A Sensor-Independent Multimodal Fusion Scheme for Human Activity Recognition

Anastasios Alexiadis[1], Alexandros Nizamis[1], Dimitrios Giakoumis[1], Konstantinos Votis[1], and Dimitrios Tzovaras[1]

Centre for Research and Technology Hellas
Information Technologies Institute (CERTH/ITI), 6km Charilaou-Thermi, Thessaloniki, Greece
talex@iti.gr, alnizami@iti.gr, dgiakoum@iti.gr, kvotis@iti.gr, Dimitrios.Tzovaras@iti.gr

Abstract. Human Activity Recognition is a field that provides the fundamentals for Ambient Intelligence and Assisted Living Applications. Multimodal methods for Human Activity Recognition utilize different sensors and *fuse* them together to provide higher-accuracy results. These methods require data for all sensors employed to operate with. In this work we present a sensor-independent, in regards to the number of sensors used, scheme for designing multimodal methods that operate when sensor-data are missing. Furthermore, we present a data augmentation method that increases the fusion model's accuracy (up to 11% increases) when operating with missing sensor-data. The proposed method's effectiveness is evaluated on the ExtraSensory dataset, which contains over 300,000 samples from 60 users, collected from smartphones and smartwatches. In addition, the methods are evaluated for different number of sensors used at the same time. However, the max number of sensors must be known beforehand.

Keywords: Automatic Human Activity Recognition · multimodal fusion · sensor fusion · sensor-independent fusion.

1 Introduction

Automatic Human Activity Recognition (HAR) is a field that constitutes the fundamentals of Ambient Intelligence (Aml) and Assisted Living Applications (AAL). It comprises the challenges of recognizing and understanding human activities and their context which are the basic pre-requisites for integrating human-aware machine decision capabilities. Human Activity Recognition can be performed using static sensors (e.g., mounted video camera [22]) or wearable sensors (e.g., smart watch or other wearable sensors [15]) or by combining both.

There is a plethora of methods to perform human activity recognition [17]. The main categories are (i) unimodal and (ii) multimodal human activity recognition methods, according to the type of sensor they employ. Unimodal methods utilize data from a single modality, such as audio signal. These methods can be

categorized into (i) space-time, (ii) stochastic, (iii) rule-based and (iv) shape-based methods. On the other hand Multimodal methods combine features from different sources (such as combining features from audio sensors with features from video sensors) [19] and can be categorized to: (i) affective, (ii) behavioral and (iii) social networking methods.

In this work we present a sensor-independent fusion method in respect to the number of sensors utilized. In addition to this, we introduce a data augmentation method that augments the collected data with sub-sets of utilized sensor data per observation, and apply these methods to fuse unimodal models for the ExtranSensory dataset [16]. This dataset served as a test-case for our proposed methods. Our methods can be used with any model that fuses a number of unimodal models for a set of sensors. The ExtraSensory dataset contains over 300,000 examples from 60 users of diverse ethnic backgrounds, collected from smartphones and smartwatches. It includes heterogeneous measurements from a variety of wearable sensors (i.e., accelerometer, gyroscope, magnometer, watch compass, audio etc.). Not all the sensors were available at all times, some phones did not have some sensors, whereas in other cases sensors were sometimes unavailable.

The remainder of the paper is structured as follows: A brief review of the related work is presented in the next section. Afterwards we present our methods for sensor-independent (in respect to the number of sensors) fusion and for data augmentation. We evaluate our methods on a set of experiments and discuss the results. Finally, we conclude the paper and present our directions for future work.

2 Related Work

In the recent years due to the rise of IoT devices and smart living environments, a lot of research has been conducted related to human activity recognition and many methods have been developed and applied. In introduction section is mentioned that the methods are distinguished in five major categories, space-time, stochastic, rule-based, shape-based and multimodal methods. In this chapter, we are presenting a brief literature review related to multimodal methods for human activity recognition in order to place our work in the current state-of-the-art and to indicate innovation and contribution of our work in the field.

As an event or action can be described by different types of data and features that provide more and useful information, several multimodal methods for human activity recognition are based on fusion techniques [4]. In [12] the multimodal fusion for human activity detection is further classified in data and feature fusion methods. In [19] a decision fusion method is also considered. The latter is not a direct fusion scheme as it firstly applies separate classifiers to obtain probability scores and combines them for final decision making [7, 21]. Since the current work is considered as a more direct type of fusion, the relevant works in this section are presented based on data and feature methods classification. Data fusion aims to increase accuracy, robustness and reliability of a system devoted

