	73.06	71.16	68.12	61.94	69.72
	0.0	2.0s	54.14	75.41	57.04	86.92	84.80	82.54	79.42	73.81	81.50
	0.05	2.0s	58.55	76.22	60.61	86.51	84.41	82.13	79.00	73.43	81.10
	0.1	2.0s	63.40	76.25	65.39	84.93	82.95	80.58	77.55	72.07	79.62
	0.15	2.0s	66.78	74.93	67.40	82.02	80.44	78.32	75.25	70.10	77.23
	0.2	2.0s	67.88	73.14	67.61	78.73	77.28	75.39	72.31	67.55	74.25
	0.25	2.0s	67.07	70.94	66.57	74.46	73.23	71.48	68.50	64.08	70.35
Inertial + Camera	0.0	0.5s	57.97	82.71	65.13	91.13	88.98	82.88	69.44	48.25	76.13
	0.05	0.5s	67.05	83.92	70.09	89.42	87.24	81.22	67.80	46.88	74.51
	0.1	0.5s	71.45	83.77	74.51	86.04	83.96	77.80	64.41	43.76	71.19
	0.15	0.5s	72.26	80.29	74.26	81.37	79.66	73.87	60.31	40.82	67.21
	0.2	0.5s	70.39	74.64	71.03	76.21	74.66	69.39	56.22	37.89	62.87
	0.25	0.5s	66.93	67.82	66.00	70.10	68.55	63.61	50.82	35.08	57.63
	0.0	1.0s	59.04	81.82	63.89	90.25	88.65	85.98	81.92	75.17	84.39
	0.05	1.0s	68.32	83.47	71.39	88.59	86.98	84.28	80.27	73.93	82.81
	0.1	1.0s	72.87	83.24	75.26	86.82	85.27	82.46	78.26	72.33	81.03
	0.15	1.0s	74.92	81.73	76.12	84.44	82.88	80.10	76.33	70.89	78.93
	0.2	1.0s	74.98	79.55	75.56	81.76	80.31	77.85	74.31	69.39	76.73
	0.25	1.0s	74.29	77.21	74.24	79.00	77.56	75.38	72.35	67.89	74.43
	0.0	2.0s	55.98	79.41	60.41	87.78	85.31	81.75	77.93	71.45	80.84
	0.05	2.0s	60.61	80.04	63.81	86.39	83.92	80.41	76.63	70.30	79.53
	0.1	2.0s	65.86	80.17	68.82	84.18	81.66	78.17	74.59	68.74	77.47
	0.15	2.0s	68.97	78.69	70.93	81.98	79.47	76.33	72.84	67.39	75.60
	0.2	2.0s	71.12	76.70	71.74	79.34	77.04	73.97	70.79	65.66	73.36
	0.25	2.0s	72.14	74.44	71.43	76.58	74.21	71.41	68.42	63.53	70.83

Table 8. TriDet score thresholding results ablation experiments on the effect of different score thresholds (0.05, 0.1, 0.15, 0.2 and 0.25) on the raw prediction results (0.0 threshold). We report results using three clip length sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Best results per modality are in **bold**.

