Neural Process Lett (2015) 42:5–26
DOI 10.1007/s11063-014-9395-0

Multi-sensor Fusion Based on Asymmetric Decision Weighting for Robust Activity Recognition

Oresti Banos · Miguel Damas · Alberto Guillen · Luis-Javier Herrera · Hector Pomares · Ignacio Rojas · Claudia Villalonga

Published online: 26 October 2014
© Springer Science+Business Media New York 2014

Abstract The recognition of human activity has been deeply explored during the recent years. However, most proposed solutions are mainly devised to operate in ideal conditions, thus not addressing crucial real-world issues. One of the most prominent challenges refers to common sensor technological anomalies. Sensor faults and failures introduce variations in the measured sensor data with respect to the equivalent observations in ideal conditions. As a consequence, predefined recognition systems may potentially fail to identify actions in the anomalous sensor data. This paper presents a novel model devised to cope with the effects introduced by sensor technological anomalies. The model builds on the knowledge gained from multi-sensor configurations, through asymmetrically weighting the decisions provided at both activity and sensor levels. Insertion and rejection weighting metrics are particularly used to eventually yield a unique recognized activity. For the sake of comparison, the tolerance to sensor faults and failures of standard activity recognition systems and the new proposed model are evaluated. The results prove classic activity-aware systems to be incapable of recognition under the effects of sensor technological anomalies, while the proposed model demonstrates to be robust against both sensor faults and failures.

Keywords Wearable sensors · Sensor anomalies · Sensor failures · Sensor faults · Decision fusion · Weighted decision · Activity recognition

1 Introduction

The automatic assessment of human behavior has gained much interest during the recent years. Also known as human activity recognition (AR), it aims at autonomously identifying human conducts from the observation of a person's actions and their interaction with the

O. Banos (✉) · M. Damas · A. Guillen · L.-J. Herrera · H. Pomares · I. Rojas · C. Villalonga
Department of Computer Architecture and Computer Technology, Research Center for Information and Communication Technologies of the University of Granada (CITIC-UGR), C/Periodista Rafael Gomez Montero 2, 18071 Granada, Spain
e-mail: oresti@ugr.es

surroundings. Sensing technologies, capable of measuring behavioral characteristics, such as human body motion, are used to observe these actions. On-body inertial sensors are on the forefront of these technologies. These sensors can measure the movements of the body parts on which they are mounted, and be easily embedded into daily use garments or items such as watches or belts. Accordingly, they can be conveniently worn and ubiquitously used, which proves to be essential for the daily usage of recognition systems. Not only are AR systems devised to help better understand human behavior, but also assist people during their daily tasks and routines. Examples of application domains explored in the past are healthcare [6,53], rehabilitation [2,45], sports [26,36], wellbeing [18,22], industrial maintenance [34,47] or gaming [24].

The activity inference process, also referred to as activity recognition chain (ARC, [43]), consists of a set of steps combining signal processing, pattern recognition and machine learning techniques to implement a specific AR system. Concretely, a set of sensors usually deliver raw unprocessed signals, which represent the measured magnitude (e.g., acceleration). The registered information is sometimes filtered to remove electronic noise and artifacts [35, 46]. To capture the dynamics of the signals, these are normally partitioned in data windows of a fixed size [9]. Subsequently, a feature extraction process is carried out on each data window to provide a handler representation of the signals for the pattern recognition stage. A wide range of heuristics [32], time–frequency domain [33,41] and other sophisticated mathematical and statistic functions [7] are commonly used. In some cases, a feature selector is used to reduce redundancy among features, as well as to minimize dimensionality [31]. The resulting feature vector is provided as input to a classifier, which ultimately yields the recognized activity or class to one of the considered for the particular problem. Extensive topical reviews on classical AR methods can be seen in [28,40].

AR approaches can be mainly categorized according to the considered sensor topology. Recognition systems may operate on data measured through a sole sensor, thus requiring from a single ARC (SARC). This is the most widely used approach by current vendors of AR applications, since they generally build on data collected through a single smartwatch [1,37], bracelet [22], fitness band [20,51] or clip [18]. Capturing the motion of different body parts normally improves the system's recognition capabilities. To do so, multi-sensor configurations are required, which are less common in commercial solutions. In that case, the data collected through each sensor node need to be fused to eventually provide a single recognized activity. This may be performed either at the feature level, i.e., by aggregating the features extracted from each sensor data stream (feature fusion multi-sensor ARC or FFMARC) or at the classification level, i.e., by combining the decisions developed on the data of each respective sensor (decision fusion multi-sensor ARC or DFMARC).

Although much effort has been put into the development of reliable AR systems, most previous approaches assume that the sensor setup remains identical during the lifelong use of the system, which has been recently proven to be an unrealistic assumption [10]. Particularly, wearable sensors are subject to an intensive use and potential harsh conditions, therefore also prone to degradation and failures. Technological anomalies may lead to changes in the sensor data streams, which are normally unforeseen during the design phase or unpredictable at runtime use. Consequently, models trained on ideal signal patterns may react in an undesired manner to imperfect or anomalous sensor data. This potentially translates into a partial or total malfunctioning of the AR system.

This work investigates the tolerance of standard AR systems to sensor technological anomalies. Based on the limitations of classic AR approaches, a novel alternate model is here proposed to cope with the effects of sensor faults and failures. This solution is particularly devised to maintain the recognition capabilities even in the event of sensor anomalies, which

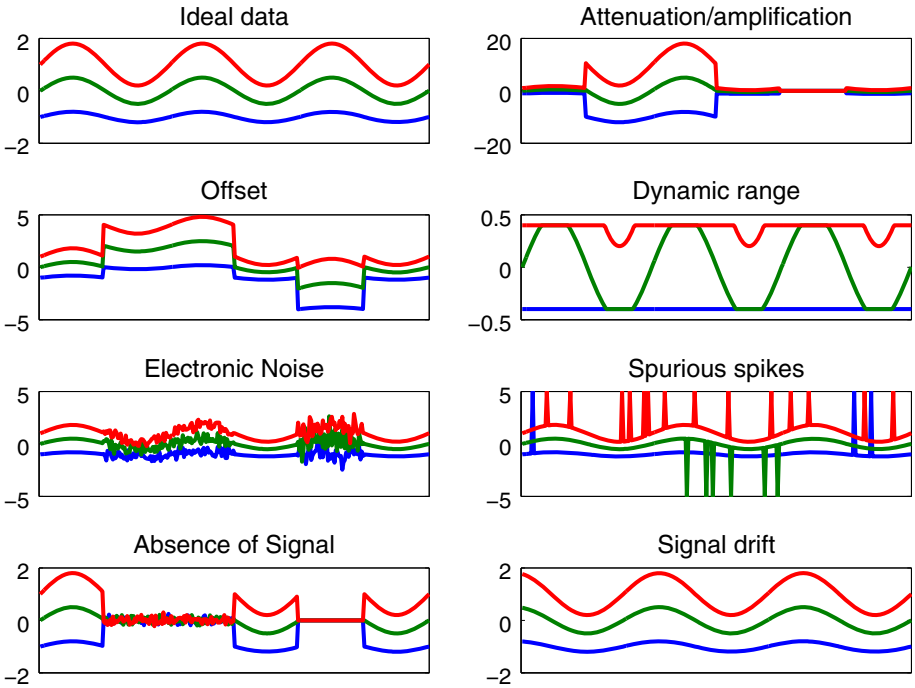

Fig. 1 Examples of the effects of some of the most prominent sensor technological anomalies

is found to be of primal necessity in critical applications where the activity-aware service must not be interrupted. The rest of the paper is organized as follows. Section 2 presents the main technological anomalies to which on-body sensors are subject. The method proposed to deal with sensor technological anomalies is described in Sect. 3. Section 4 evaluates the robustness of standard AR models and the proposed new model to the effects of sensor technological anomalies. The results are extensively discussed in Sect. 5, while final conclusions are summarized in Sect. 6.

2 Sensor Technological Anomalies

2.1 Signal Effects

Electronic devices are subject to diverse technological anomalies. On-body inertial sensors are particularly prone to changes in the bias, scale factors, non-linearity or electronic noise, among other effects (see Fig. 1), normally due to decalibration or battery failures. Some of these anomalies have been extensively studied in the past, and important improvements have been certainly made to minimize their effects. Nevertheless, some others are more difficult to handle from a hardware perspective and therefore more likely to appear during the use of these devices.

