The *MARBLE* dataset: Multi-Inhabitant Activities of Daily Living combining Wearable and Environmental Sensors Data

Luca Arrotta, Claudio Bettini, and Gabriele Civitarese

University of Milan, Milan, Italy
{luca.arrotta, claudio.bettini, gabriele.civitarese}@unimi.it

Abstract. While the sensor-based recognition of Activities of Daily Living (ADLs) is a well-established research area, few high-quality labeled datasets are available to compare the results of different approaches. This is especially true for multi-inhabitant settings, where multiple residents live in the same home performing both individual and collaborative ADLs. The reference multi-inhabitant datasets consider only environmental sensors data and two residents in the same home. In this paper, we present *MARBLE*: a novel multi-inhabitant ADLs dataset that combines both smart-watch and environmental sensors data. *MARBLE* includes sixteen hours of ADLs considering scripted but realistic scenarios where up to four subjects live in the same home environment. Twelve volunteers participated in data collection. We describe *MARBLE* also providing details on the design of data collection and tools. We also present initial benchmarks of ADLs recognition on *MARBLE*, obtained by applying state-of-the-art deep learning methods. Our goal is to share the result of a complex and time consuming data acquisition and annotation task, hoping that the challenge of improving the current baselines on *MARBLE* will contribute to the progress of the research in multi-inhabitant ADLs recognition.

Keywords: activity recognition · smart-home · multi-inhabitant.

1 Introduction

The recognition of Activities of Daily Living (ADLs) in smart-home environments is a well-known research area in pervasive computing enabling intelligent context-aware services [7]. Accurately recognizing ADLs is also crucial for complex health-care systems that continuously monitor the behavior of fragile elderly subjects in their homes. For instance, the sequence of ADLs performed by a subject and their execution modalities may reveal early symptoms of cognitive decline [19]. Among other methods, ADL recognition has been shown to be feasible through the intelligent analysis of data generated by unobtrusive sensors deployed in the home environment and/or sensors on wearable devices.

Any new approach in this research area requires an empirical evaluation on labeled datasets, i.e., datasets in which the stream of timestamped sensor values

has been annotated with the actual ADLs performed by a subject specifying the interval of time for each ADL. However, collecting these labeled datasets is costly, time consuming and intrusive [6]. Moreover, publishing a dataset is often constrained by privacy motivations [13]. Indeed, sensor and activity data can be considered sensitive, and sometimes can even be used to re-identify a subject even if explicit identifiers have been substituted by pseudonyms in the dataset. For these reasons, there are only a few high-quality and publicly available ADLs datasets. However, public datasets are necessary to make research more transparent through experiments reproducibility, to speed up new research contributions, and to provide reference benchmarks.

A limitation of most of the existing ADLs datasets is that they only include data from single-inhabitant settings, where only one subject is living in the home [13]. This scenario is actually realistic considering the large amount of elderly subjects living alone in their homes. However, multiple subjects may live in the same home (e.g., married couples of elderly subjects, an elderly and her caregiver, a whole family). In these settings it is often necessary to identify ADLs performed by specific residents as well as those performed collaboratively.

Multi-inhabitant ADLs recognition is still a poorly explored research area [4,13]. The main reference datasets are CASAS [8] and ARAS [2]. These datasets have been collected in real home environments inhabited by two subjects. However, only environmental sensors were considered for data collection.

Wearable sensors can provide important additional information to significantly improve ADLs recognition. By associating the physical movements of the subjects to environmental sensor events it is possible to accurately discriminate a larger number of activities (e.g., sitting at the kitchen table, eating at the kitchen table and drinking at the kitchen table). Moreover, wearable sensors can also monitor ADLs not captured only by environmental sensors. This is especially important considering that it can be too costly to deploy a significant amount of environmental sensors that can capture all the possible household items. Most importantly for multi-inhabitant settings, wearable sensors can be used to address the *data association* problem [4]: how to associate each environmental sensor event (e.g., the fridge has been opened) to the inhabitant that actually triggered it? In this context, a wearable, being a personal device, identifies the subject and can also reveal the proximity to the environmental sensor that was triggered. While constantly wearing devices may be considered unrealistic, smartwatches and wristbands nowadays are becoming quite common and they represent a non-intrusive technology that can be continuously worn in home environments.

Hence, in this paper we present *MARBLE*: a new publicly available dataset of ADLs performed in multi-inhabitant settings. Differently from existing datasets, *MARBLE* includes data from both wearable and environmental sensors. Moreover, *MARBLE* includes scenarios where up to four subjects perform activities in the same home environment. Overall, *MARBLE* includes data from 12 different subjects performing 13 types of ADLs. *MARBLE* contains around 16 hours of labeled multi-inhabitant ADLs data.

We believe that *MARBLE* can be used by the activity recognition community to evaluate novel approaches both for single-inhabitant and multi-inhabitant ADLs recognition. Moreover, *MARBLE* can be used to investigate novel data association strategies.

The contributions of this paper are the following:

- We present a novel publicly available[1] dataset of multi-inhabitant ADLs that includes both environmental and wearable sensors data, where up to four subjects perform ADLs both jointly and independently.
- We describe in details how we designed the data acquisition/annotation tools and the collection of labeled data.
- We provide some benchmarks on the performance of state-of-the-art deep learning approaches on *MARBLE* that could be used as baselines for future work in this area.

2 Related Work

Single-inhabitant ADLs recognition has been extensively studied in the last decades [7]. Results have been validated on several public datasets collected in single-inhabitant settings, like the well-known OPPORTUNITY [14], CASAS [8], and Amsterdam [11] datasets.