	Threshold	CL	P	R	F1	mAP					
						0.3	0.4	0.5	0.6	0.7	Avg
Inertial	0.0	0.5s	54.18	77.70	60.50	83.96	80.03	73.27	63.93	53.05	70.85
	0.05	0.5s	62.04	78.53	64.27	82.20	78.38	72.16	62.94	52.06	69.55
	0.1	0.5s	68.58	78.30	70.41	79.66	76.07	69.78	60.57	49.95	67.21
	0.15	0.5s	69.86	74.77	70.41	75.63	72.10	66.24	57.34	47.48	63.76
	0.2	0.5s	69.84	69.83	68.18	71.70	67.99	62.28	54.17	44.77	60.18
	0.25	0.5s	65.54	62.54	62.40	65.46	62.05	56.51	48.87	40.75	54.73
	0.0	1.0s	52.39	77.82	59.66	85.73	84.20	82.08	77.27	70.71	80.00
	0.05	1.0s	60.94	78.74	64.08	83.90	82.43	80.42	76.13	70.06	78.59
	0.1	1.0s	68.18	78.58	70.36	81.89	80.46	78.47	74.59	68.91	76.86
	0.15	1.0s	71.92	76.62	72.17	79.80	78.36	76.42	72.58	67.29	74.89
	0.2	1.0s	72.22	73.95	71.52	76.61	75.31	73.60	70.11	65.20	72.17
	0.25	1.0s	71.83	70.79	69.94	73.17	71.93	70.29	67.19	62.71	69.06
	0.0	2.0s	49.96	76.24	57.44	83.66	81.72	78.63	75.15	69.00	77.63
	0.05	2.0s	57.17	76.55	60.68	81.23	79.35	76.36	72.88	66.92	75.35
	0.1	2.0s	63.14	75.64	65.94	78.69	77.02	73.89	70.68	65.31	73.12
	0.15	2.0s	66.58	72.86	67.38	75.45	73.97	70.93	68.13	63.23	70.34
	0.2	2.0s	67.75	70.27	67.12	72.66	71.22	68.27	65.86	61.48	67.90
	0.25	2.0s	66.75	67.33	65.61	68.75	67.40	64.53	62.50	58.67	64.37
Camera	0.0	0.5s	48.68	72.06	54.16	85.34	81.82	77.26	68.62	59.33	74.47
	0.05	0.5s	58.86	73.47	59.99	84.71	81.15	76.52	67.78	58.51	73.73
	0.1	0.5s	64.69	72.89	64.84	80.46	76.95	72.62	64.46	55.52	70.00
	0.15	0.5s	65.10	68.81	63.72	74.67	71.32	67.34	59.43	51.23	64.80
	0.2	0.5s	62.64	63.16	60.19	67.93	64.64	61.17	53.63	46.58	58.79
	0.25	0.5s	57.95	57.43	55.79	59.97	57.09	53.79	47.40	41.57	51.96
	0.0	1.0s	49.99	73.89	55.56	87.72	86.46	84.83	82.18	78.56	83.95
	0.05	1.0s	61.68	75.51	62.67	87.11	85.80	84.14	81.43	77.87	83.27
	0.1	1.0s	67.42	75.21	66.88	84.88	83.49	82.12	79.75	76.24	81.30
	0.15	1.0s	70.30	73.96	68.32	81.95	80.62	79.36	77.23	74.00	78.63
	0.2	1.0s	70.76	71.98	67.95	78.68	77.46	76.21	74.22	71.38	75.59
	0.25	1.0s	69.58	69.64	66.67	74.53	73.54	72.40	70.70	68.33	71.90
	0.0	2.0s	50.21	74.28	56.32	85.93	84.86	82.71	80.34	77.16	82.20
	0.05	2.0s	59.34	75.22	60.83	85.25	84.21	82.09	79.75	76.59	81.58
	0.1	2.0s	65.53	75.36	65.91	82.94	82.00	80.22	78.13	75.20	79.70
	0.15	2.0s	68.58	74.02	67.66	79.67	78.85	77.10	75.28	72.66	76.71
	0.2	2.0s	69.12	72.05	67.61	76.02	75.27	73.60	71.95	69.39	73.25
	0.25	2.0s	67.22	69.05	65.74	71.38	70.70	69.42	68.16	65.70	69.07
Inertial + Camera	0.0	0.5s	59.15	82.23	65.24	90.07	87.62	82.36	74.05	61.73	79.17
	0.05	0.5s	69.69	83.84	72.14	87.86	85.55	80.58	72.50	60.50	77.40
	0.1	0.5s	75.74	82.79	76.83	84.45	82.07	77.12	68.81	57.16	73.92
	0.15	0.5s	76.74	79.06	76.20	80.72	78.37	73.26	65.01	54.01	70.28
	0.2	0.5s	74.38	73.15	72.15	75.13	73.42	68.33	60.81	50.38	65.61
	0.25	0.5s	68.44	64.87	65.10	67.25	65.71	61.35	54.47	45.76	58.91
	0.0	1.0s	55.95	80.96	62.89	90.17	89.16	87.12	84.07	79.12	85.93
	0.05	1.0s	67.87	82.58	70.67	88.63	87.63	85.55	82.44	77.64	84.38
	0.1	1.0s	73.39	82.00	75.09	86.42	85.48	83.35	80.39	75.64	82.26
	0.15	1.0s	75.96	80.49	76.38	83.94	83.07	80.87	78.01	73.62	79.90
	0.2	1.0s	76.64	78.38	75.95	80.89	80.09	78.20	75.57	71.80	77.31
	0.25	1.0s	75.76	76.03	74.64	77.79	77.09	75.29	73.02	69.64	74.56
	0.0	2.0s	53.20	80.55	61.50	88.39	86.47	84.31	81.56	76.90	83.52
	0.05	2.0s	62.39	81.35	66.42	86.78	84.86	82.72	80.09	75.59	82.01
	0.1	2.0s	69.19	81.09	71.98	84.66	82.79	80.66	78.11	74.15	80.07
	0.15	2.0s	72.36	78.89	73.41	81.72	79.92	77.93	75.41	71.94	77.38
	0.2	2.0s	72.78	75.99	72.69	78.24	76.71	74.79	72.53	69.45	74.34
	0.25	2.0s	72.94	74.01	71.94	75.85	74.47	72.71	70.58	67.84	72.29

Table 9. Results using only inertial sensors placed on the right wrist (RW) and right wrist + ankle (RW + RA) compared with using all sensors for different clip lengths (CL) 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. Best results per modality are in **bold**.

	Model	CL	RW		RW + RA		all	
			F1	mAP	F1	mAP	F1	mAP
Inertial	Shallow D.	0.5s	60.50	27.56	72.42	47.26	72.26	55.99
	A-and-D	0.5s	65.65	32.38	69.60	45.66	72.10	52.15
	ActionFormer	0.5s	59.64	61.87	70.67	66.22	66.80	64.04
	TriDet	0.5s	62.05	64.64	74.06	71.02	70.41	67.21
	Shallow D.	1s	61.97	29.56	73.54	49.49	75.44	59.14
	A-and-D	1s	66.64	33.92	72.14	46.20	74.67	56.44
	ActionFormer	1s	61.62	72.80	73.13	77.88	68.40	75.01
	TriDet	1s	63.62	73.80	74.91	81.01	70.36	76.86
	Shallow D.	2s	60.74	29.26	74.48	50.72	75.43	61.25
	A-and-D	2s	66.19	34.50	73.98	51.36	77.10	59.04
	ActionFormer	2s	58.38	68.15	68.06	75.90	64.57	73.04
	TriDet	2s	60.48	68.22	70.90	76.81	65.94	73.12
I + C	ActionFormer	0.5s	71.75	70.48	76.54	72.80	74.51	71.19
	TriDet	0.5s	74.37	74.91	79.36	77.45	76.83	73.92
	ActionFormer	1s	73.49	79.46	78.18	83.55	75.26	81.03
	TriDet	1s	74.64	82.26	79.34	85.02	75.09	82.26
	ActionFormer	2s	69.21	79.69	73.37	81.13	68.82	77.47
	TriDet	2s	70.68	80.14	75.30	83.33	71.98	80.07

C.3 Ablation Study on the Selection of Inertial Sensors

Adding to results reported in the main paper, Table 9 provides extended results on the influence of using only a subset of the inertial sensors, specifically 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 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) and dominant hand movement being overall more consistent.