One of the most difficult to overcome, and limiting, data losing effect is associated to changes in the sensor dynamic range. These variations may appear due to a misconfiguration of the sensor or when the system is not adequately supplied. Sensor misconfigurations rarely

take place; however, battery malfunctions are more frequently observed. In the event of a battery fault, internal amplifiers and analog-to-digital converters are not appropriately supplied. In turn, the amplification and conversion process is not correctly performed, leading to a new reduced dynamic range. This translates into distortions in the signals gleaned from the sensor, such as flattening or skew, which basically correspond to an underrepresentation of the measured data, mainly for those samples that fall out of the bounds of the new dynamic range (see Fig. 1, second row, right column).

Another key shortcoming of on-body sensors, which applies to any wireless technology, refers to energy limitations. Sensor batteries are not of unlimited capacity and therefore need to be recharged from time to time. Moreover, sensor batteries lose charge over time, and its capacity reduces as they are charged and drained. Hence, there exist potential situations in which a sensor is not supplied, and consequently no data delivered. Permanent critical failures may appear in more extreme situations, for example, when the sensor device falls to the ground, is accidentally hit or physically damaged in any other way. In those situations, not only could the sensor get broken, but simply an essential electronic component be destroyed (e.g., communication interface, processing unit). As a consequence, it is normal to expect an absence of signal or data (see Fig. 1, bottom, left corner).

2.2 Tolerance of Standard Activity Recognition Models

Although some sensor anomalies may be removed during the preprocessing stage of the ARC (e.g., electronic noise and spurious spikes filtering), others are more difficult to overcome or avoid, especially at runtime. In case of a critical sensor failure or misfunctioning (e.g., sensor out of battery or totally broken), systems based on a single sensor unit demonstrate futile. As a matter of fact, SARC models rely on a single data stream, thus they cannot operate since no data is available. In these circumstances, the use of redundant or multi-sensor configurations seems to be a reasonable alternative. However, not all multi-ARC models are seen to cope with the problem of a discharged or broken sensor. In fact, FFMARC models suffer from similar limitations to SARC approaches to this respect. FFMARC aggregates the features extracted from each sensor node into a single vector, which is used as input to a classifier or reasoner. Therefore, if data from a sensor is missing, no features can be obtained from that particular node; consequently, the feature vector turns to be incomplete and no activity-aware capabilities are supported. Conversely, this problem is not seen to occur to DFMARC models. DFMARC approaches are based on the aggregation of the decisions computed from the processing of each individual sensor data stream. Therefore, even whether a sensor data stream is not available, a decision may be made by combining the decisions obtained from the remaining active sensors.

Most DFMARC approaches proposed in past work are based on two main techniques, namely, hierarchical decision (HD) and majority voting (MV). HD is based on favoring those sensor entities that generally behave better, thus allowing them to decide first. Accordingly, decisions are made in strict order of classification capabilities, with the ranking normally set according to performance criteria. In MV, the eventual recognized class is simply the one endorsed by a plurality of sensor classifiers. Although highly used in activity recognition, HD and MV models show significant weaknesses when dealing with sensor technological anomalies. An example is used next to illustrate this. Let us consider a sensor setup consisting of seven sensors worn in diverse body parts. A classifier is learned for each sensor, and their corresponding performance metrics, e.g., accuracy in %, obtained after evaluation. Let us assume the resulting performance values are, e.g., $S_1 \rightarrow 99\,\%$, $S_2 \rightarrow 85\,\%$, $S_3 \rightarrow 82\,\%$, $S_4 \rightarrow 49\,\%$, $S_5 \rightarrow 36\,\%$, and $S_6 \rightarrow 31\,\%$. Now, if the HD model is used, the overall accuracy

of the system will principally depend on the performance of the decisor ranked on top of the hierarchy (i.e., S_1). If a low-ranked sensor (e.g., S_4–S_6) becomes unavailable, no variation is expected in the system performance. However, if any sort of sensor anomaly affects the top-ranked decisors (e.g., S_1, S_2) the recognition capabilities may seriously worsen. For example, if S_1 gets out of battery, the bulk of the decisions would rely on S_2, thus a drop on the recognition performance may be expected. Conversely to HD, MV does benefit from the disappearance of sensors with poor performance. Thus for the above example, if S_4, S_5 or S_6 shutdown, an improvement in the global performance can be expected. Unfortunately, this also applies the other way around. Then, if high-rated sensors become unavailable, the probability of misrecognition increases, especially when it results into a majority of low-rated decision makers. For example, if S_1 gets out of the pool of decision makers, a plurality of low-rated sensors (S_4, S_5, S_6) would outnumber a minority of higher-rated sensors (S_2, S_3), thus potentially leading to an overall low performance. According to this, HD and MV models are qualitatively shown to have limited capabilities to deal with sensor failures. Moreover, from the previous example, it can be easily ascertained that the most beneficial approach would consist in determining the activity from the decisions provided by the highest-rated sensors (i.e., S_2 and S_3), which roughly corresponds to the combination of both HD and MV models.

3 Hierarchical Weighted Classifier

Taking into account the advantages and drawbacks of HD and MV approaches, a new ensemble model is here presented to cope with the effects of sensor technological anomalies. The model combines the decisions obtained from each individual sensor, making them all part of the decision process. However, the decisions are first ranked based on their relative importance, by means of weights derived from the individual performance of each classification entity. Moreover, the decisions are not only combined at the sensor level but also at the activity level, which is devised to increase both reliability and robustness of recognition systems, and also considered of special importance to support flexible sensor setups.

The proposed model, hereafter called hierarchical weighted classifier (HWC), is composed by three decision making levels or stages (see Fig. 3). Given an scenario with M nodes of information (sensors) and N classes (activities), a set of M by N base classifiers (c_{mn}, $\forall m = 1, \ldots, M, n = 1, \ldots, N$) are defined. These are binary classifiers specialized in the discrimination of the activity or class n by using the information obtained from the sensor or node m. Each base classifier applies an one-versus-rest binary classification strategy,[1] which further allows for the use of any type of standard classification paradigm. This defines the first layer of the model, here identified as base, class or *activity level*. The second classification level, so-called *sensor level*, is defined through M node or sensor classifiers (S_m, $\forall m = 1, \ldots, M$). Sensor classifiers are not machine learning-type entities, but rather decision making structures. Each sensor classifier consists of N base classifiers (one per class), whose decisions are combined through an activity-dependent weighting scheme. Finally, the last layer, so-named *network level*, is in charge of the weighting and aggregation of the decisions given by each sensor classifier, eventually providing the recognized activity or class. The weights used in the network level depend on the recognition capabilities of each individual sensor classifier.

[1] Other approaches as the one-versus-one may be similarly applied; however, the one-versus-rest model is particularly recommended here to minimize the number of classification entities.

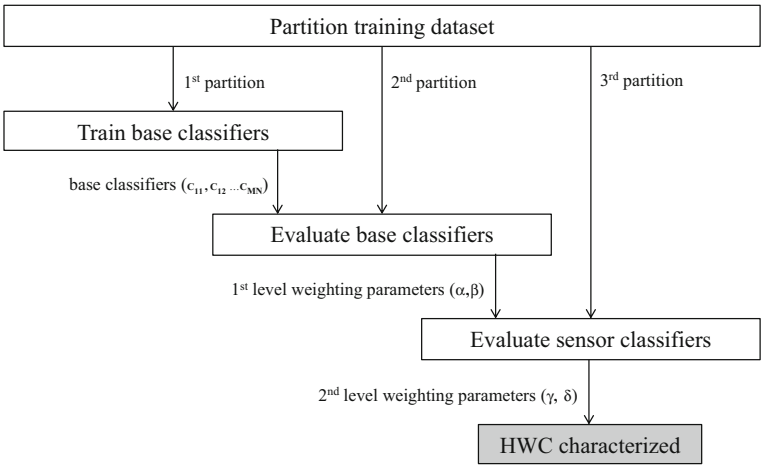

Fig. 2 Training steps of the HWC model

The training of the HWC model requires just a few steps (Fig. 2). Firstly, the training dataset is partitioned into three equally-distributed parts. One of these partitions is used to train all base classifier entities. After training, a second partition is used to test the performance of each base classifier. From here, statistical metrics are obtained and further used to define the first level of weighting parameters. Sensor classifiers are completely defined at this point. Then, the weighting parameters of the network level are assessed. To that end, the third yet unused part of the dataset is utilized to evaluate the performance of each individual sensor classifier. Then, the network level weights are extracted from the performance statistics of each sensor classifier. The HWC is, at this point, completely characterized.

