Supplementary Material:
Weak-Annotation of HAR Datasets using Vision Foundation Models

MARIUS BOCK, University of Siegen, 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**.

Additional Key Words and Phrases: Data Annotation, Human Activity Recognition, Body-worn Sensors

A ANNOTATION PIPELINE

The following provides figures and tables supplementary to the annotation pipeline experiments of the main paper.

A.1 Number of Samples per Cluster Clip Length

As detailed in the experiments section of the main paper, we used only the centroid clip to propagate labels to all other instances within the cluster. Table 1 provides supplementary results comparing our approach with using not only the centroid, but also the labels of four other neighboring clips to determine the label of the cluster. In this approach, the four other instances are the closest neighbors to their respective centroid. We determine the label by majority vote, meaning the label of the centroid can be overridden if the majority of the other four clips is labeled otherwise. One can see that throughout all datasets, our approach remains stable compared to using more instances, suggesting that our label propagation based on the centroid clip is stable and that the centroid clip is indeed representative of other clips within the cluster.

Similarly, Table 2 provides supplementary results when applying different clip lengths during clustering. As we mention in our main paper, we deem a clip length of four seconds to be adequate for an annotator to correctly label the activity present in the clip. While using longer clips would cause clips to potentially contain multiple activities, shorter clips would make it harder for a human annotator to correctly identify ongoing activities. Nevertheless, as also shown in Table 2, a longer clip also equates to higher labeling accuracies across all datasets, suggesting activities can be more easily grouped together during clustering.

A.2 Number of Cluster Per-Class Results

As mentioned in the main paper, we noticed that even though a dataset might not contain 100 activities, applying a higher number of clusters results in higher labeling accuracies. As Figure 1 shows, when applying only a small number of clusters, similar activities tend to not be well differentiated (e.g., different styles of push-ups within the WEAR dataset).

Authors' Contact Information: Marius Bock, Ubiquitous Computing & Computer Vision, University of Siegen, Siegen, Germany, marius.bock@uni-siegen.de; Kristof Van Laerhoven, Ubiquitous Computing, University of Siegen, Siegen, Germany, kvl@eti.uni-siegen.de; Michael Moeller, Computer Vision, University of Siegen, Siegen, Germany, michael.moeller@uni-siegen.de.

Table 1. Comparison of average labeling accuracy and standard deviation across study participants propagating labels based on only the human-annotated label of the centroid clip ($s = 1$) versus based on the 5 most-centered clips ($s = 5$). Each setting applied a GMM-based clustering using 100 clusters. One can see that both approaches yield the same results across all datasets [2–4].

