sensors

MDPI

Article

Discovering Daily Activity Patterns from Sensor Data Sequences and Activity Sequences

Mirjam Sepesy Maučec * and **Gregor Donaj**

Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroška Cesta 46, SI-2000 Maribor, Slovenia; gregor.donaj@um.si

* Correspondence: mirjam.sepesy@um.si; Tel.: +386-2-220-7225

Abstract: The necessity of caring for elderly people is increasing. Great efforts are being made to enable the elderly population to remain independent for as long as possible. Technologies are being developed to monitor the daily activities of a person to detect their state. Approaches that recognize activities from simple environment sensors have been shown to perform well. It is also important to know the habits of a resident to distinguish between common and uncommon behavior. In this paper, we propose a novel approach to discover a person's common daily routines. The approach consists of sequence comparison and a clustering method to obtain partitions of daily routines. Such partitions are the basis to detect unusual sequences of activities in a person's day. Two types of partitions are examined. The first partition type is based on daily activity vectors, and the second type is based on sensor data. We show that daily activity vectors are needed to obtain reasonable results. We also show that partitions obtained with generalized Hamming distance for sequence comparison are better than partitions obtained with the Levenshtein distance. Experiments are performed with two publicly available datasets.

Keywords: activities of daily living; sensors; Hamming distance; clustering; entropy

check for updates

Citation: Sepesy Maučec, M.; Donaj, G. Discovering Daily Activity Patterns from Sensor Data Sequences and Activity Sequences. *Sensors* **2021**, *21*, 6920. https://doi.org/10.3390/s21206920

Academic Editor: Lars Donath

Received: 9 September 2021
Accepted: 14 October 2021
Published: 19 October 2021

1. Introduction

The number and proportion of elderly people in the population are increasing. In 2019, the number of people aged 60 years and older was 1 billion. This number will increase to 1.4 billion by 2030 and 2.1 billion by 2050 (https://www.who.int/health-topics/ageing#tab=tab_1, accessed on 1 August 2021). The world's aging population is placing increasing pressure on health and social systems, and healthcare providers are struggling to care for elderly people efficiently. In addition, the cost of caring for the elderly in nursing homes is much higher than the cost of in-home care. All these facts forced the fast development of new technologies that can help seniors to stay at home and remain independent for longer [1,2]. Smart home environments are environments that attempt to make the life of their residents more comfortable by using technology that monitors the residents' activities. Monitoring can be performed using video cameras—these approaches are called vision-based approaches [3]. They are problematic with regard to the security and privacy concerns of the residents. The alternative is sensor-based approaches, in which home environments are equipped with several sensors and smart devices. Sensors gather information of different types. The approaches differ based on sensor deployment, which can be wearable or environmental [4,5]. The major problem with wearable sensors is that wearing a tag is sometimes not feasible [6].

For example, in the case of elderly persons or patients, they may forget to wear the tags or they may resist wearing the tags at all. On the other hand, environmental sensors are attached to objects in a house or apartment, and the resident does not need to care about them, except for occasional battery changes. Environmental sensors have many advantages, such as being low cost, less intrusive, and more privacy-preserving [5].

Activity recognition infers a person's activities from monitoring the environment. Approaches that recognize activities from simple environment sensors have been shown to perform well with an accuracy exceeding 90% [7–9]. However, the problem lies in how to interpret the gathered information and what to infer from it. The goal is to detect changes in the activities of daily living (ADL) as they might indicate deteriorating health or mental condition [10]. The ability to detect an emergency situation and set off an alarm is crucial in such environments.

Two commercial solutions are available today: A wearable alarm button to call for help and wearable systems based on accelerometers for automatic fall detection. Such systems require the user to be involved actively by wearing the device, pushing the button, charging the batteries, etc. Automated detection of the unusual behavior of the resident could assist in earlier diagnosis of physical or mental decline and timely treatment. However, the high level of complexity in activity patterns and a large amount of noise stemming from real-life behaviors pose great challenges in achieving this task.

What is unusual behavior of a resident? It is behavior that deviates from their routine [11]. For example, if a resident leaves home frequently but suddenly is at home almost all the time, it could indicate social isolation. On the contrary, if a resident hardly ever leaves home and suddenly the frequency of leaving and returning home increases, it could indicate dementia. Another example is a significant change in personal hygiene practices. For example, if we notice that a resident is bathing infrequently, but usually he was bathing frequently and for a long time, this could indicate a fear of falling in the shower or bath.

Some behavioral patterns could be typical for one person and unusual for another, or could be typical for weekdays and unusual for weekends. For this reason, we define our research problem to discover several different usual behavior patterns of a resident. Our starting point is the claim that a resident, whose activities are recorded in the dataset, is healthy and behaves normally. We define usual behavior patterns as partitions in a clustering algorithm used with recorded data. Later, changes in those patterns, such as frequent new patterns that do not fit in any partition, could be declared as unusual and may be indicators of declining health.

The remainder of the paper is organized as follows. In Section 2, we present an overview of related work. In Section 3, we detail the descriptions of two basic metrics for sequence comparison. Section 4 presents the proposed framework, which consists of a newly proposed sequence comparison and clustering. First, mathematical definitions are given for different comparisons of sensor sequences and activity sequences. Afterward, the clustering method is explained, based on proposed comparisons. Section 5 presents the results of our experiments. We conclude the paper with a final discussion in Section 6.

2. Related Work

Due to remarkable improvements in sensor technology, interest in activity recognition has increased significantly in the last decade [5,12]. Recently, ADL recognition systems were proposed that utilize the sensor data from smartphones [13,14]. There are three main groups of approaches for sensor-based activity recognition: Data-driven, knowledge-driven, and hybrid approaches. Data-driven approaches use various machine learning techniques to learn activities from collected sensor data. The most frequently used are: Naive Bayes classifier [15], Hidden Markov Models [8,16,17], Support Vector Machines [3], dictionaries of patterns [18], and neural networks [6,19,20]. These approaches require a great amount of annotated data to train the models accurately.

For that reason, the scientific community has developed and provided a considerable amount of data sets [21]. The idea of knowledge-driven approaches is to use prior knowledge to create rough activity models. Ontology-based activity recognition was shown to perform comparably well to the data-driven approaches [22]. Hybrid approaches take advantage of the positive features of data-driven and knowledge-driven approaches [23].

Supervised training poses a problem when applying the models on a large scale. Due to differences in monitored environments, a model trained in one environment cannot be used for another environment. Transfer learning was studied to avoid the need to gather a labeled dataset for each new environment [17]. The use of a prior distribution over the model parameters has proved to be efficient with probabilistic models. The prior distribution provides an initial estimate of the model parameters for the target environment and is learned from the source environment. The influence of the prior distribution decreases as more training data are observed for the target environment.

Another problem in ADL recognition is unseen activities. Machine learning algorithms classify activities whose instances have already been seen during training. Very recently, zero-shot learning methods were proposed, which can extend the learning model to detect unseen activities without prior knowledge regarding sensor readings about those previously unseen activities [24].

A literature review has shown that many problems regarding ADL recognition were addressed, and the proposed solutions demonstrated good results [7–9]. However, the question remains what to infer from the recognized sequence of activities. Clustering the sequences could identify typical types of patterns in activity sequences. Noticing unusual patterns of activity sequences could indicate changes in a person's behavior.

Clustering and sequence comparison is studied widely in Bioinformatics, where similarity between protein sequences is sought in order to cluster them into groups of sequences with similar functionality or structure [25]. Biological sequences are greatly different from activity sequences with respect to the timing and duration of the sequence elements. Sequence analysis is also a key method in Social Sciences, where it is used to study the spans of life trajectories and careers [26,27].

Measuring the similarity between sequences depends highly on the choice of similarity measure. Different measures were studied in [28,29]. The first group of analyzed measures is based on distances between probability distributions. The second is based on counts of common attributes, and the third group of measures looks for optimal matching between sequences [30]. In the framework of ADL, we expect the measure to reflect differences in the timing, duration, and sequencing. From the theoretical knowledge, no measure dominates all others in all three dimensions of interest.

Discovering the ADL patterns performed in a day has been a relatively unexplored research area. Activities were discovered by clustering [7]. They employed activity clustering to group the patterns into activity definitions, where the partition centroids represented the activities that were tracked and recognized afterwards. The k-nearest neighbors algorithm is the most widely used clustering algorithm for ADL recognition [4]. In [31], a two-stage ADL recognition was defined, where, in the first stage, activity records were clustered into two partitions by regarding temporal features, and, in the second stage, the classifiers were used to recognize the daily activities in each partition according to the spatial features. Recently, a self-organizing neural network model was presented that considers the following ADL features: The ADL start time, duration, and spatial information [32].

Until recently, research works were focused on ADL recognition with the aim to increase the accuracy of recognition results [33]. However, these works did not analyze recognized activities to determine behavior patterns. Contextual behavior patterns were studied in [34]. Context features were the day of the week, weather, season, noise levels, visitor presence, etc. In [11], normal behavior patterns were defined as lists of activities that a resident performs in their house, with the time of the day and duration. Lists were made from recorded data. Deviations from those definitions were discovered by a decision-support system and may indicate unusual behavior.

In [35,36], an activity-dependent anomaly detection approach was defined, and "sleeping" was selected as the activity of interest. As data similarity measures, Euclidean, Chebyshev, and Canberra distances were studied. A literature review demonstrated that behavior patterns always corresponded to time intervals of one activity.

A summary of selected references from the reviewed literature is presented in Table 1.

Table 1. Studies related to our work.

Refs	Aim	Data or Methods Used
[21]	review	ADL datasets
[15]	ADL recognition	Naive Bayes classifier
[8,16,17]	ADL recognition	HMM
[3]	ADL recognition	Support Vector Machines
[6,19,20]	ADL recognition	neural networks
[31]	ADL recognition	clustering and classification
[4]	ADL recognition	clustering
[30]	review	alignment based similarity measures
[28]	review	similarity measures and distances
[26,27]	life trajectories study	sequence comparison
[25]	analysis of biological sequences	sequence comparison
[7]	ADL definition	clustering
[34]	discovering ADL patterns	similarity adapted to selected features
[32]	discovering ADL patterns	neural network model wit ADL features
[33]	discovering ADL patterns and anomaly detection	HMM
[35]	anomaly detection	numerical distances
[36]	anomaly detection	numerical Euclidean distance
[37]	anomaly detection	Channel State Information

Aim and Research Contribution

This paper aims to analyze a person's daily activities that are usual and to identify their patterns. Our proposed framework differs from the research in the literature in that we are looking for patterns based on activity vectors of whole days. In contrast, in the research work in the literature, methods were examined for pattern extraction from shorter interval sequences [7]. Consequently, clustering of the extracted sensor data patterns is studied in the literature, where the obtained clusters define single activities. In our research, clusters present groups of activity patterns for whole days. Most related research is aimed toward correct ADL recognition, whereas our research aims to use the recognition results as a starting point for discovering the usual behavior of residents.

