Efficient Human Activity Recognition Based on Grouped Representations of Multimodal Wearable Data

Guillaume Habault(✉) and Shinya Wada

KDDI Research Inc., Fujimino, Japan
{xgu-habault,sh-wada}@kddi.com

Abstract. Human Activity Recognition (HAR) is a vast and complex research domain that has multiple applications, such as healthcare, surveillance or human-computer interaction. Several sensing technologies exist to record data later used to recognize people's activity. This paper aims to linger over the specific case of HAR based on multimodal wearable sensing devices. Corresponding HAR datasets provide multiple sensors information collected from different body parts. Previous approaches consider each information separately or altogether. Vision HAR methods consider each body segment and their position in space in order to perform activity recognition. This paper proposes a similar approach for Multimodal Wearable HAR (MW-HAR). Datasets are first re-sampled at a higher sampling rate (i.e., lower frequency) in order to both decrease the overall processing time and facilitate interpretability. Then, we propose to group sensing features from all the sensors corresponding to the same body part. For each group, the proposal determines a different representation realm of the group information. This abstracted representation depicts the different states of the corresponding body part. Finally, activity recognition is performed based on these trained abstractions of each considered body part. We tested our proposal on three benchmark datasets. Our evaluations first confirmed that a re-sampled dataset offers similar or even better performance for activity recognition than usual processing. But the primary advantage is to decrease significantly the training time. Finally, results show that a grouped abstraction of the sensors features is improving the activity recognition in most cases, without increasing training time.

Keywords: Human Activity Recognition · Data Abstraction · Data Processing

1 Introduction

Human Activity Recognition (HAR) is a complex research topic [29], as it has several sensing technologies and multiple applications.

Depending on the sensing technology used, data collected for HAR can be divided in three groups: (i) Ambient-based: where sensing devices placed at

R. Hou et al. (Eds.): BDTA 2022, LNICST 480, pp. 232–254, 2023.
https://doi.org/10.1007/978-3-031-33614-0_16

fixed location monitor the environment variables (such as smart-homes with actuators for lights or other appliances, temperature sensors, etc.); (ii) Vision-based: where sensing devices are camera or radar either continuously monitoring an environment (such as surveillance) or set-up for specific events (such as live games); and (iii) Wearable-based: where sensing devices, embedded in our clothes or equipment we carry, purposely monitor our conditions (such as vitals, body part movements, temperature, etc.).

Some research [24] might even combine different sensing technologies together in order to cross information and gain in knowledge.

Nevertheless, HAR research proves to be useful in various domains, such as:

1. Healthcare: preventing domestic accidents or detecting anomaly with elderly people [28] as well as assisting people with disabilities or in rehabilitation after an accident [5,14];
2. Live monitoring: preventing crimes or threats [25] or it could also be used to assist live sport events;
3. Human Computer Interaction: interacting with people through games (for fitness game or for rehabilitation of people with disabilities) [10,18]. We can even imagine future applications in virtual worlds, such as in the meta-verse.

Among all these sensing technologies, wearable-based is the most challenging [29]. In fact, placement of body-worn sensors plays an important role in the efficiency of activity recognition. For instance, a smartphone may monitor different body parts at different times, as it can be alternatively positioned in our pocket, hand, against our head (e.g., during a call) or on another object (e.g., while charging). This position uncertainty makes patterns recognition more difficult without external information or ways to determine the position [19].

In Multimodal Wearable HAR (MW-HAR) scenarios [2,23], volunteers are equipped with several sensors placed on different areas of the body. These fixed positions are usually located close to the body joints (i.e., wrist, ankle, hip, etc.) as illustrated on the bottom left part of Fig. 1. This scheme allievates the position uncertainty, but such scenarios are for now unrealistic. Indeed, these days people are carrying/wearing at most two devices (i.e., a smartphone and perhaps an activity tracker). However, with the progress of Internet-of-Things (IoT) (e.g., smart-clothing) and our life being progressively more digital, more of these monitoring devices might be available in the future. In the past years, we have already witnessed the adoption by a large part of the population of activity trackers (such as smart-watch). Therefore, it is fair to assume that in a few years from now, we might carry more of these *wearable sensors*.

According to [7], features extraction and data interpretability are some of the remaining challenges in Wearable HAR. Indeed, sensors' data are noisy [8], (i) presenting lots of fluctuations because of sensors' high precision (capturing micro movements); or (ii) having outliers because of sensors imperfections or calibrations. These variations make it difficult for us to accurately interpret the data, especially when recording long sequences. In addition, machines might interpret these fluctuations as sub-activities and, at the end, reduce their efficiency. Therefore, there is a need for a higher level of abstraction of the sensing

features in order to simplify both (a) the training and learning process for the machine; and (b) the understanding and interpretation for human.

Several researches have been conducted in order to improve recognition efficiency. Most of the proposed solutions are either based on novel Deep Learning (DL) architectures [29] or data augmentation [30] (some solution even use both [17]). Nevertheless, data handling and pre-processing are often neglected in favor of more complex and more efficient architectures.

This paper aims to determine whether a different approach could improve both recognition efficiency and interpretability. For this purpose, we propose to focus on data handling of MW-HAR scenarios. We posit that proper representation of the data associated with current state-of-the-art (SOTA) architectures could further enhance activity recognition. Our approach is twofold:

1. Decrease the amount of samples used during training; and,
2. Accordingly group modalities and generate an abstracted representation.

The contributions of this paper are the following:

1. Compare the relevance in terms of recognition accuracy and training time of different pre-processing methods;
2. Initiate a unified labelling of human activities;
3. Evaluate four grouping strategies and two abstraction techniques to produce a higher level representation of a given group of body-worn features;
4. Visually analyze the impact of the proposal on the data.

The rest of this paper is organized as follows. The subsequent section presents how multimodal wearable data is handled in the literature. In Sect. 3, we describe our approach that aims at (i) re-sampling the dataset to a lower frequency; and (ii) grouping sensing features and defining another representation of such grouped data. Then, we explain our methodology and how our proposal is evaluated, as well as the metrics used. Section 4 describes in details the selected datasets and a start on proposing a unified labelling of human activities. Results of our evaluations are described in Sect. 6. In this section, we also analyze the impact and visual interpretability of our proposal. Before concluding this paper, we further discusses these results in Sect. 7 and we provide new fields of endeavor.