to human activity recognition as this type of fusion involves integration of data collected by multiple mobile and wearable sensor devices. The work of [20] introduces a deep learning framework named DeepSense, targeting to overcome noise and feature customization challenges in order to increase recognition accuracy. The framework exploits interactions among different sensory modalities by integrating convolutional and recurrent neural networks to sensors. To the same aim [13] proposes a deep learning framework for activity recognition based on convolutional and LSTM recurrent units. The framework is suitable for multimodal wearable sensors as it can perform sensor fusion naturally without requiring expert knowledge in designing features. In [14] the authors introduce an activity recognition system to be used for elderly people monitoring. The system collects and combines data from sensors such as state sensors at doors, movement sensors in rooms and sensors attached to appliances or furniture pieces. However, the activity recognition is enabled just by using some multiple Hidden Markov Models (HMMs). A combination of HMMs and neural networks for multi-sensor fusion for human daily activity recognition has been introduced in [23]. The solution was based on wearable sensors and tested in a robot-assisted living system for elderly people environment. Other similar approaches such as [10] are based on information fusion of features extracted from experimental data collected by different sensors, such as a depth camera, an accelerometer and a micro-Doppler radar. The authors create combinations of the aforementioned sensors data for classification of the activity. They found that the addition of more sensors was continuously improving the accuracy of classification. In particular, the authors have measured the accuracy of quadratic-kernel SVM classifier and of an Ensemble classifier.

In order to further increase the accuracy and to improve performance of activity recognition systems some fusion methods at the features level have been developed in previous years. Feature fusion techniques enable the combination of features extracted from sensor data with machine learning algorithms. This type of fusion is used in the current work as well. Regarding the sensors used for activity recognition they are lay on various categories such as (a) 3D sensors [9, 1, 18] for recognising activities such as walking, running or sitting, (b) thermal cameras [3, 11] for household activity recognition and (c) event cameras [8] for event-based activity detection or even activity tracking applications [2]. In [6] a multimodal feature-level fusion approach for robust human action recognition was proposed. The approach utilizes data from multiple sensors such as depth sensor, RGB camera, and wearable inertial sensors. The recognition framework was tested on a publicly available dataset including over 25 different human actions. For training and testing the proposed fusion model, SVM classifiers and K-nearest neighbor were used. The authors observed that better results were produced by using more sensors in the fusion. However, the achieved accuracy improvement had some significant loses (over 10%) in terms of performance comparing to fusion approaches with less sensors combinations. In another approach [19] a human action recognition with multimodal feature selection and fusion based on videos was introduced. The authors extracted both audio and visual

features from a public dataset/video and used them as input for a set of SVM classifiers. The outputs were fused to obtain a final classification score through fuzzy integral and two-layer SVM. The authors observed that audio context is more useful than visual one but the audio is not always helpful for some actions due to its high diversity. Recently, in [5] the authors proposed an intelligent sensor fusion strategy for activity recognition in body sensor networks that very often have uncertain or even incomplete data. Their approach was based on the Dezert-Smarandache theory. In this, as training dataset they employed kernel density estimation (KDE)-based models for sensors readings and they selected the best discriminative model of them. A testing dataset was also used in order to calculate basic belief assignments based on KDE models for each activity. Finally, the calculated belief assignments were combined with redistribution rules for final decision-making. The authors concluded that their approach outperformed state-of-the-art methods in accuracy, as it was tested and compared in two public datasets.

In this work, a novel feature fusion method that provides high accuracy and robustness in the human activity recognition, in comparison to aforementioned data fusion techniques, is introduced. Furthermore, the introduced approach provides a fusion method that is sensor-independent, in terms of sensors' number. Opposite to other fusion approaches that were introduced in the previous paragraph, our approach does not require to recreate the fusion model in the case that less sensors are available. Another improvement that the method introduces in comparison with the above-mentioned fusion methods, is the data augmentation technique that was applied. This technique augments the collected data with sub-sets of utilized sensor data per observation. By adding combinations of existing sensor data the amount of the available data is increased, so the method's accuracy with regards to human activity recognition is increased as well.

3 Methods

Seven unimodal classifiers were given, which are considered as black boxes for the scope of this paper, where each one classifies human daily activities according to a specific sensor from the ExtraSensory dataseset. Each classifier applies to one of the following sensors:

- Watch Accelerometer (WA)
- Watch Compass (WC)
- Phone Accelerometer (PA)
- Phone Gyroscope (PG)
- Phone Magnet (PM)
- Phone State (PS)
- Audio (A)

In Table 1 the F1 scores of the seven classifiers in regards to their respective test sets (which are sub-sets of the test set containing only the observations that