The research contribution of the present study is the definition of simple similarity metrics, adapted to vectors of sensor data and vectors of daily activities, and the application of clustering to both types of vectors, with the idea to identify days with similar patterns of resident behavior. Such partitions could be used to detect days with unusual patterns of activities.

Our similarity metrics differ from those in the literature in certain aspects. First, they are applied to vectors representing whole days with one entry for one second. The exception are vectors used in the case of the Levenshtein distance. Those vectors are shorter, with one entry for one activity in the sequence. Similarity metrics in literature are derived from numerical distances, whereas our aim was to define a metric applicable to original data. The essential difference in vector comparison is also its sensitivity to the adjacency of activities.

Clustering in literature is applied to specific activity records or is used in the scope of ADL recognition [4]. We apply clustering to ADL sequences representing whole days.

Previous research works do not analyze all activities of the resident performed during the day. They mainly focus on the behavior changes related to one activity only. For example, [35] focuses on the behavior changes related to sleeping, whereas the authors in [37] developed the solution to quickly detect "a fall" of the monitored person.

3. Preliminary

We chose two distance metrics in our research. The Hamming distance was chosen because it is the basic metric for comparing sequential data of equal length. We can use it to compare full-length sensor and activity data.

To compare daily activity vectors, we may also consider that the duration of activities may vary. If we later discharge this duration by merging repetition of the same activity, the Hamming distance cannot be used, because the vectors are now shorter and not of the same length anymore. The Levenshtein distance can be used instead, to determine if two vectors could be considered variations of the same pattern. If a resident would shift his daily routine, such as waking up later, the Levenshtein distance would not be affected, since the sequence of activities would not change.

We wanted to compare these two distances to find which metric was more appropriate for detecting unusual behavior.

3.1. Hamming Distance

In general, the Hamming distance between two vectors $\vec{x}$ and $\vec{y}$ is the number of positions in which the two vectors are different:

$$H(\vec{x},\vec{y}) = \sum_{i=1}^{n} \mathit{diff}(x_i, y_i), \tag{1}$$

where n is the dimension of the vectors, x_i and y_i are the i-th components of vectors $\vec{x}$ and $\vec{y}$, respectively. The difference function *diff* gives a result of 1 if x_i and y_i differ, and 0 if they are the same. This distance can only be applied to sequences of equal length.

In order to use the Hamming distance in our research, we needed to make some generalizations to the *diff* function, which we present in the next section.

3.2. Levenshtein Distance

The Levenshtein distance is given by the smallest number of edit operations needed to turn one sequence into another. The Levenshtein distance between two vectors $\vec{x}$ and $\vec{y}$ is defined as:

$$L(\vec{x},\vec{y}) = \begin{cases} |\vec{x}| \cdot cost_D, & \text{if } |\vec{y}| = 0, \\ |\vec{y}| \cdot cost_I, & \text{if } |\vec{x}| = 0, \\ L(tail(\vec{x}), tail(\vec{y})), & \text{if } x_1 = y_1, \\ min \begin{cases} L(tail(\vec{x}), \vec{y}) + cost_D, \\ L(\vec{x}, tail(\vec{y})) + cost_I, \\ L(tail(\vec{x}), tail(\vec{y})) + cost_S \end{cases} & \text{otherwise.} \end{cases} \tag{2}$$

Here, $|\vec{x}|$ denotes the length of vector $\vec{x}$, and $tail(\vec{x})$ is vector $\vec{x}$ without the first element. Edit operations (insertion, deletion, or substitution) are penalized by costs (i.e., $cost_I$, $cost_D$, or $cost_S$), which are all equal to one in the original version of the distance. As one substitution equals one deletion and one insertion, its cost could be the sum of the cost of deletion and the cost of insertion. Using more than one substitution cost allows even more flexibility in the comparison.

4. Proposed Framework

Our research framework addresses the problem of discovering the daily activity patterns of the resident. Identifying patterns is difficult, as the duration and the order of activities may vary from one occurrence to another. In some pattern occurrences, certain activities may be missing.

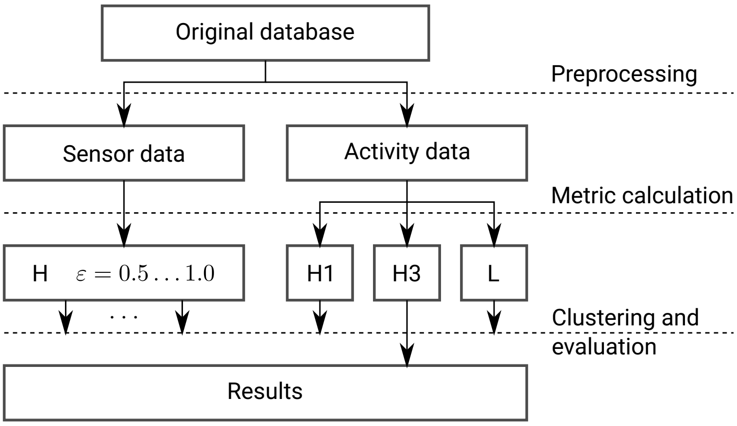


Figure 1. Flowchart of the proposed framework.

The proposed framework is based on the premise that time annotated sensor readings and daily activity sequences of a resident are available. The entire framework is presented in Figure 1 and is composed of five tasks: Data preprocessing, comparing sequences by metric calculation on sensor and activity data separately, clustering based on pairwise comparison methods with evaluation of the clustering results, visual representation of clusters, and visual representation of daily activity vectors within clusters. The first task, data preprocessing, generates a set of vectors of active sensors and a set of vectors of daily activities. This task is described in the next section. The second task, comparing sequences, provides definitions of distance metrics, adapted to sensor data and activity sequences, respectively. The third task, clustering, provides clusters of similar vectors based on distance metrics. These two tasks are described in the continuation of this section. The fourth and the fifth tasks visualize the results of clustering. They are given in the next section. The visual representation of daily activities within clusters shows similar daily activity patterns of the resident.

4.1. Comparing Sequences of Active Sensors

The basic source of information about a resident's activities is sensor data. The sensor data stream is organized into a vector of active sensors sets over time:

$$\vec{s} = [s_1, \dots, s_n], \quad s_i \subset \mathcal{S}, \tag{3}$$

where $\mathcal{S}$ is a finite alphabet of sensors, and s_i is a set of sensors active in time slot i. The dimension of the vector n depends on the time scale. For example, if time slots correspond to seconds, and the vector corresponds to a full day, the dimension n is 86,400.

Considering vectors of active sensors $\vec{s}$ and $\vec{q}$, at each time slot i, we have sets of active sensors s_i and q_i. In this case, we define the difference function as:

$$diff_S(s_i, q_i, \varepsilon) = \begin{cases} 0, & 2 \cdot |s_i \cap q_i| > \varepsilon \cdot (|s_i| + |q_i|), \\ 1, & \text{otherwise}. \end{cases} \tag{4}$$

The parameter $\varepsilon \in [0, 1]$ allows us to consider an incomplete matching between the sets of active sensors, since the data captured through sensor devices tend to be noisy. The parameter determines the ratio of matching active sensors from both vectors to the sum of active sensors in both vectors, which is needed to call an agreement.

4.2. Comparing Sequences of Activities

Using ADL recognition techniques, sensor data are transformed into an activity sequence. We consider daily activity sequence $\vec{a}$ as a vector of activities from the finite alphabet $\mathcal{A}$ in successive time slots.

$$\vec{a} = [a_1, \ldots, a_n], \quad a_i \in \mathcal{A}. \tag{5}$$

We call this a daily activity vector. Since the position in the sequence conveys time information, the difference between two positions defines a duration. As in the case of sensors, the dimension of the vector n depends on the time scale.

4.2.1. Entropy

Some individuals, if they have a routine, perform their daily activities in a predictive way. Others have very diverse timelines. Shannon entropy can be used to measure the uncertainty of activity at time slot i. It is calculated using:

$$h_i = -\sum_{j=1}^{A} p(a_{i,j}) \cdot \log_2 p(a_{i,j}), \quad A = |\mathcal{A}|, \tag{6}$$

where $p(a_{i,j})$ denotes the probability of activity a_j at time slot i. Entropy is 0 or close to 0 if the activity at the considered time slot is predictable, with the probability close to 1. Entropy is close to its maximum, which is $\log_2 A$, if all activities are equiprobable.

In Equation (6), the activity at a considered time slot is selected independent of previous activities. However, in real-life scenarios, activities are not independent. If we consider that the selection of an activity at a considered time slot i is dependent on the activity at the immediately preceding time slot $i-1$ (i.e., a first-order Markov source), the conditional entropy is calculated as:

$$h_i^* = -\sum_{k=1}^{A} p(a_{i-1,k}) \cdot \sum_{j=1}^{A} p(a_{i,j}|a_{i-1,k}) \cdot \log_2 p(a_{i,j}|a_{i-1,k}), \quad A = |\mathcal{A}|, \tag{7}$$

where $p(a_{i-1,k})$ denotes the probability of activity a_k at time slot $i-1$ and $p(a_{i,j}|a_{i-1,k})$ is the conditional probability of activity a_j at time slot i if the activity a_k was performed at time slot $i-1$.

Using entropy, we estimate how difficult it is to predict the daily activity vector for a given resident.

4.2.2. Generalized Hamming Distance

The simple Hamming distance between two daily activity vectors $\vec{a}$ and $\vec{b}$ is the number of positions in which the two vectors of daily activities are different (see Equation (1)), where the difference function is defined as:

$$\mathit{diff}_A(a_i, b_i) = \begin{cases} 0, & a_i = b_i, \\ 1, & a_i \neq b_i. \end{cases} \tag{8}$$

We denote the Hamming distance based on Equation (8) with H1.