2 Background

To the best of our knowledge, in HAR scenarios, the recognition operations can be summarized as follows:

1. Handle missing data
2. Normalize data
3. Extract additional information (interval-based, frequency-based or using other transformations, such as shapelet)
4. Select important features (either manually or in an automated way)

5. Defining inputs (based on points or sampling, for instance with a sliding window method)
6. Train the model

Wearable HAR (W-HAR) research follows all or some of these steps in order to recognize human activity [26]. Based on the literature, missing measurements are usually ignored or interpolated [6]. Similarly to other Machine Learning (ML) researches, raw sensors' data might lead to issues such as vanishing gradient [31]. Normalization enables to limit these problems. The most commonly employed methods in W-HAR are *min-max* and *standardization*. As mentioned previously, sensors' data present lots of fluctuations. These irregular variations – occurring even within the same activity – make it more difficult for models to differentiate activities only based on raw data. Extracting information or transforming data can therefore provide additional information on a feature or a set of features. Chen et al. [7] surveyed the different methods to produce such additional information. This step commonly uses all raw features, but it can also be sensor-based or feature-based with either a dedicated or shared generation process. Statistical methods (such as determining minimum, maximum, mean and variance over a sliding window) are the basis for such a generation of additional features. However, adding multiple statistics increases the computational cost and often requires domain knowledge. To cope with this issue, Qian et al. [22] proposed a method for generating only relevant statistics. Another way to reduce the impact of these additional features is to select the most relevant ones. Several strategies defined in time-series studies (such as regularization, correlation or even more complex one [13]) exist to perform this step.

The final set of features is either used altogether or each of them separately to perform the recognition task. We believe that such an approach is the main reason for the lack of interpretability. In fact, even though the number of features is decreased, considering all of them together requires very complex architecture [20]. But with such architectures, we let *black-box* models discover any correlations – more generally, any relations – between activities and data from all these features. But the more complex the architecture is, the less it provides tools to interpret its decisions [1]. For instance, a multimodal-based architecture, which investigates both auto- and cross-modal relationships, such as [32], will see its complexity exponentially increase with the number of modalities. As a consequence, the usage of very complex architectures hinders understanding or visualizing the interpretations of these models.

All these reasons motivate us to seek for a different approach to the pre-processing phase when dealing with multiple body-worn sensing devices.

3 Proposed Approach

Based on the previously mentioned observations, we adopted a different point-of-view on human activity data when targeting recognition. Our approach takes place at the pre-processing level and can be divided into two parts.

3.1 Re-sampling Dataset

We consider to re-sample the measurements in order to decrease the number of inputs in the dataset. This preliminary step basically changes the resolution of the measurements to a coarser level (i.e., sampling frequency). The main effect of this step is to reduce the impact of micro fluctuations and any outliers by acting as a low-pass filter. In addition, it will take care of dealing with most missing data within the dataset. As a result, we hypothesize that with fewer samples to manage, the model will find relations between features and activities more easily. Consequently, it decreases the associated cost when manipulating them (processing and training). Moreover, when plotting the data, there will be fewer points to visualize, therefore simplifying both analysis and interpretation.

However, these benefits mitigate the coarser the selected re-sampling gets. Therefore, this rate needs to be selected depending on the targeted applications in order to maintain most of the measured information.

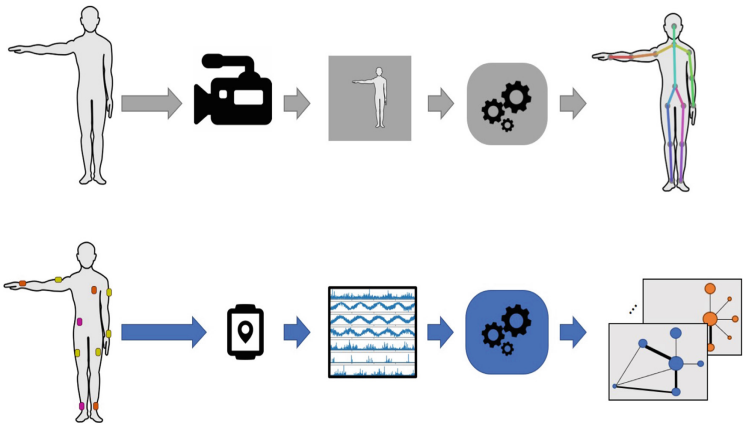


Fig. 1. Illustration for our proposal (bottom pipeline) compared to the camera-based process (top pipeline). Camera-based produces body graph from pictures. In the same way, we propose to produce abstractions of groups of sensing features (depicted here as state transition graph).

3.2 Specific Features Grouping and Data Abstraction

In other HAR studies, information directly related to some activities can be extracted. In smart-home scenarios, some ambient sensors monitor objects explicitly linked to an activity or a set of activities. For instance, when the entrance door is opened, the model can limit the potential activities to *leaving* or *entering* home [11]. Information from other sensors will then settle the decision. For vision-based studies, recognition of surrounding objects can also give hints and restrict the potential activities [15]. Such information about the

environment and surrounding objects is highly beneficial for activity recognition. However, for W-HAR, these types of information are usually not available (unless combined with others [6]). Moreover, as aforementioned, wearables' position uncertainty hinders the learning process of relations between activities and sensors' data.

On the other hand, camera-based HAR has an interesting way of treating images in order to recognize activities performed. From an image, a coarse body representation is produced [3]. In this graph-like representation –referred as body-graph in the rest of the paper–, each segment represents a body part, while nodes depict the body joints. Finally, a sequence of body graphs (i.e., sequence of stances/postures) can relate to an activity. In camera-based HAR, body graph information is used for activity recognition task, but not raw image information. On the contrary, in MW-HAR studies, raw measurements of the body parts are used to perform the activity recognition task.

Therefore, the key idea of our proposal is to process raw sensors' data in order to produce a mid-level abstraction. As illustrated in the bottom part of Fig. 1, the pipeline of our idea is similar to the one used in camera-based HAR, (top part of Fig. 1). This idea is based on the observation that, at a fixed time-step, models should be efficient in recognizing the state of different body parts, but not really the activity, or at least not directly. The activity will be efficiently recognized when considering a sequence of body parts' states. For example, with a person performing a single arm lateral raise, there are two main states (i.e., *arm along the body* and *arm raised*) and multiple transition states. Therefore, instead of using raw data (acceleration, rotation, etc.), if data can represent that the wrist is moving in-between two states, while other body parts remain in the same state (standing still), it should be easier for the model to recognize the lateral raise.

To summarize, this part consists of the following steps:

1. Group sensors that monitor the same body part
2. (optional) Extract additional information on the set of features
3. Produce abstracted representations of monitored body parts

Finally, obtained abstractions become modalities of the model tackling with activity recognition.

Table 1. Mapping of activity standardized names and original ones for the targeted datasets.

Category	ID	Activity	mHealth	PAMAP2 Protocol	LDPR
1	10	Passive Standing	Standing still		
	11	Active Standing		Standing	
	12	Passive Sitting	Sitting and relaxing		Sitting
	13	Active Sitting		Sitting	
	14	Passive Lying	Lying down		Lying
	15	Active Lying		Lying	
	16	Passive Walking	Walking	Walking	Walking
	17	Active Walking			
	18	Falling			Falling
	19	On All Fours			On all fours
2	20	Standing Up			Standing up from sitting Standing up from sitting on the ground Standing up from lying
	21	Sitting Down			Sitting down Sitting on the ground
	22	Lying Down			Lying Down
3	30	Ironing		Ironing	
	31	Vacuuming		Vacuum cleaning	
4	40	Ascending Stairs	Climbing stairs	Ascending stairs	
	41	Descending Stairs		Descending stairs	
5	50	Nature Walk		Nordic walking	
	51	Jogging	Jogging	Running	
	52	Running	Running		
	53	Cycling	Cycling	Cycling	
6	60	Jumping	Jump front and back	Rope jumping	
	61	Waist Bend	Waist bends forward		
	62	Elevation Arm	Frontal elevation of arms		
	63	Knee Bend	Knees bending (crouching)		

4 Datasets

In order to test our proposal properly, we focus on MW-HAR datasets that (a) provide data from sensors monitoring different body parts, (b) if possible, include a diversity of sensors, and (c) target Activities of Daily Living (ADL) or at least full body activities (e.g., standing, sitting, etc.). Finally, we selected three datasets to test our proposal efficiency when performing ADL recognition.