Two weighting schemes are proposed for the HWC model. The first approach, firstly introduced in [8], consists in weighting the decisions provided by each classifier through a single weight. Thus, two weights are used in this model: an α_{mn} weight for each base classifier c_{mn}, and a γ_m weight for each sensor classifier S_m. These weights are used to ponder both insertions and rejections[2] in a similar way (i.e., *unified weighting*). This model is generically identified as $\text{HWC}_{\alpha\gamma}$. The second novel approach (Fig. 3) corresponds to an improved version of the former model, in which two independent weights are used to ponder insertions and rejections (i.e., *insertion–rejection weighting*). Concretely, α_{mn} and β_{mn} weights are respectively used to ponder insertions and rejections of each base classifier c_{mn}, while γ_m and δ_m weights serve the same purpose for each sensor classifier S_m. In this way, it is possible to leverage the potential of all classifiers even when they are either accurate inserters or good rejecters. This new model, hereafter identified as $\text{HWC}_{\alpha\beta\gamma\delta}$, is neatly described next.

As stated before, two weights are obtained at the activity level. These parameters are defined as α_{mn} and β_{mn}, and respectively represent the insertion and rejection weights for c_{mn}. The values of α_{mn} and β_{mn} are obtained from the performance assessment of each base classifier. In particular, α_{mn} corresponds to the sensitivity of c_{mn}, whilst β_{mn} relates to

[2] In machine learning, insertions (hits) and rejections (deletions) respectively refer to positive and negative classifications. For the one-versus-rest decision strategy, an insertion is observed when the classifier recognizes a class to belong to its class of specialization, while a rejection is generated when the class is identified to belong to any of the rest of the classes.

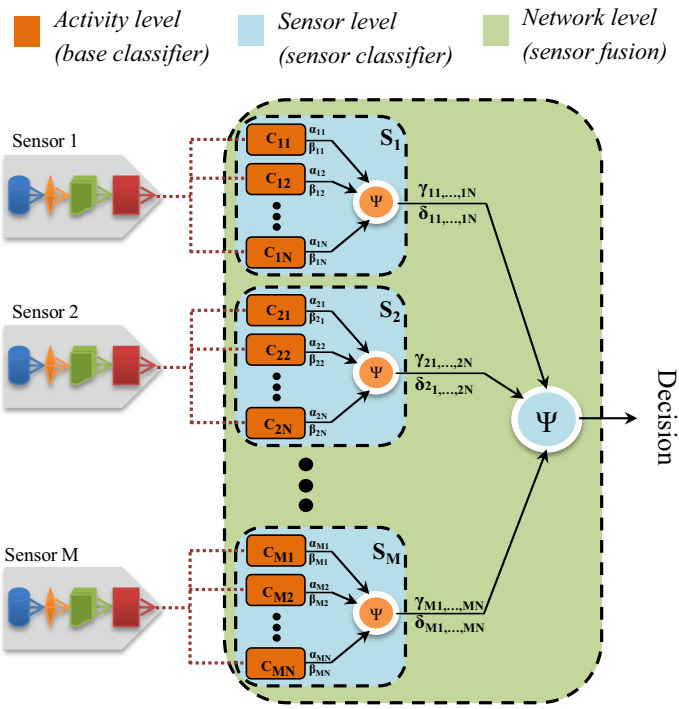

Fig. 3 Structure of the HWC for the insertion–rejection weighting model ($\text{HWC}_{\alpha\beta\gamma\delta}$). The features extracted from each sensor are used as inputs to a set of N by M base classifiers (c_{mn}). Classifiers' insertions and rejections are respectively α- and β-weighted. These decisions are aggregated through a combiner function (Ψ), thus yielding a set of decisions for each sensor classifier (S_m). The decisions made across all sensor classifiers are γ-weighted (insertions) and δ-weighted (rejections), and once again combined to provide the eventual recognized activity

its specificity. We have selected these performance parameters since they represent well the insertion and rejection capabilities of the classifier. Given TP_{mn} (true positives) the number of correctly identified samples, FP_{mn} (false positives) the incorrectly identified samples, TN_{mn} (true negatives) the number of correctly rejected samples and FN_{mn} (false negatives) the incorrectly rejected samples, all specifically computed from the evaluation of the classifier c_{mn}, α_{mn} and β_{mn} are obtained as follows:

$$\alpha_{mn} = \text{Sensitivity}\,(c_{mn}) = \frac{TP_{mn}}{TP_{mn} + FN_{mn}} \tag{1}$$

$$\beta_{mn} = \text{Specificity}\,(c_{mn}) = \frac{TN_{mn}}{TN_{mn} + FP_{mn}} \tag{2}$$

A voting method is at this point considered to fuse all the decisions provided by the base classifiers for each corresponding sensor classifier. For a sensor m, given a window instance s_{mk} characterized through the corresponding feature vector $f_m(s_{mk})$, and being q the activity or class predicted by c_{mn} for that instance, if such class belongs to the class of specialization of c_{mn} (i.e., $q = n$), the classifier will set its decision to α_{mn} for the class n and 0 for the rest of classes. Otherwise (i.e., $q \neq n$), the decision is set to 0 for the class n and to β_{mn} for the others. In summary, the weighted decision for c_{mn} (i.e., WD_{mn}) may be defined as $(\forall\,\{q, n\} = 1, \ldots, N)$:

$$WD_{mn}\left(f_m(s_{mk})\right) = \begin{cases} \alpha_{mn}, & f_m(s_{mk}) \text{ classified as } q \\ 0, & f_m(s_{mk}) \text{ not classified as } q \\ \beta_{mn}, & f_m(s_{mk}) \text{ not classified as } q \\ 0, & f_m(s_{mk}) \text{ classified as } q \end{cases} \begin{matrix} (\forall q = n) \\ \\ (\forall q \neq n) \\ \\ \end{matrix} \tag{3}$$

The aggregation of the weighted decisions provided by each base classifier for the mth sensor classifier (S_m) may be computed as follows:

$$O_m\left(f_m(s_{mk})\right) = \sum_{n=1}^{N} WD_{mn}\left(f_m(s_{mk})\right) \tag{4}$$

The class predicted by S_m is the class q for which the sensor classifier output is maximized:

$$q_m\left(f_m(s_{mk})\right) = \underset{q}{\operatorname{argmax}}\left(O_m\left(f_m(s_{mk})\right)\right) \tag{5}$$

For the next level, similar parameters to α_{mn} and β_{mn} are obtained, here defined as γ_m (insertions) and δ_m (rejections). Nonetheless, the way these are computed varies slightly with respect to the formers. At the network level the classifiers are not binary but multiclass models. Therefore, the evaluation of each sensor classifier requires to extend sensitivity and specificity concepts to the multiclass case (see details in [52]).

According to this generalization, γ_m and δ_m may be described as:

$$\gamma_m = \langle \gamma_{m1}, \gamma_{m2}, \ldots, \gamma_{mn} \rangle = \left\langle \frac{TP_{m1}}{TP_{m1} + FN_{m1}}, \frac{TP_{m2}}{TP_{m1} + FN_{m2}}, \ldots, \frac{TP_{mn}}{TP_{mn} + FN_{mn}} \right\rangle \tag{6}$$

$$\delta_m = \langle \delta_{m1}, \delta_{m2}, \ldots, \delta_{mn} \rangle = \left\langle \frac{TN_{m1}}{TN_{m1} + FP_{m1}}, \frac{TN_{m2}}{TN_{m1} + FP_{m2}}, \ldots, \frac{TN_{mn}}{TN_{mn} + FP_{mn}} \right\rangle \tag{7}$$

where $\{TP/TN/FP/FN\}_{mn}$ refer to the classification counting values. Conversely to the binary case, these values are here computed for each class k across the confusion matrix results obtained from the evaluation of S_m ($\forall\, m = 1, \ldots, M, n = 1, \ldots, N$).