A generalization will allow for state-dependent costs of mismatching. The generalized Hamming distance is defined as the sum of activity-dependent position-wise mismatches between two daily activity vectors by using the difference function:

$$\mathit{diff}_G(a_i, b_i) = \begin{cases} 0, & a_i = b_i, \\ \mathit{cost}, & a_i \sim b_i, \\ 1, & a_i \neq b_i. \end{cases} \tag{9}$$

Here, *cost* is a fixed value in the interval $[0,1]$, and $a \sim b$ denotes adjacency of a and b, which means that the activities are different, but a transition exists from activity a to activity b or from activity b to activity a, i.e., the two activities are consecutive to each other at least once in the dataset. A typical example would be the activity pair "meal preparation" and "eating".

The last option, $a \neq b$, denotes that activities a and b are neither the same nor adjacent. We denote the Hamming distance based on Equation (9) with H2.

Interleaved or concurrent activities may occur. If there is a possibility of two concurrent activities, the generalized Hamming distance must take all possible transitions into account. We now use a_i and b_i as a notation for sets of concurrent activities at time slot i. We limit the number of concurrent activities to two. The sets can now have one or two elements, e.g., $a_i = \{a_{i,1}\}$ or $a_i = \{a_{i,1}, a_{i,2}\}$.

The generalized Hamming distance for this case uses the difference function:

$$diff_{G*}(a_i, b_i) = \begin{cases} 0, & a_i = b_i, \\ cost, & a_{i,1} \sim b_{i,1} \wedge |a_i| = |b_i| = 1, \\ cost, & (a_{i,1} \sim b_{i,1} \vee a_{i,2} \sim b_{i,1}) \wedge |a_i| = 2 \wedge |b_i| = 1, \\ cost, & (a_{i,1} \sim b_{i,1} \vee a_{i,1} \sim b_{i,2}) \wedge |a_i| = 1 \wedge |b_i| = 2, \\ cost, & (a_{i,1} \sim b_{i,1} \vee a_{i,1} \sim b_{i,2} \vee a_{i,2} \sim b_{i,1} \vee a_{i,2} \sim b_{i,2}) \wedge |a_i| = |b_i| = 2, \\ 1, & a_i \neq b_i. \end{cases} \quad (10)$$

In Equations (9) and (10), the costs of mismatches (denoted as *cost*) can be fixed, or they could be derived from the observed transition rates. The probability of transition from activity a to activity b in the sequence is estimated as:

$$p(b|a) = \frac{\sum_{j=1}^{d} C_j(a \to b)}{\sum_{j=1}^{d} C_j(a)}, \quad (11)$$

where d denotes the number of observed days, $C_j(a \to b)$ counts the number of transitions from activity a to activity b in the daily activity vector of day j, and $C_j(a)$ counts the number of transitions from activity a to any other activity in the vector for day j. The symmetrical cost is defined as:

$$cost = 1 - 0.5 \cdot p(a|b) - 0.5 \cdot p(b|a). \quad (12)$$

We denote the Hamming distance with costs from Equations (11) and (12) with H3.

The Hamming distance is symmetrical $(H(\vec{a},\vec{b}) = H(\vec{b},\vec{a}))$, and the symmetry is preserved in generalized Hamming distances H2 and H3. The similarity measure expresses the similarity between two vectors on a scale from 0 to 1. For the Hamming distance, it is defined as:

$$sim_H(\vec{a},\vec{b}) = 1 - \frac{H(\vec{a},\vec{b})}{n}. \quad (13)$$

4.2.3. Levenshtein Distance

Daily activity vectors could be compared as sequences of activities irrespective of their duration. The Levenshtein distance measures the distance in this sense. The Levenshtein distance between two daily activity vectors $\vec{a}$ and $\vec{b}$ is given in Equation (2). In our experiments, we set $cost_I = cost_D = 1$ and $cost_S = 2$.

The similarity measure, defined with the Levenshtein distance, is:

$$sim_L(\vec{a},\vec{b}) = 1 - \frac{L(\vec{a},\vec{b})}{\max(|\vec{a}|,|\vec{b}|)}. \quad (14)$$

The Hamming distance can only be applied to sequences of equal length. On the contrary, the Levenshtein distance can be computed between sequences of different lengths. By shrinking the activity sequence to transitions between activities, the time span of each activity is lost. Where the timing of activities is crucial, Hamming distance should be used. Where it is less important, the Levenshtein distance could be more appropriate.

4.3. Clustering

Based on the above-described distance metrics of sensor or activity data, we can form a distance matrix for all the days in our datasets. Hereafter, the days in the datasets are our data points for clustering, which is used to divide the data points into partitions. Since the data points are not in a vector space, we cannot calculate means for partitions. Therefore, clustering is based on medoids instead. A medoid is a representative data point, and serves as the "center" of the partition to which distances from other data points are used.

Clustering is performed using the Partition Around Medoids (PAM) algorithm [38], which works in two phases. In the first phase, a predetermined number of elements k from the set is randomly selected as possible medoids—one for each cluster. All data points are then associated with the closest medoid candidate, and a cost function is calculated. The cost function is the sum of distances from all data points in the dataset to their medoids.

In the second phase, medoids are swapped with other data points, and the cost function is recalculated. This swap is repeated for pairs of metoids and non-metoid data points. Only the swap, which results in the best new cost function value, is then applied for the next iteration. Iterations are repeated as long as the cost function improves.

The clustering algorithm results are k partitions, each containing a number of the data points, one of which is the medoid. For a better graphical representation, the distance matrix can be rearranged so that data points belonging to the same partitions form consecutive rows and columns in the matrix.

We consider two different clusterings to be similar if the data points are clustered in similar partitions. In order to determine such similarities, we calculate the frequently used Rand index. This index is calculated as a value between 0 and 1, where higher values mean higher similarity and lower values mean lower similarity. The value of the Rand index is 1 if and only if the two partitionings are identical.

However, it was shown that typically values of the Rand index are in an interval close to 1. Even in the case of statistical independence, the index values can be rather high, and must, therefore, be interpreted carefully [39].

The Rand index is calculated using the formula:

$$R = \frac{N_s + N_d}{\binom{m}{2}}, \tag{15}$$

where N_s is the number of data-point pairs that are in the same partition in both partitionings, N_d is the number of data-point pairs that are in different partitions in both partitionings, and m is the total number of data points in the dataset.

5. Experiments

5.1. Data Sets

The problem tackled in our research is the discovery of daily activity patterns from vectors of active sensors and vectors of activity sequences. Currently, many sensor-based datasets for ADL recognition are available to end-users and researchers [21]. However, to the best of our knowledge, there is no dataset with ADL recognition results available directly. For this reason, the main guiding principle in the selection of the dataset were the published ADL results (classification accuracy, F-measure), which ensures that the activities can be identified correctly from the sensor readings in the dataset.

Table 2. Datasets used in experiments.

Dataset	Kasteren	CASAS 11
Occupancy	1 resident	2 residents
Capture	28 days	232 days
Number of sensors	14	88
Number of binary sensors	14	82
Number of activities	7	13 + 12
Maximum no. of concurrent activities	2	2

The experiments were performed on two different datasets. The Kasteren dataset (http://casas.wsu.edu/datasets/kasterenDataset.zip, accessed on 31 May 2021) was recorded in an apartment with three rooms, where one 27-year old male lived. The dataset was annotated using a handwritten diary of activities made by the resident. At certain time intervals, two activities were annotated as concurrent (for example, ”use toilet” and ”go to bed”). ADL recognition accuracy of 95.6 % was reported for the Kasteren dataset [40]. The CASAS center (http://casas.wsu.edu/datasets/, accessed on 31 May 2021) collected several datasets of sensor and activity data [41]. We selected the CASAS 11 dataset. This dataset was collected in an apartment with two residents with spontaneous activities and no predetermined scenarios. It also has concurrent activities. A comparable accuracy for the Kasteren and CASAS datasets was reported in [42]. Several other studies reported high recognition accuracy in CASAS datasets with daily living data. An overview of these studies can be found in [21]. Details of the datasets used in experiments are collected in Table 2. We followed the rule that datasets should be of comparable sizes, and used only the first 30 days of the CASAS 11 dataset.

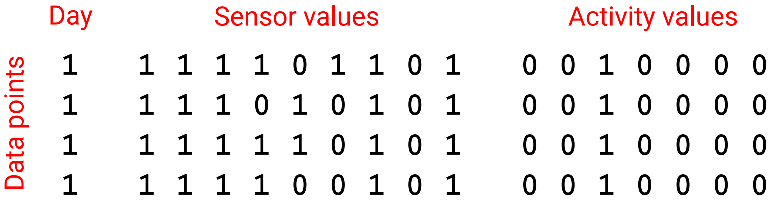
Since the CASAS 11 dataset has two residents and later in the paper, we will analyze the activities for both residents separately, we will refer to a total of three datasets.

5.2. Data Preprocessing

We considered only binary sensor data and, therefore, excluded non-binary sensor data, such as temperature, electric power consumption, etc. In the CASAS 11 dataset, additionally, we had to make some minor corrections (e.g., sensor value “OF” is replaced with “OFF,” the year 22009 was corrected to 2009, etc.).

All datasets were then reformatted into a new format—a text file, where each line corresponds to one time slot (a second), and all binary sensor and activity data are written in columns with numeric values 0 and 1. Timestamps for events (changes in sensor value or activity transitions) were rounded to the nearest second, where needed.

An examination of the daily activities in both datasets revealed that the residents were performing different activities on different days at midnight. In order to have the same activity at the start and end of each day—sleep—we decided to shift the start. Therefore, we decided to start a day in our experiments at 4 a.m. on one calendar day and end the day at 4 a.m. on the next calendar day. The format of the preprocessed datasets is presented in Figure 2.



```
     Day      Sensor values          Activity values
D    1    1 1 1 1 0 1 1 0 1    0 0 1 0 0 0 0
a    1    1 1 1 0 1 0 1 0 1    0 0 1 0 0 0 0
t    1    1 1 1 1 1 0 1 0 1    0 0 1 0 0 0 0
a    1    1 1 1 1 0 0 1 0 1    0 0 1 0 0 0 0
points
```

Figure 2. Excerpt from the preprocessed dataset. The first column denotes the day in the dataset, the following columns denote sensor values (one column per sensor), and the last columns denote activity values (one column per activity). Each line represents one data point and corresponds to one time slot. Value 1 denotes active sensor or present activity.