Table 1. Unimodal classifiers F1-scores

classes	WA F1	WC F1	PA F1	PG F1	PM F1	PS F1	A F1
SITTING-TOILET	0.23	0.0	0.0	0.0	0.0	0.08	0.13
SITTING-EATING	0.7	0.38	0.064	0.0	0.0	0.26	0.42
STANDING-COOKING	0.09	0.0	0.0	0.13	0.13	0.0	0.33
SITTING-WATCHING TV	0.05	0.0	0.67	0.57	0.57	0.72	0.83
LYING DOWN-WATCHING TV	0.63	0.0	0.28	0.18	0.18	0.37	0.58
STANDING-EATING	0.24	0.49	0.47	0.22	0.22	0.62	0.69
STANDING-CLEANING	0.63	0.0	0.4	0.19	0.19	0.69	0.77
WALKING-EATING	0.43	0.0	0.27	0.38	0.38	0.43	0.6
STANDING-WATCHING TV	0.33	0.46	0.3	0.27	0.27	0.54	0.59
STANDING-TOILET	0.29	0.31	0.59	0.5	0.5	0.7	0.79
WALKING-WATCHING TV	0.49	0.0	0.1	0.08	0.08	-	0.63
WALKING-COOKING	0.42	0.0	0.39	0.11	0.11	0.37	0.51
SITTING-COOKING	0.59	0.0	0.35	0.38	0.38	0.67	0.68
WALKING-CLEANING	0.05	0.0	0.19	0.1	0.1	0.11	0.48
LYING DOWN-EATING	0.16	0.0	0.62	0.58	0.58	0.24	0.73
SITTING-CLEANING	0.15	0.0	0.0	0.0	0.0	0.16	0.27
accuracy	0.55	0.4	0.5	0.45	0.45	0.65	0.74
macro avg	0.34	0.1	0.29	0.23	0.23	0.4	0.57
weighted avg	0.55	0.34	0.52	0.43	0.43	0.64	0.75

contain data for their respective sensors) are presented. A dash "-" denotes that there were no observations of the respective sensor for that class. There are 16 daily activity classes. Each of the seven classifiers contains a 32-node dense layer followed by a softmax layer at their end.

3.1 Sensor Independent Fusion Model

For the fusion model a feed-forward Artificial Neural Network was used. The softmax layer for each unimodal model was discarded so when simulating each of the models we obtain a feature vector of length 32 (the output of the penultimate layer of the unimodal models). These feature vectors provide the inputs to their respective unimodal models' softmax layers, and thus contain the activity information extracted from the data before being converted to a probability distribution by the softmax layer. Each unimodal model is simulated using its design features from observations of its respective sensor.

The input layer is defined with a length of 32 multiplied by the number of sensors. To provide sensor independence in respect to the number of sensors used we add a binary vector to the input layer of the network with length equal to the number of sensors. So for our specific case the input layer has a size of $32 \cdot 7 + 7 = 231$.

$F_{s1_1}F_{s1_2}...F_{s1_{32}}...F_{s7_1}F_{s7_2}...F_{s7_{32}}...F_{acts7}$,$F_{acts6}$ ... F_{acts1}

The diagram above illustrates the input layer of the fusion model. F_{s1_1} denotes the first feature of the first sensor, $F_{s1_{32}}$ denotes the last feature of the first

Algorithm 1 Create Fusion model training set

procedure CREATE_DATASET($train, feature\_sensors, unimodal\_models, sensors$)
$T \leftarrow [0]_{len(train)\times 32\cdot len(sensors)+len(sensors)}$ {Matrix of zeros}
for each sensor k ranging from 0 to $len(sensors) - 1$ **do**
  $train\_s \leftarrow train[feature\_sensors[sensors[k]]]$
  $train\_sn \leftarrow train\_s.dropna()$
  $idxs \leftarrow train\_sn.index$
  $feature\_matrix \leftarrow simulate\_model(unimodal\_models[sensors[k]], train\_sn)$
  for each observation i ranging from 0 to $len(T) - 1$ **do**
    if $i \in idxs$ **then**
      $T[i, T.no\_of\_cols–k–1] \leftarrow 1$
    end if
  end for
  for each observation i ranging from 0 to $len(feature\_matrix) - 1$ **do**
    $T[idxs[i], k \cdot 32 : (k + 1) \cdot 32] \leftarrow feature\_matrix[i]$
  end for
end for
return T

sensor, whereas F_{acts1} denotes the binary feature that actives/deactivates sensor 1 input for the fusion model. When F_{acts1} is set to 0, the features corresponding to sensor 1 input, that is the features from F_{s1_1} to $F_{s1_{32}}$ are set to 0 too. When F_{acts1} is set to 1, the input features corresponding to sensor 1 on the input of the fusion model are set to the 32 values of the feature vector which is the output of the respective unimodal model for sensor 1. In a similar manner we set the other sensor inputs. The fusion model performs feature-level fusion.

Algorithm 1 computes the dataset for training the fusion model. It is given the training set ($train$) containing the features for all sensors, a dictionary ($feature\_sensors$) with mappings of the form $sensor\_name \rightarrow [feature\_indeces]$, for each sensor providing the indeces for each sensor's features in the training set, a dictionary of the unimodal models ($unimodal\_models$) of the form $sensor\_name \rightarrow model$ and a list of the sensor names ($sensors$). For each observation in the training set, the feature vectors of the unimodal models are computed, in the cases when there are data available for their respective sensors, and the activating feature for these sensors is set to 1. When data is missing for a sensor, the respective features for that sensor are set to 0, as well as the activating feature.