On the contrary, the literature on the same problem in multi-inhabitant settings is less advanced. Only a few approaches have been proposed to tackle this problem (e.g., [1, 3, 20, 22, 24, 25]). The lack of public datasets for multi-inhabitant ADLs recognition is indeed one of the major issues in this research area [13]. Some of the existing works validated their methods on datasets that are not publicly available. Some public datasets [10,12,17] have been acquired from video or audio streams, like the BEHAVE dataset [5]. However, those sensing approaches are often perceived as too intrusive for home environments (especially considering elderly subjects), even if data is processed locally to preserve residents' privacy.

The public CASAS dataset is actually a collection of datasets, including some that have been acquired in multi-resident settings [23]. For this reason, these datasets have often been considered as the reference benchmark datasets also for multi-resident ADLs recognition. The experimental setup in those datasets mainly includes simple enviromental sensors, like PIR sensors, and magnetic sensors. Activities have been performed by the residents both individually and jointly. For instance, the Kyoto dataset ("WSU Smart Apartment ADL Multi-Resident Testbed") includes 15 different types of ADLs performed by two residents, including *reading a magazine*, *watering plants*, *playing a game of checkers*, and *setting dining room table*. Among the CASAS datasets, we also mention PUCK [9], that combines wearable and environmental sensors similarly to our dataset, but in a single-inhabitant setting.

Another public dataset that has been considered as a benchmark is ARAS [2], that was collected in two different home environments, each one inhabited by

[1] The dataset can be downloaded here: `tinyurl.com/marbledataset`

two residents. Several environmental sensors have been used for data collection, including photocells, pressure mats, contact sensors, proximity sensors, float sensors, and infrared receivers. Overall, the dataset includes 27 different ADLs types, including *taking shower*, *brushing teeth*, *sleeping*, *having conversation*, and *watching tv*.

The major drawback of the two datasets described above is that they do not include wearable sensors data, which is very informative for the data association problem and to detect activities at a finer granularity, as described in the introduction. Moreover, those datasets are limited to two residents in the same home.

On the other hand, there are public datasets that only consider wearable sensors. For instance, the DyadHAR dataset [21] includes inertial sensor data from two subjects in an indoor environment wearing smart-phones on the belt and performing ADLs (e.g., participating in a meeting, coffee-break, work, lunch). The dataset also contains RSSI values from iBeacons in the environment.

The main advantage of *MARBLE* with respect to the described datasets is that it combines environmental and wearable sensors in a multi-inhabitant setting to capture a wide set of activities, and that it includes scenarios with up to four participants.

3 *MARBLE*: Data collection design and tools

In this section, we describe in details the *MARBLE* dataset. We present our design choices, the experimental setup, the data collection process and tools, and the dataset format.

3.1 Dataset design

The design of *MARBLE* was driven by the multi-inhabitant ADLs recognition problem, and in particular by *data association*. Indeed, during the design phase, we realized that monitoring ADLs with a combination of environmental sensors and wearable sensors is a promising but poorly explored direction [22]. Wearable devices have the potential of: a) collecting data about the physical movements of the subject, b) taking advantage of indoor positioning systems, and c) associating an identity to each subject. On the other hand, wearable sensors alone can not capture complex ADLs, while environmental sensors can provide precious information about the interaction of the residents with the home environment.

Hence, the *MARBLE* dataset includes both data from wearable devices and environmental sensors. We opted for smart-watches as wearables since they have low obtrusiveness, they are becoming very common, and they can capture hand gestures useful to reveal ADLs (e.g., washing dishes). Among environmental sensors we include magnetic sensors to detect open/close of drawers and doors, mat (pressure) sensors to detect when residents are sitting on chairs/sofa, plug sensors to detect the usage of home appliances. We also planned to deploy BLE beacons and WiFi APs to enable indoor positioning.

Due to privacy concerns, we were not able to acquire long term data from actual inhabitants in real homes. Nonetheless, based on our previous experience

in real world deployments and in-the-lab data collections [18], we designed a new multi-inhabitant dataset acquisition campaign in a smart-home lab with significant efforts in making it realistic and diverse. Moreover, annotations are complete and very accurate.

Based on applications of interest for our lab, we planned the acquisition of the following activities: *Answering Phone*, *Clearing Table*, *Cooking*, *Eating*, *Entering Home*, *Leaving Home*, *Making Phone Call*, *Preparing Cold Meal*, *Setting Up Table*, *Taking Medicines*, *Using PC*, *Washing Dishes*, and *Watching TV*.

We carefully designed several single- and multi-inhabitant scenarios for data acquisition. Each scenario is a template that describes the type of activities that subjects should perform and their order. As we will explain later, each scenario has been performed several times by different subjects. We did not specify in details how each activity should be actually performed, allowing subjects to freely execute activities with the goal of introducing high variability in the dataset.

In the following, we represent the *MARBLE* scenarios through several tables. In these tables, the flow of time is represented vertically, from top to bottom. Except from Table 1 where each column describes a single-inhabitant scenario, each of the other tables describes a single scenario with a column for each resident. Horizontal dashed lines indicate transitions between subsequent activities. When residents collaboratively perform an activity the vertical line is suppressed. Each designed scenario is identified by a letter followed by the number of residents involved during the data acquisition for that scenario.

We designed four single-inhabitant scenarios graphically represented in Table 1.