Since the decisions are made in a multiclass fashion, γ_m and δ_m are used to reward or penalize each identified class. Accordingly, given q_m the decision of S_m for the sample s_{mk}, the set of weighted decisions for this classifier are defined as:

$$WD_m\left(q_m\left(f_m(s_{mk})\right)\right) = \begin{cases} \gamma_{mn}, & n = q_m\left(f_m(s_{mk})\right) \\ -\delta_{mn}, & n \neq q_m\left(f_m(s_{mk})\right) \end{cases} (\forall n = 1, \ldots, N) \tag{8}$$

The output at the network level is now calculated taking into account the individual outputs obtained from each sensor classifier. Given a sample s_k, defined through the corresponding data windows obtained from each respective sensor ($\{s_{1k}, s_{2k}, ..., s_{Mk}\}$), and being characterized through the corresponding feature vectors ($\{f_1(s_{1k}), f_2(s_{2k}), ..., f_M(s_{Mk})\}$), the aggregated output is:

$$O\left(x_k\right) = O\left(\{f_1(s_{1k}), f_2(s_{2k}), ..., f_M(s_{Mk})\}\right) = \sum_{p=1}^{M} WD_p\left(q_p\left(f_p(s_{pk})\right)\right) \tag{9}$$

Finally, the class q yielded on top of the hierarchy is obtained as:

$$q = \underset{q}{\operatorname{argmax}}\left(O\left(x_k\right)\right) \tag{10}$$

4 Evaluation of the Tolerance Against Sensor Technological Anomalies

This section aims to quantitatively analyze the capabilities of the proposed HWC model to deal with sensor technological anomalies. To that end, the HWC is first compared to standard AR models in ideal conditions, in order to prove that it provides similar recognition capabilities to them. Next, the model is evaluated for the case in which critical sensor failures are assumed. Finally, the recognition capabilities under the effects of moderate sensor faults are assessed.

4.1 Benchmark Dataset

Sensor technological anomalies normally appear in a random and occasional manner, thereby it may be complicated to find them during experimental recordings. Nevertheless, an interesting characteristic of these anomalies is that their effects may be reasonably easily modeled. Therefore, the approach followed in this work consists in synthetically introducing sensor hardware anomalies into activity data, which is experimentally recorded in a daily living setting.

The activity dataset used in this work was first introduced in [11] and has been widely used to benchmark AR models. This dataset has been considered especially interesting for evaluation, since not only includes data collected in a natural out-of-lab settings but for a diverse sample population and different activities. Moreover, this is one of the few publicly available datasets.[3]

The dataset comprises acceleration data collected for twenty subjects, aged 17–48, while performing a set of daily living activities. Concretely, five bi-axial accelerometers are employed to register the motion experienced by the subjects' right hip (H), dominant wrist (W), non-dominant arm (A), dominant ankle (K) and non-dominant thigh (T), respectively. From the complete activity set, the most representative nine were selected covering from intense activities such as *running* and *cycling* to fitness exercises like *stretching* and *strength-training*, moderate routines such as *walking* and *climbing stairs* or sedentary activities like *sitting*, *standing* and *lying down*.

4.2 Experimental Setup

For the AR process, the diverse stages of the ARC are implemented. Raw acceleration signals are acquired through the on-body inertial sensors. The recorded signals are affected by spurious spikes and electronic noise, which are removed through a 20 Hz cutoff low pass elliptic FIR filter. This is demonstrated to not eliminate valid information for daily physical activity assessment [12,32]. The signals are subsequently partitioned into windows of data of approximately 6 s, as suggested in [11]. Next, features are extracted to characterize each window data. To analyze the computational complexity required for this problem, diverse feature vector dimensions are tested (1, 5, 10 and 20 features). Here, a subset of the complete group of features proposed in a previous work is considered [7]. These features correspond to the combination of statistical functions such as median, kurtosis, mode, range, and magnitudes or functions obtained from a domain transformation of the original data such as energy spectral density, spectral coherence or wavelet coefficients ("a1 to a5" and "d1 to d5" Daubechies levels of decomposition), among others. The best features ranked through the use of a receiver operating characteristic feature selector [50] are chosen until complete the

[3] Dataset files and description could be obtained at http://architecture.mit.edu/house_n/data/Accelerometer/BaoIntilleData04.htm.

feature vector lengths defined for each case. Specific of each ARC model, SARC builds on a single feature vector extracted from the data of the corresponding used sensor. FFMARC aggregates all feature vectors computed across the five sensor streams into a single feature vector. DFMARC uses the feature vector extracted from each sensor as input to each respective sensor classifier.

The classification stage is different for each ARC model. SARC and FFMARC employ standard multi-class classifiers. Four machine learning paradigms, which have been widely and successfully used in past AR approaches, are here considered. They comprise, decision trees (DT, [17]), which were proved to perform well in [11,27,38]; naive Bayes (NB, [50]) utilized in [29,33,41]; k-nearest neighbor (KNN, [14]), used in [3,5,21,39] and support vector machines (SVM, [15]) employed in [19,21,23,30,49]. Concretely, the k value is empirically tuned for the KNN models while SVM implements a radial basis function (RBF) kernel with automatically tuned (grid search) hyper-parameters γ and C. These standard classifiers are also used as core of the base classifiers used by the DFMARC approaches (here, HD, MV and HWC).

Systems evaluation is carried out through a cross validation process. Although leave-one-subject-out cross validation (LOOXV) has been used in the literature, it is here rather chosen a ten-fold cross-validation (10-fold XV) process to compare the diverse models. In fact, as summarized in [4] and according to [13,25], LOOXV is the best technique for risk estimation, whereas 10-fold XV is the most accurate approach for model selection. Moreover, this process is repeated 100 times to ensure statistical robustness as well as to procure an asymptotic convergence to a correct estimation of the systems performance [48].

4.3 Performance in Ideal Conditions

The HWC is particularly devised to deal with the effects of sensor failures and faults. However, not only should the model be useful to overcome these limitations, but also be valid for AR tasks in normal circumstances. Therefore, the HWC recognition capabilities in invariant setups is first evaluated. Moreover, the performance of the model is compared to the baseline given by standard AR systems.

The HWC was defined upon the observation of the pros and cons of HD and MV models. Therefore, the first comparison is performed for these three models. In Fig. 4, the confusion matrices computed from the assessment of the performance of HD, MV and HWC models are shown. For the HWC, the weighting model proposed in [8] (unified weighting, $\text{HWC}_{\alpha\gamma}$) and the novel model proposed in this work (insertion–rejection weighting, $\text{HWC}_{\alpha\beta\gamma\delta}$) are respectively evaluated. From the results, HD and MV models demonstrate worst recognition capabilities. A high misclassification rate is observed when simple feature sets are employed. The performance is nevertheless enhanced when the feature sets are significantly enriched (Fig. 4c, d). This is motivated by the improvement of the recognition capabilities of each individual sensor classifier. This improvement translates into more accurate decisions on top of the hierarchy, thus reducing the errors made by the HD model. For the MV model, the aggregated decisions build then on a potential plurality of accurate sensor classifiers. HWC largely exceeds the recognition capabilities shown by HD and MV models. Moreover, this happens to occur for all classification paradigms and independently of the complexity of the feature vector. Indeed, very promising results are already obtained when a sole feature is used for each base classifier. The performance turns to be practically absolute (confusion matrices almost diagonal) when richer feature vectors are used. Both $\text{HWC}_{\alpha\gamma}$ and $\text{HWC}_{\alpha\beta\gamma\delta}$ provide outstanding results, which proves the potential of the HWC structure. However, the latter model surpasses the performance of the former. This proves the importance of the