3.2 Data augmentation method

A data augmentation method was designed and implemented based on the following premise: The dataset can be expanded by adding more observations with all possible subsets of activated sensors from the sensors containing data, in each observation of the original dataset. As an example, consider the case of adding more data based on a single observation, where only five of the seven

Algorithm 2 Data augmentation method

procedure AUGMENT_DATA($data, Y, no\_of\_sensors, C$)
$aug\_data \leftarrow ()$ {Empty Sequence}
$aug\_Y \leftarrow ()$ {Empty Sequence}
$sensor\_ids \leftarrow \{z : \exists n \in \mathbb{Z}$ such that $z = 0 + 1 \times n$, and $z \in [0, no\_of\_sensors - 1)\}$
for each observation i ranging from 0 to $len(data) - 1$ **do**
  $sources\_used \leftarrow \sum_{x=data.no\_of\_cols-no\_of\_sensors}^{data.no\_of\_cols} data[i, x]$
  for each sensor k ranging from $no\_of\_sensors - 1$ to 0 in steps of -1 **do**
    if $k < sources\_used$ **then**
      if $k < C$ **then**
        break
      end if
      $v \leftarrow \binom{sensor\_ids}{k}$ {k-length tuples with no repetition}
      for l ranging from 0 to $len(v) - 1$ **do**
        $OBV \leftarrow [0]_{data.no\_of\_cols}$ {Vector of zeros}
        for m ranging from 0 to $len(v[l]) - 1$ **do**
          $OBV[v[l][m] \cdot 32 : (v[l][m]+1) \cdot 32] \leftarrow data[i, v[l][m] \cdot 32 : (v[l][m]+1) \cdot 32]$
          $OBV[data.no\_of\_cols - v[l][m] - 1] \leftarrow 1$
        end for
        $aug\_data \leftarrow aug\_data^\frown(OBV)$ {Sequence Concatenation}
        $aug\_Y \leftarrow aug\_Y^\frown(Y[i])$
      end for
    else
      $aug\_data \leftarrow aug\_data^\frown(data[i, :])$
      $aug\_Y \leftarrow aug\_Y^\frown(Y[i])$
    end if
  end for
end for
return $as\_matrix(aug\_data), as\_vector(aug\_Y)$

sensors are utilized. The following combinations are available with only the five sensors activated:

```
[('s1', 's2', 's3', 's4', 's5'),
 ('s1', 's2', 's3', 's4', 's6'),
 ...,
 ('s2', 's4', 's5', 's6', 's7'),
 ('s3', 's4', 's5', 's6', 's7')]
```

For each of these cases we know the target activity, it holds the same label as the original observation which contains data for all sensors, so we can augment the dataset with a new observation per case, where data is provided for the respective sensors above and the features of the missing sensor data, as well as their respective activation features, are set to 0 in the fusion model training set.

Table 2. Fusion Model with no data augmentation in training/validation sets, no data augmentation in test set. Xs F1 denotes F1 scores for X sensors, Xs S denotes support

classes	2s F1	2s S	3s F1	3s S	4s F1	4s S	5s F1	5s S	6s F1	6s S	7s F1	7s S	As F1	As S
SITTING-TOILET	0.50	2	0.50	2	0.84	22	0.82	145	0.69	64	0.67	18	0.38	17
SITTING-EATING	0.90	14	0.67	4	0.70	18	0.89	69	0.52	9	0.94	154	0.55	30
STANDING-COOKING	0.67	2	1.00	2	0.51	17	0.00	1	0.71	104	0.71	12	0.55	14
SITTING-WATCHING TV	-	-	0.80	4	0.83	47	0.87	611	0.50	2	0.43	7	0.87	2294
LYING DOWN-WATCHING TV	0.92	6	0.89	4	0.44	4	0.71	119	0.85	137	0.59	59	0.70	228
STANDING-EATING	0.97	18	0.95	30	0.50	8	0.85	328	0.65	51	0.70	90	0.78	479
STANDING-CLEANING	-	-	-	-	0.83	5	0.25	3	0.53	9	0.88	942	0.89	401
WALKING-EATING	-	-	-	-	0.67	2	0.59	15	0.86	588	0.65	27	0.73	523
STANDING-WATCHING TV	-	-	-	-	0.00	1	0.77	22	0.76	76	0.89	830	0.68	227
STANDING-TOILET	0.00	1	1.00	2	0.88	109	0.72	61	0.83	340	0.72	234	0.86	1522
WALKING-WATCHING TV	-	-	0.67	3	0.38	9	0.75	104	0.73	117	-	-	0.70	209
WALKING-COOKING	-	-	-	-	0.00	1	0.84	26	0.50	3	0.78	208	0.72	80
SITTING-COOKING	-	-	-	-	-	-	0.50	1	0.81	36	0.73	60	0.74	59
WALKING-CLEANING	-	-	-	-	0.67	3	0.67	13	0.76	29	0.87	44	0.52	57
LYING DOWN-EATING	-	-	-	-	-	-	0.64	59	0.67	5	0.60	23	0.79	96
SITTING-CLEANING	-	-	-	-	-	-	-	-	0.51	35	0.46	8	0.58	9
accuracy	0.88	43	0.88	51	0.77	246	0.82	1577	0.80	1605	0.84	2716	0.82	6245
macro avg	0.66	43	0.81	51	0.56	246	0.66	1577	0.68	1605	0.71	2716	0.69	6245
weighted avg	0.88	43	0.88	51	0.78	246	0.82	1577	0.80	1605	0.84	2716	0.82	6245