4.1 Target Activities

Each dataset used different terminology for activities that often are similar. In order to simplify comparison in between these datasets, we propose a mapping between a unified terminology and the original ones in Table 1. In this table, empty cells suggest that the considered dataset does not target corresponding activities. Note that such a unified labelling could be extended for future usage.

Similarly to [20], in the first category, "Passive" refers to people being in a given stance doing nothing in particular (e.g., walking, standing or sitting still). While "Active" refers to people performing a specific activity while maintaining a stance (e.g., sitting and typing or standing and talking to someone). Therefore, "Passive" activity recognition targets recognition of a stance, while "Active" targets recognition of an activity performed in a given stance. Even though not present in these datasets, some activities can be carried out in different stances. The second category lists stance transition (targets spontaneous dataset, as for scripted one, transitions between activities are ignored). The third one presents household chores. The fourth category is related to stairs movements. The fifth one describes activities that may target transportation and/or exercise. The last category comprises exercise and fitness activities.

4.2 Datasets Presentation

Mobile Health (mHealth) [2] and Physical Activity Monitoring (PAMAP2) [23] are scripted datasets, meaning that volunteers perform each activity for a given duration (e.g., 1 min) or a given number of repetition (e.g., 20 times). Both datasets monitor the volunteer's heart from the chest, mHealth with a 2-lead electrocardiogram (ECG) and PAMAP2 with a heart-rate monitor. For both datasets, volunteers are wearing Inertial Measurement Units (IMUs) on the chest, one wrist and one ankle. IMUs comprise at least a 3D-Accelerometer, a 3D-Gyroscope and 3D-Magnetometer (except for the chest one in mHealth). Each IMU of PAMAP2 is also equipped with a temperature sensor and an additional 3D-Accelerometer. In contrast, Localization Data for Posture Reconstruction (LDPR) [16] is a spontaneous dataset. Volunteers perform activities in a given environment without specific order or duration for the activities. In this dataset, volunteers are wearing position tag on the chest, the belt and both ankles. Table 2 summarizes the main parameters of these datasets.

Table 2. Summary of the different datasets parameters: number of volunteers, sensors positions and types, as well as measurement rate.

Dataset	Num. subject	Sensors position	Sensors type	Record rate
mHealth	10	Chest	2-lead ECG	50 Hz (0.02 s)
			3D-Accelerometers	
		Right Wrist	3D-Acc.,3D-Gyroscope and 3D-Magnetometer	
		Left Ankle		
PAMAP2	9	Chest	Heart-rate Monitor	9 Hz
			Temperature, 2 3D-Accelerometers, 3D-Gyroscope, 3D-Magnetometer	100 Hz (0.01 s)
		Dominant Wrist		
		Dominant Ankle		
LDPR	5	Chest	Position	250 Hz (0.004 s)
		Belt		
		Left ankle		
		Right ankle		

Spontaneous datasets are more difficult to handle because of imbalance record among activities. In fact, scripted datasets ensure to have very similar amount of records, assuming that volunteer followed the script. Conversely, a spontaneous dataset is subject to the volunteer behavior. As a result, some activities might be under-represented (i.e., not enough or no repetitions at all). Without proper handling, it can lead to overfit the trained model to the most represented activities. Although some solutions exist to solve potential imbalance in the activity distribution, we do not consider them in this investigation. In fact, except for transition activities (category 2 in Table 1), LDPR has enough records per activity. In addition, for this study, we do not consider sub-activities, as all the corresponding records are labelled with their main activity. For instance, we labelled *Standing up from sitting* and *Standing up from lying* as *Standing up*. This labelling trick resolves the imbalance in this category. Moreover, it enables us to focus on our proposal and its impact on input recognition.

mHealth. consists of records from 10 volunteers. In this dataset, all actions performed between two scripted activities (labelled *Null*) are ignored in our study. This dataset is composed of one file per volunteer. Therefore, in order to create training, validation and evaluation sets, we randomly sampled unique data from each considered activity based on a 0.6/0.15/0.25 ratio.

PAMAP2. comprises records from 9 volunteers for the main protocol, among which 5 of them performed optional activities. In this study, we focus only on the main protocol. As shown in Table 5, the number of inputs of the ninth volunteer is extremely low compared to the others. In fact, this volunteer only performed *Jumping* in the main protocol. Thus, we removed this volunteer from

our experiments, as it is pointless to perform activity recognition in this situation. In this dataset, break time actions –performed between two scripted activities– are also labelled as *Null*. Similarly to mHealth, and as suggested by the dataset's creator, we ignored them in our study. In addition, we use the same training/validation/evaluation split ratio.

LDPR. consists of records from 5 volunteers. Each volunteer performed 5 runs. For this dataset, we used runs 1 and 2 to train models. Run 3 is used for validation, while runs 4 and 5 are used for evaluating the trained models.

5 Evaluation Method

In this paper, we aim to investigate the efficiency of our proposal in recognizing ADL. In order to achieve this goal, we target to decrease the quantity of samples and group features by body parts. Associating the latter with an abstraction technique will produce a new representation of the data. We infer that the obtained abstraction of the data will represent the different states a body part can fall into (along with transitions between these states).

Table 3. Summary of the distinct steps performed within each pre-processing method.

	Basic (P1)	Common (P2)	Re-sampled (P3)
Re-sampling	X	X	✓
Handling missing data	Ignored	Linearly interpolated	
Normalization	Z-Score		

5.1 Evaluation of Preliminary Re-sampling

This evaluation consists of determining the number of inputs generated with different pre-processing as well as assessing the accuracy of the resulting dataset for inputs-based recognition task. These results are then compared in order to establish their advantages and disadvantages. As detailed in Table 3, we investigate three types of data pre-processing. The first one limits the processing to a minimum, i.e., ignoring missing data and normalizing the rest. This method is referred as the basic pre-processing (P1). Conversely, the pre- processing (P2) linearly interpolates missing data and normalizes the dataset. This pre-processing is frequently found in W-HAR publications and is considered as the baseline of this evaluation. Finally, the re-sampled pre-processing (P3) method begins with re-sampling the data to a coarser resolution. Then, similarly to P2, we interpolate the remaining missing data and normalize the dataset.