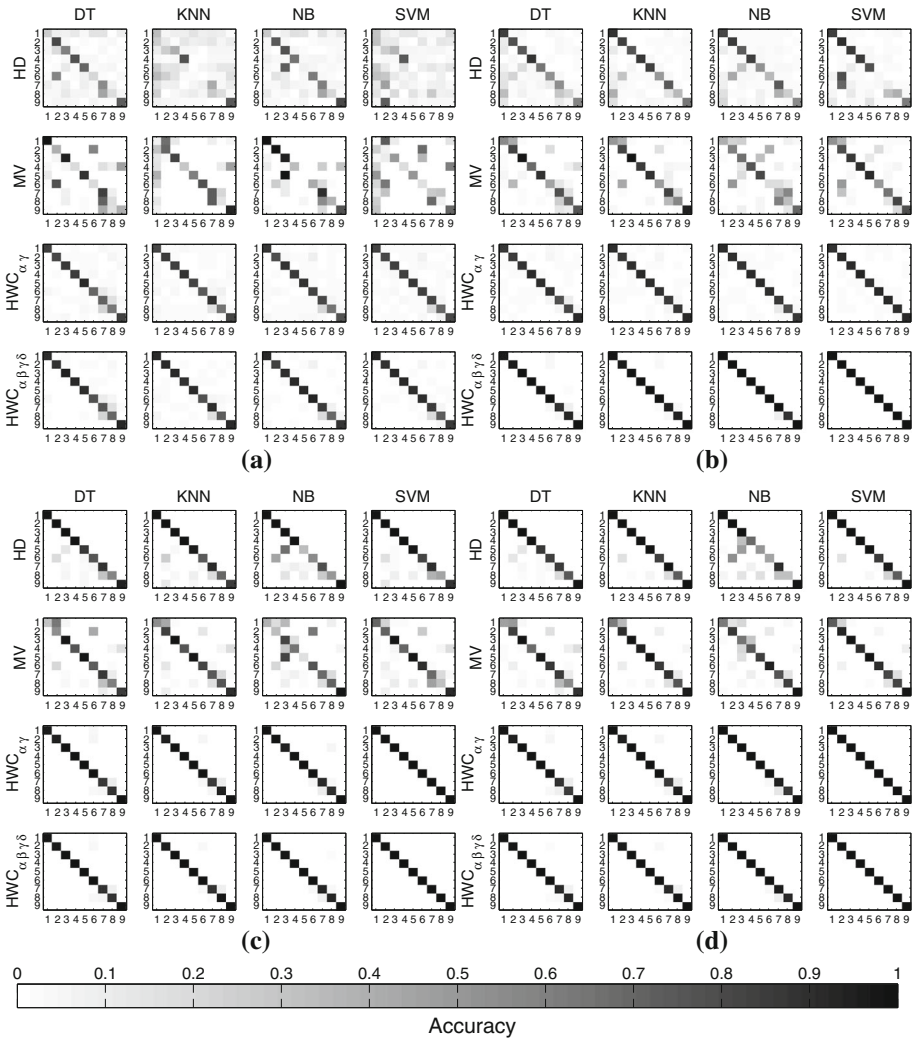

Fig. 4 Confusion matrices obtained from the experimental evaluation of each DFMARC modality (HD, MV, HWC$_{\alpha\gamma}$, and HWC$_{\alpha\beta\gamma\delta}$) and machine learning paradigm (DT, KNN, NB, and SVM). Diverse feature vector lengths are considered; (**a** 1, **b** 5, **c** 10, and **d** 20 features). Confusion matrix legend (activities): *1* walking, *2* running, *3* cycling, *4* sitting, *5* standing, *6* lying down, *7* stretching, *8* strength-training, *9* climbing stairs

considered weighting scheme. Accordingly, and for the sake of simplicity, the HWC$_{\alpha\beta\gamma\delta}$ will be used in advance as the predominant HWC model.

Most of the contributions in the AR domain are based on a SARC or FFMARC model. These approaches have shown good recognition capabilities in a wide sort of AR problems. Accordingly, SARC and FFMARC models are here used to define the performance baseline for the recognition in ideal circumstances. For the considered experimental setup, five SARC models are devised (one per sensor), whilst the fusion of the features extracted from each sensor data stream is implemented for the FFMARC approach. In Fig. 5, the performance results obtained from the evaluation of SARC, FFMARC and HWC models are depicted. Clearly, FFMARC and HWC stand out as the most accurate models. In general, FFMARC

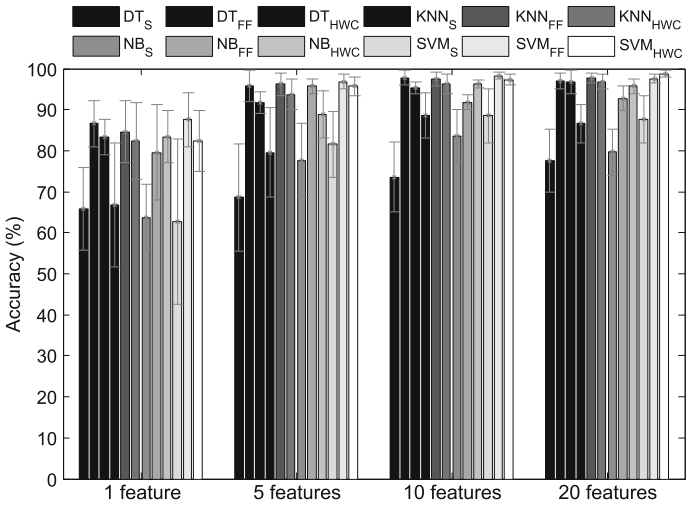

Fig. 5 Accuracy (mean and standard deviation) results from the evaluation of SARC (S), FFMARC (FF) and $\text{HWC}_{\alpha\beta\gamma\delta}$ (HWC) approaches. Results are averaged across all sensors for the SARC model. Diverse feature vector lengths are considered for evaluation (1, 5, 10, and 20 features). Legend: <classification paradigm $><AR\ approach>$

outperforms the HWC model for simple feature sets (1 and 5 features). This is quite reasonable since the aggregated feature vector used in FFMARC is richer than the ones used for each HWC base classifier. For example, when a single feature is extracted from each sensor stream, base classifiers operate on a 1-dimensional (1-D) feature space, while the classification model used in FFMARC operates in a 5-D feature space. Anyway, the differences are not higher than 7 % accuracy, at worst case. The gap between both models reduces as more features are employed, with HWC the most accurate approach for some cases, and performance levels close to absolute. Conversely to FFMARC and HWC, SARC models provide fair results, especially for reduced feature sets. Here again, enriching the feature vector translates into an improvement in the performance of the model, yet providing accuracies below 90 %.

To put it in a nutshell, the HWC proves as reliable as other standard AR approaches under ideal circumstances, while theoretically surpassing them in terms of tolerance to sensor technological anomalies. This is now evaluated in the next two sections.

4.4 Tolerance to Sensor Failures

In the event of a sensor failure or shutdown, SARC and FFMARC models were proved to not be capable of operating. SARC models cannot process the information because of a practical lack of information. FFMARC may utilize the information of the remaining active sensors; however, the aggregated feature vector cannot be built since no values can be gleaned from the affected sensor. Conversely to SARC and FFMARC models, the HWC was devised to deal with the situation of having missing sensors.

In the following, the recognition capabilities under the event of critical sensor failures are analyzed. To that end, a HWC model designed for the predefined sensor setup (i.e., all five sensors) is tested on diverse setup configurations, respectively corresponding to the cases in which data from a sensor or various sensors are not available. Concretely, the $\text{HWC}_{\alpha\beta\gamma\delta}$