The proposed method starts with k equals the activated sensors of each observation and loops, decreasing the number of used sensors by 1 in each iteration and computes the combinations, that is the k-length tuples with no repetition for each value of k until $k < C$ where C is a constant defining the minimum number of sensors that can be utilized. No interpolation or estimation techniques are performed to augment the dataset, the labels for the generated data are already known as well as the sensor data used for the new observations are the feature vectors computed by the unimodal models for the measured sensor data.

Algorithm 2 augments the dataset based on the above method. It is given the dataset for the fusion model (*data*), which is computed using Algorithm 1, the vector of labels containing the class for each observation (Y), the number of sensors ($no\_of\_sensors$) and the constant C, defining the minimum number of sensors utilized for the augmented dataset. For each observation in the training set, all $k-tuple$ combinations without repetition are computed, ranging from k equal to the number of activated sensors down to C.

4 Evaluation

To evaluate the proposed methods a set of experiments was devised to investigate the fusion model's improvement of F1-score per class, for the whole test set, as well as specific sub-sets of the test-set split according to the number of sensors used per observation. The dataset was split 70%–30% into a train and a test set. The resulting train and test sets were the ones used to train and test the unimodal models. The training set was farther split 80%–20% during training for the final train and validation sets. The fusion model was trained using the Adam optimizer with $lr = 0.001$, using a $batch\_size = 64$ for 200 epochs using early-stopping with $patience = 50$ for validation accuracy, while

Table 3. Fusion Model with no data augmentation in training/validation sets, with data augmentation in test set. Xs F1 denotes F1 scores for X sensors, Xs S denotes support

classes	2s F1	2s S	3s F1	3s S	4s F1	4s S	5s F1	5s S	6s F1	6s S	7s F1	7s S	As F1	As S
SITTING-TOILET	0.08	357	0.10	595	0.14	595	0.22	337	0.54	135	0.63	18	0.14	1949
SITTING-EATING	0.19	630	0.25	1050	0.34	1050	0.81	6180	0.90	1215	0.92	154	0.32	3453
STANDING-COOKING	0.24	294	0.32	490	0.40	490	0.56	294	0.68	86	0.80	12	0.40	1666
SITTING-WATCHING TV	0.65	47894	0.73	78780	0.79	75044	0.82	32741	0.23	58	0.32	7	0.75	242583
LYING DOWN-WATCHING TV	0.36	4748	0.46	7808	0.54	7533	0.59	3295	0.85	7182	0.60	59	0.49	24168
STANDING-EATING	0.50	10059	0.59	16765	0.66	16017	0.68	6697	0.73	1560	0.71	90	0.62	51306
STANDING-CLEANING	0.46	8040	0.61	13234	0.73	12618	0.47	570	0.67	694	0.88	942	0.64	41442
WALKING-EATING	0.41	10943	0.50	18099	0.56	16497	0.62	7490	0.87	6150	0.68	27	0.52	55018
STANDING-WATCHING TV	0.40	4626	0.49	7632	0.57	7494	0.55	2414	0.67	1755	0.90	830	0.51	22690
STANDING-TOILET	0.57	31899	0.67	53029	0.75	52447	0.81	24898	0.58	464	0.71	234	0.70	169256
WALKING-WATCHING TV	0.38	4327	0.48	7142	0.56	6834	0.64	2915	0.68	496	-	-	0.52	21776
WALKING-COOKING	0.19	1680	0.28	2800	0.40	2732	0.51	1198	0.57	218	0.76	208	0.32	8655
SITTING-COOKING	0.41	1239	0.51	2065	0.59	1895	0.59	614	0.83	344	0.71	60	0.53	6002
WALKING-CLEANING	0.21	1197	0.28	1995	0.35	1961	0.40	916	0.63	166	0.84	44	0.31	6168
LYING DOWN-EATING	0.50	2016	0.59	3360	0.68	3326	0.78	1695	0.37	91	0.62	23	0.64	10785
SITTING-CLEANING	0.06	189	0.10	247	0.14	143	0.18	64	0.11	3	0.43	8	0.10	646
accuracy	0.53	130138	0.63	215091	0.70	206676	0.76	92318	0.81	20617	0.84	2716	0.65	667563
macro avg	0.35	130138	0.43	215091	0.51	206676	0.58	92318	0.62	20617	0.70	2716	0.47	667563
weighted avg	0.54	130138	0.63	215091	0.70	206676	0.76	92318	0.81	20617	0.84	2716	0.66	667563