C.4 Ablation study on second execution of workout sessions

Table 10 provides additional results which compare repeated sessions for participants sbj_0 and sbj_14. 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). Unlike our prior experiments, each algorithm is trained using the data of all but the validation subjects' recordings, ensuring the validation subjects (sbj_0 and sbj_14) remain unseen during the training of each algorithm.

provides additional results which compare validation results obtained on the original, first recording of sbj_0 and sbj_14 with their second execution of the workout plan for different window lengths.

C.5 Ablation Study on Influence of Frequency of Inputs

With the frequencies both the camera (60 FPS) and inertial sensors (50 HZ) being set fairly high, the WEAR dataset allows to explore lower frequency experiments and their effect fewer datapoints per second might have

Table 10. Comparison of obtained results of repeated sessions for participants sbj_0 and sbj_14 for different clip lengths (CL) on our WEAR dataset evaluated in terms of F1-score and mean average precision (mAP). These figures are, as in the earlier results, averaged across 3 runs using 3 different random seeds. For the first recording, both subjects' best results per modality are in underlined. For the second recording, both subjects' best results per modality are in **bold**. Unlike our prior experiments, each algorithm is trained using the data of all but the validation subjects' recordings, ensuring the validation subjects (sbj_0 and sbj_14) remain unseen during the training of each algorithm. All results are postprocessed as reported in the main paper.

	Model	CL	sbj_0				sbj_14			
			1st Recording		2nd Recording		1st Recording		2nd Recording	
			F1	mAP	F1	mAP	F1	mAP	F1	mAP
Inertial	Shallow D.	0.5s	71.82	42.54	89.09	78.46	85.14	64.03	81.92	50.32
	A-and-D	0.5s	72.35	36.98	86.34	67.69	78.36	48.64	73.14	43.13
	ActionFormer	0.5s	78.36	72.34	81.82	65.79	74.24	77.74	68.47	71.76
	TriDet	0.5s	80.36	69.34	84.73	76.12	78.99	87.08	73.86	80.56
	Shallow D.	1s	77.20	47.43	88.94	79.88	85.45	65.96	81.97	54.89
	A-and-D	1s	75.63	41.19	87.02	71.75	83.32	59.02	79.82	51.98
	ActionFormer	1s	75.48	72.70	78.83	87.22	76.36	92.02	71.29	84.42
	TriDet	1s	77.67	78.87	79.31	89.44	79.90	93.18	71.27	84.75
	Shallow D.	2s	76.57	46.99	87.57	79.69	86.37	71.02	84.88	59.60
	A-and-D	2s	68.48	39.71	87.49	77.10	85.75	65.61	85.48	51.92
	ActionFormer	2s	66.28	65.83	71.15	80.38	70.47	89.54	63.12	79.97
	TriDet	2s	71.27	64.33	72.94	86.30	75.54	91.81	65.01	88.61
C	ActionFormer	0.5s	62.17	57.52	75.29	68.31	63.06	65.84	69.57	77.35
	TriDet	0.5s	60.67	64.37	78.28	69.20	62.12	65.95	76.13	83.55
	ActionFormer	1s	62.97	63.64	73.86	88.70	58.67	78.13	71.27	89.42
	TriDet	1s	64.71	67.99	76.04	87.78	64.26	80.25	72.81	88.30
	ActionFormer	2s	54.72	55.36	75.03	87.78	56.66	81.11	70.52	91.94
	TriDet	2s	53.03	58.03	72.97	86.56	63.05	80.73	67.39	91.43
I + C	ActionFormer	0.5s	84.38	74.86	85.02	74.69	78.19	81.74	73.28	76.28
	TriDet	0.5s	84.07	66.72	82.42	74.63	82.46	87.74	78.34	80.04
	ActionFormer	1s	84.78	79.79	82.52	94.07	79.05	91.91	73.38	87.82
	TriDet	1s	83.03	82.55	83.54	93.18	78.35	97.66	70.78	86.76
	ActionFormer	2s	65.06	74.76	79.31	87.59	74.81	93.57	66.96	87.44
	TriDet	2s	76.16	77.61	76.34	88.40	76.89	96.23	68.73	84.22

on the predictive quality of the trained models. Table 11 summarizes experiments conducted using only 50% and 20% of the available frequency for both types of sensors. Note that a clip length of 0.5 seconds was not explored during experiments as it was not possible anymore to extract two-stream I3D feature embeddings [3] as the amount of frames was lower than the required minimum input frames. Looking at results presented in Table 11 one can see that all models trained using only inertial data suffered from lower frequency inputs with both classification and mAP scores decreasing. Contrarily, models trained using camera-based improved when using features extracted from videos with a lower FPS, which might be caused by Kinetics-400 [6], which was used for pretraining the I3D extraction method, on consisting of videos with a lower FPS than the WEAR dataset.

C.6 Attend-and-Discriminate Improvements

Instead of employing a plain Attend-and-Discriminate model as proposed by Abedin et al. [1], we incorporate architecture improvements suggested by Bock et al. [2]. Said architecture improvements are (1) using one instead of two recurrent layers, (2) increasing the amount of hidden units in the recurrent layer from 128 to 1024 and (3) scaling the convolutional kernel by the same factor the window size increases or decreases. Table 12 shows performance difference gained from employing the improved Attend-and-Discriminate architecture by comparing it to the original architecture. Note that results are reported without having applied any postprocessing.

Table 11. Results of evaluating different frequencies (Freq.) as input for different clip lengths (CL) on our WEAR dataset. Features were downsampled to be only 50% (30 FPS and 25 Hz) and 20% (12 FPS and 10 Hz) of the original frequency input. While predictive performance of inertial models decreases with a lower frequency input, camera and combined models increase in performance when lower frequency inputs. Best results per modality are in **bold**.