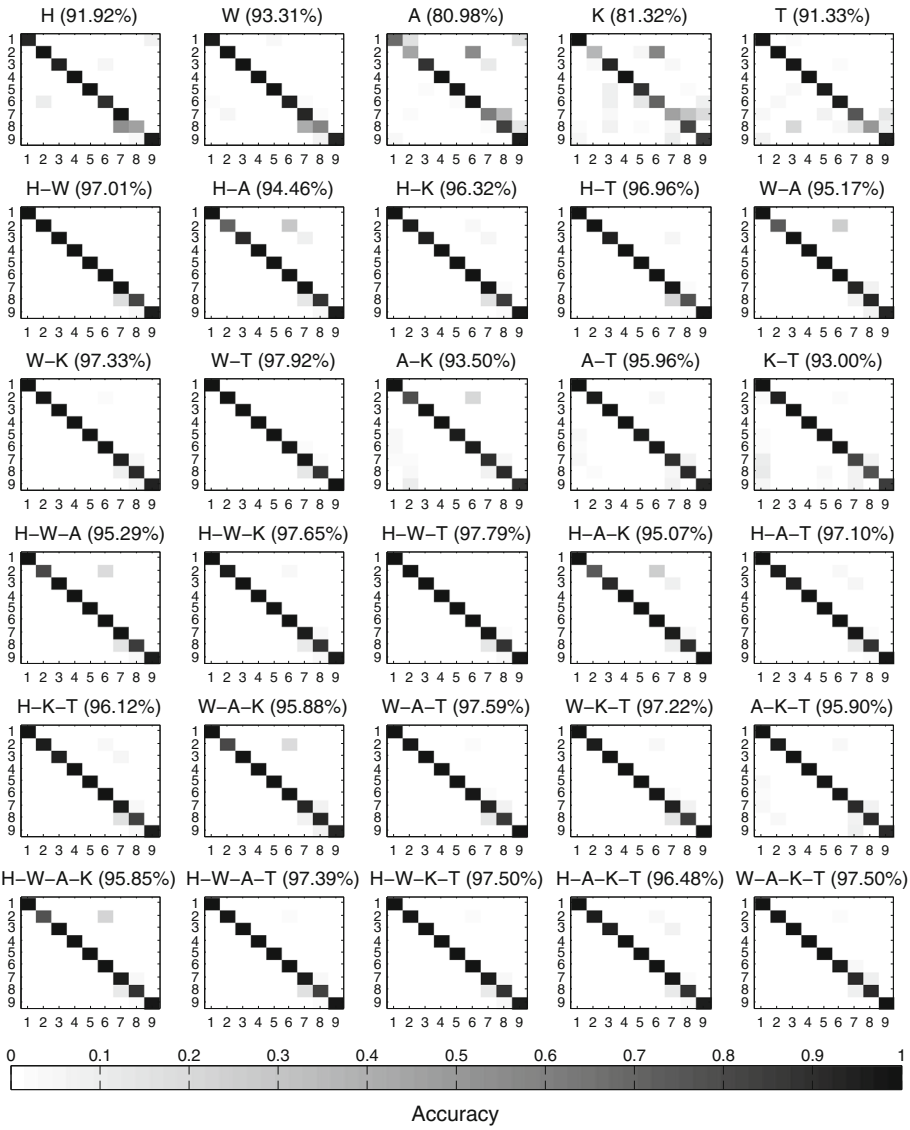

Fig. 6 Confusion matrices for the $HWC_{\alpha\beta\gamma\delta}$ model for all possible sensor setup configurations after the effect of sensor failures. KNN and the 10 features setting is used for the base classifiers. *Top* title of each confusion matrix identifies active sensors and overall accuracy (in brackets). Sensors legend *H* hip, *W* wrist, *A* arm, *K* ankle, *T* thigh. Confusion matrix legend (activities) *1* walking, *2* running, *3* cycling, *4* sitting, *5* standing, *6* lying down, *7* stretching, *8* strength-training, *9* climbing stairs

with KNN and ten features is employed, given its remarkable recognition capabilities in ideal conditions (Figs. 4, 5). The results of this evaluation are summarized in Fig. 6.

In case one sensor stops delivering data, the performance remains almost similar to what is observed when all sensors are available (>97 % accuracy). For example, only a 2 % drop is seen when the sensors placed on the thigh (T) or wrist (W) are missing. A high

tolerance is also observed for two non-operating sensors, subtly lower than in case of having one missing sensor. Not only may the HWC cope with failures on two sensors, but also on three of them. In fact, there is almost no significant drop on the performance for some combinations of active sensors (W–T, W–K, H–W), while less than a 4 % is seen for the rest. In the worst case scenario, the AR system must rely on data captured through a single active sensor. As a consequence, the recognition capabilities are seen to reduce, although in a different way for each sort of sensor. For example, the accuracy is superior to 91 %, i.e., less than 6 % from baseline, for setups in which only W, H or T remain operative. The performance decays to approximately 81 % when A or K are the only functioning sensors. This is strictly related to the informativeness of the body parts on which these sensors are correspondingly mounted. Obviously, in the rare case of becoming all sensors unavailable, e.g., given a simultaneous battery discharge, no operation may be expected. Therefore, it has not been explicitly evaluated here.

4.5 Tolerance to Sensor Faults

Conversely to the case of sensor failures, a faulty sensor is capable of delivering data. Nevertheless, sensor faults generally entail signal degradation. As introduced in Sect. 2, when the sensor circuitry is not adequately supplied a reduction of the sensor dynamic range may be observed. This shortening translates into a change in the boundaries of the signal space, thus potentially leading to a misrepresentation of the measured motion data (see Fig. 1). For example, the sensors used in this study are capable of converting all measured accelerations within the range $[-10g, 10g]$, with $g = 9.8$ m/s^2 the gravitational constant and the sign representing the direction of the acceleration. Therefore, if the dynamic range reduces to a tenth of the original interval (i.e., $[-1g, 1g]$), most of the digitized signals will likely not represent the actual measured body motion.

The tolerance of HWC and standard AR models to this phenomena is here analyzed. To that end, the performance of the recognition systems is evaluated for two scenarios. In the first case, the dynamic range is reduced to a 30 % of the original one (i.e., $[-3g, 3g]$). In this new signal space, active exercises leading to abrupt movements and high accelerations, are expected to be misrepresented, whereas, low intensity activities will in principle not suffer relevant variations. In the second more challenging scenario, the dynamic range is reduced to a 10 % of the predefined interval (i.e., $[-1g, 1g]$). Here, changes in all measured activity patterns are envisioned. It must be noted that these scenarios are neatly selected after inspecting the considered dataset: highest acceleration values are around $5g$, therefore no relevant influence is expected when changing the dynamic range beyond $\pm 5g$.

For experimental purposes, the changes in the dynamic range are simply modeled through a thresholding process. Thus, those measurements that fall out of the bounds of the new considered dynamic range are set to the extreme values of this new range. Similar parameters to the considered for the study of the tolerance to sensor failures are here used for the AR models (i.e., KNN and ten best ranked features, which render highest average performance for all AR models in absence of anomalies, Fig. 5). For the SARC model, the dynamic range of the corresponding utilized sensor is modified. For multi-sensor approaches (i.e., FFMARC, HD, MV and HWC), various configurations with an increasing number of faulty sensors are evaluated. The anomalous sensors are randomly selected from one iteration to another, but for the case in which all sensors are faulty. The evaluation results are presented in Table 1.

A considerable performance worsening is seen for most AR systems when the dynamic range is reduced to a 30 % of the original one. SARC models are clearly the most sensitive to sensor faults. Nonetheless, practical differences are observed among the sensors considered

Table 1 Average (standard deviation) accuracy values obtained for each ARC approach for diverse number of anomalous sensors. KNN and 10 features is used for both standard and base classifiers

AR model\#faulty sensors	0	1	2	3	4	5
New dynamic range = 30 % original dynamic range						
SARC (hip)	81.52 ± 4.56	66.15 ± 3.78	–	–	–	–
SARC (wrist)	87.52 ± 5.12	53.51 ± 6.31	–	–	–	–
SARC (arm)	80.16 ± 3.16	57.83 ± 7.23	–	–	–	–
SARC (ankle)	82.52 ± 3.79	58.16 ± 8.12	–	–	–	–
SARC (thigh)	88.52 ± 2.03	71.98 ± 4.21	–	–	–	–
FFMARC	**97.39 ± 1.69**	88.31 ± 4.01	76.14 ± 4.79	61.15 ± 8.36	42.39 ± 11.15	39.15 ± 13.16
HD	89.84 ± 2.57	85.35 ± 4.16	79.77 ± 8.96	68.21 ± 13.15	59.17 ± 16.14	52.75 ± 20.07
MV	82.07 ± 6.17	79.29 ± 5.36	66.74 ± 7.12	43.21 ± 10.11	36.29 ± 14.79	31.47 ± 19.02
$HWC_{\alpha\beta\gamma\delta}$	96.34 ± 2.34	**95.68 ± 2.17**	**92.77 ± 3.78**	**86.21 ± 5.22**	**73.34 ± 8.16**	**65.36 ± 13.98**
New dynamic range = 10 % original dynamic range						
SARC (hip)	81.52 ± 4.56	21.36 ± 11.18	–	–	–	–
SARC (wrist)	87.52 ± 5.12	17.78 ± 9.37	–	–	–	–
SARC (arm)	80.16 ± 3.16	26.31 ± 14.13	–	–	–	–
SARC (ankle)	82.52 ± 3.79	21.16 ± 7.18	–	–	–	–
SARC (thigh)	88.52 ± 2.03	19.98 ± 6.41	–	–	–	–
FFMARC	**97.39 ± 1.69**	69.16 ± 5.39	41.26 ± 8.12	17.23 ± 15.13	21.07 ± 10.86	18.19 ± 8.94
HD	89.84 ± 2.57	79.96 ± 6.31	59.49 ± 13.12	41.92 ± 11.69	**29.75 ± 17.25**	21.16 ± 15.78
MV	82.07 ± 6.17	77.16 ± 6.01	46.19 ± 11.16	38.21 ± 9.98	27.18 ± 12.87	**26.36 ± 8.37**
$HWC_{\alpha\beta\gamma\delta}$	96.34 ± 2.34	**94.23 ± 1.79**	**86.77 ± 6.03**	**53.21 ± 21.84**	27.18 ± 16.65	25.12 ± 19.21

Bold values represent the top accuracy for each category

in this study. Concretely, those sensors placed on body locations that are subject to lower accelerations demonstrate more robust, here hip and thigh, with a performance drop around 15 % with respect to baseline. On the contrary, wrist, arm and ankle sensors suffer from a higher reduction on their performance, which is of 35 % at worst. The reason is that sensors worn on the extremities are normally subject to higher accelerations, especially during the execution of intense activities such as running or cycling, and also walking to a lesser extent. These accelerations values are prone to fall out of the bounds defined by the new dynamic range. More tolerance to sensor faults is gained when using multi-sensor configurations. Nevertheless, once again not all models behave similarly. FFMARC appears to be the most vulnerable multi-sensor model to changes in the dynamic range, especially when two or more sensors malfunction. More than 20 % drop with respect to the baseline is observed when two sensors are affected, 35 % for three anomalous sensors and more than 50 % for four or more faulty sensors. DFMARC approaches turn to be the most tolerant. MV demonstrates capable of dealing with changes on a sole sensor, but low tolerance to anomalies in a plurality of sensors (>40 % performance drop from baseline). More robustness is shown by the HD model when top-ranked sensor classifiers are not affected by anomalies; however, as the number of erroneous sensors increases, also grows the possibility of having a faulty high-ranked sensor. At worst conditions, the performance of the HD model is observed to drop up to 40 % from baseline. The high standard deviation for the HD results may be explained since the anomalous sensors are randomly selected from one iteration to another, thus leading to diverse configurations in which the designated faulty sensors may be either low-ranked (good HD performance) or high-ranked (poor HD performance). From all evaluated models, the HWC clearly stands out as the most robust approach to sensor faults. In fact, almost no worsening is detected when a minority of the sensors are affected, and only a 10 % drop is seen when three faulty sensors are considered. The performance reduces to approximately 70 % accuracy when the complete set of sensors function defectively.