Table 1: Single-inhabitant scripted scenarios

	A1	B1	C1	D1
morning	cook	set table	enter home	
	set table	cook	watch tv	answer call
	eat	eat	prepare meal	prepare meal
	clear table	clear table	answer call	watch tv
	wash dishes	wash dishes	make call	answer call
	use pc	watch tv	leave home	take meds
	answer call	make call	enter home	leave home
afternoon	prepare meal	take meds	take meds	enter home
	set table	make call	set table	wash dishes
	take meds	cook	prepare meal	use pc
	eat	set table	eat	make call
	make call	eat	clear table	use pc
	clear table	clear table	wash dishes	cook
	use pc	leave home	watch tv	leave home
evening	leave home	enter home	cook	enter home
	enter home	prepare meal	eat	wash dishes
	eat	eat	take meds	watch tv
	watch tv	wash dishes	use pc	take meds
	make call	answer call	answer call	
	take meds	use pc	leave home	
		take meds		

We designed three different scenarios involving two subjects concurrently performing both independently and jointly the activities. These scenarios are shown in Table 3. Finally, we also designed four different scenarios of ADLs concurrently performed by four inhabitants, presented in Table 2.

Table 2: Multi-inhabitant scripted scenarios involving four subjects

A4

Subject 1	Subject 2	Subject 3	Subject 4
cook	set table	use pc	watch tv
	eat		
wash dishes	watch tv	clear table	use pc

B4

Subject 1	Subject 2	Subject 3	Subject 4
watch tv	enter home		use pc
	watch tv		
prepare meal	watch tv	eat	leave home

C4

Subject 1	Subject 2	Subject 3	Subject 4
set table	prepare meal	enter home	
	eat	use pc	make call
watch tv			

D4

Subject 1	Subject 2	Subject 3	Subject 4
enter home			set table
	eat		
clear table		watch tv	
wash dishes	cook	watch tv	answer call

Note that, despite scenarios describe the transition from an activity to another as instantaneous, this will not be the case for their executions since transitions will have a duration. Moreover, activities specified as concurrent for different subjects may begin and end at slightly different times with also different duration of transitions. For instance, Table 4 shows an execution of the scenario *A*4 that we acquired during data collection. Since subjects freely executed the ADLs, activities and transitions are not perfectly aligned as specified in *A*4.

3.2 Experimental setup

Figure 1 illustrates how the smart-home lab is divided into six semantic areas, each representing a different room (hall, kitchen, dining room, medicine area, living room, and office).

Different environmental sensors were deployed to monitor the interaction of the subjects with their surrounding environment: five magnetic sensors, nine pressure mats, and two smart-plugs. Figure 1 shows how these sensors were deployed in the environment. Magnetic sensors monitored the interactions with the pantry, the cutlery drawer, the pots drawer, the medicines cabinet, and the fridge. Pressure mats monitored the interactions with four dining room chairs,

Table 3: Scenarios involving two inhabitants

(a) A2 scenario

A2

	Subject 1	Subject 2
morning	set table	cook
	eat	
	clear table	wash dishes
	use pc	watch tv
		make call
	watch tv	
	answer call	take meds
afternoon	prepare meal	cook
		make call
	take meds	set table
	eat	
	use pc	clear table
	make call	
	leave home	
evening	enter home	
	eat	
	eat	answer call
	take meds	use pc
	make call	take meds
	watch tv	

(b) B2 scenario

B2

	Subject 1	Subject 2
morning	set table	enter home
	cook	watch tv
	eat	
	watch tv	
	clear table	prepare meal
	wash dishes	make call
	watch tv	answer call
	make call	leave home
	take meds	enter home
afternoon	cook	take meds
	make call	prepare meal
	set table	
	eat	
	wash dishes	clear table
	watch tv	
	leave home	watch tv
evening	enter home	
	watch tv	
	prepare meal	cook
	eat	
	wash dishes	take meds
	answer call	use pc
	use pc	answer call
	take meds	leave home

(c) C2 scenario

C2

	Subject 1	Subject 2
morning	cook	enter home
	set table	watch tv
	eat	
	clear table	prepare meal
	wash dishes	answer call
		make call
	use pc	leave home
	answer call	enter home
afternoon	prepare meal	take meds
	set table	prepare meal
	take meds	set table
	eat	
	use pc	wash dishes
	make call	clear table
	watch tv	
evening	leave home	
	enter home	set table
	eat	
	eat	take meds
	watch tv	use pc
	take meds	watch tv
	make call	
	watch tv	
	watch tv	answer call
		leave home

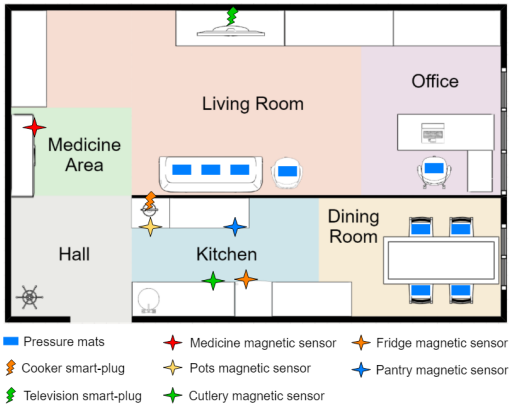

Fig. 1: The smart-home lab used in the dataset collection

the office chair, and four different seats in the living room (i.e., a couch and an armchair). Smart-plugs monitored the interactions with the electric cooker and the television. The environmental sensors communicated their readings through the Z-Wave protocol to a Linux server in charge of storing data into a MongoDB database.