Re-sampling. A record rate close to the millisecond or tenth of milliseconds (as usually found in MW-HAR datasets) is too detailed for describing coarse body-parts states. Indeed, for ADL, our movements do not significantly change at such low time intervals. Therefore, in this study, we selected a re-sampling at 50 ms (i.e., every 0.05 s). This preliminary step averages the values of each feature over a non-overlapping 50 ms sliding window. As an example, with a dataset that recorded measurements every 0.01 s, 5 measurements are used to create one measurement of the re-sampled dataset. Note that for this pre-processing method, we ignore $50ms$ inputs in between two activities. It ends up only removing a few inputs, which is marginal compared to the tenth of thousands of remaining inputs. Finally, comparing different re-sampling rate in order to determine the optimal one is out-of-scope of this study, as optimal rate might vary depending on the application.

Normalization follows the z-score principle: $x_{norm} = \frac{x-\mu}{\sigma}$, where μ and σ are the training values mean and standard deviation, respectively, for each feature.

Table 4. Summary of the different parameters used for each grouped abstraction.

	G0	G1	G2	G3	G4
Grouping Strategy	None	Feature	Sensor	Body part	All
Data Abstraction	X	✓	✓	✓	✓

5.2 Evaluation of Grouped Abstraction

In this second evaluation, we compare the accuracy obtained with different combination of grouping strategy and data abstraction technique. As detailed in Table 4, we consider five grouping strategies. First, recognition is performed on all features together and without data abstraction (G0). Then, we test one abstracted representation for each feature (G1). This strategy may uncover if a specific and more efficient projection of a single feature can be found. Moreover, we consider one abstracted representation method for each sensor (G2) (e.g., an 3D-accelerometer representation is based on values of its three components). As explained in Sect. 3.2, we proposed to group sensing features corresponding to the same body part and to produce a representation of the different states of such a body part (G3). Finally, we also test one abstracted representation for all features together (G4). We consider G0 and G4 as baselines because previous works have already used these techniques. In addition, G2 and G4 enables us to further evaluate the relevance and assess if one abstracted representation method for each body part (G3) can be more efficient for recognition task.

Data Abstraction. The second part of our proposal relies on representing the data in such a way that it will better describe the state of different body parts.

This transformation could be performed manually by listing the different states of a given body part and perform a mapping between sensors' measurements and corresponding states. However, to the best of our knowledge, no HAR datasets provide such type of information. Furthermore, processing such information from current datasets requires a significant amount of knowledge and effort. As a result, in this paper, we investigated dimension reduction techniques. As the name implies, they enable to reduce the dimension or at least project the data into another space with the same dimension. In this evaluation, we considered two techniques: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). We focus on them here, as they do not require fine tuning and their computational cost is limited compared to other methods.

For group strategies G1 to G4, we trained a dedicated abstraction for each group. For instance, in G1 configuration, a representation is trained for each feature, while for G4, we trained one representation for all the considered features.

5.3 Classification Method and Evaluation Metrics

For this study, only one classification method is used: k Nearest Neighbors (k-NN) ($k = 11$ in our experiments). We omit DL methods because two of the selected datasets are scripted ones (cf. Sect. 4). For them, we assumed that sequenced inputs are unnecessary. However, SOTA activity recognition models are based on Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Transformers, which require such sequenced inputs.

We evaluate the classification performance with two metrics: (1) the *score* of the model on the evaluation set (i.e., the mean accuracy as defined in *scikit-learn* library *KNeighborsClassifier*; and (2) the duration for training the model.

Table 5. Number of inputs for all datasets after applying the different pre-processing methods (P). mHealth has originally no missing values, so the number of inputs is the same for the first two pre-processing. LDPR has missing values on several time-steps as sensors records are not synchronized. Therefore, P1 ends up ignoring the majority of inputs. The re-sampling method decreases the number of inputs by 60%, 80% and 37% for mHealth, PAMAP2 and LDPR, respectively.

Dataset	P	Volunteer									
		1	2	3	4	5	6	7	8	9	10
mHealth	1	35174	35532	35380	35328	33947	32205	34253	33332	34354	33690
	2										
	3	14066	14210	14150	14130	13578	12880	13698	13328	13739	13476
PAMAP2	1	22590	23691	15841	20863	24592	22659	21055	23624	583	
	2	249957	263349	174338	231421	272442	250096	232776	262102	6391	
	3	49980	52661	34861	46275	54478	50013	46547	52414	1278	
LDPR	1	0	0	0	1	1					
	2	26285	29024	30695	30912	42665					
	3	16507	18432	19355	19472	26916					

6 Results

6.1 Re-sampling Performance

This evaluation aims to determine if the re-sampled pre-processing (P3) achieved similar performance to the common pre-processing (P2). For this evaluation, for each considered dataset, we first performed recognition per volunteer. Then, we carry out recognition when training a unique model for all volunteers (referred in the rest as the *1-for-all model*). The training set in this scenario is composed of the training data from all the volunteers. Finally, each volunteer's evaluation set is classified by this trained model. Because we intend to test the pre-processing performance, we fixed the grouping strategy to G0.

Table 6. Activity recognition accuracy with different pre-processing methods (P). Values in italic represent the percentage decrease (▼) or increase (▲) when using a 50 ms re-sampled pre-processing (P3) compared to the common one (P2).

Dataset	P	Volunteer										All
		1	2	3	4	5	6	7	8	9	10	
mHealth	1	0.9811	0.9600	0.9891	0.9889	0.9915	0.9948	0.9971	0.9960	0.9934	0.9954	0.9863
	2	0.9811	0.9600	0.9891	0.9889	0.9915	0.9948	0.9971	0.9960	0.9934	0.9954	0.9863
	3	0.9773	0.9468	0.9774	0.9816	0.9838	0.9926	0.9948	0.9928	0.9910	0.9911	0.9798
		▼0.39%	*▼1.375%*	*▼1.18%*	*▼0.74%*	*▼0.78%*	*▼0.22%*	*▼0.23%*	*▼0.32%*	*▼0.24%*	*▼0.43%*	*▼0.66%*
PAMAP2	1	0.9800	0.9490	0.9614	0.9701	0.9603	0.9645	0.9728	0.9746			0.9581
	2	0.9980	0.9952	0.9972	0.9978	0.9982	0.9968	0.9977	0.9979			0.9962
	3	0.9905	0.9789	0.9838	0.9857	0.9838	0.9830	0.9894	0.9868			0.9800
		▼0.75%	*▼1.64%*	*▼1.34%*	*▼1.21%*	*▼1.44%*	*▼1.38%*	*▼0.83%*	*▼1.11%*			*▼1.63%*
LDPR	1	–	–	–	–	–						–
	2	0.7367	0.7315	0.7550	0.5780	0.6817						0.7150
	3	0.7388	0.7462	0.7711	0.5784	0.6897						0.7253
		▲ 0.28%	*▲ 2.01%*	*▲ 2.13%*	*▲ 0.07%*	*▲ 1.17%*						*▲ 1.44%*

Table 7. Training time (in seconds) with different pre-processing methods (P) using different datasets (D). Values in italic represent the percentage decrease (▼) or increase (▲) when comparing a $50ms$ re-sampled pre-processing (P3) with the common one (P2).