Due to this shift, we disregarded the first 4 h of the first day from all datasets. In the CASAS 11 datasets, we then also used 4 h from the first three days to obtain the full 30 periods of 24 h. In the Kasteren dataset; however, there were no more data for the following day. The last day in the dataset ended with no activity at all, indicating that the resident was away for the night. We decided to extend this state for another 4 h to complete the reformatted dataset.

We found that in both datasets, two activities could occur at the same time. In the Kasteren dataset, the activity ″use toilet″ can occur during the activities ″prepare dinner″ and ″go to bed,″ which—judging from the data—also means staying in bed and sleeping. Similarly, in the CASAS 11 datasets, concurrent activities are possible for each of the two residents, e.g., ″eating″ and ″watching TV″.

We were interested in the residents' daily habits. Can we define their routine directly from sensor data, or do we need ADL recognition first? To this end, we performed two types of transformations from the new file format. We created a file where each line corresponded to active sensors in one day. In the second file, each line corresponded to activities performed on one day. In both files, each column corresponded to one second. These files were then used to perform distance metric calculations and for a graphical representation of activity patterns.

5.3. Results and Discussion

5.3.1. Entropy

First, we were interested in how uncertain the activities were at different times. Entropy plots for all datasets were calculated (see Equation (6)) and are given in Figure 3. Entropy was calculated every half minute. For this purpose, we resized the daily activity vectors from the dimension n = 86,400 to the dimension $n = 2880$. In this case, i denotes the i-th time slot with the duration of half a minute. Data for the time slots in the resized vector were obtained by merging data from an interval of time slots in the original vector.

An activity is marked as present if it is present in at least one time slot within the corresponding interval. Entropy is always lower than 2.8 bits. In the Kasteren dataset, the average entropy is 0.95 bits, whereas in the CASAS 11 datasets, it is 0.51 and 0.33 bits. Entropy between the Kasteren and CASAS datasets cannot be compared directly quantitatively, as the CASAS datasets have more activities than the Kasteren dataset.

In the Kasteren dataset, uncertainty at night was caused by the activity ″use toilet″, surrounded by the activity ″go to bed.″ Uncertainty in the morning resulted from occasionally skipping the activities ″prepare breakfast″ or ″take shower.″ In the middle of the day, the entropy falls to a small value as the resident has almost always left home. Uncertainty in the evening is a consequence of the activities ″prepare dinner,″ ″get drink,″ ″use toilet,″ and ″take shower,″ which were not taken consistently.

In both CASAS 11 datasets, we can see a long period with the entropy equal to zero. This indicates that, at this time, the activity is always the same. Each day, both residents left home for a long period of time, as is evident from the the activity sequence ″leave home,″ ″no activity,″ and ″enter home.″ On certain days, activity "leave home" or activity "enter home" is missing, but its occurrence is evident from the previous activity or the following activity.

Plots in Figure 4 show the conditional entropy (see Equation (7)). In all plots, it is lower than 1.3 bits. Conditional entropy is always lower than unconditional entropy. As expected, the activities in adjacent time slots are not independent. All calculated entropies are also much lower than their upper limits, being $log_2 7 = 2.8$ bits for the Kasteren dataset and $log_2 13 = 3.7$ bits ($log_2 12 = 3.58$ bits) for the CASAS 11 dataset, which confirms that activities are not selected randomly. All observations for unconditional and conditional entropy show patterns of normal behavior of the resident.

Given that the entropies for both residents in the CASAS 11 datasets are higher than the entropy in the Kasteren dataset despite having more possible activities, we can conclude that the residents in the CASAS 11 datasets are more consistent in their daily activities.

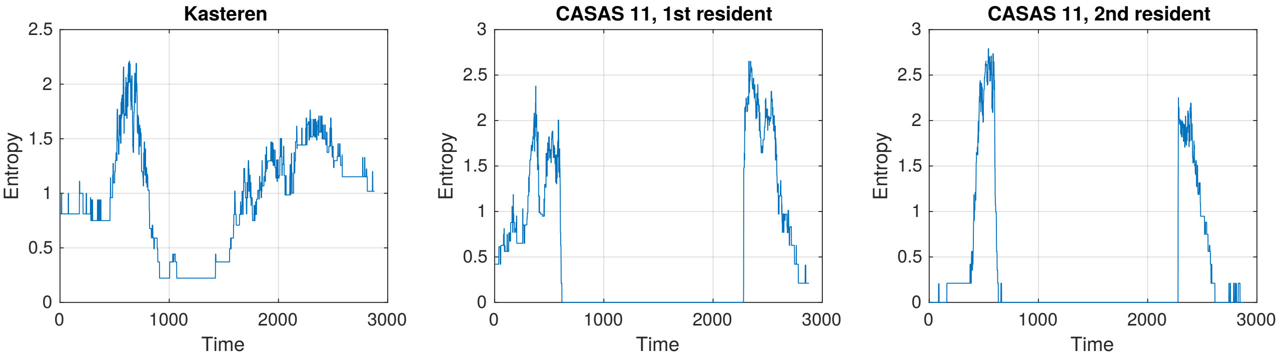


Figure 3. The entropy of activities at different times of the day for the Kasteren dataset, the first resident of the CASAS 11 dataset, and the second resident of the CASAS 11 dataset. A day starts at 4 a.m. on one calendar day and ends at 4 a.m. on the next calendar day. In the CASAS 11 dataset, entropy was calculated separately for each of the two residents. It was calculated every half minute.

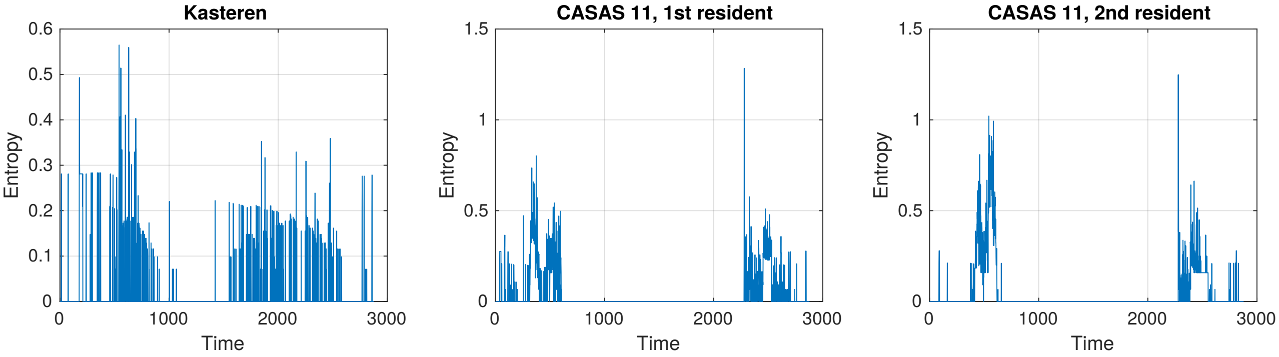


Figure 4. The conditional entropy of activities at different times of the day for the Kasteren dataset, the first resident of the CASAS 11 dataset, and the second resident of the CASAS dataset.

5.3.2. Distances between Daily Activity Vectors

From the reformatted datasets, the distances between days were calculated based on the metrics described in Sections 4.1 and 4.2. We chose to use distances rather than similarities. This decision does not influence the final results, as they are equivalent. The distances can be calculated based on sensor data or daily activity vectors.

First, we were interested in activity vectors. Hamming distance and Levenshtein distance were examined. When using the Hamming distance, the activities of each day are presented with a daily activity vector of constant size $n = 86{,}400$, where one component in the vector corresponds to one second. For the Levenshtein distance, the vector sizes are smaller and diverse, where one component in the vector corresponds to one activity regardless of its duration.

We were interested in the distances between consecutive days in the datasets. All three types of Hamming distance between activity vectors of consecutive days for all datasets are shown in Figure 5. In the Kasteren dataset, distances ranged from 10,000 to 50,000, whereas, in the CASAS 11 datasets, they were between 0 and 30,000. As data points correspond to seconds, the distance 50,000 means that activities in two daily vectors do not match for almost 14 h. Distance 0 corresponds to two days, on which the second resident in the CASAS 11 dataset left home. As expected, H3 distances are shorter than H1 distances and larger or equal to H2 distances, as H2 and H3 use *cost* values smaller than 1. Our general observation is that the Hamming distance between activity vectors of consecutive days can vary a great deal, and we cannot infer unusual behavior of the resident from them.

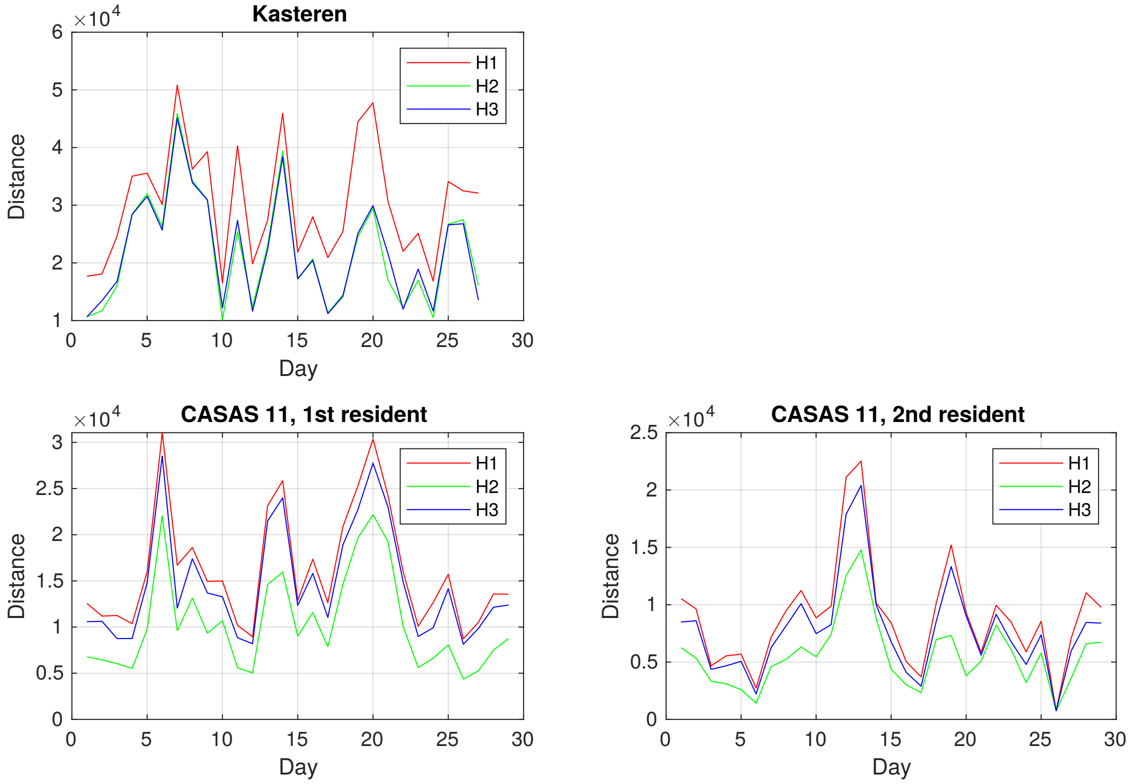


Figure 5. Hamming distances between daily activity vectors for consecutive days for the Kasteren dataset, the first resident of the CASAS 11 dataset, and the second resident of the CASAS dataset.