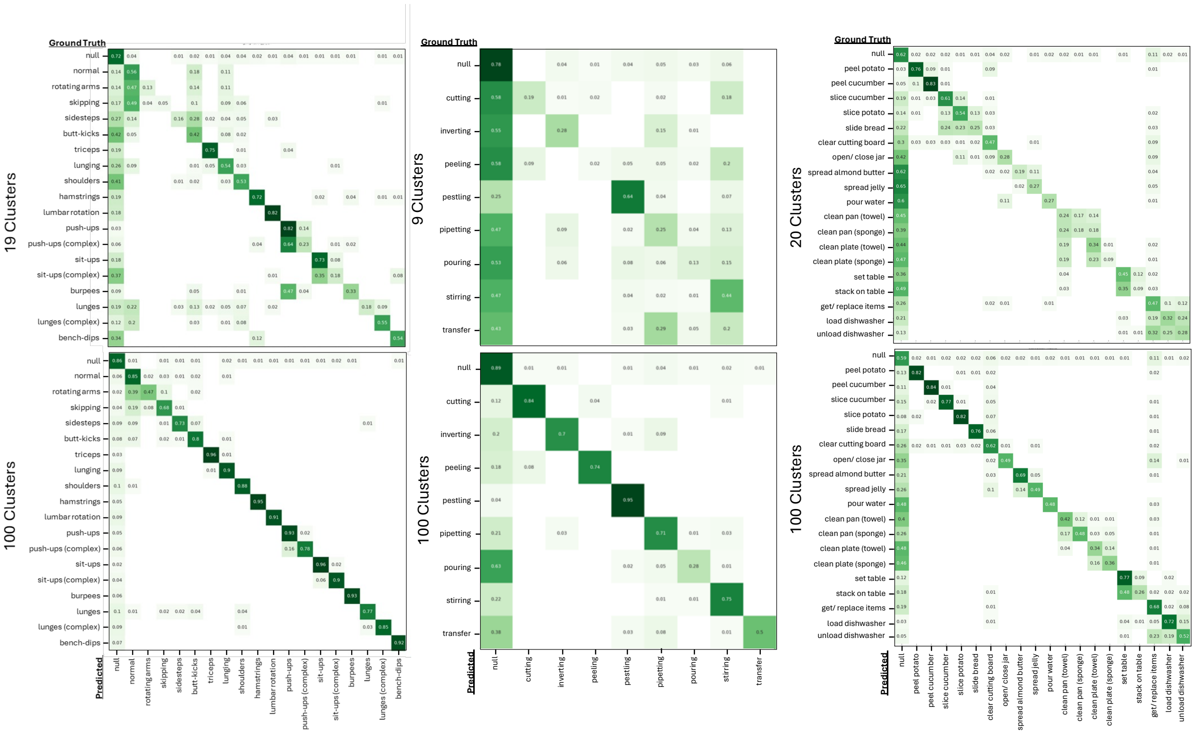
		$c = 19$	$c = 50$	$c = 100$
WEAR	s	L. Acc	L. Acc	L. Acc
	1	50.02 (± 14.69)	75.28 (± 7.97)	83.96 (± 4.48)
	5	50.15 (± 14.67)	75.62 (± 7.94)	84.25 (± 4.34)
		$c = 9$	$c = 50$	$c = 100$
Wetlab	s	L. Acc	L. Acc	L. Acc
	1	27.23 (± 6.45)	54.30 (± 8.79)	66.23 (± 7.86)
	5	27.21 (± 6.47)	54.53 (± 8.94)	66.32 (± 8.19)
		$c = 20$	$c = 50$	$c = 100$
ActionS.	s	L. Acc	L. Acc	L. Acc
	1	36.13 (± 7.53)	50.50 (± 5.09)	57.29 (± 5.64)
	5	36.15 (± 7.46)	50.45 (± 5.26)	57.95 (± 5.78)

Table 2. Comparison of average labeling accuracy and standard deviation across study participants based on different clip lengths. One can see that 4 second video clips result in the best labelling accuracy across (almost) all datasets [2–4]. The table further provides the relative amount of dataset an annotator would need to manually label.

	Clip Length	L. Acc	% Data
WEAR	1 sec.	75.15 (± 5.93)	2.66
	2 sec.	79.11 (± 4.85)	5.32
	4 sec.	83.96 (± 4.48)	10.65
	Clip Length	L. Acc	% Data
Wetlab	1 sec.	64.80 (± 7.90)	0.36
	2 sec.	67.15 (± 6.54)	3.95
	4 sec.	66.23 (± 7.86)	15.79
	Clip Length	L. Acc	% Data
ActionS.	1 sec.	49.66 (± 3.84)	0.66
	2 sec.	53.37 (± 6.74)	3.32
	4 sec.	57.29 (± 5.64)	13.27

A.3 Thresholding

As detailed in the main paper, we apply distance-based thresholding to eliminate outlier instances and increase the labeling accuracy within the cluster. Table 3 details the results of applying three different thresholds (8, 6, and 4). One can see that the smaller the thresholded distance, the more data is being discarded. Nevertheless, we are capable of increasing the average labeling accuracy by a significant margin across all datasets. As shown in the weakly-supervised training section of the main paper, the thresholding significantly boosted the classifiers' performances (especially when applying only a small number of clusters).