Table 4: One of the A4 multi-inhabitant scenario instances

A4 - Instance 1

Time	Subject 1	Subject 2	Subject 3	Subject 4
15:26:43				
15:26:55				
15:26:58		set table		
15:27:38				
15:28:37	cook		use pc	watch tv
15:29:34				
15:29:35		transition		
15:30:04			transition	
15:30:43				transition
15:30:46	transition		eat	
15:30:48		eat		
15:30:50			eat	
15:32:30				
15:32:34		eat		
15:32:35			eat	
15:32:37	transition	eat		
15:32:52				transition
15:32:53		transition		
15:32:57			clear table	
15:32:58				
15:34:25	wash dishes			use pc
15:35:21		watch tv		
15:35:24				
15:35:35				

In order to acquire sensor data from wearables, we developed a WearOS application in charge of continuously transmitting the stream of inertial sensors data to our Linux server. As wearable devices, we used smart-watches running the WearOS operating system[2].

Since we planned to monitor answering and receiving phone call activities, the subjects also carried an Android smartphone in their pockets. We developed an Android application in charge of communicating in real-time the phone events to our Linux server (i.e., start/end of receiving/making phone calls).

As we discussed in Section 3.1, we planned to take advantage of smart-watches also to collect data from indoor positioning systems to detect the semantic location of each subject. However, indoor localization is an orthogonal problem with respect to activity recognition. Hence, while we actually deployed a specific microlocalization infrastructure[3], *MARBLE* only includes the ground-truth about the semantic areas of the residents.

3.3 Data collection

MARBLE includes data from 12 different volunteers that contributed to the data collection by performing several instances of the scenarios described in Section 3.1. The volunteers' age was 27 ± 5, and they had no connection with our research team. Ten volunteers contributed both to single- and multi-inhabitants

[2] We used Huawei Sport 2 and other brands with similar features.

[3] In our experimental setup, we used machine learning methods to analyse RSSI signal of BLE beacons and WiFi APs in order to classify the semantic location of each subject in real-time.

scenarios, while the other two participated to single-inhabitant acquisitions only. Each volunteer contributed to multiple scenarios. Considering privacy concerns, each volunteer is only identified with a numeric pseudo-identifier in the dataset. Hence, *MARBLE* does not contain any explicit identifier and it is very unlikely that any re-identification can be performed based on sensor data. Before the acquisition, we showed the smart-home environment and tools to the volunteers, and we instructed them about the scenario they had to perform. As explained before, the volunteers were free to execute each ADL as they felt more appropriate. Since we had time restrictions for data collection (due to the availability of volunteers), we limited the execution time of each performed ADL to a duration that in some cases does not reflect the actual time a person would actually need, but long enough to obtain a significant amount of labeled data. For instance, considering *Eating* or *Cooking*, we asked our volunteers to perform the ADL only for a few minutes.

As we previously mentioned, each instance of a scenario was performed by different volunteers in order to guarantee sufficient variability and robustness. Overall, we acquired 12 single-inhabitant scenario instances (two instances for $D1$; three instances for $A1$ and $C1$; four instances for $B1$) and 20 multi-inhabitant scenarios instances (three instances for $A2$, $B2$, $B4$, and $D4$; two instances for $A4$, and $C4$; four instances for $C2$).

Table 5 shows, for each ADL type, the amount of recorded labeled data (in minutes), the average duration (in seconds), and the number of collected instances. Finally, Table 6 shows the overall amount of recorded labeled data (in minutes) and the average duration (in minutes) for single-, 2-, and 4-inhabitants scenarios.

Table 5: Statistics on labeled activities

	recorded minutes	average duration (s)	instances
ANSWERING PHONE	68.6	67.5	61
CLEARING TABLE	38.5	39.9	58
COOKING	80.5	81.9	59
EATING	150.2	28.2	320
ENTERING HOME	19.3	12.2	95
LEAVING HOME	13.7	16.1	51
MAKING PHONE CALL	63.6	53.8	71
PREPARING COLD MEAL	53.0	59.9	53
SETTING UP TABLE	53.9	39.4	82
TAKING MEDICINES	36.3	28.3	77
TRANSITION	276.1	12.9	1282
USING PC	94.1	86.9	65
WASHING DISHES	54.6	48.2	68
WATCHING TV	267.6	90.2	178

Table 6: Statistics on scripted scenarios

type of scenarios	recorded minutes	average duration (min)
single-inhabitant	307.5	25.6 ± 4.0
2-inhabitants	315.5	31.5 ± 7.7
4-inhabitants	84.0	8.4 ± 1.8

3.4 Data annotation

In order to make data acquisition as realistic as possible, annotation was performed by a different team that was watching live video streams of what was happening in each area of the smart-home lab.

The members of this team used a dedicated software that we implemented to easily annotate in real-time: a) the ADL being performed by each subject, b) the semantic area in which the subject is performing the ADL, and c) the associations between environmental sensor events and the subjects that triggered them. The last type of annotation is particularly useful to evaluate the effectiveness of novel data association strategies. Moreover, it also can be used to isolate the environmental events triggered by each subject in order to evaluate single-inhabitant approaches. Clearly, sensor data collected from the smart-watches are automatically associated with the correct subject by the WearOS application. Since annotating multi-inhabitant scenarios turned out to be a very hard task, each member of the annotation team was in charge of annotating data for a single subject.

In order to obtain accurate annotations, both the environmental sensors and the annotation software communicated with the same gateway that was in charge of providing the timestamps both to data and annotations, before storing them in a MongoDB database. At the same time, the clocks of the smartwatches were synchronized with the one of the gateway.