In Table 3, we collected the average values for all Hamming distances between consecutive days for all datasets. We can see that the average distances are quite large, even using a generalization of the Hamming distance. The large average distances between consecutive days show that the behavior of residents changes from day to day.

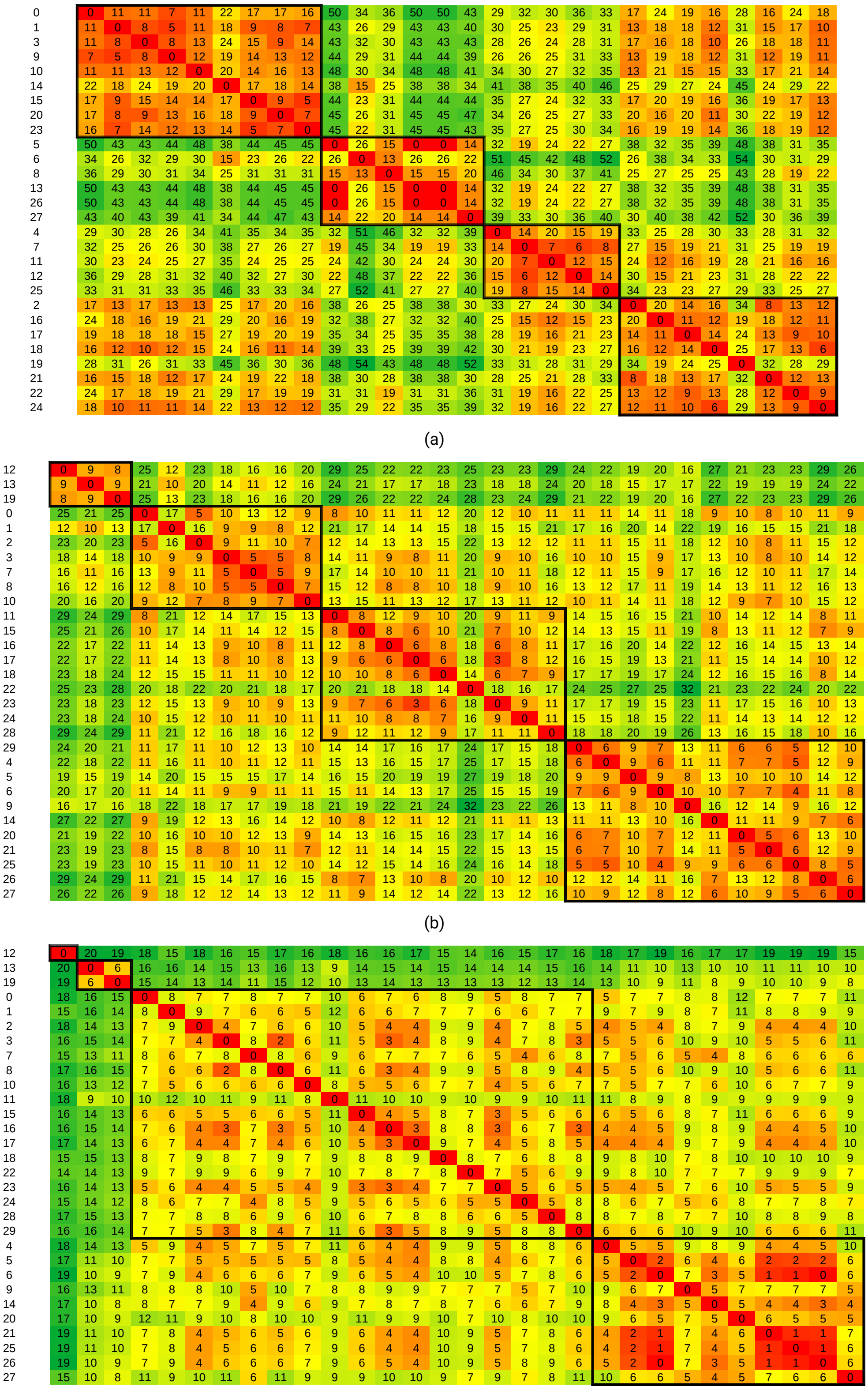
Table 3. Average Hamming distances between consecutive days.

Dataset	H1	H2	H3
Kasteren	30,341.63	21,888.31	22,134.49
CASAS 11, first resident	16,218.69	10,378.21	14,588.89
CASAS 11, second resident	8910.83	5555.45	7720.90

5.3.3. Clustering of Daily Activity Vectors

In the continuation of our study, we were interested to discover if we could define partitions of common daily activity patterns by grouping daily activity vectors having shorter distances. Thus, we wanted to identify days with similar behavior of the residents. For further experiments, we used the metrics H1 and H3. The graphs in Figure 5 show similar behavior for all three metrics.

We formed distance matrices containing the distances, taken pairwise, between activity vectors for all days. These matrices were used for clustering, which was performed as described in Section 4.3. The resulting matrices, in which activity vectors were reorganized according to the obtained partitions, are given in Figure 6.



(a)

(b)

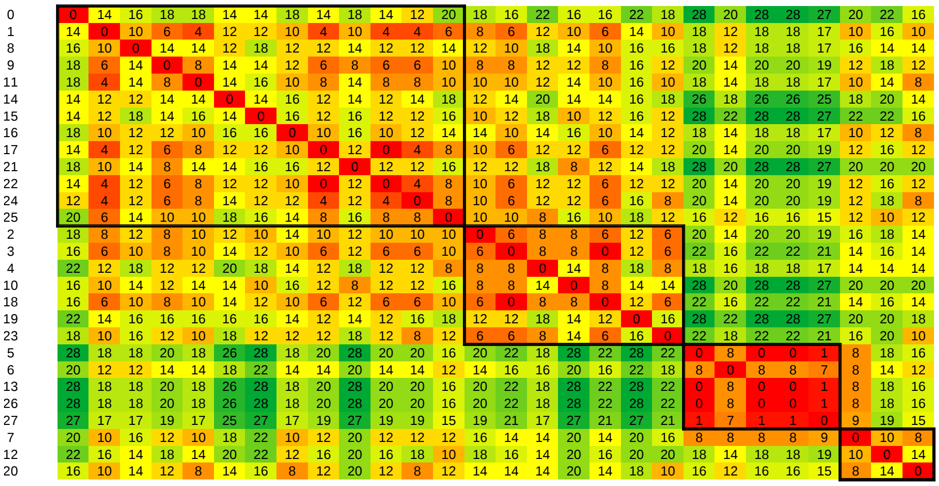
(c)

Figure 6. Distance matrices of the H3 metric between daily activity vectors for (**a**) Kasteren dataset, (**b**) The first resident of the CASAS 11 dataset, and (**c**) The second resident of the CASAS dataset. Values are in thousands. The background color shows gradient changes in values, with red tones indicating low values and green tones indicating high values.

In Figures 6–8, the color scheme indicates distances, with green tones representing large distances and red tones representing small distances. The numeric distance values are in thousands. Bolted borders enclose distances between data points from the same partition. The numbers on the left side denote consecutive numbers for the days in the dataset (starting with 0). Figures 6–8 were obtained using the H3 metric. Similar figures can be obtained using the H1 metric.

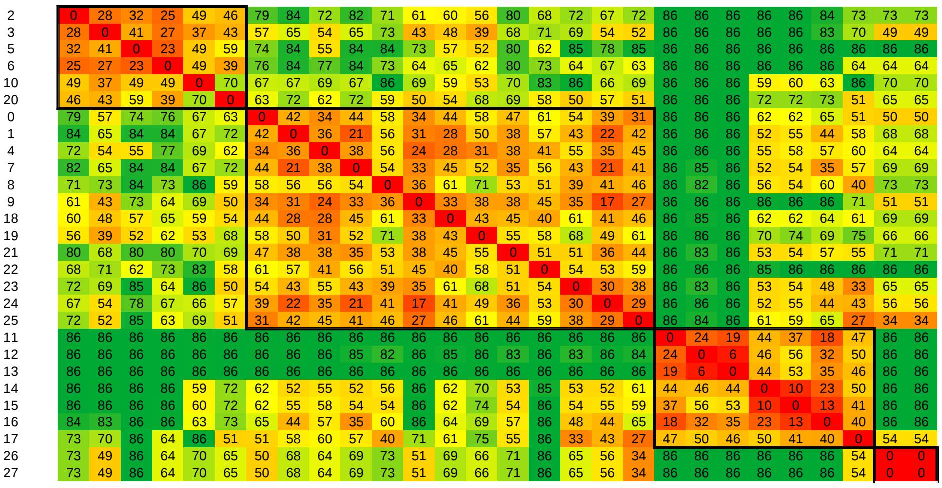
From Figure 6, partitions can be seen clearly. For example, in the Kasteren dataset, we see that distances within the first partition (top left to bottom right) are significantly smaller than the distances between days from the first partition and days from the second or third partitions. Distances between days from the first partition and days from the fourth partition are smaller than distances between other combinations of partitions, but still larger than distances within the first partition. This observation indicates that the first and fourth partitions contain daily activity vectors with some degree of similarity. We can also extend such interpretations to other pairs of partitions in all three datasets.