		volunteer										**All**
D	**P**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
mHealth	**2**	21.37	21.92	22.25	20.28	19.97	17.91	19.32	18.80	21.79	19.91	1928.08
	3	3.92	4.02	3.92	3.86	3.64	3.25	3.72	3.64	4.08	3.72	310.92
		▼81.7%	*▼81.7%*	*▼82.4%*	*▼81.0%*	*▼81.8%*	*▼81.9%*	*▼80.8%*	*▼80.6%*	*▼81.3%*	*▼81.3%*	*▼83.9%*
PAMAP2	**2**	1184.67	1128.58	533.80	972.34	1419.79	1006.05	945.23	1155.67			102537.39
	3	46.07	46.24	22.00	40.32	51.46	44.16	39.75	48.78			4226.43
		▼96.1%	*▼90.8%*	*▼91.2%*	*▼90.5%*	*▼90.9%*	*▼90.9%*	*▼90.4%*	*▼90.3%*			*▼91.9%*
LDPR	**2**	5.42	5.52	5.58	5.78	10.02						75.89
	3	2.75	3.02	3.16	3.52	5.16						39.72
		▼49.3%	*▼45.3%*	*▼43.4%*	*▼39.1%*	*▼48.5%*						*▼47.7%*

Table 5 lists the total number of inputs obtained after each pre-processing. Because of the asynchronicity of the measurements in LDPR, P1 ends up ignoring most of the inputs, as shown in this table. Nonetheless, from this table, we can notice that *not ignoring missing data* (P2) significantly increases the total number of inputs for PAMAP2 – there is no impact on mHealth as it has no missing data. However, this number is greatly decreased with the preliminary re-sampling step (P3). At the end, mHealth [resp. PAMAP2, LDPR] obtained on average a drop of 60% [resp. 80%, 37%] compared with the common method.

Table 6 describes the accuracy obtained for activity recognition with the different pre-processing methods for the targeted datasets. From these results, we notice that the common pre-processing (P2) is performing better than the others for mHealth and PAMAP2. However, results obtained with the re-sampled pre-processing (P3) are fairly close, with on average a drop in accuracy of 0.59% and 1.21% for mHealth and PAMAP2 respectively. But as shown in Table 7, in returns, the training time is decreased by approximately 81.4% and 90.7% for mHealth and PAMAP2 respectively. Even though not presented in this paper, a re-sample rate of 25 ms can further improve the accuracy performance for mHealth and PAMAP2, while still substantially reducing training time. Conversely, P3 is the best for LDPR and improve accuracy on average by 1.13%. Besides, it still decreases the training time by approximately 45.1% (cf. Table 7).

As a result, the re-sampled pre-processing (P3) is worth considering for MW-HAR scenarios. Its advantages (reducing the number of inputs, while smoothing the data) will prove to be especially important for configuration tackling activity recognition task with more complex architectures and large datasets.

As a conclusion, this evaluation shows that the re-sampled pre-processing (P3) can achieve similar and even better performance than the common pre-processing (P2). However, there is a trade-off to be considered between the re-sampling rate (which affect the number of inputs and the training time) and recognition accuracy.

6.2 Grouped Abstraction Performance

In this evaluation, in order to simplify the analysis, we do not integrate P1 and concentrate on the other pre-processing methods. In addition, as shown in Table 8, we listed only the recognition accuracy of the best- and worst-performing volunteers from previous evaluations. We also included the performance of the *1-for-all model*. Based on our previous experiment, volunteer 2 [resp. 2, 4] and 7 [resp. 1, 3] are the ones with respectively the worst and best performance for mHealth [resp. PAMAP2, LDPR]. Table 8 presents the performance of different group strategies using either PCA or LDA as the abstraction technique. In this table, bold colored [resp. italic] value represents the best [resp. the second best] performance for a given scenario (i.e., combination of a pre-processing, a grouping strategy and an abstraction technique for a given volunteer).

As one could expect, we can first conclude that LDA is generally better than PCA for producing a different representation of the data. For scripted datasets

Table 8. Activity recognition accuracy for all datasets (D) with different (i) pre-processing (P), (ii) grouping strategies (G) and (iii) abstraction techniques (A = {PCA, LDA}). Results are presented for the best- and worst-performing volunteers on each dataset, as well as the scenario with a unique model for all volunteers (*1-for-all*). Results in bold [resp. italic] depicts the best [resp. second best] performance for the considered scenario (i.e., one pre-processing, one grouping, and one abstraction).

		Volunteer											
D	G	Worst	Best	All	Worst	Best	All	Worst	Best	All	Worst	Best	All
mHealth	0	**0.9600**	**0.9971**	**0.9863**	0.9600	0.9971	**0.9863**	**0.9468**	**0.9948**	**0.9798**	0.9468	0.9948	***0.9798***
	1	***0.9581***	***0.9956***	***0.9830***	0.9632	***0.9982***	0.9846	***0.9460***	***0.9918***	***0.9770***	0.9542	***0.9962***	0.9796
	2	***0.9581***	***0.9956***	***0.9830***	***0.9633***	**0.9992**	0.9850	***0.9460***	***0.9918***	***0.9770***	**0.9564**	**0.9977**	**0.9802**
	3	***0.9581***	***0.9956***	***0.9830***	**0.9653**	**0.9992**	***0.9852***	***0.9460***	***0.9918***	***0.9770***	***0.9556***	**0.9977**	**0.9802**
	4	0.9488	0.9916	0.9702	0.9541	0.9975	0.9773	0.9272	0.9892	0.9627	0.9454	0.9959	0.9714
PAMAP2	0	**0.9952**	**0.9982**	**0.9974**	0.9952	0.9982	**0.9974**	**0.9789**	**0.9905**	**0.9852**	0.9789	0.9905	**0.9852**
	1	***0.9937***	***0.9978***	***0.9954***	***0.9993***	**0.9995**	***0.9968***	***0.9755***	***0.9885***	***0.9810***	0.9949	***0.9971***	***0.9843***
	2	***0.9937***	***0.9978***	***0.9954***	**0.9994**	**0.9995**	***0.9968***	***0.9755***	***0.9885***	***0.9810***	0.9951	0.9968	0.9841
	3	***0.9937***	***0.9978***	***0.9954***	**0.9994**	**0.9995**	***0.9968***	***0.9755***	***0.9885***	***0.9810***	***0.9957***	**0.9974**	0.9839
	4	0.9796	0.9891	0.9737	**0.9994**	***0.9994***	0.9895	0.9566	0.9723	0.9480	**0.9967**	***0.9971***	0.9705
LDPR	0	***0.5780***	**0.7550**	**0.7150**	0.5780	0.7550	0.7150	***0.5784***	***0.7711***	**0.7253**	0.5784	0.7711	0.7253
	1	***0.5780***	**0.7550**	**0.7150**	***0.5994***	**0.7917**	0.7242	***0.5784***	***0.7711***	**0.7253**	***0.6062***	**0.8032**	0.7354
	2	***0.5780***	**0.7550**	**0.7150**	**0.6179**	***0.7795***	**0.7282**	***0.5784***	***0.7711***	**0.7253**	**0.6281**	***0.7818***	**0.7396**
	4	**0.5863**	***0.7541***	***0.7127***	0.5807	0.7614	***0.7258***	**0.5807**	**0.7749**	***0.7199***	0.6040	0.7676	***0.7366***
		PCA			LDA			PCA			LDA		A
		2						3					P

(mHealth and PAMAP2) with any grouping strategy, PCA always failed to provide improvement compared to the configuration where all features are considered without data abstraction (G0). For LDPR (the spontaneous dataset), considering all features (G4) with PCA data abstraction improved only the accuracy of the worst-performing volunteer when using the common pre-processing (P2). While with the re-sampled pre-processing (P3), it improved the accuracy of both volunteers. Nevertheless, PCA abstraction always failed to improve accuracy for the *1-for-all* scenario. As a consequence, the rest of our analysis focuses on LDA.