4 Experimental Evaluation

In this section we provide some benchmarks on *MARBLE* that could be used as baselines for future work on multi-inhabitant ADL recognition methods. For the sake of this work, we assume that data association can be computed perfectly i.e., we assume that the association between each environmental sensor event and the resident that triggered it, is always correct.[4] We compare the performance of different deep learning solutions that we have adapted to be applied to *MARBLE* data.

4.1 Data pre-processing

In order to provide sensor data as input for deep learning networks, we apply some simple pre-processing steps. Inertial sensors data are smoothed using a

[4] We proposed in [3] a data association method evaluated on *MARBLE*. However, the dataset was not public yet and it was not described in detail.

median filter to reduce the intrinsic noise of inertial sensors. Then, inertial and environmental sensors data are temporally aligned and segmented into windows of w seconds, with an ov overlap factor. The two types of data are provided as separate inputs to the networks.

For each window of inertial sensors data, we extract a matrix of shape $(9, L_w)$, where L_w is the average number of measurements collected by inertial sensors (according to the sampling rate) when the segmentation window size is equal to w seconds[5]. Each of the nine rows of the matrix encodes the measurements of one of the three axes of a specific inertial sensor.

Regarding environmental sensors, for each window we generate a binary matrix with shape $(25, w)$, where w is the window size. Each of the 25 rows represents a specific environmental sensor or a specific semantic location. Each column represents a specific second within the window (e.g., column 3 is the third second inside the window). The value of the matrix at row i and column j is 1 if sensor/location i was active at second j, 0 otherwise.

4.2 Considered approaches

In the following, we describe the methods that we implemented as benchmarks. We warmly invite the researchers in this area to take advantage of this dataset to validate more sophisticated solutions and compare them with the provided baselines. We empirically determined the architecture of each network.

Fully Connected Deep Learning (DNN) The first method we evaluated is a simple fully connected Deep Neural Network (we will refer to this approach as DNN). We use DNN as a baseline to compare it with more advanced methods in the literature. The flow of inertial sensors data is composed of two Fully Connected (FC) layers of 64 neurons, two FC layers of 128 neurons, and four FC layers of 256 neurons interleaved by a Dropout layer (with 0.5 dropout rate). On the other hand, the flow of environmental sensors data is composed of four FC layers of 64, 32, 128, 32 layers, respectively. Within both the data flows, we flatten the output of the last layer with a Flatten layer. The two flows are then merged using a Concatenation layer. Then, the DNN has a FC layer with 64 neurons followed by a Softmax layer used for classification.

Convolutional Neural Network (CNN) This approach is quite popular in the literature, possibly due to its good performance, especially when multiple types of sensors are considered [15]. Inertial measurements are provided as input to a stack of two Convolutional layers, each one composed of 64 filters with a $2x2$ kernel, followed by two Convolutional layers composed of 128 filters with a $2x2$ kernel. Then, we flatten the output of the last convolutional layer with a Flatten layer. The flow continues with two FC layers (64 and 32 neurons, respectively) interleaved by a Dropout layer (0.5 as dropout rate). On the other hand, the flow of environmental sensors data is composed of a Convolutional layer of 16 filters with a $2x2$ kernel, a Flatten layer, and two FC layers (128 and 32 neurons,

[5] Since the number of measurements in a window may slightly differ from L_w, we interpolate missing values or downsample measurements when needed.

respectively). The two flows are then merged using a Concatenation layer. Then, the CNN has a FC layer with 32 neurons. Finally, a Softmax layer is used for classification.

Convolutional and Recurrent Deep Learning (CNN-LSTM) Finally, we implemented an approach that combines convolutional and recurrent layers (we will refer to this approach as CNN-LSTM). In particular, we slightly adapted the method presented in [16] to include both inertial and environmental sensors. Inertial measurements are provided as input to two Convolutional layers composed of 64 filters with a $2x2$ kernel, followed by two Convolutional layers composed of 128 filters with a $2x2$ kernel. Hence, the output of the last Convolutional layer is flattened with a Flatten layer. The network continues with a LSTM layer of 256 units, followed by a Dropout layer with a 0.5 dropout rate and a 64 neurons FC layer. On the other hand, environmental sensor data are provided to a Convolutional layer of 8 filters with a $2x2$ kernel, followed by a Flatten layer, a 128 units LSTM, a Dropout layer with a 0.5 dropout rate, and a FC layer with 32 neurons. Hence, the two flows are merged with a Concatenation layer. The network then continues with two FC layers with 64 and 32 neurons. Finally, a Softmax layer is used for classification.

4.3 Results

In the following, we show the performance on *MARBLE* of the approaches described above. For each approach, we trained the corresponding neural network with the data collected both in single- and multi-inhabitant scenarios. In this way, the evaluation is affected by the interactions between the subjects of multi-inhabitant scenarios. We chose the optimal segmentation parameters using a grid search approach. In particular the best parameters we found are $w = 6$, and $ov = 0.8$. Each approach was evaluated by considering an ideal perfect association between the environmental events and the subjects that triggered them.

We adopted three well-known evaluation methodologies and a new one that is particularly significant for a multi-inhabitant dataset. The first methodology simply consists of splitting the dataset in 70% for training, 10% for validation, and 20% for testing. The second one is a 10-fold cross validation. The third one is a *leave-one-subject-out* cross-validation: at each fold, one subject is used as test set and the remaining subjects as training set. The leave-one-subject-out is generally used to test the generalization capability of the classifier on subjects that did not contribute with labeled data. Finally, we propose a new evaluation methodology that we call *leave-one-scenario-out* cross-validation: at each fold, an instance of one of the *MARBLE* scenarios is used as test set, while the training set excludes both data from instances of the scenario considered in the test set as well as data related to subjects that contributed to the test set. This last methodology is the most restrictive one since it aims at assessing the generality of the classifier over unseen users and also over sequences of activities not included in the training set.