Graphical depictions of the distances within partitions show that a four partition solution achieves a balance between internal partition distances and the interpretability of the partitions.



0	0	14	16	18	18	14	14	18	14	18	14	12	20	18	16	22	16	16	22	18	28	20	28	28	27	20	22	16
1	14	0	10	6	4	12	12	10	4	10	4	4	6	8	6	12	10	6	14	10	18	12	18	18	17	10	16	10
8	16	10	0	14	14	12	18	12	12	14	12	12	14	12	10	18	14	10	16	16	18	12	18	18	17	16	14	14
9	18	6	14	0	8	14	14	12	6	8	6	6	10	8	8	12	12	8	16	12	20	14	20	20	19	12	18	12
11	18	4	14	8	0	14	16	10	8	14	8	8	10	10	10	12	14	10	16	10	18	14	18	18	17	10	14	8
14	14	12	12	14	14	0	14	16	12	14	12	14	18	12	14	20	14	14	16	18	26	18	26	26	25	18	20	14
15	14	12	18	14	16	14	0	16	12	16	12	12	16	10	12	18	10	12	16	12	28	22	28	28	27	22	22	16
16	18	10	12	12	10	16	16	0	10	16	10	12	14	14	10	14	16	10	14	12	18	14	18	18	17	10	12	8
17	14	4	12	6	8	12	12	10	0	12	0	4	8	10	6	12	12	6	12	12	20	14	20	20	19	12	16	12
21	18	10	14	8	14	14	16	16	12	0	12	12	16	12	12	18	8	12	14	18	28	20	28	28	27	20	20	20
22	14	4	12	6	8	12	12	10	0	12	0	4	8	10	6	12	12	6	12	12	20	14	20	20	19	12	16	12
24	12	4	12	6	8	14	12	12	4	12	4	0	8	10	6	12	12	6	16	8	20	14	20	20	19	12	18	8
25	20	6	14	10	10	18	16	14	8	16	8	8	0	10	10	8	16	10	18	12	16	12	16	16	15	12	10	12
2	18	8	12	8	10	12	10	14	10	12	10	10	10	0	6	8	8	6	12	6	20	14	20	20	19	16	18	14
3	16	6	10	8	10	14	12	10	6	12	6	6	10	6	0	8	8	0	12	6	22	16	22	22	21	14	16	14
4	22	12	18	12	12	20	18	14	12	18	12	12	8	8	8	0	14	8	18	8	18	16	18	18	17	14	14	14
10	16	10	14	12	14	14	10	16	12	8	12	12	16	8	8	14	0	8	14	14	28	20	28	28	27	20	20	20
18	16	6	10	8	10	14	12	10	6	12	6	6	10	6	0	8	8	0	12	6	22	16	22	22	21	14	16	14
19	22	14	16	16	16	16	16	14	12	14	12	16	18	12	12	18	14	12	0	16	28	22	28	28	27	20	20	18
23	18	10	16	12	10	18	12	12	12	18	12	8	12	6	6	8	14	6	16	0	22	18	22	22	21	16	20	10
5	28	18	18	20	18	26	28	18	20	28	20	20	16	20	22	18	28	22	28	22	0	8	0	0	1	8	18	16
6	20	12	12	14	14	18	22	14	14	20	14	14	12	14	16	16	20	16	22	18	8	0	8	8	7	8	14	12
13	28	18	18	20	18	26	28	18	20	28	20	20	16	20	22	18	28	22	28	22	0	8	0	0	1	8	18	16
26	28	18	18	20	18	26	28	18	20	28	20	20	16	20	22	18	28	22	28	22	0	8	0	0	1	8	18	16
27	27	17	17	19	17	25	27	17	19	27	19	19	15	19	21	17	27	21	27	21	1	7	1	1	0	9	19	15
7	20	10	16	12	10	18	22	10	12	20	12	12	12	16	14	14	20	14	20	16	8	8	8	8	9	0	10	8
12	22	16	14	18	14	20	22	12	16	20	16	18	10	18	16	14	20	16	20	20	18	14	18	18	19	10	0	14
20	16	10	14	12	8	14	16	8	12	20	12	8	12	14	14	14	20	14	18	10	16	12	16	16	15	8	14	0

Figure 7. Distance matrix of the Levenshtein metric between daily activity vectors for the Kasteren dataset. Values are in thousands. The background color shows gradient changes in values, with red tones indicating low values and green tones indicating high values.



2	0	28	32	25	49	46	79	84	72	82	71	61	60	56	80	68	72	67	72	86	86	86	86	86	84	73	73	73
3	28	0	41	27	37	43	57	65	54	65	73	43	48	39	68	71	69	54	52	86	86	86	86	86	83	70	49	49
5	32	41	0	23	49	59	74	84	55	84	84	73	57	52	80	62	85	78	85	86	86	86	86	86	86	86	86	86
6	25	27	23	0	49	39	76	84	77	84	73	64	65	62	80	73	64	67	63	86	86	86	86	86	86	64	64	64
10	49	37	49	49	0	70	67	67	69	67	86	69	59	53	70	83	86	66	69	86	86	86	59	60	63	86	70	70
20	46	43	59	39	70	0	63	72	62	72	59	50	54	68	69	58	50	57	51	86	86	86	72	72	73	51	65	65
0	79	57	74	76	67	63	0	42	34	44	58	34	44	58	47	61	54	39	31	86	86	86	62	62	65	51	50	50
1	84	65	84	84	67	72	42	0	36	21	56	31	28	50	38	57	43	22	42	86	86	86	52	55	44	58	68	68
4	72	54	55	77	69	62	34	36	0	38	56	24	28	31	38	41	55	35	45	86	86	86	55	58	57	60	64	64
7	82	65	84	84	67	72	44	21	38	0	54	33	45	52	35	56	43	21	41	86	85	86	52	54	35	57	69	69
8	71	73	84	73	86	59	58	56	56	54	0	36	61	71	53	51	39	41	46	86	82	86	56	54	60	40	73	73
9	61	43	73	64	69	50	34	31	24	33	36	0	33	38	38	45	35	17	27	86	86	86	86	86	86	71	51	51
18	60	48	57	65	59	54	44	28	28	45	61	33	0	43	45	40	61	41	46	86	85	86	62	62	64	61	69	69
19	56	39	52	62	53	68	58	50	31	52	71	38	43	0	55	58	68	49	61	86	86	86	70	74	69	75	66	66
21	80	68	80	80	70	69	47	38	38	35	53	38	45	55	0	51	51	36	44	86	83	86	53	54	57	55	71	71
22	68	71	62	73	83	58	61	57	41	56	51	45	40	58	51	0	54	53	59	86	86	86	85	86	86	86	86	86
23	72	69	85	64	86	50	54	43	55	43	39	35	61	68	51	54	0	30	38	86	83	86	53	54	48	33	65	65
24	67	54	78	67	66	57	39	22	35	21	41	17	41	49	36	53	30	0	29	86	86	86	52	55	44	43	56	56
25	72	52	85	63	69	51	31	42	45	41	46	27	46	61	44	59	38	29	0	86	84	86	61	59	65	27	34	34
11	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	0	24	19	44	37	18	47	86	86
12	86	86	86	86	86	86	86	86	86	85	82	86	85	86	83	86	83	86	84	24	0	6	46	56	32	50	86	86
13	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	86	19	6	0	44	53	35	46	86	86
14	86	86	86	86	59	72	62	52	55	52	56	86	62	70	53	85	53	52	61	44	46	44	0	10	23	50	86	86
15	86	86	86	86	60	72	62	55	58	54	54	86	62	74	54	86	54	55	59	37	56	53	10	0	13	41	86	86
16	84	83	86	86	63	73	65	44	57	35	60	86	64	69	57	86	48	44	65	18	32	35	23	13	0	40	86	86
17	73	70	86	64	86	51	51	58	60	57	40	71	61	75	55	86	33	43	27	47	50	46	50	41	40	0	54	54
26	73	49	86	64	70	65	50	68	64	69	73	51	69	66	71	86	65	56	34	86	86	86	86	86	86	54	0	0
27	73	49	86	64	70	65	50	68	64	69	73	51	69	66	71	86	65	56	34	86	86	86	86	86	86	54	0	0

Figure 8. Distance matrix based on sensor data with $\varepsilon = 0.7$ for the Kasteren dataset. Values are in thousands. The background color shows gradient changes in values, with red tones indicating low values and green tones indicating high values.

The same procedure can be repeated using other distance metrics. However, we find that partitions are most clearly distinguishable using H3, as shown when comparing Figure 6a with Figure 7, in which the Levenshtein distance was used. Although we can still recognize partitions, aside from the third partition, they are not as distinguished as in the clustering with the H3 metric. The Rand index between these two clustering results was 0.67, indicating a very loose agreement between them.

In the next set of experiments, we performed clustering based on the distances from sensor data alone (see Equations (1) and (4)). If we obtain comparable clustering results, ADL recognition would not be necessary to investigate residents' daily living.

The resulting distance matrix is presented in Figure 8. These results were obtained by setting the parameter ϵ to 0.70. Similar results were obtained with values from 0.50 to 0.90. Although partitions were now well distinguishable, the clustering result was not in agreement with the clustering based on activity data—the Rand index between them was 0.60. This result indicates that clustering should be done on activity data and not on sensor data. Therefore, we can argue that ADL recognition is necessary.

Although ADL recognition does not give perfect results, an accuracy of 95%, if it is based on time-slots, would give a Hamming distance H1 of 4320 between the recognition results and the reference data. The metrics H2 and H3 would be even lower. Since the typical distances between days and clusters are significantly greater, we can argue that such an ADL recognition accuracy may be sufficient for the purpose of clustering daily activity vectors, although further studies would be necessary to confirm this.

The average Hamming distances H1 and H3 between days are given in Tables 4 and 5. In the first row, we have the average over the distances between all possible pairs of days. The following rows show distances of all possible pairs within a given partition, and the last row gives the average over all four partitions.