We directly noticed that, for worst-performing volunteers, LDA abstractions with higher level of grouping strategies (G3 and G4) perform better. While, for best-performing volunteers, lower-level ones are to be considered (G1 to G3).

We draw from this observation that volunteer's data that do not have obvious patterns benefit from a higher level of grouped abstraction. Nonetheless, for scripted datasets, LDA associated with any grouping strategies mostly failed to improve the accuracy of the *1-for-all* scenario. This observation implies that it is difficult to produce a generic abstraction from data collected from different volunteers – some might have particularities preventing from such a generalization.

When considering scripted datasets and common pre-processing (P2), we noticed that the accuracy obtained with grouped abstractions was almost the same despite the grouping strategy. Conversely, with re-sampled pre-processing (P3) there are more fluctuations between the performance of each grouping strategy.

Thus, for mHealth, G3 is the best grouping strategy with P2. However, with P3, G3 is slightly less efficient than G2 for the worst-performing volunteer. Otherwise, their performances are the same. Considering that G3 is computationally less expensive, G3 combined with LDA offers the best performance for this dataset for both pre-processing methods.

Focusing on PAMAP2 results, we observe that with P2, G2 and G3 have the same performances. However, G3 and G4 have better recognition accuracy with P3. Therefore, determining which one is better in this dataset depends on the computational cost requirement. For a faster training time, higher levels of grouping should be preferred, otherwise lower level ones should be selected.

When considering LDPR, readers can notice that there is no row for the grouping strategy G3. As mentioned previously, LDPR used only one type of sensor that monitors the spatial coordinate of the considered body part. Therefore, grouping by sensor or by body part is the same for this dataset. Then, only G2 appears in Table 8. For this dataset, G2 [resp. G1] is the best grouping strategy for the worst-performing [resp. best-performing] volunteer as well as the *1-for-all* scenario. This observation is valid for both pre-processing methods, but the re-sampled one (P3) is providing the best results in all scenarios (i.e., worst- and best-performing volunteers as well as *1-for-all*). Therefore, G2 combined with LDA is offering the best performance. Based on previous results, we can suppose that G3, combined with an abstraction technique, will also perform well in an MW-HAR spontaneous dataset.

As a conclusion, this evaluation shows that the re-sampled pre-processing (P3) associated with a grouped abstraction of the sensing features offers better performance compared to a scenario not using grouped abstraction (G0). Unless a unique model for all volunteers is necessary, as for such a scenario, it will depend on the type of dataset. In addition, grouping sensing features is not necessarily increasing the overall computational complexity. Indeed, the bigger the number of features in a group, the fewer the number of abstractions. Therefore, bigger groups limit the cost and impact of the abstraction technique. We assume that more complex abstraction techniques will achieve even greater activity recognition accuracy. However, there will be a trade-off to consider between the computational cost (the overall training time) and recognition accuracy.

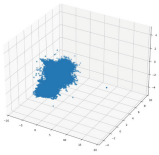
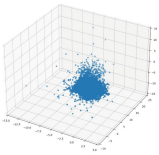
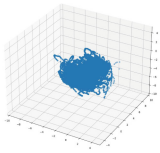
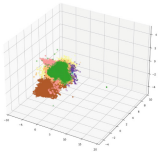
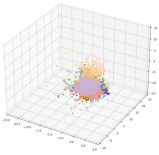
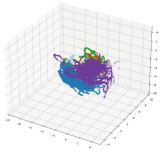
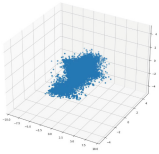
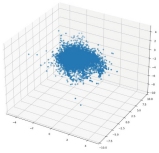
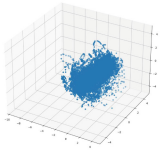
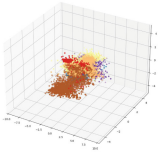
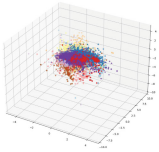
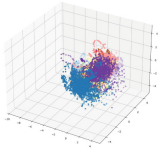
Based on these evaluations, for all datasets and any pre-processing method, one abstracted representation method for each body part (G3) appears to be a suitable compromise for efficient computational performances (accuracy and processing time). In addition, worst- and best-performing volunteers are unknown in a real scenario, therefore G3 appears as the optimal default grouping strategy.

6.3 Visualization of Grouped Abstraction

This section aims to visualize the effect of both the re-sampled step and the grouped abstraction. Thus, we focus on scripted datasets that have more sensors. Due to space limitation, we do not show plots of all sensors and all volunteers.

Table 9 shows 3D scatter plots of training measurements from three sensors worn by the volunteer 4 of PAMAP2. We selected this volunteer as it presents average performances in the results of Sect. 6.1.

Table 9. 3D scatter plots of 3 sensors (Chest and wrist accelerometers and wrist gyroscope) for volunteer 4 from PAMAP2 after either P2 or P3 pre-processing. P2c and P3c rows represent the same plots with colors per activity showing that, even with colors, re-sample only is not sufficient to visually differentiate activities.

P	Chest Accelerometer	Wrist Accelerometer	Wrist Gyroscope
2			
2c			
3			
3c			

In order for recognition models to perform efficiently, measurements corresponding to the same activity [resp. different activities] have to be close to [resp. distant from] each other.