Figure 2 shows that the evaluation methodology has a significant impact on the measured F1 score. We observed that the 70/10/20 methodology over-

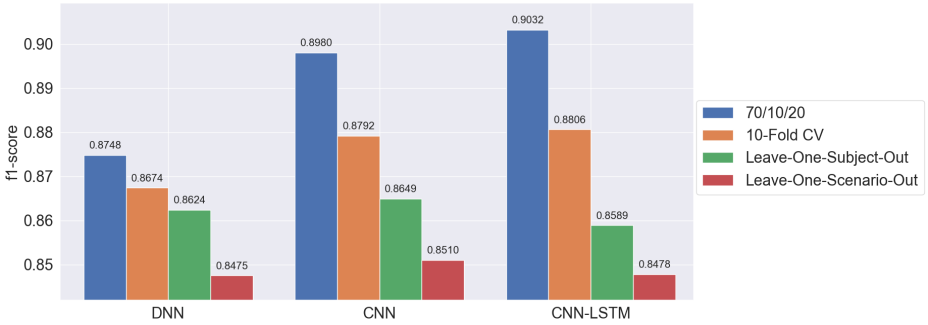

Fig. 2: Overall recognition rate based on the evaluation method

estimates the recognition rate. By using this evaluation methodology, it emerges that CNN-LSTM outperforms the other approaches. However, both the training and the test sets contain data samples related to the same subjects and scenarios, thus it is likely that this evaluation methodology suffers from overfitting problems.

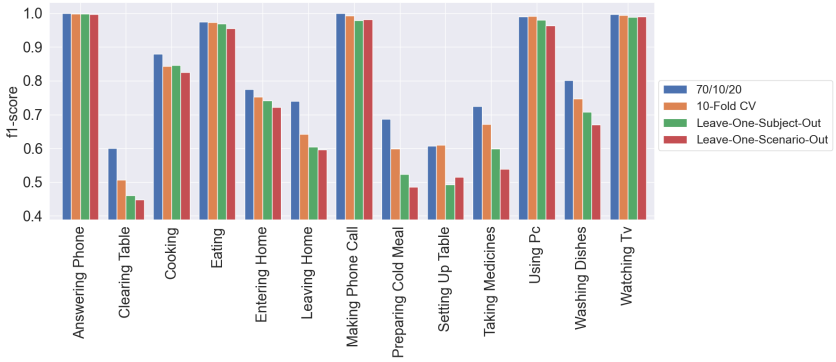

Fig. 3: Recognition rate of CNN-LSTM for each activity

The 10-fold cross-validation methodology is more robust and it provides a better estimate of the recognition rate. However, at each fold, training and test sets may still include data from the same subjects or scenarios. This type of evaluation confirms that CNN-LSTM reaches the highest recognition rates.

Leave-one-subject-out and leave-one-scenario-out methodologies provide a more robust assessment of the recognition rate than the other evaluation methodologies. By using these methodologies, we observed that all the considered approaches reach similar recognition rates. As expected, the leave-one-scenario-out methodology is the one that estimates the recognition rate with the lowest F1 score values. The lower recognition rates reached by these methodologies is due

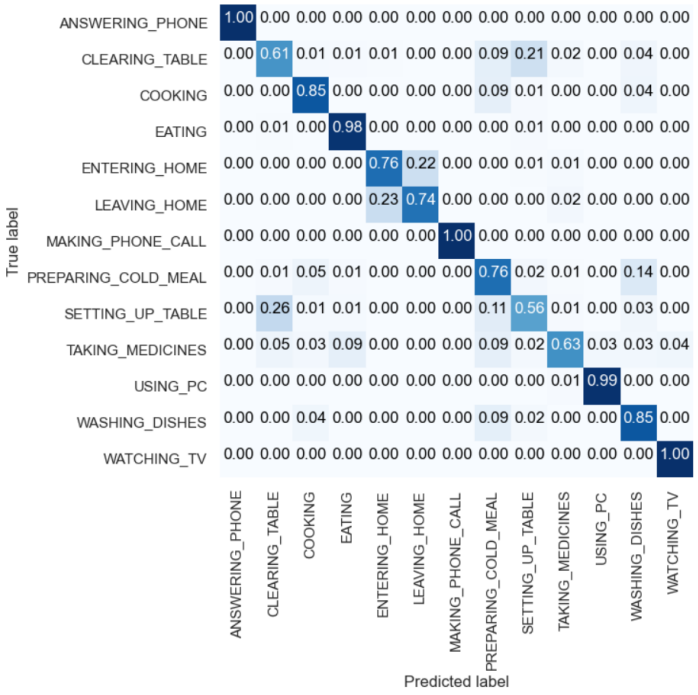

Fig. 4: Confusion Matrix of CNN-LSTM (70/10/20)

to the fact that our dataset includes activities that are particularly difficult to recognize on subjects/scenarios that were not observed during the training phase. Consider Figure 3, that shows how the evaluation methodology affects the recognition rate of each activity.