Table 4. Average Hamming distances H1 between all days and days in the same partitions.

Dataset	Kasteren	CASAS 11, First Resident	CASAS 11, Second Resident
All days	33,704.20	16,354.60	9232.97
Partition 1	19,629.49	14,408.00	/
Partition 2	15,353.48	8643.72	7252.16
Partition 3	19,380.11	/	7547.24
Partition 4	/	14,219.90	7382.67
Partition average	18,121.03	12,423.87	7394.02

Table 5. Average Hamming distances H3 between all days and days in the same partitions.

Dataset	Kasteren	CASAS 11, First Resident	CASAS 11, Second Resident
All days	25,881.38	14,335.95	8092.51
Partition 1	12,896.39	8640.04	/
Partition 2	14,480.30	9345.59	5611.50
Partition 3	13,203.07	10,680.56	6641.08
Partition 4	16,308.99	9300.99	4235.39
Partition average	14,222.19	9491.80	5495.99

Missing data indicates only one day in the given partition. The data show that distances within partitions are significantly smaller than over all days, confirming that our clustering method indeed generates partitions with similar days according to the H1 and H3 metric for activities.

However, the ratio of the distances is larger for H3 than H1, giving a better differentiation between partitions. When using the H1 metric, we obtain more partitions with a single day. Therefore, we present our research results using the H3 metric.

5.3.4. Graphical Presentation of Daily Activity Vectors

Having partitions, we were interested in activity patterns that were common to daily activity vectors in the same partition. We made a graphical representation of the activity clusters so that we could obtain a more intuitive view of them. Activity patterns are evident from Figure 9, where we compare the daily activity vectors for consecutive days with the daily activity vectors grouped according to partitions.

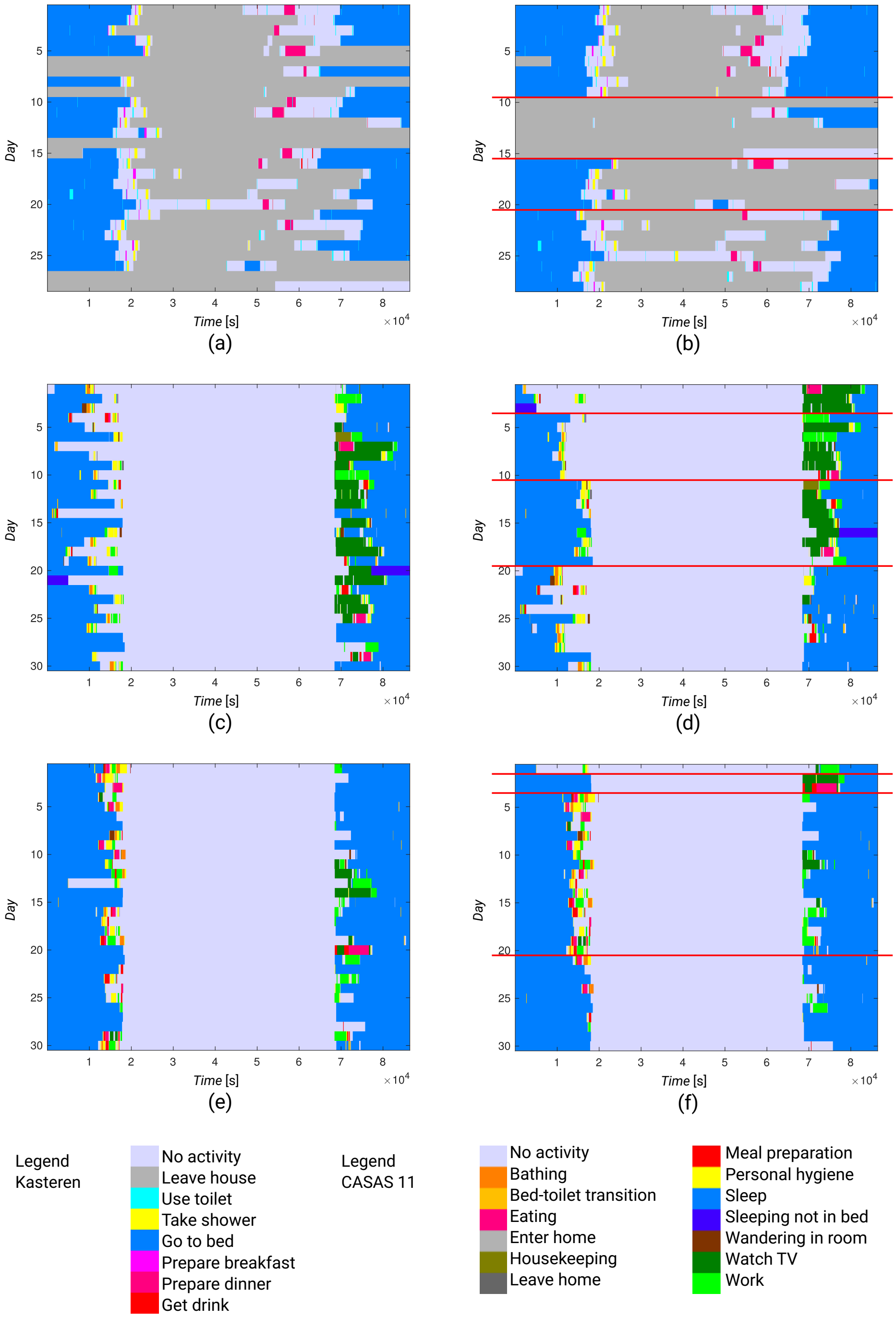

(a)
(b)
(c)
(d)
(e)
(f)

Figure 9. Daily activity representations of the resident in the (**a**) Kasteren dataset, consecutive days; (**b**) Kasteren dataset, partitioned on daily activity vectors; (**c**) CASAS 11 dataset, first resident, consecutive days; (**d**) CASAS 11 dataset, first resident, partitioned on daily activity vectors; (**e**) CASAS 11 dataset, second resident, consecutive days; and (**f**) CASAS 11 dataset, second resident, partitioned on daily activity vectors.

By comparing the daily activity vectors for consecutive days (Figure 9a,c,e), we can see dissimilarities between vectors for consecutive days. This observation is consistent with the high values in Figure 5 and Table 3.

On the contrary, we can examine the graphical presentation for the partitioned daily activity vectors. For example, in the Kasteren dataset (Figure 9b), we can see similarities between vectors within partitions. We see that the second and third partitions contain vectors that are very dissimilar to the vectors in the other two partitions. In the second partition, the early hours do not contain any activity (light blue), which might mean that the resident was not in the apartment at this time.

In the third partition, this same lack of activities is shown in the evening and the night hours. The differences between the first and fourth partitions are smaller. However, in the first partition, we can see more activities in the early evening hours (time between 50,000 and 60,000) and earlier transition to bed (green) than in the fourth partition. These observations are consistent with our earlier interpretation of the distance matrix in Figure 6a.

Similarly, we can examine the graphical presentation for the partitioned daily activity vectors for both residents in the CASAS 11 dataset (see Figure 9d,f). However, we can also see that both residents in this dataset had a more consistent daily routine than the resident in the Kasteren dataset.

In Figure 10, daily activity vectors from the Kasteren dataset are clustered according to sensor data (see the distance matrix in Figure 8). The Figure shows that the daily activity vectors within partitions are far more varied than the results from clustering based on activity data, showing the need for activity recognition.

From Figure 9f, we can easily recognize a single day with unusual behavior in the first partition when compared to the other days. Thus, we may say that such a graphical representation can also be used in practice. It may be useful for a physician or caregiver, as they may be able to look quickly at a resident's normal behavior, their recent behavior, and recognize changes without the need for a lengthy conversation.

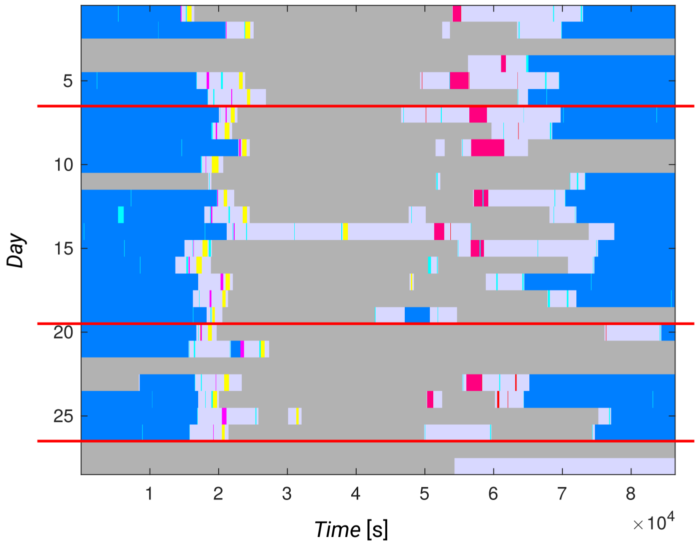


Figure 10. Daily activity representations of the resident in the Kasteren dataset, partitioned according to the clustering results based on sensor data.

6. Conclusions

Elderly people often have a regular daily routine that gives structure and a natural flow to the day. Deviations from this routine may indicate health problems. On the other hand, the everyday life of the younger population often seems unpredictable and disorganized. In this paper, we have shown that, even for younger people, we can identify common patterns regarding how individuals sequence their everyday activities during the day. We define a framework that involves a definition of a generalized Hamming distance to quantify the degree of similarity between each pair of daily activity sequences in the dataset, and then to use all of this pairwise information stored in a distance matrix to identify partitions of similar daily patterns. The obtained partitions identify the normal routines of the residents.

This research was performed on annotated data from the original datasets. In the future, a research interest may be in comparing the clustering results from annotated data and different automatic recognition results. For this purpose, a dataset with publicly available reference recognition results would be useful.

Our future work will focus on a method to detect abnormal signs in a daily activity sequence by comparing it with recognized partitions of normal routines. Having such a method will also provide valuable information in the current outbreak of Coronavirus (COVID-19) disease [43]. Detecting the deterioration of ADL performance of COVID-19 patients after the acute phase of infection will contribute toward identifying the need for rehabilitation. On the other hand, investigating the routines of a person's life across days also provides valuable information for intelligent recommendation systems for healthy living. For example, lack of sleep affects the ability to do everyday activities. Too many snacks can also lead to health issues. Both cases can be identified by the framework described in this paper.