A much higher impact is seen when the dynamic range is shorten to a 10 % of the default interval. Here, the performance of SARC models plummet to negligible values. This is also seen for FFMARC. Just for only one anomalous sensor the accuracy drops more than 25 %, above 50 % for two affected sensors and nearly 80 % when all sensors are faulty. The performance also declines more abruptly for the DFMARC approaches. HD shows a similar tendency to what was seen for the former fault scenario. MV provides a reasonable tolerance to one faulty sensor, but no practical utility when two or more behave anomalously. Again, the most robust approach is the HWC, which shows almost no worsening for one erroneous sensor, and an acceptable drop for two faulty sensors. Yet, the model is not capable of overcoming the effects of a majority of faulty sensors in this complex scenario.

5 Discussion

5.1 Performance in Ideal Conditions

This work does not aim at delving into the capabilities of AR systems devised for ideal settings. In fact, much research has been already performed to this respect during the last years, and good models are available. The evaluation in ideal conditions was rather planned to compare the recognition capabilities of the proposed HWC with respect to other well-known standard AR approaches. Moreover, these results serves to this work as a baseline for the performance of this model in absence of sensor anomalies. From the evaluation, it can be concluded that the use of multi-sensor configurations is especially recommended to

ensure a high quality of recognition. SARC models demonstrate limited recognition capabilities, particularly when the systems are designed to be simple. On the other hand, FFMARC proves to be the most reliable approach from tested, especially when reduced feature sets are just considered. Traditional DFMARC models, such as HD and MV, also show interesting capabilities when building on complex feature sets, but not of much utility when kept simple. Through combining the main advantages of HD and MV approaches, the HWC model manages to achieve recognition characteristics similar to FFMARC, even for simple feature sets. This makes from the HWC a valid approach for AR in idealistic conditions.

5.2 Tolerance to Sensor Failures

Major changes in on-body sensor setups are normally produced by critical sensor failures. At worst, sensors may get broken or damaged to an extent they stop delivering data. Similar situations may be observed when a user leaves the sensor behind, it gets out of battery or it is powered down. Under these circumstances, standard AR models devised for steady sensor configurations are prone to fail to provide recognition capabilities. In fact, SARC and FFMARC models cannot strictly operate, thus they present no other option than stopping the monitoring process until the setup is recovered to its original or default state. Halting the recognition process could be unacceptable for some applications (e.g., elderly fall detection, freezing of gait in Parkinson, or epileptic seizures detectors), and especially burdensome and discouraging for general users of AR applications.

Decision fusion models are seen to be a valid solution to help not interrupt the AR process. In fact, since DFMARC models operate on the individual decisions provided by each sensor classifier or entity, modifications in the sensor network are in principle supported. Although applicable, HD and MV were qualitatively shown to be sensitive to these changes. Conversely, the HWC demonstrates very robust to sensor failures. In fact, the model practically maintains the recognition capabilities even when a majority of the sensors are missing. At worst, when a sole sensor remains operational, the performance is similar or even higher to the obtained through a SARC approach, thus demonstrating the potential of the HWC even for single sensor setups, as well as its notable scalability.

Not only should the AR models be capable of coping with the effects of occasional sensor failures but to facilitate user maintenance tasks. Thus for example, to provide a means to continue operating while a discharged sensor is being recharged is especially important in realistic applications. As concluded before, the HWC may help to provide a seamless recognition capabilities, thus supporting quotidian real-world situations in which part of the sensing infrastructure is temporarily unavailable.

5.3 Tolerance to Sensor Faults

Although less damaging than critical failures, sensor faults may also lead to a potential malfunctioning of AR systems. Conversely to the formers, faulty sensors are capable of delivering data, albeit this information is subject to degradation. This is the case of changes in the sensor dynamic range due to an inadequate energy supply of the device. The most reduced the dynamic range becomes, the higher the impact this anomaly has. Nevertheless, changes in the dynamic range produce a different impact on each activity pattern. Intense activities that involve a high body motion are primarily distorted, since their acceleration values may potentially fall outside the bounds of the new data range. This is the case of running, cycling or walking for the activity set considered in this work. On the other hand,

sedentary or low motion activities could remain unaffected if the variation of the dynamic range is not much important, although can be also distorted if this reduces dramatically.

The performance of SARC models considerably declines when the considered sensor suffer from a moderate reduction in its predefined dynamic range. The highest performance worsening is seen for those SARC models operating on data collected from body parts subject to intense accelerations (i.e., body extremities), which fits in well with previous conclusions. The use of various sensors may help overcome the effects of sensor faults; however, not all multi-sensor models demonstrate similar robustness. FFMARC models are capable of partially coping with changes in one of the sensors, but show low tolerance to two or more faulty sensors. Artifacts introduced by individual faulty sensors contaminate the complete aggregated feature vector, therefore leading to misclassifications. The effects are more prominent as the number of anomalous sensors increases. DFMARC models benefit from the independent processing of each sensor classifier. HD and MV demonstrate capable of facing the challenge of one faulty sensor; however, their recognition capabilities considerably drop when two or more sensors behave anomalously. From all tested models, the HWC shows the best fault-tolerance. In fact, almost no worsening is observed when three or less sensors are affected. The performance reduces when most sensors are distorted, although it is still higher to what is achieved through other AR models in better circumstances.

When the dynamic range is more severely reduced, SARC and FFMARC models demonstrate almost as useless as for the case of sensor failures. HD and MV also show little resilience to the effects of critical sensor faults, even when a single sensor is affected, thus of doubtful utility. Only the HWC demonstrates a strong tolerance to sensor faults, perfectly dealing with the situation of a faulty sensor and moderately coping with the effects of two anomalous sensors. Nonetheless, when a plurality of sensors are affected the HWC approach neither overcomes the effects of severe changes in the sensor dynamic range.

5.4 HWC Advantages

The HWC model was originally devised to cope with the effects of sensor failures and faults. To this respect, promising capabilities have been already demonstrated along this discussion. Nevertheless, the HWC possesses other remarkable properties of ensemble models [16], which are especially required in AR systems for the real-world. These properties are discussed next.

SARC and FFMARC models cannot function when a sensor breaks or disappears from the original sensor topology, whereas DFMARC models may keep the recognition process by using the information provided by the remaining active sensors. In this way, DFMARC models in general and the HWC in particular allow for an uninterrupted AR. However, returning the system to its initial performance require to replace or substitute the failure sensor with a new one, possibly of different characteristics (e.g., different calibration or signal modality). In this case, SARC and FFMARC could be newly utilized, but first a complete retraining of the model is needed. Conversely, decision fusion models only require to train the sensor classifier that operates on the new sensor. This is a very valuable characteristic in the AR domain, since depending on the complexity of the recognition problem, the retraining of the systems may take a significant time.

The HWC also demonstrates to scale well to the number of used sensors. It has been shown that the HWC provides good results for diverse sensor setups, even for combinations of a reduced set of sensors or a sole single device. From this, the HWC proves to not be only useful for multi-sensor configurations but also applicable in single sensor setups.