(a) WEAR

(b) Wetlab

(c) ActionSense

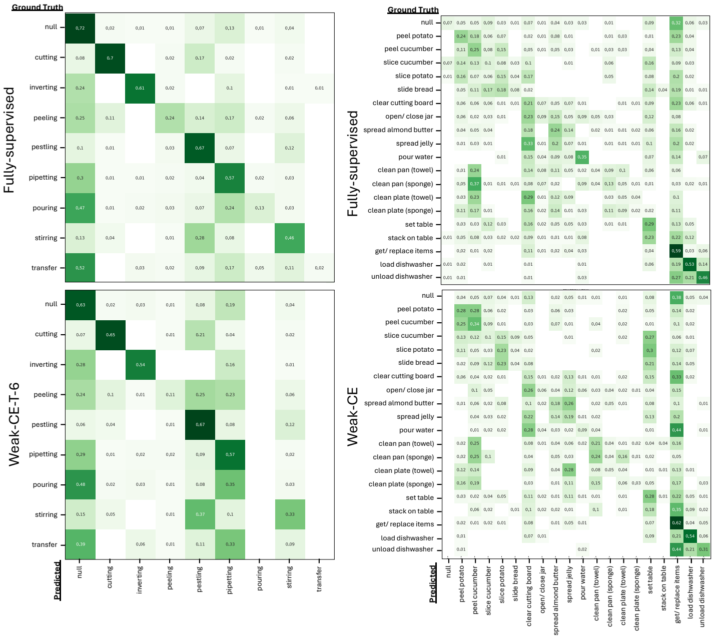
Fig. 1. Confusion matrices comparing the weak-labelling of the (a) WEAR [2], (b) Wetlab [4] and (c) Actionsense dataset [3] applying only (a) 19 (b) 9 (c) 20 versus (a-c) 100 clusters. One can see that with an increasing number of clusters, also similar activities can be differentiated.

B WEAKLY-SUPERVISED TRAINING

Figure 2 provides a comparison of confusion matrices for the Wetlab and ActionSense datasets. As in the main paper, we compare the baseline results with the best-performing weak-supervision approach. One can see that, similar to the WEAR dataset, with the exception of the NULL class, all activities were classified close to the performance of the fully-supervised approach.

Table 3. Average labeling accuracy and standard deviation across study participants applying different degrees of distance-based thresholding. Each setting applied a GMM-based clustering using 100 clusters. Though thresholding decreases amount of available training data, one can achieve a significant increase in labelling accuracy across all datasets [2–4].

		$c = 19$		$c = 50$		$c = 100$	
	t	L. Acc	% Data	L. Acc	% Data	L. Acc	% Data
WEAR	-	50.02 (± 14.69)	100	75.28 (± 7.97)	100	83.96 (± 4.48)	100
	8	53.80 (± 11.87)	64.89	76.57 (± 7.32)	86.89	84.24 (± 4.50)	95.25
	6	62.41 (± 9.02)	33.83	81.70 (± 4.71)	58.78	86.50 (± 4.57)	74.84
	4	87.15 (± 8.56)	11.19	91.07 (± 4.50)	22.51	93.01 (± 3.69)	33.76
		$c = 9$		$c = 50$		$c = 100$	
	t	L. Acc	% Data	L. Acc	% Data	L. Acc	% Data
Wetlab	-	27.23 (± 6.45)	100	54.30 (± 8.79)	100	66.23 (± 7.86)	100
	8	31.89 (± 10.09)	75.31	54.86 (± 9.00)	96.34	66.47 (± 8.04)	99.19
	6	47.94 (± 19.86)	37.97	62.31 (± 12.78)	72.62	70.03 (± 9.98)	85.79
	4	77.98 (± 20.72)	14.02	76.92 (± 15.15)	37.67	77.37 (± 9.83)	51.31
		$c = 20$		$c = 50$		$c = 100$	
	t	L. Acc	% Data	L. Acc	% Data	L. Acc	% Data
ActionSense	-	36.13 (± 7.53)	100	50.50 (± 5.09)	100	57.29 (± 5.64)	100
	8	40.20 (± 8.88)	56.27	53.37 (± 5.63)	80.87	58.51 (± 5.99)	85.14
	6	57.75 (± 10.58)	10.58	64.70 (± 8.05)	34.42	66.53 (± 6.16)	50.66
	4	77.40 (± 9.52)	3.73	80.42 (± 5.74)	8.23	79.84 (± 7.13)	7.13



(a) Wetlab

(b) ActionSense

Fig. 2. Confusion matrices comparing the shallow DeepConvLSTM [1] baseline results compared to that of the best performing weak-labelling approach on the (a) Wetlab [4] and (b) ActionSense dataset [3] . With exception of the NULL-class, all activities were able to be classified close to the performance of the fully-supervised approach.

REFERENCES

[1] 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

[2] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. 2023. WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition. *CoRR* abs/2304.05088 (2023). https://arxiv.org/abs/2304.05088

[3] 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

[4] Philipp M. Scholl, Matthias Wille, and Kristof Van Laerhoven. 2015. Wearables in the Wet Lab: A Laboratory System for Capturing and Guiding Experiments. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*. https://doi.org/10.1145/2750858.2807547