Author Contributions: Conceptualization, M.S.M.; methodology, M.S.M.; software, M.S.M. and G.D.; writing, M.S.M. and G.D.; visualization, G.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Slovenian Research Agency (research core funding No. P2-0069-Advanced Methods of Interaction in Telecommunications).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: http://casas.wsu.edu/datasets/ (accessed on 31 May 2021) and http://casas.wsu.edu/datasets/kasterenDataset.zip (accessed on 31 May 2021).

Acknowledgments: We want to thank the authors of the Kasteren and CASAS datasets for making their data publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Thakur, N.; Han, C.Y. A review of assistive technologies for activities of daily living of elderly. *arXiv* **2021**, arXiv:2106.12183.
2. Camp, N.; Lewis, M.; Hunter, K.; Johnston, J.; Zecca, M.; Di Nuovo, A.; Magistro, D. Technology used to recognize activities of daily living in community-dwelling older adults. *Int. J. Environ. Res. Public Health* **2021**, *18*, 163. [CrossRef] [PubMed]
3. Brdiczka, O.; Crowley, J.L.; Reignier, P. Learning situation models in a smart home. *IEEE Trans. Syst. Man, Cybern. Part B (Cybernetics)* **2008**, *39*, 56–63. [CrossRef] [PubMed]
4. Ariza Colpas, P.; Vicario, E.; De-La-Hoz-Franco, E.; Pineres-Melo, M.; Oviedo-Carrascal, A.; Patara, F. Unsupervised human activity recognition using the clustering approach: A review. *Sensors* **2020**, *20*, 2702. [CrossRef] [PubMed]
5. Debes, C.; Merentitis, A.; Sukhanov, S.; Niessen, M.; Frangiadakis, N.; Bauer, A. Monitoring activities of daily living in smart homes: Understanding human behavior. *IEEE Signal Process. Mag.* **2016**, *33*, 81–94. [CrossRef]
6. Xu, C.; Chai, D.; He, J.; Zhang, X.; Duan, S. InnoHAR: A deep neural network for complex human activity recognition. *IEEE Access* **2019**, *7*, 9893–9902. [CrossRef]
7. Rashidi, P.; Cook, D.J.; Holder, L.B.; Schmitter-Edgecombe, M. Discovering activities to recognize and track in a smart environment. *IEEE Trans. Knowl. Data Eng.* **2010**, *23*, 527–539. [CrossRef] [PubMed]
8. Donaj, G.; Maučec, M.S. Extension of HMM-Based ADL Recognition With Markov Chains of Activities and Activity Transition Cost. *IEEE Access* **2019**, *7*, 130650–130662. [CrossRef]
9. Rahman, S.; Irfan, M.; Raza, M.; Moyeezullah Ghori, K.; Yaqoob, S.; Awais, M. Performance analysis of boosting classifiers in recognizing activities of daily living. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1082. [CrossRef]
10. Dahmen, J.; Cook, D.J. Indirectly Supervised Anomaly Detection of Clinically Meaningful Health Events from Smart Home Data. *ACM Trans. Intell. Syst. Technol. (TIST)* **2021**, *12*, 1–18. [CrossRef]
11. Zekri, D.; Delot, T.; Thilliez, M.; Lecomte, S.; Desertot, M. A Framework for Detecting and Analyzing Behavior Changes of Elderly People over Time Using Learning Techniques. *Sensors* **2020**, *20*, 7112. [CrossRef] [PubMed]
12. Hussain, Z.; Sheng, M.; Zhang, W.E. Different approaches for human activity recognition: A survey. *arXiv* **2019**, arXiv:1906.05074.

13. Wu, J.; Feng, Y.; Sun, P. Sensor fusion for recognition of activities of daily living. *Sensors* **2018**, *18*, 4029. [CrossRef]
14. Ferreira, J.M.; Pires, I.M.; Marques, G.; Garcia, N.M.; Zdravevski, E.; Lameski, P.; Flórez-Revuelta, F.; Spinsante, S.; Xu, L. Activities of daily living and environment recognition using mobile devices: a comparative study. *Electronics* **2020**, *9*, 180. [CrossRef]
15. van Kasteren, T.; Krose, B. Bayesian activity recognition in residence for elders. In Proceedings of the 2007 3rd IET International Conference on Intelligent Environments, Ulm, Germany, 24–25 September 2007; pp. 209–213.
16. Wei, H.; He, J.; Tan, J. Layered hidden Markov models for real-time daily activity monitoring using body sensor networks. *Knowl. Inf. Syst.* **2011**, *29*, 479–494. [CrossRef]
17. Ordóñez, F.J.; Englebienne, G.; De Toledo, P.; Van Kasteren, T.; Sanchis, A.; Kröse, B. In-home activity recognition: Bayesian inference for hidden Markov models. *IEEE Pervasive Comput.* **2014**, *13*, 67–75. [CrossRef]
18. Liu, L.; Peng, Y.; Wang, S.; Liu, M.; Huang, Z. Complex activity recognition using time series pattern dictionary learned from ubiquitous sensors. *Inf. Sci.* **2016**, *340*, 41–57. [CrossRef]
19. Steven Eyobu, O.; Han, D.S. Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network. *Sensors* **2018**, *18*, 2892. [CrossRef] [PubMed]
20. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognit. Lett.* **2019**, *119*, 3–11. [CrossRef]
21. De-La-Hoz-Franco, E.; Ariza-Colpas, P.; Quero, J.M.; Espinilla, M. Sensor-based datasets for human activity recognition—A systematic review of literature. *IEEE Access* **2018**, *6*, 59192–59210. [CrossRef]
22. Noor, M.H.M.; Salcic, Z.; Kevin, I.; Wang, K. Enhancing ontological reasoning with uncertainty handling for activity recognition. *Knowl.-Based Syst.* **2016**, *114*, 47–60. [CrossRef]
23. Azkune, G.; Almeida, A. A scalable hybrid activity recognition approach for intelligent environments. *IEEE Access* **2018**, *6*, 41745–41759. [CrossRef]
24. Al Machot, F.; R Elkobaisi, M.; Kyamakya, K. Zero-Shot Human Activity Recognition Using Non-Visual Sensors. *Sensors* **2020**, *20*, 825. [CrossRef] [PubMed]
25. Mantaci, S.; Restivo, A.; Sciortino, M. Distance measures for biological sequences: Some recent approaches. *Int. J. Approx. Reason.* **2008**, *47*, 109–124. [CrossRef]
26. Abbott, A.; Tsay, A. Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociol. Methods Res.* **2000**, *29*, 3–33. [CrossRef]
27. Vagni, G.; Cornwell, B. Patterns of everyday activities across social contexts. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 6183–6188. [CrossRef] [PubMed]
28. Studer, M.; Ritschard, G. What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *J. R. Stat. Soc. Ser. (Stat. Soc.)* **2016**, *179*, 481–511. [CrossRef]
29. Lupiani, E.; Sauer, C.; Roth-Berghofer, T. Implementation of similarity measures for event sequences in myCBR. In Proceedings of the 18th UK Workshop on Case-Based Reasoning, Cambridge, UK, 10 December 2012.
30. Luu, V.T.; Forestier, G.; Weber, J.; Bourgeois, P.; Djelil, F.; Muller, P.A. A review of alignment based similarity measures for web usage mining. *Artif. Intell. Rev.* **2020**, *53*, 1529–1551. [CrossRef]
31. Liu, Y.; Ouyang, D.; Liu, Y.; Chen, R. A novel approach based on time cluster for activity recognition of daily living in smart homes. *Symmetry* **2017**, *9*, 212. [CrossRef]
32. Gao, S.; Tan, A.H.; Setchi, R. Learning ADL daily routines with spatiotemporal neural networks. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 143–153. [CrossRef]
33. Sánchez, V.G.; Lysaker, O.M.; Skeie, N.O. Human behaviour modelling for welfare technology using hidden Markov models. *Pattern Recognit. Lett.* **2020**, *137*, 71–79. [CrossRef]
34. Lago, P.; Jiménez-Guarín, C.; Roncancio, C. Contextualized behavior patterns for change reasoning in Ambient Assisted Living: A formal model. *Expert Syst.* **2017**, *34*, e12163. [CrossRef]
35. Yahaya, S.W.; Lotfi, A.; Mahmud, M. Detecting anomaly and its sources in activities of daily living. *SN Comput. Sci.* **2021**, *2*, 1–18. [CrossRef]
36. Yahaya, S.W.; Lotfi, A.; Mahmud, M. Towards a data-driven adaptive anomaly detection system for human activity. *Pattern Recognit. Lett.* **2021**, *145*, 200–207. [CrossRef]
37. Wang, Y.; Wu, K.; Ni, L.M. Wifall: Device-free fall detection by wireless networks. *IEEE Trans. Mob. Comput.* **2016**, *16*, 581–594. [CrossRef]
38. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 344.
39. Warrens, M.J.; van der Hoef, H. Understanding the Rand Index. In *Advanced Studies in Classification and Data Science*; Imaizumi, T., Okada, A., Miyamoto, S., Sakaori, F., Yamamoto, Y., Vichi, M., Eds.; Springer: Singapore, 2020; pp. 301–313.
40. Van Kasteren, T.; Noulas, A.; Englebienne, G.; Kröse, B. Accurate activity recognition in a home setting. In Proceedings of the 10th International Conference on Ubiquitous Computing, Seoul, Korea, 21–24 September 2008; pp. 1–9.
41. Cook, D.J.; Schmitter-Edgecombe, M. Assessing the quality of activities in a smart environment. *Methods Inf. Med.* **2009**, *48*, 480–485. [PubMed]

42. Fahad, L.G.; Tahir, S.F.; Rajarajan, M. Feature selection and data balancing for activity recognition in smart homes. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 512–517.
43. Pizarro-Pennarolli, C.; Sánchez-Rojas, C.; Torres-Castro, R.; Vera-Uribe, R.; Sanchez-Ramirez, D.C.; Vasconcello-Castillo, L.; Solís-Navarro, L.; Rivera-Lillo, G. Assessment of activities of daily living in patients post COVID-19: A systematic review. *PeerJ* **2021**, *9*, e11026. [CrossRef] [PubMed]