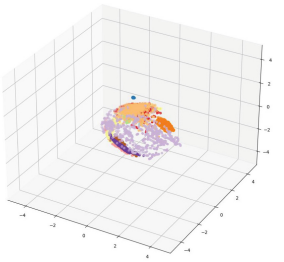
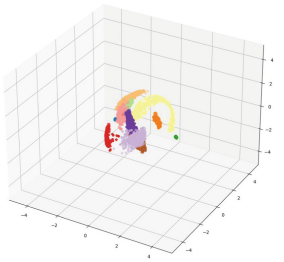
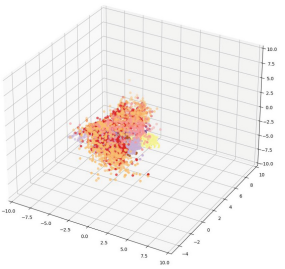
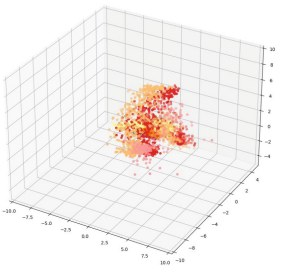
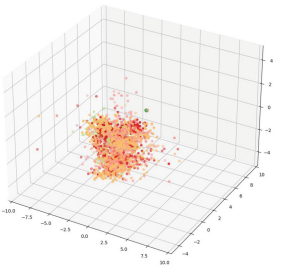
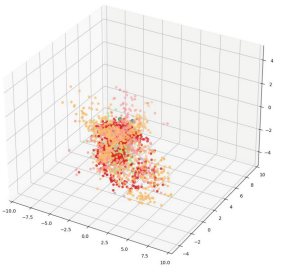
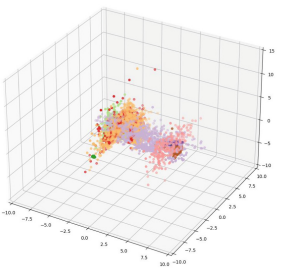
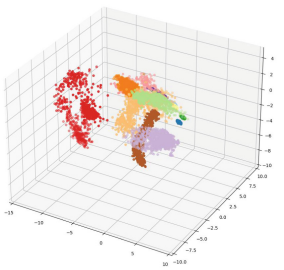
From these figures, we observe that it is impossible to distinctively identify activities. The re-sampled pre-processing (P3), even though removing some outliers and decreasing the total number of points, does not make this identification easier. Finally, associating a color per activity, can sometimes simplify this task, but it mostly remains difficult. Therefore, with such a raw data/representation, activity recognition is subject to errors (especially with spontaneous datasets).

Table 10 compares 3D scatter plots without (G0) and after grouped abstraction (G3 combined with LDA). In this table, we concentrate on mHealth and its worst- and best-performing volunteers according to the experiment of Sect. 6.1. From these plots, we first notice that raw measurements (G0) from the wrist gyroscope have already a good per activity separation for volunteer 2 (i.e., the worst-performing one). These separations are even more distinct for volunteer 7 (i.e., the best-performing one). Nonetheless, the proposal ends up losing this specificity. On the other hand, such demarcations are not clear from the raw measurements of the ankle accelerometer. However, the considered grouped abstraction improves the separation for the volunteer 2. The demarcations are significantly better for the volunteer 7 with a clear cloud per activity.

As a result, these plots show that our proposal is not necessarily beneficial for all features of a given group. Some features would benefit from the knowledge of the group, while other should remain in their raw format. For instance, as shown in Table 10, for the best-performing volunteer, a recognition model would benefit more from using the raw representation (G0) of the wrist gyroscope and the G3 representation of the ankle accelerometer. Another way to enhance the performance of grouped abstraction could be to use feature selection mechanisms (such as the ones presented in [4]). Such an additional step will enable the model to select the most significant abstracted features and probably to achieve a better recognition. Finally, dimension reduction could also be a solution to solve the degradation of some features resulting from our proposal. With such a reduction, data from multiple sensors belonging to the same body part could be projected in a lower dimensional space. For instance, a three-dimensional space for easy interpretation. The resulting data would represent different body part states and provide higher-level knowledge of sensors' data.

Last but not least, these visualizations of raw sensor data also help us determine that activities with the same name from different datasets could potentially be different. For instance, we assume that *Running* and *Cycling* were performed either indoor (i.e., on a machine) or on a straight and flat path for PAMAP2. While for mHealth, the same activities were probably realized outdoor, on a path that has elevation gains and curves. Indeed, in the former chest acceleration mostly follows a single axis. When in the latter, measurements present variations on all three axes. Further investigations are then required in order to clarify these disparities and update accordingly the terminology work initiated in Sect. 4.1. Such a task would improve the global recognition knowledge.

Table 10. 3D scatter plots of 2 sensors (Ankle accelerometer and wrist gyroscope) for volunteers 2 (Worst performance) and 7 (Best performance) from mHealth without (G0) and with grouped abstraction (G3 + LDA). Grouped abstraction can, depending on the measurements (sensor and user), worsen or improve the separation of activities.

S	GA	Worst (2)	Best (7)
Wrist Gyroscope	No (G0)		
	G3 + LDA		
Ankle Accelerometer	No (G0)		
	G3 + LDA		

7 Discussion

These experiments show promising results, but open up to several questions and remarks.

As we demonstrated in this paper, re-sampling datasets may lessen the recognition accuracy. But on the other hand, this additional step significantly decreases the number of inputs. This curtailment will definitely be an advantage when combining such a re-sampling with SOTA recognition models. Indeed, it will greatly mitigate their training time (as there will be fewer inputs to handle). However, it is important to determine how to define the optimal re-sampling rate.

In addition, we notice that the recognition accuracy varies from one dataset to another. Besides, among dataset, there are also discrepancies in performance between volunteers. Therefore, we suppose that both the type and the calibration of used sensors play a significant part in the performance. Moreover, Fong et al. [9] showed that classification accuracy can increase significantly when a specific processing is applied on bio-signals, such as ECG. The noteworthy good performance of the per feature abstraction (G1) lets us suppose that other IMU features might also require a specific processing.

These results also pave the way for more future investigations. Indeed, as mentioned previously, there are several options to produce a different representation of a feature or set of features. For that reason, additional experiments should be conducted in order to determine the most appropriate one. For instance, such experiments could consider more sophisticated techniques, such as Neighborhood Components Analysis (NCA) [12] or t-distributed Stochastic Neighbor Embedding (tSNE) [27]. Another option would be to use the latent representation of DL models as the abstracted representation. In fact, architectures such as Variational Auto Encoder (VAE) could be trained to learn the latent representation of a group of body-worn sensors data. Once trained, the obtained latent representations will become the inputs for training the classifier.

Furthermore, in this paper, either all features have a new representation or none of them have. We did not investigate a more complex architecture, as illustrated in Fig. 2. An architecture where some features will be grouped in order to produce abstracted representations (such as features $\{f_2, f_3\}$ or $\{f_4, f_5, f_6\}$ in Fig. 2); some features will be unchanged (such as f_1); and others will be projected into another space (f_N). Indeed, as shown by the data plots in Sect. 6.3, some sensors' data are already well represented. Conversely, for others, computing a new representation significantly improves the interpretability and thus enhances the activity recognition. Therefore, there might be further research to pursue in this direction. We imagine that a multimodal architecture such as the one proposed in [21] could be particularly adapted in this situation. An architecture that could automate whether a new representation is required, based on some prior analysis of the features, their auto- and cross-dependencies.

Finally, by projecting the representation on lower dimensions, it would probably be possible to better interpret the impact of each body part state on the final decision. Coupled with some attention mechanisms, it could help us determine which features are the most important and when to use them.

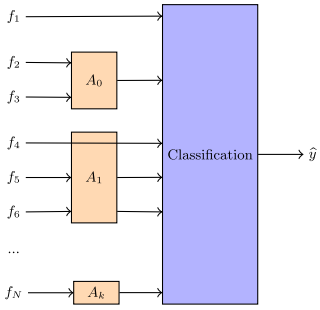


Fig. 2. Generic illustration of the grouped representation proposal. In a generic version, some features are grouped and have a high level representation, while others might only be pre-processed. $A_i, i \in \{0, ..., k\}$ is the technique used to transform a group of features into a higher level representation.