	Threshold	Freq.	CL	P	R	F1	mAP					
							0.3	0.4	0.5	0.6	0.7	Avg
Inertial	Shallow D.	Orig	1s	77.71	77.41	75.44	63.37	61.35	58.91	57.02	55.05	59.14
	Shallow D.	50%	1s	75.84	76.37	73.60	61.08	58.90	57.02	55.21	52.97	57.04
	Shallow D.	20%	1s	76.02	76.78	73.89	61.61	59.77	57.57	55.88	53.75	57.72
	A-and-D	Orig	1s	79.97	75.92	74.67	61.72	58.77	56.29	53.83	51.59	56.44
	A-and-D	50%	1s	79.00	75.33	74.35	59.69	57.12	54.46	51.98	48.96	54.44
	A-and-D	20%	1s	77.53	**79.12**	76.00	62.55	60.65	58.75	56.94	54.51	58.68
	ActionFormer	Orig	1s	65.88	78.44	68.40	**81.91**	79.75	76.99	72.09	64.33	75.01
	ActionFormer	50%	1s	66.09	77.89	68.24	81.74	79.30	75.85	69.88	62.68	73.89
	ActionFormer	20%	1s	64.43	77.30	67.22	81.39	79.33	76.25	71.49	63.39	74.37
	TriDet	Orig	1s	68.18	78.58	70.36	81.89	**80.46**	**78.47**	**74.59**	**68.91**	**76.86**
	TriDet	50%	1s	67.47	77.09	69.07	81.59	80.02	77.89	73.83	67.38	76.14
	TriDet	20%	1s	66.80	76.46	68.50	80.74	79.32	77.19	73.04	65.89	75.24
	Shallow D.	Orig	2s	78.00	77.19	75.43	65.14	63.58	60.98	59.15	57.39	61.25
	Shallow D.	50%	2s	78.60	75.23	74.07	60.99	57.86	54.78	52.53	50.18	55.27
	Shallow D.	20%	2s	75.02	76.76	73.38	60.42	58.43	55.87	54.11	52.13	56.19
	A-and-D	Orig	2s	**80.96**	78.42	**77.10**	63.52	61.57	58.82	56.88	54.41	59.04
	A-and-D	50%	2s	78.60	75.23	74.07	60.99	57.86	54.78	52.53	50.18	55.27
	A-and-D	20%	2s	79.25	76.13	74.76	61.43	59.55	56.85	54.80	52.55	57.04
	ActionFormer	Orig	2s	61.56	76.63	64.57	80.14	77.24	74.41	70.24	63.18	73.04
	ActionFormer	50%	2s	61.84	76.28	64.67	79.30	76.10	73.53	69.20	62.58	72.14
	ActionFormer	20%	2s	60.40	74.65	62.99	78.29	74.95	71.67	67.22	59.82	70.39
	TriDet	Orig	2s	63.14	75.64	65.94	78.69	77.02	73.89	70.68	65.31	73.12
	TriDet	50%	2s	63.22	74.52	65.52	77.99	76.18	73.49	69.94	63.06	72.13
	TriDet	20%	2s	63.01	73.43	64.84	77.21	75.22	72.71	69.75	62.99	71.58
Camera	ActionFormer	Orig	1s	65.82	75.34	66.40	85.19	83.31	80.94	77.22	69.96	79.32
	ActionFormer	50%	1s	69.37	78.23	70.09	**87.20**	**85.40**	83.05	79.54	72.58	81.55
	ActionFormer	20%	1s	67.40	77.26	68.73	86.03	84.37	82.55	78.23	70.94	80.43
	TriDet	Orig	1s	67.42	75.21	66.88	84.88	83.49	82.12	79.75	76.24	81.30
	TriDet	50%	1s	**70.92**	77.88	**70.19**	86.62	85.22	83.42	80.75	**77.10**	**82.62**
	TriDet	20%	1s	69.52	76.68	68.71	84.58	83.49	82.12	80.12	75.14	81.09
	ActionFormer	Orig	2s	63.40	76.25	65.39	84.93	82.95	80.58	77.55	72.07	79.62
	ActionFormer	50%	2s	67.55	**79.65**	69.36	86.99	85.34	83.10	80.54	75.00	82.19
	ActionFormer	20%	2s	65.70	76.30	66.59	84.90	83.02	81.09	77.77	71.33	79.62
	TriDet	Orig	2s	65.53	75.36	65.91	82.94	82.00	80.22	78.13	75.20	79.70
	TriDet	50%	2s	69.19	78.70	69.81	86.31	85.30	**83.64**	**80.99**	76.86	**82.62**
	TriDet	20%	2s	68.36	76.60	68.28	84.02	82.48	80.62	78.08	73.92	79.82
Inertial + Camera	ActionFormer	Orig	1s	72.87	83.24	75.26	86.82	85.27	82.46	78.26	72.33	81.03
	ActionFormer	50%	1s	74.60	**84.85**	77.19	87.88	86.87	84.34	79.84	73.00	82.39
	ActionFormer	20%	1s	74.73	84.27	76.79	87.99	86.63	83.78	80.32	72.86	82.31
	TriDet	Orig	1s	73.39	82.00	75.09	86.42	85.48	83.35	80.39	75.64	82.26
	TriDet	50%	1s	76.60	84.44	**78.00**	87.80	86.87	85.43	82.37	76.98	83.89
	TriDet	20%	1s	**76.66**	83.32	77.00	**88.59**	**87.69**	**86.16**	**83.63**	**78.53**	**84.92**
	ActionFormer	Orig	2s	65.86	80.17	68.82	84.18	81.66	78.17	74.59	68.74	77.47
	ActionFormer	50%	2s	68.74	82.46	71.65	86.73	84.12	80.99	77.49	71.53	80.17
	ActionFormer	20%	2s	69.56	82.26	71.97	86.58	83.99	80.78	76.56	69.49	79.48
	TriDet	Orig	2s	69.19	81.09	71.98	84.66	82.79	80.66	78.11	74.15	80.07
	TriDet	50%	2s	69.84	80.94	72.17	84.73	83.09	80.22	78.31	74.06	80.08
	TriDet	20%	2s	71.84	82.26	73.46	86.01	84.11	81.92	79.78	75.73	81.51

Table 12. Results demonstrating the effectiveness of made modifications to the Attend-and-Discriminate model. We compare the plain original model with an optimised version (1-layered LSTM with 1024 hidden units and an adjusted convolutional kernel sizes). We report results on the three employed window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Note that results are reported with no postprocessing applied.