Table 4. Fusion Model with data augmentation in training/validation sets, with data augmentation in test set. Xs F1 denotes F1 scores for X sensors, Xs S denotes support

classes	2s F1	2s S	3s F1	3s S	4s F1	4s S	5s F1	5s S	6s F1	6s S	7s F1	7s S	As F1	As S
SITTING-TOILET	0.13	357	0.19	595	0.23	595	0.31	337	0.60	135	0.63	18	0.21	1949
SITTING-EATING	0.31	630	0.39	1050	0.46	1050	0.86	6180	0.91	1215	0.93	154	0.43	3453
STANDING-COOKING	0.44	294	0.53	490	0.61	490	0.71	294	0.80	86	0.86	12	0.57	1666
SITTING-WATCHING TV	0.72	47894	0.78	78780	0.82	75044	0.85	32741	0.29	58	0.44	7	0.79	242583
LYING DOWN-WATCHING TV	0.45	4748	0.54	7808	0.60	7533	0.64	3295	0.87	7182	0.60	59	0.56	24168
STANDING-EATING	0.59	10059	0.66	16765	0.69	16017	0.71	6697	0.74	1560	0.71	90	0.66	51306
STANDING-CLEANING	0.65	8040	0.75	13234	0.81	12618	0.56	570	0.68	694	0.89	942	0.77	41442
WALKING-EATING	0.48	10943	0.57	18099	0.62	16497	0.66	7490	0.89	6150	0.59	27	0.58	55018
STANDING-WATCHING TV	0.45	4626	0.54	7632	0.61	7494	0.57	2414	0.69	1755	0.91	830	0.55	22690
STANDING-TOILET	0.69	31899	0.76	53029	0.81	52447	0.85	24898	0.58	464	0.72	234	0.78	169256
WALKING-WATCHING TV	0.49	4327	0.57	7142	0.63	6834	0.71	2915	0.73	496	-	-	0.60	21776
WALKING-COOKING	0.38	1680	0.49	2800	0.57	2732	0.59	1198	0.58	218	0.76	208	0.51	8655
SITTING-COOKING	0.60	1239	0.67	2065	0.72	1895	0.68	614	0.86	344	0.76	60	0.67	6002
WALKING-CLEANING	0.24	1197	0.29	1995	0.32	1961	0.32	916	0.72	166	0.91	44	0.29	6168
LYING DOWN-EATING	0.66	2016	0.72	3360	0.76	3326	0.82	1695	0.31	91	0.76	23	0.74	10785
SITTING-CLEANING	0.35	189	0.41	247	0.39	143	0.35	64	0.21	3	0.33	8	0.38	646
accuracy	0.64	130138	0.71	215091	0.76	206676	0.79	92318	0.83	20617	0.85	2716	0.72	667563
macro avg	0.48	130138	0.55	215091	0.60	206676	0.64	92318	0.65	20617	0.72	2716	0.57	667563
weighted avg	0.63	130138	0.71	215091	0.75	206676	0.79	92318	0.83	20617	0.85	2716	0.72	667563

Table 5. Fusion Model with data augmentation in training/validation sets, no data augmentation in test set. Xs F1 denotes F1 scores for X sensors, Xs S denotes support

classes	2s F1	2s S	3s F1	3s S	4s F1	4s S	5s F1	5s S	6s F1	6s S	7s F1	7s S	As F1	As S
SITTING-TOILET	0.80	2	0.80	2	0.86	22	0.80	145	0.63	64	0.60	18	0.31	17
SITTING-EATING	0.86	14	0.67	4	0.73	18	0.83	69	0.60	9	0.93	154	0.52	30
STANDING-COOKING	0.80	2	1.00	2	0.55	17	0.00	1	0.68	104	0.73	12	0.57	14
SITTING-WATCHING TV	-	-	0.89	4	0.76	47	0.87	611	0.44	2	0.50	7	0.88	2294
LYING DOWN-WATCHING TV	1.00	6	0.75	4	0.43	4	0.67	119	0.85	137	0.63	59	0.66	228
STANDING-EATING	0.97	18	0.95	30	0.43	8	0.84	328	0.64	51	0.68	90	0.76	479
STANDING-CLEANING	-	-	-	-	0.67	5	0.00	3	0.33	9	0.89	942	0.88	401
WALKING-EATING	-	-	-	-	0.57	2	0.56	15	0.87	588	0.62	27	0.70	523
STANDING-WATCHING TV	-	-	-	-	0.00	1	0.78	22	0.74	76	0.91	830	0.72	227
STANDING-TOILET	0.00	1	0.80	2	0.84	109	0.69	61	0.83	340	0.70	234	0.87	1522
WALKING-WATCHING TV	-	-	0.67	3	0.29	9	0.82	104	0.71	117	-	-	0.68	209
WALKING-COOKING	-	-	-	-	0.00	1	0.84	26	0.67	3	0.76	208	0.73	80
SITTING-COOKING	-	-	-	-	-	-	1.00	1	0.80	36	0.74	60	0.74	59
WALKING-CLEANING	-	-	-	-	0.67	3	0.60	13	0.82	29	0.89	44	0.48	57
LYING DOWN-EATING	-	-	-	-	-	-	0.60	59	0.55	5	0.69	23	0.80	96
SITTING-CLEANING	-	-	-	-	-	-	-	-	0.46	35	0.43	8	0.78	9
accuracy	0.91	43	0.88	51	0.74	246	0.81	1577	0.80	1605	0.85	2716	0.82	6245
macro avg	0.74	43	0.82	51	0.52	246	0.66	1577	0.66	1605	0.71	2716	0.69	6245
weighted avg	0.90	43	0.88	51	0.74	246	0.81	1577	0.80	1605	0.84	2716	0.82	6245