We observed that activities like *Eating*, *Watching TV*, and *Using PC* reach high recognition rates independently from the evaluation methodology. This is due to the fact that these activities can be performed only in specific smart-home areas, triggering environmental sensors that are not involved in other activities. Some activities significantly decrease their recognition rate with more restrictive evaluation methodologies. This is also reflected by the confusion matrix in Figure 4. For instance, *Clearing Table* and *Setting Up Table* are often confused between them since they share similar inertial signals and the same environmental sensors. *Preparing a Cold Meal* is sometimes confused with *Cooking* or *Washing Dishes* since all these actions are performed within the *Kitchen* semantic location.

It is important to note that the overall recognition rate reached by these baselines is relatively high. However, this is likely due to the fact that we considered an unrealistic perfect data association. Figure 5 shows a comparison between perfect data association (i.e., based on ground truth) and a naive data association strategy. In particular, we considered a naive method that associates each environmental sensors event with each subject in the home environment. In order to better highlight the impact of data association, we only considered environmental sensors and we discarded the activities that are not captured by those sensors in the dataset (i.e., *answering phone*, *entering home*, *leaving home*, *making phone call*, and *washing dishes*). We also grouped the activities *clearing table* and *setting up table* since it is not possible to discriminate them only by using environmental sensors in our dataset. A perfect data association of the environmental sensors events dramatically affects the recognition rate of some activities ($\approx +40\%$ in terms of F1 score), like *cooking* and *using pc*. On the other hand, the improvement is lower for those activities that are often performed at the same time by all the subjects of the scenario (e.g., *eating* and *watching tv*). We believe that researchers may use *MARBLE* to investigate novel data association strategies that outperform the recognition rate of the naive approach we presented as a baseline. At the same time, the recognition rate of the perfect data association can be considered as an upper bound while evaluating more realistic strategies.

Finally, since one of the contributions of our dataset is the combination of wearable and environmental sensors, we show in Figure 6 the impact of the inertial sensors data provided by wearable sensors on the recognition rate of each activity. We observed that some activities like *clearing table*, *preparing a cold meal*, and *washing dishes* significantly benefit from wearable sensor data. Indeed, those activities include significant hand gestures that are typical for those activities. As expected, wearable sensor data do not have impact on those activities that are monitored by distinctive environmental sensors (e.g., *watching TV*).

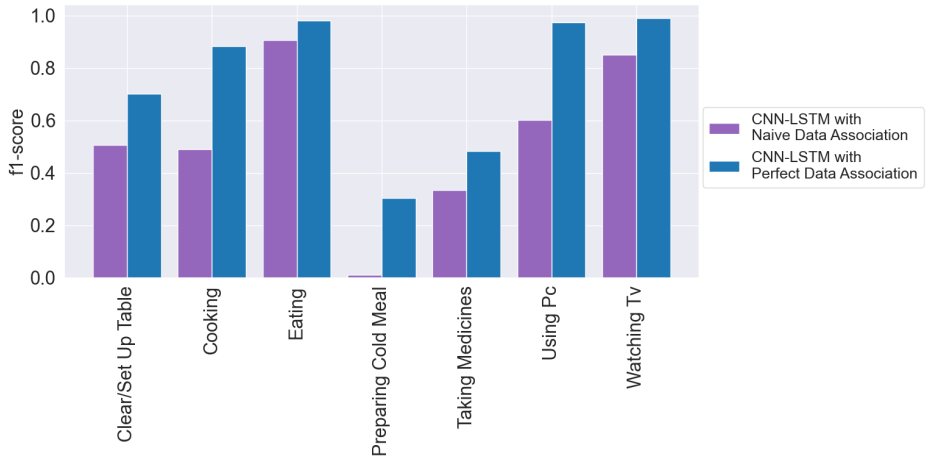

Fig. 5: Comparison between a naive data association strategy and a perfect data association with CNN-LSTM (70/10/20)

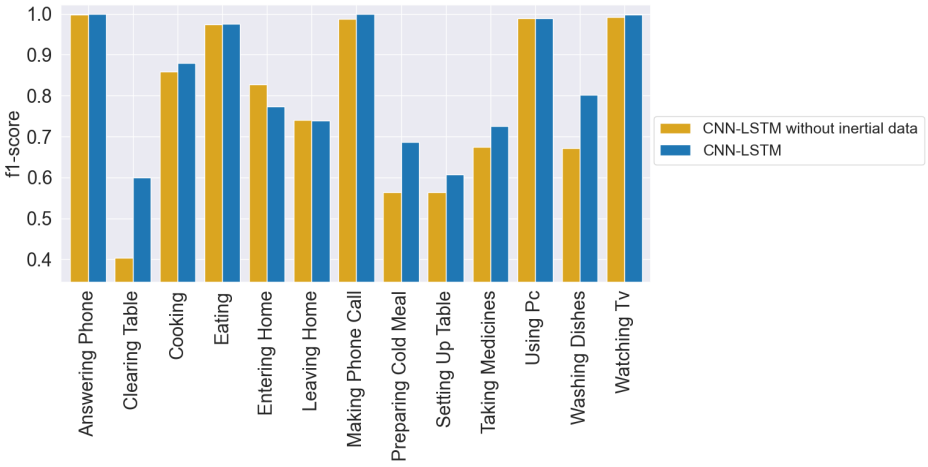

Fig. 6: Impact of inertial data on CNN-LSTM (70/10/20)

5 Conclusion

In this paper we address the need of more public multi-inhabitant ADLs datasets by the research community. We present *MARBLE*, a dataset that includes both wearable and environmental sensors data collected in scenarios where up to four residents concurrently and jointly perform activities in the same smart-home environment. The major limitation of *MARBLE* is that it has not been acquired by continuous monitoring of residents in real homes. Nonetheless, we dedicated a significant effort in designing realistic scenarios, in leaving freedom in activities execution, and in an accurate data acquisition, making the dataset as realistic as possible. We believe that *MARBLE* can be used in the future by several research groups to propose new approaches for single- and multi-inhabitant ADLs recognition. Moreover, *MARBLE* can be used to evaluate novel methods for data association, that is still one of the main open challenges of multi-inhabitant settings.