	Model	P	R	F1	mAP					
					0.3	0.4	0.5	0.6	0.7	Avg
0.5s	Original A-and-D	65.72	71.04	65.91	4.98	4.13	3.16	2.72	2.24	3.45
	Optimised A-and-D	70.61	65.85	65.32	3.60	3.04	2.42	1.95	1.60	2.52
1.0s	Original A-and-D	66.04	71.72	66.21	8.02	6.84	5.28	4.69	4.12	5.79
	Optimised A-and-D	73.54	69.83	68.95	9.18	7.66	5.97	5.26	4.56	6.53
2.0s	Original A-and-D	68.29	74.05	68.42	16.54	14.33	12.23	10.91	9.74	12.75
	Optimised A-and-D	76.51	73.63	72.36	18.80	16.00	13.31	10.78	9.76	13.73

Table 13. Results demonstrating the effectiveness of longer training times on the inertial-based models. Compared are the shallow DeepConvLSTM and improved Attend-and-Discriminate model using either a short training time (30 epochs and no step-wise learning rate schedule (LRS)) or long training time (100 epochs and LRS). We report results on the three employed window sizes (0.5, 1.0 and 2.0 seconds) each with a 50% overlap. Note that results are reported with no postprocessing applied.

	Model	Epochs	LRS	P	R	F1	mAP					
							0.3	0.4	0.5	0.6	0.7	Avg
0.5s	Shallow D.	30		65.18	71.14	65.49	5.73	4.79	3.97	3.45	3.06	4.20
	Shallow D.	100	✓	68.70	68.42	65.96	5.62	4.53	3.55	3.01	2.66	3.87
	A-and-D	30		67.36	68.64	65.48	4.10	3.12	2.57	2.13	1.87	2.76
	A-and-D	100	✓	70.61	65.85	65.32	3.60	3.04	2.42	1.95	1.60	2.52
1.0s	Shallow D.	30		68.68	74.60	68.92	12.39	10.83	9.37	8.06	7.24	9.58
	Shallow D.	100	✓	73.54	69.83	68.95	9.18	7.66	5.97	5.26	4.56	6.53
	A-and-D	30		70.49	72.15	68.74	9.38	7.86	6.18	5.26	4.67	6.67
	A-and-D	100	✓	73.54	69.83	68.95	9.18	7.66	5.97	5.26	4.56	6.53
2.0s	Shallow D.	30		71.45	77.04	71.55	22.54	20.30	18.11	16.42	14.84	18.44
	Shallow D.	100	✓	75.85	73.61	72.08	23.54	20.86	17.93	16.04	14.37	18.55
	A-and-D	30		73.58	75.94	72.28	18.88	16.35	13.72	11.91	10.65	14.30
	A-and-D	100	✓	76.51	73.63	72.36	18.80	16.00	13.31	10.78	9.76	13.73

C.7 Longer vs. Shorter Training Runs

As mentioned in the main paper, all inertial-based architectures are trained for 100 epochs as compared to 30 epochs. To compensate for longer training times we employ a step-wise learning rate schedule with a step size of 10 epochs and a decay rate of 0.9 [1]. Table 13 shows results of the longer training runs remain stable compared to a shorter training time of 30 epochs.

C.8 Additional Visualizations of LOSO Experiments

In addition to the visualisations supplied in the main paper, the following provides supplementary visualizations of the LOSO cross-validation experiments for further analysis. All models mentioned in this section were trained using a clip length of 1.0 second with a 50% overlap. Predictions made by the TAL models were filtered using a score threshold of 0.1 and predictions made by inertial-based architectures were filtered using a majority

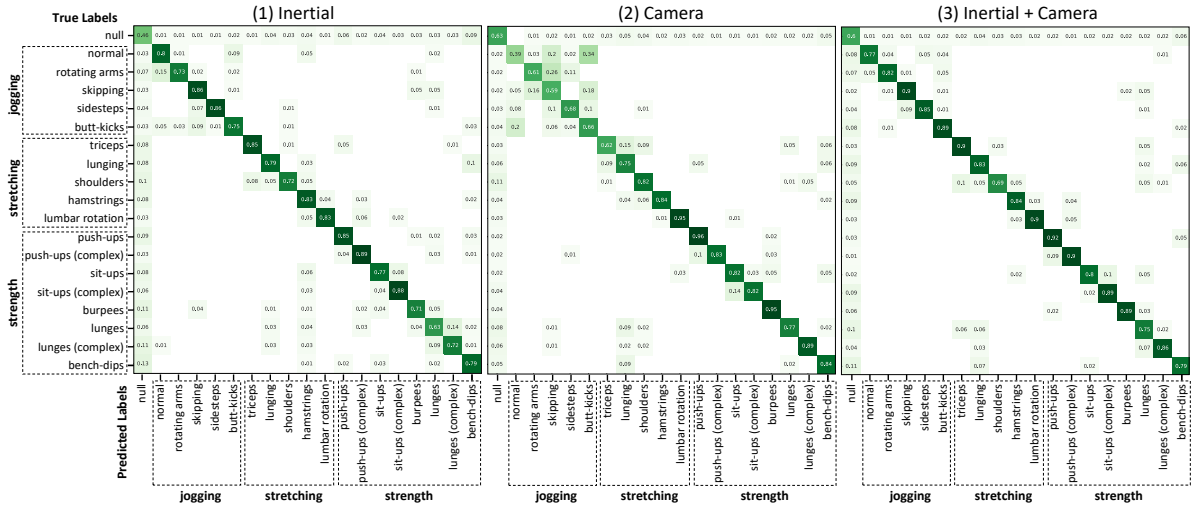


Fig. 5. Confusion matrices of the ActionFormer being applied using only inertial, vision (camera) and both combined (inertial + camera).