reducing the learning rate when the validation accuracy has stopped improving using $factor = 0.1, patience = 2$. The training set was shuffled for the training process.

Table 2 presents the results for these experiments. The F1 columns denote the F-1 scores per number of sensors used, whereas the S columns denote the corresponding samples provided in the test set (*support* metric), e.g., *3s F1*, denotes the F1-scores for the sub-set holding the observations which utilized data from exactly 3 sensors, whereas *3s S* denotes the number of samples provided for that sub-set (*support*). *As* denotes that the whole test-set was used (*all sensor data*). The resulting model provided an accuracy of 82% on the whole test set, whereas the sub-set containing the observations which hold data for all 7 sensors provided an accuracy of 84%. There were fewer observations that provided data for less than 5 sensors, according to the support metrics. The classes with larger sample sizes (e.g., *SITTING-WATCHING TV* and *STANDING-TOILET*) had accuracy scores of at least 86%, whereas on the opposite spectrum classes with too few samples (e.g., *SITTING-TOILET* and *STANDING-COOKING*) had low scores of 38% and 55% respectively denoting balancing issues between the classes.

The next set of experiments investigates the proposed data augmentation's method performance. In Table 3 we present the result of the experiments with the test set augmented with a value of $C = 2$ by using Algorithm 2. In Table 4 both the training/validation sets are augmented with a value of $C = 2$, as well as the test set. Finally in Table 5 only training and validation sets are augmented with a value of $C = 2$. As there are more combinations of 2-sensor data than 3-sensor data, and 3-sensor data than 4-sensor data etc., in Table 3 we can observe a sharp drop on the accuracy of the new augmented data set, as it drops to 65%, whereas the accuracy on the sub-set containing data for all 7-sensors remains at 84%. In Table 4 we observe an increase of total accuracy to 72%. When providing data for at least 4 sensors the accuracy is higher, 76% for 4-sensor observations, 79% for 5-sensor observations and 83% for 6-sensor observations. The largest increase of performance was 11% for 2-sensor observations. Notice that according to Table 1 the sensor unimodal models do not perform equally. From this premise it follows logically that not all sensor sub-sets will perform equally (especially the 2 sensor sub-sets). In Table 5 we can observe that the fusion model trained using the augmented training and validation sets performs equally to the non-augmented one on the original test set with an accuracy score of 82%. Its 2-sensor accuracy is higher though (91% versus 88% when trained without the proposed data augmentation method).

5 Conclusions & Future Work

In this work, we proposed a sensor-independent fusion method in respect to the number of sensors utilized for Automatic Human Activity Recognition. It utilizes feature-level fusion. This method allows the design of fusion models that

can operate with fewer data sources than the ones the model was designed to operate on. However, the max number of sensors must be known beforehand.

Furthermore, to increase the fusion model's performance when operating on observations with fewer sensor data we proposed a data augmentation method that uses no interpolation or estimation techniques to augment the dataset. Instead it generates all possible combinations of utilized sensors for recorded observations, with a minimum number of sensor data required defined by a constant. The results showed an increase in all sub-sets of the test set, split according to the number of sensors used per observation, indicating the method's effectiveness.

For the future we will investigate class balancing methods to improve the scores for classes with fewer samples. Moreover, we will also investigate the individual sensor contributions to the results, as well as the accuracy differences between different combinations of sensors totaling to the same number, e.g., what is the difference of accuracy when using 3-sensor data between *Audio, Watch Compass, Phone Magnet* and *Phone State, Watch Accelerometer, Phone Accelerometer*. The goal is to generate a general method of evaluating a sensor's performance in the fusion model, while the unimodal models for the sensors are considered as black boxes.