At the end, all the knowledge acquired from multimodal wearable sensing devices could benefit to other HAR research. As mentioned previously, there is uncertainty of the sensor's location at a given time in uni-modal HAR research (e.g., using a smartphone). Analysis based on MW-HAR scenarios provides knowledge of the different range of motion/values from all monitored body parts. This knowledge could help uni-modal HAR models estimate where the sensor is placed on the body. Such a transfer learning will eventually improve the understanding of what is monitored during a given period. As a result, models will better grasp associated gestures and consequently recognize more efficiently corresponding activities.

8 Conclusion

This paper describes an alternative approach to process multimodal wearable sensors' data. The presented results attempt to answer whether this proposal is the key for solving HAR task. Although it is difficult to be unequivocal on this point, this proposal depicts interesting aspects and performance over the tested datasets. These results would need to be automated and generalized to other datasets in order to settle this question. The first part of this proposal is to re-sample the data to a coarser resolution. After some pre-processing, sensing features monitoring the same body-part are grouped. For each resulting group, we train a new representation of its data. This proposal reduces the computational cost and improves the comprehension of multimodal sensors' data.

This study opens up to a multitude of possibilities. In the future, we will extensively study this proposal with more complex recognition models, different abstraction techniques, and other spontaneous Multimodal Wearable HAR datasets. We also plan to further investigate the visualization of the generated abstraction in order to better understand the advantages of this proposal.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Banos, O., et al.: mHealthDroid: a novel framework for agile development of mobile health applications. In: Pecchia, L., Chen, L.L., Nugent, C., Bravo, J. (eds.) IWAAL 2014. LNCS, vol. 8868, pp. 91–98. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13105-4_14
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields (2019)
4. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**(1), 16–28 (2014)
5. Chang, Y.J., Chen, S.F., Huang, J.D.: A Kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities. Res. Dev. Disabil. **32**(6), 2566–2570 (2011)
6. Chavarriaga, R., et al.: The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. Pattern Recogn. Lett. **34**(15), 2033–2042 (2013)
7. Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., Liu, Y.: Deep learning for sensor-based human activity recognition: overview, challenges, and opportunities. ACM Comput. Surv. **54**(4) (2021)
8. Ferrari, A., Micucci, D., Mobilio, M., Napoletano, P.: Trends in human activity recognition using smartphones. J. Reliable Intell. Environ. **7**(3), 189–213 (2021)
9. Fong, S., Lan, K., Sun, P., Mohammed, S., Fiaidhi, J.: A time-series pre-processing methodology for biosignal classification using statistical feature extraction. In: Proceedings of the IASTED International Conference on Biomedical Engineering, BioMed 2013 (2013)
10. Gerling, K., Livingston, I., Nacke, L., Mandryk, R.: Full-body motion-based game interaction for older adults. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2012, pp. 1873–1882. Association for Computing Machinery, New York (2012)
11. Gochoo, M., Tan, T.H., Liu, S.H., Jean, F.R., Alnajjar, F.S., Huang, S.C.: Unobtrusive activity recognition of elderly people living alone using anonymous binary sensors and DCNN. IEEE J. Biomed. Health Inform. **23**(2), 693–702 (2019)
12. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Proceedings of the 17th International Conference on Neural Information Processing Systems, NIPS 2004, pp. 513–520 (2004)
13. Gu, K., Vosoughi, S., Prioleau, T.: Feature selection for multivariate time series via network pruning. In: 2021 International Conference on Data Mining Workshops (ICDMW). IEEE (2021)
14. Hayes, A.L., Dukes, P.S., Hodges, L.F.: A virtual environment for post-stroke motor rehabilitation (2011)
15. Joshi, A., Parmar, H.R., Jain, K., Shah, C.U., Patel, V.R.: Human activity recognition based on object detection. IOSR J. Comput. Eng. **19**, 26–32 (2017)
16. Kaluža, B., Mirchevska, V., Dovgan, E., Luštrek, M., Gams, M.: An agent-based approach to care in independent living. In: de Ruyter, B., et al. (eds.) AmI 2010. LNCS, vol. 6439, pp. 177–186. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16917-5_18
17. Kwon, H., Abowd, G.D., Plötz, T.: Complex deep neural networks from large scale virtual IMU data for effective human activity recognition using wearables. Sensors **21**(24), 8337 (2021)

18. Lawrence, E., Sax, C., Navarro, K.F., Qiao, M.: Interactive games to improve quality of life for the elderly: towards integration into a WSN monitoring system. In: 2010 Second International Conference on eHealth, Telemedicine, and Social Medicine, pp. 106–112 (2010)
19. Miyamoto, S., Ogawa, H.: Human activity recognition system including smartphone position. Procedia Technol. **18**, 42–46 (2014)
20. Münzner, S., Schmidt, P., Reiss, A., Hanselmann, M., Stiefelhagen, R., Dürichen, R.: CNN-based sensor fusion techniques for multimodal human activity recognition. In: Proceedings of the 2017 ACM International Symposium on Wearable Computers, ISWC 2017, pp. 158–165 (2017)
21. Perez-Rua, J.M., Vielzeuf, V., Pateux, S., Baccouche, M., Jurie, F.: MFAS: multimodal fusion architecture search. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6959–6968 (2019)
22. Qian, H., Pan, S.J., Da, B., Miao, C.: A novel distribution-embedded neural network for sensor-based activity recognition. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp. 5614–5620. International Joint Conferences on Artificial Intelligence Organization (2019)
23. Reiss, A., Stricker, D.: Creating and benchmarking a new dataset for physical activity monitoring. In: Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2012 (2012)
24. Rossi, S., Capasso, R., Acampora, G., Staffa, M.: A multimodal deep learning network for group activity recognition. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–6 (2018)
25. Ryoo, M.S.: Human activity prediction: early recognition of ongoing activities from streaming videos. In: 2011 International Conference on Computer Vision, pp. 1036–1043 (2011)
26. Straczkiewicz, M., James, P., Onnela, J.P.: A systematic review of smartphone-based human activity recognition methods for health research. NPJ Digit. Med. **4**(11), 1–15 (2021)
27. Van Der Maaten, L.: Accelerating T-SNE using tree-based algorithms. J. Mach. Learn. Res. **15**(1), 3221–3245 (2014)
28. Vo, Q.V., Lee, G., Choi, D.: Fall detection based on movement and smart phone technology. In: 2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future, pp. 1–4 (2012)
29. Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L.: Deep learning for sensor-based activity recognition: a survey. Pattern Recogn. Lett. **119**, 3–11 (2019)
30. Wang, J., Chen, Y., Gu, Y., Xiao, Y., Pan, H.: SensoryGANs: an effective generative adversarial framework for sensor-based human activity recognition. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2018)
31. Zeng, M., et al.: Convolutional Neural Networks for human activity recognition using mobile sensors. In: 6th International Conference on Mobile Computing, Applications and Services, pp. 197–205 (2014)
32. Zhang, L., Zhang, X., Pan, J., Huang, F.: Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In: Proceedings of the AAAI Conference on Artificial Intelligence (2022)