vote filter of 10 seconds. Figure 5 provides confusion matrices of the ActionFormer [7] being applied using inertial, camera and combined (inertial + camera) features. Figure 6 provides Confusion matrices of the shallow DeepConvLSTM [2] and improved Attend-and-Discriminate [1] applied on inertial data. Figure 7 shows a color-coded visualisation of predictions streams of all models mentioned in the results table of the main paper. Figure 8 delivers a side-by-side comparison of the confusion matrices of prediction streams of all models involved in the *Oracle*-late-fusion-approach analysis mentioned in the main paper. Figure 8 shows that a joint learning of both modalities particularly improves differentiation between the NULL-class and the activity classes resulting in better action boundaries, i.e. higher mAP scores, and classification scores.

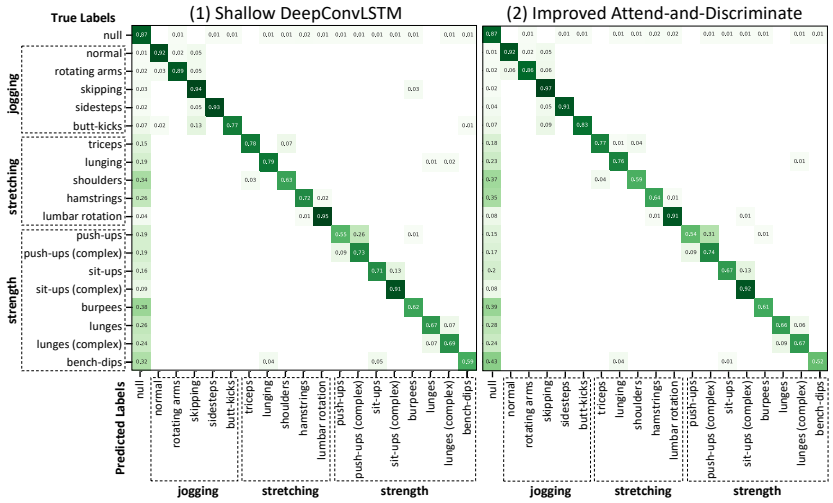


Fig. 6. Confusion matrices of the shallow DeepConvLSTM and improved Attend-and-Discriminate.