Acknowledgments

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Name: ACTIVE, Project Number: HFRI-FM17-2271)

References

1. Aggarwal, J.K., Xia, L.: Human activity recognition from 3d data: A review. Pattern Recognition Letters **48**, 70–80 (2014)
2. Barrios-Avilés, J., Iakymchuk, T., Samaniego, J., Medus, L.D., Rosado-Muñoz, A.: Movement detection with event-based cameras: comparison with frame-based cameras in robot object tracking using powerlink communication. Electronics **7**(11), 304 (2018)
3. Batchuluun, G., Nguyen, D.T., Pham, T.D., Park, C., Park, K.R.: Action recognition from thermal videos. IEEE Access **7**, 103893–103917 (2019)
4. Chandrasekaran, B., Gangadhar, S., Conrad, J.M.: A survey of multisensor fusion techniques, architectures and methodologies. In: SoutheastCon 2017. pp. 1–8. IEEE (2017)
5. Dong, Y., Li, X., Dezert, J., Khyam, M.O., Noor-A-Rahim, M., Ge, S.S.: Dezert-smarandache theory-based fusion for human activity recognition in body sensor networks. IEEE Transactions on Industrial Informatics **16**(11), 7138–7149 (2020)
6. Ehatisham-Ul-Haq, M., Javed, A., Azam, M.A., Malik, H.M., Irtaza, A., Lee, I.H., Mahmood, M.T.: Robust human activity recognition using multimodal feature-level fusion. IEEE Access **7**, 60736–60751 (2019)

7. Grabisch, M., Raufaste, E.: An empirical study of statistical properties of the choquet and sugeno integrals. IEEE Transactions on Fuzzy Systems **16**(4), 839–850 (2008)
8. Innocenti, S.U., Becattini, F., Pernici, F., Del Bimbo, A.: Temporal binary representation for event-based action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 10426–10432. IEEE (2021)
9. Lee, Y.S., Cho, S.B.: Activity recognition using hierarchical hidden markov models on a smartphone with 3d accelerometer. In: International conference on hybrid artificial intelligence systems. pp. 460–467. Springer (2011)
10. Li, H., Shrestha, A., Fioranelli, F., Le Kernec, J., Heidari, H., Pepa, M., Cippitelli, E., Gambi, E., Spinsante, S.: Multisensor data fusion for human activities classification and fall detection. In: 2017 IEEE SENSORS. pp. 1–3. IEEE (2017)
11. Naik, K., Pandit, T., Naik, N., Shah, P.: Activity recognition in residential spaces with internet of things devices and thermal imaging. Sensors **21**(3), 988 (2021)
12. Nweke, H.F., Teh, Y.W., Mujtaba, G., Al-Garadi, M.A.: Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. Information Fusion **46**, 147–170 (2019)
13. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)
14. Sebestyen, G., Stoica, I., Hangan, A.: Human activity recognition and monitoring for elderly people. In: 2016 IEEE 12th international conference on intelligent computer communication and processing (ICCP). pp. 341–347. IEEE (2016)
15. Uddin, M.Z., Soylu, A.: Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning. Sci Rep **11**(1), 16455 (Aug 2021)
16. Vaizman, Y., Ellis, K., Lanckriet, G.: Recognizing detailed human context in the wild from smartphones and smartwatches. IEEE Pervasive Computing **16**(4), 62–74 (2017). https://doi.org/10.1109/MPRV.2017.3971131
17. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Frontiers in Robotics and AI **2**, 28 (2015). https://doi.org/10.3389/frobt.2015.00028
18. Wang, L., Huynh, D.Q., Koniusz, P.: A comparative review of recent kinect-based action recognition algorithms. IEEE Transactions on Image Processing **29**, 15–28 (2019)
19. Wu, Q., Wang, Z., Deng, F., Chi, Z., Feng, D.D.: Realistic human action recognition with multimodal feature selection and fusion. IEEE Transactions on Systems, Man, and Cybernetics: Systems **43**(4), 875–885 (2013). https://doi.org/10.1109/TSMCA.2012.2226575
20. Yao, S., Hu, S., Zhao, Y., Zhang, A., Abdelzaher, T.: Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In: Proceedings of the 26th International Conference on World Wide Web. pp. 351–360 (2017)
21. Zeng, Z., Zhang, Z., Pianfetti, B., Tu, J., Huang, T.S.: Audio-visual affect recognition in activation-evaluation space. In: 2005 IEEE International Conference on Multimedia and Expo. pp. 4–pp. IEEE (2005)
22. Zhang, S., Wei, Z., Nie, J., Huang, L., Wang, S., Li, Z.: A review on human activity recognition using Vision-Based method. Journal of Healthcare Engineering **2017**, 3090343 (Jul 2017)
23. Zhu, C., Sheng, W.: Multi-sensor fusion for human daily activity recognition in robot-assisted living. In: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction. pp. 303–304 (2009)