References

1. Alemdar, H., Ersoy, C.: Multi-resident activity tracking and recognition in smart environments. Journal of Ambient Intelligence and Humanized Computing **8**(4), 513–529 (2017)
2. Alemdar, H., Ertan, H., Incel, O.D., Ersoy, C.: Aras human activity datasets in multiple homes with multiple residents. In: 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops. pp. 232–235. IEEE (2013)
3. Arrotta, L., Bettini, C., Civitarese, G., Presotto, R.: Context-aware data association for multi-inhabitant sensor-based activity recognition. In: 2020 21st IEEE International Conference on Mobile Data Management (MDM). pp. 125–130. IEEE (2020)
4. Benmansour, A., Bouchachia, A., Feham, M.: Multioccupant activity recognition in pervasive smart home environments. ACM Computing Surveys (CSUR) **48**(3), 34 (2016)
5. Blunsden, S., Fisher, R.: The behave video dataset: ground truthed video for multi-person behavior classification. Annals of the BMVA **4**(1-12), 4 (2010)
6. Calatroni, A., Roggen, D., Tröster, G.: Collection and curation of a large reference dataset for activity recognition. In: 2011 IEEE International Conference on Systems, Man, and Cybernetics. pp. 30–35. IEEE (2011)
7. Chen, L., Hoey, J., Nugent, C.D., Cook, D.J., Yu, Z.: Sensor-based activity recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**(6), 790–808 (2012)
8. Cook, D.J., Crandall, A.S., Thomas, B.L., Krishnan, N.C.: Casas: A smart home in a box. Computer **46**(7), 62–69 (2012)
9. Das, B., Cook, D.J., Schmitter-Edgecombe, M., Seelye, A.M.: Puck: an automated prompting system for smart environments: toward achieving automated prompting—challenges involved. Personal and ubiquitous computing **16**(7), 859–873 (2012)
10. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome: Real-world activities of daily living. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 833–842 (2019)

11. van Kasteren, T.L., Englebienne, G., Kröse, B.J.: Human activity recognition from wireless sensor network data: Benchmark and software. In: Activity recognition in pervasive intelligent environments, pp. 165–186. Springer (2011)
12. Kong, Q., Wu, Z., Deng, Z., Klinkigt, M., Tong, B., Murakami, T.: Mmact: A large-scale dataset for cross modal human action understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8658–8667 (2019)
13. Li, Q., Gravina, R., Li, Y., Alsamhi, S.H., Sun, F., Fortino, G.: Multi-user activity recognition: Challenges and opportunities. Information Fusion **63**, 121–135 (2020)
14. Lukowicz, P., Pirkl, G., Bannach, D., Wagner, F., Calatroni, A., Förster, K., Holleczek, T., Rossi, M., Roggen, D., Tröster, G., Doppler, J., Holzmann, C., Riener, A., Ferscha, A., Chavarriaga, R.: Recording a complex, multi modal activity data set for context recognition. In: Proceedings of ARCS '10 - 23th International Conference on Architecture of Computing Systems. pp. 161–166. VDE Verlag (2010)
15. 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. pp. 158–165 (2017)
16. Ordóñez, F., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)
17. Rai, N., Chen, H., Ji, J., Desai, R., Kozuka, K., Ishizaka, S., Adeli, E., Niebles, J.C.: Home action genome: Cooperative compositional action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11184–11193 (2021)
18. Riboni, D., Bettini, C., Civitarese, G., Janjua, Z.H., Bulgari, V.: From lab to life: Fine-grained behavior monitoring in the elderly's home. In: 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops). pp. 342–347. IEEE Computer Society, Washington, D.C. (2015)
19. Riboni, D., Bettini, C., Civitarese, G., Janjua, Z.H., Helaoui, R.: SmartFABER: Recognizing fine-grained abnormal behaviors for early detection of mild cognitive impairment. Artificial Intelligence in Medicine **67**, 57–74 (2016)
20. Riboni, D., Murru, F.: Unsupervised recognition of multi-resident activities in smart-homes. IEEE Access **8**, 201985–201994 (2020)
21. 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. IEEE (2018)
22. Roy, N., Misra, A., Cook, D.: Ambient and smartphone sensor assisted adl recognition in multi-inhabitant smart environments. Journal of ambient intelligence and humanized computing **7**(1), 1–19 (2016)
23. Singla, G., Cook, D.J., Schmitter-Edgecombe, M.: Recognizing independent and joint activities among multiple residents in smart environments. Journal of ambient intelligence and humanized computing **1**(1), 57–63 (2010)
24. Tran, S.N., Nguyen, D., Ngo, T.S., Vu, X.S., Hoang, L., Zhang, Q., Karunanithi, M.: On multi-resident activity recognition in ambient smart-homes. Artificial Intelligence Review **53**(6), 3929–3945 (2020)
25. Wang, T., Cook, D.J.: Toward unsupervised multiresident tracking in ambient assisted living: methods and performance metrics. In: Assistive Technology for the Elderly, pp. 249–280. Elsevier (2020)