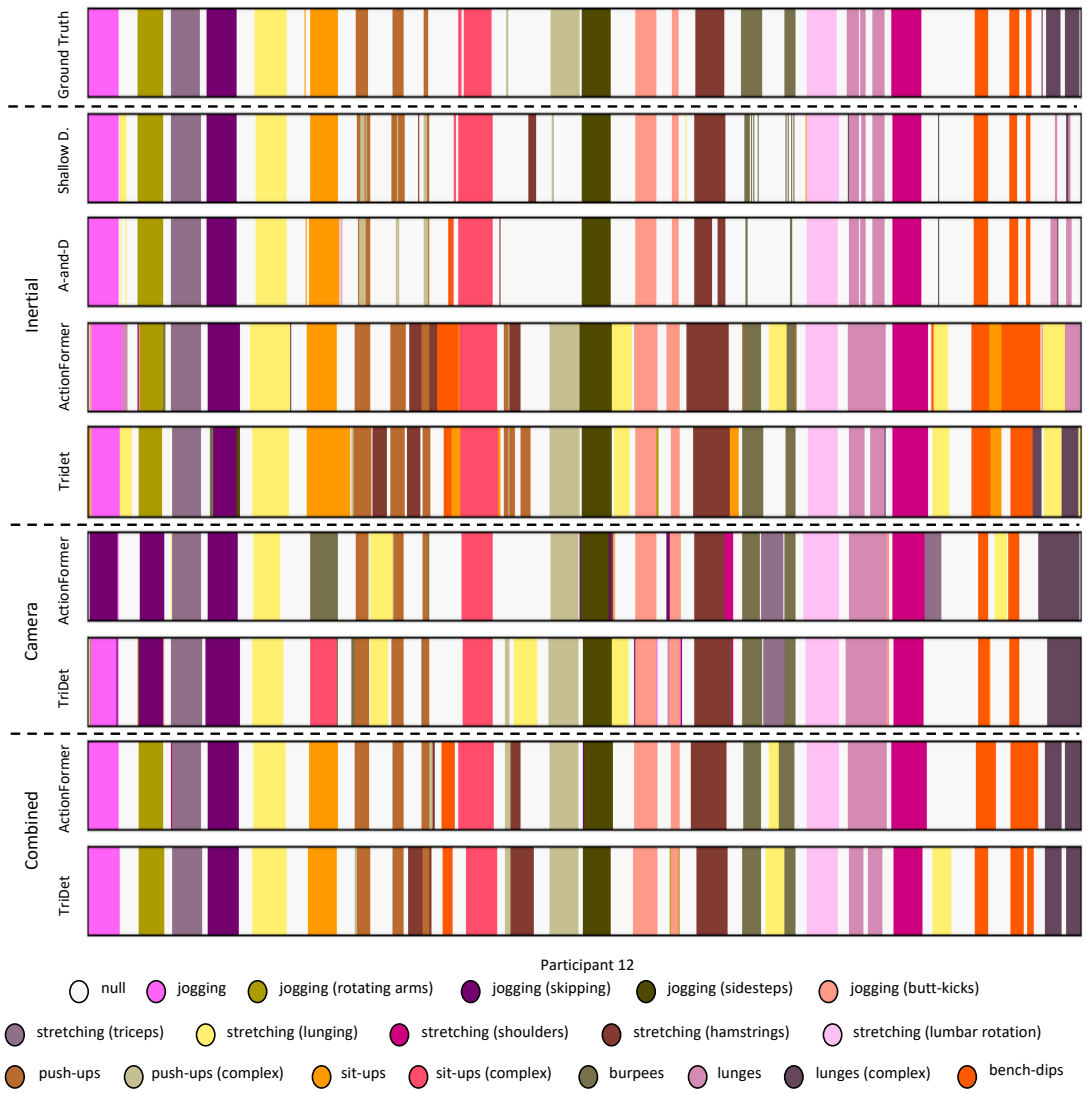


Fig. 7. Color-coded comparison of the ground truth data (top row) with the shallow DeepConvLSTM, improved Attend-and-Discriminate, ActionFormer and TriDet model on varying input modalities.

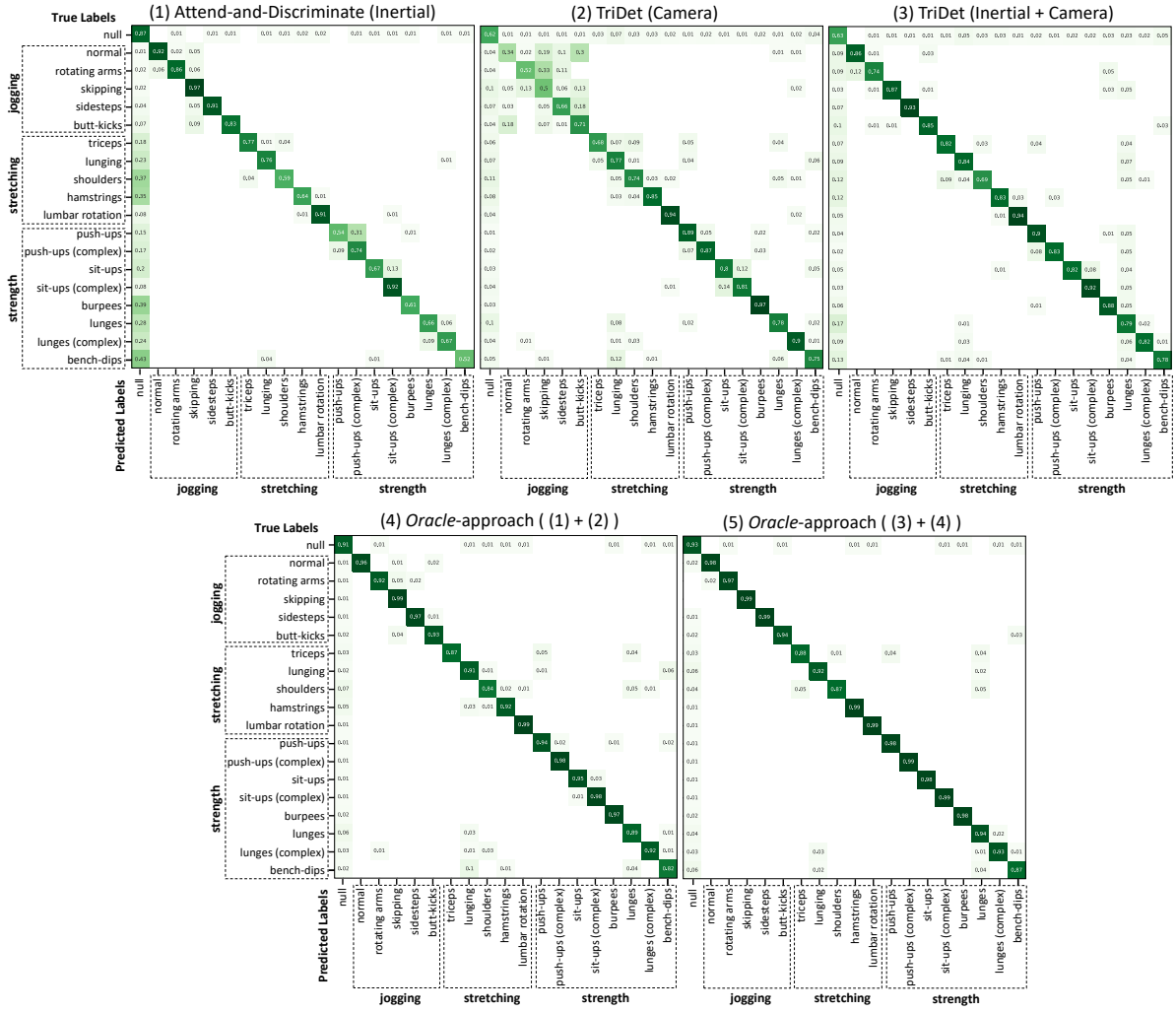


Fig. 8. Confusion matrices of the best (1) inertial model (Attend-and-Dicriminate), (2) vision model (TriDet) and (3) combined model (vision + inertial) compared with (4) an *Oracle*-combination of the inertial and camera as well (5) *Oracle*-combination of the previous oracle with the combined approach.

D FULL-TEXT RECORDING PLAN

WEAR Dataset Recording Plan

Setup:

- 4 Bangle.js Smartwatches on each limb
  - Sampling rate at 50 Hz with ±8g
- 2 GoPros
  - Static (GoPro Hero 5):
    - Mounted on tripod; observing whole scene
    - For labeling purposes (not part of final dataset)
    - 1080p 30 FPS Superview FOV (no video stabilization)
  - Head-mounted (GoPro Hero 8):
    - Hands & feet centrally visible in frames; roughly 45° downward
    - 1080p 60 FPS Superview FOV (auto low light + video stabilization)

Equipment List:

- 4 Bangle.js Smartwatches
- 2 GoPros + Head-Mount + Tripod
- Bench/ Box/ Chair
- Optional:
  - Stopwatch
  - Yoga Mat

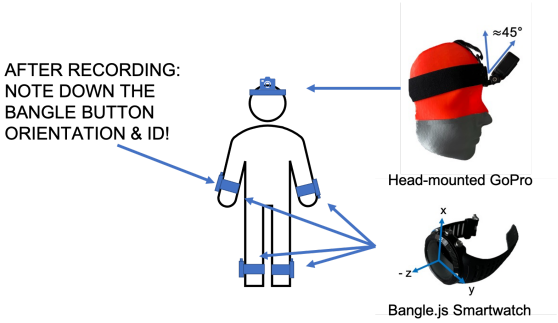


Fig. 9. First page of the recording plan of the WEAR dataset.

1st Session (ca. 30 min)

➔ **Each activity 3 sets à repetitions at will (ca. 30 sec); short break after each set (ca. 30 sec)**

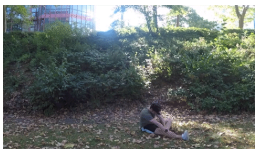
Activity	Sub-Activity	Description	Normal Variant	Easy Variant
Running	Sidestep	Start with both feet together; jump on one foot to the side and repeat same motion for other foot to create a jumping motion; repeat		
	Butt-Kicks	Fold hands on butt; jog while trying to lift alternating each heel as close to butt as possible ("kicking" it)		
Stretching	Shoulder	Start by standing straight; stretch left arm to the right while keeping it parallel to the ground; use lower right arm to press against left arm upper left arm (close to the elbow) trying to move it closer to the body; hold stretch; repeat by switching arms		
	Hamstrings	Start by sitting down; have left leg stretched out straight in a 45° angle to the left side and keep right leg as sitting cross-legged with right foot touching the left knee's side; try to reach for left foot; hold stretch; repeat by switching legs		
	Lumbar Rotation	Start by laying on back; reach out with both arms to the side; raise legs; move legs to the left side as close as possible to the ground while keeping them straight; do not move torso; hold stretch; repeat by moving legs to opposite side		

Fig. 10. Second page of the recording plan of the WEAR dataset.

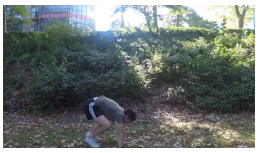
Burpees	Normal	Start by standing straight; put your hands on the ground and jump back with your feet into a push-up position; do a push-up; jump forward with your feet into the starting position and jump up with raised arms; repeat		No Push Up, but lay flat on ground; get up by standing up instead of jumping
Walking Lunges	Normal	Stand straight; Keep your hands crossed in front of your torso; move your right foot forward, bending your left knee to the ground and right knee into a 90° angle; step your left foot forward going back into the position your started in; repeat to create walking motion		
	Complex	Stand straight; keep your arms raised in front of you pointing forward parallel to the floor; do a lunge, but when at the bottom turn your torso to the side of the foot being forward; repeat to create walking motion		
Bench Dips	Normal	Stand with your back facing a chair; go into dip position with your hands on the chair and your legs straight in front of you; move your body downwards by bending your arms into a 90° angle; keep legs and back as straight as possible; move up again; repeat		

Fig. 11. Third page of the recording plan of the WEAR dataset.

2nd Session (ca. 30 min)

➔ **Each activity 3 sets à repetitions at will (ca. 30 sec); short break after each set (ca. 30 sec)**

Activity	Sub-Activity	Description	Normal Variant	Easy Variant
Running	Jogging	Normal jogging		
	Jogging with rotating arms	Jogging while rotating arms backwards; first only left arm; then only right arm; then both arms simultaneously		
	Skipping	Alternate between jumping from left leg while lifting the right knee high; land on the left leg; repeat by switching legs to create running motion		
	Triceps	Start by standing straight; raise left arm behind head and try to touch the right shoulder; use right arm to grab left elbow trying to move left hand further down the shoulder; hold stretch; repeat by switching arms		
	Lunging	Start by standing in a split stance with right front forward and left foot straight back; bend right knee about 90°; place hands on your forward knee; hold stretch; repeat by switching legs		

Fig. 12. Fourth page of the recording plan of the WEAR dataset.

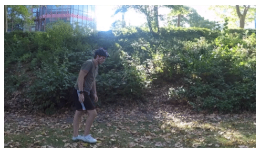
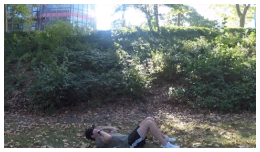
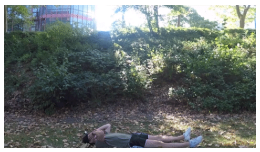
Push-ups	Normal	Normal Push-Up		On knees
	Complex	Move into a push-up position; do a push-up by lowering your body to the ground; after moving it back up, reach out with the right arm to the sky, opening your torso so that it faces to the right; move back into push-up; repeat for left arm; repeat sequence		On knees
Sit-ups	Normal	Lay on your back; have your hands touch the sides of your head; move your legs into a 90° angle with your feet on the ground; move your torso towards your knees while keeping the legs in place; repeat		Straight legs
	Complex	Lay on your back with your hands touching the sides of your head; move your legs into a 90° angle (feet on the ground) while also (!) moving your torso towards your knees; when reaching highest point, touch first your right heel with your right hand; and left heel with left hand; repeat		First move up with upper body; then with lower body so that legs are straight during situp

Fig. 13. Fifth page of the recording plan of the WEAR dataset.

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] 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

[3] 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

[4] 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

[5] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. *CoRR* abs/1805.03677 (2018). https://arxiv.org/abs/1805.03677

[6] 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

[7] 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