The flexibility of the HWC does not only applies at the sensor level but also at the activity level. AR systems are normally devised for a set of particular activities; however, the actions of interest may change in the course of time depending on the particular user and application needs. For example, additional activities to the originally planned may be required when a new exercise routine is considered or a workout plan modified. These changes are not only seen to add new activities but also remove some of these at the point of need. This is found of special interest to reduce the complexity of the systems and increase their recognition performance, as well as to procure systems personalization to subjects. Standard AR model require a complete redefinition of the system when the activity set is varied. Conversely, the HWC may support this sort of reconfiguration. For the inclusion of new activities, only new base classifiers must be trained for the added activities, and their associated weights computed. If an activity is rather removed, an update of the model weights is only required. These properties are eligible to support important requirements of real-world AR systems such as self-configuration, auto-adaptation and evolvability.

Two weighting models were evaluated in this work, a unified weighting for both insertions and rejections ($\text{HWC}_{\alpha\gamma}$) and an asymmetric model to weight them independently ($\text{HWC}_{\alpha\beta\gamma\delta}$). Although both models showed good classification properties, the second weighting approach demonstrates a higher potential. Through independently weighting insertions and rejections the HWC becomes more problem-sensitive, therefore capable of leveraging all base classifiers, even when their classification or rejection capabilities might be unbalanced. According to the weights, these could be defined through diverse criteria. In this work, accuracy (α,γ) and sensitivity-specificity (α,β,γ,δ) metrics has been particularly considered; however, an important asset of this model is that other performance metrics may be likewise used.

5.5 Open Issues

The comparison with previous work turns to be difficult since the effects of sensor failures and faults have been seldom investigated in the AR domain. Moreover, there is no gold standard and also a clear lack of datasets for benchmarking AR models. To compensate all this, a comparison of the capabilities of the HWC with the most widely used AR solutions has been provided. Moreover, in order to ensure the reproducibility of our experiments, the models have been evaluated on a dataset extensively employed in past works. Anyway, a strong effort must be put in the wearable AR domain to collect new datasets that may serve to validate these and future contributions.

The models used here to emulate the effects of sensor technological anomalies represent quite precisely what can be observed in a realistic setting. In fact, no differences are expected for critical anomalies or sensor failures. However, it would be valuable to not only simulate their impact but further observe them in a real-world scenario.Unfortunately, this is not an easy task since sensor failures appear in a random and occasional manner. An approximation to this could be the dataset collected in [42]. Here the authors gathered multimodal AR data, in which sensors are sometimes switched off but principally for energy saving reasons. Packets loss are also reported in this dataset; however, these are normally associated to missing data from turned-off sensors. Again, long-term AR datasets including realistic sensor anomalies could be worth to benchmark these and new models.

The HWC has been clearly demonstrated as the most robust approach. Nevertheless, for a plurality of faulty sensors the system reduces its performance, which may be more or less critical depending on the magnitude of the fault. To overcome this, a error detection procedure could facilitate to exclude the decisions yielded by faulty sensors. In this line, a

recent work [44] proposed the use of distance measures and information theory techniques to identify erroneous measurements in a multi-sensor setup. The HWC could leverage this type of mechanism to not only identify the damaged sensors, but also update the corresponding weights (γ_m and/or δ_m), thus reducing their impact on the eventual yielded decision.

Not only could changes in the sensor setup be incorporated in the HWC model but also at the activity level. Sensor anomalies may affect the recognition of part of the activities (e.g., intense activities when the dynamic range is reduced), but not alter the identification capabilities for others. Then, instead of reducing the decision weight at the network level (γ_m, δ_m) this could be rather performed at the sensor level (α_{mn}, β_{mn}), so only base classifiers weights from those unrecognizable activities are modified. This updating procedure is not only devised to overcome the limitations imposed by sensor technological anomalies but may be also utilized to dynamically adapt the AR system to people changing conditions.

Finally, it is worth noting that the complexity of the HWC model depends on the magnitude of the AR problem. Therefore, if several activities and sensors are considered, the model may require from a considerable set of decision entities or base classifiers. Nevertheless, these are very simple models that may potentially benefit from parallel computing, something that cannot be that easily applied to other standard models.

6 Conclusions

Classic AR systems assume steady sensor setups that remain invariant during the lifelong use of the system. Nevertheless, as other electronic devices, on-body sensors are subject to faults and failures. These technological anomalies lead to changes in the sensor data streams, which are normally unforeseen during the design phase and unpredictable at runtime. Consequently, models trained on ideal signal patterns may react in an undesired manner to anomalous sensor data. This potentially translates into a partial or total malfunctioning of the AR systems.

This work extensively explored the effects of both sensor failures and faults. Standard AR approaches based on single sensor configurations and multi-sensor feature fusion demonstrate low tolerance against sensor technological anomalies. Classic multi-sensor decision fusion models show a higher robustness, although they are of limited utility when more than one sensor behave abnormally.

Taking into consideration the limitations of current AR approaches, a novel model, the hierarchical weighted classifier (HWC), has been presented. The model implements for each sensor a set of base or activity classifiers, whose decisions are asymmetrically weighted according to their recognition capabilities and further fused. In ideal conditions, the HWC renders a performance similar to the best standard AR models. More importantly, it also proves to remarkably deal with sensor failures, and a high fault-tolerance when a minority of the sensors are affected. Nonetheless, when a plurality of sensors are affected the HWC approach neither overcomes the effects of severe sensor faults. In such case, identifying the anomalous sensors may be of much utility to temporarily leave them out of the recognition process. The flexibility of the HWC may also help to overcome these extreme situations through the dynamic reconfiguration of the model.

Acknowledgments This work was partially supported by the HPC-Europa2 project (no. 228398), the Spanish CICYT Project SAF2010-20558, Junta de Andalucia Project P09-TIC-175476 and the FPU Spanish Grant AP2009-2244.

References

1. Adidas® (2013) Adidas micoach. http://micoach.adidas.com/
2. Albert MV, Toledo S, Shapiro M, Kording K (2012) Using mobile phones for activity recognition in Parkinsons patients. Front Neurol 3:158
3. Altun K, Barshan B (2010) Human activity recognition using inertial/magnetic sensor units. In: Salah AA, Ruiz-del-Solar J, Çetin M, Oudeyer P-Y (eds) Human behavior understanding. Proceedings of the third international workshop, HBU 2012, Vilamoura, Portugal, October 7, 2012. Lecture notes in computer science, pp. 38–51. Springer, Berlin
4. Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Statist Surv 4:40–79
5. Atallah L, Lo B, King R, Yang GZ (2011) Sensor positioning for activity recognition using wearable accelerometers. IEEE Trans Biomed Circ Syst 5(4):320–329
6. Avci A, Bosch S, Marin-Perianu M, Marin-Perianu R, Havinga P (2010) Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: a survey. In: 23rd international conference on architecture of computing systems, pp 1–10
7. Banos O, Damas M, Pomares H, Prieto A, Rojas I (2012) Daily living activity recognition based on statistical feature quality group selection. Expert Syst Appl 39(9):8013–8021
8. Banos O, Damas M, Pomares H, Rojas F, Delgado-Marquez B, Valenzuela O (2013) Human activity recognition based on a sensor weighting hierarchical classifier. Soft Comput 17:333–343
9. Banos O, Galvez JM, Damas M, Pomares H, Rojas I (2014) Window size impact in human activity recognition. Sensors 14(4):6474–6499
10. Banos O, Toth MA, Damas M, Pomares H, Rojas I (2014) Dealing with the effects of sensor displacement in wearable activity recognition. Sensors 14(6):9995–10023
11. Bao L, Intille SS (2004) Activity recognition from user-annotated acceleration data. Perv Comput 23:1–17
12. Bouten C, Koekkoek K, Verduin M, Kodde R, Janssen J (1997) A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity. IEEE Trans Biomed Eng 44(3):136–147
13. Breiman L, Spector P (1992) Submodel selection and evaluation in regression. the x-random case. Int Statist Rev 60(3):291–319
14. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theor 13(1):21–27
15. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines: and other kernel-based learning methods. Cambridge University Press, New York
16. Dietterich TG (2000) Ensemble methods in machine learning. In: Proceedings of the first international workshop on multiple classifier systems, pp 1–15. Springer, London
17. Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley-Interscience, New York
18. Fitbit® (2013) Fitbit products. http://www.fitbit.com/es/store
19. He Z, Jin L (2009) Activity recognition from acceleration data based on discrete consine transform and svm. In: IEEE international conference on systems, man and cybernetics, pp 5041–5044
20. Hidalgo® (2012) Equivital eq02. http://www.equivital.co.uk/products/tnr/sense-and-transmit
21. Huynh T, Blanke U, Schiele B (2007) Scalable recognition of daily activities with wearable sensors. In: Location-and context-awareness. Springer, Berlin, pp 50–67
22. Jawbone® (2013) Jawbone up. https://jawbone.com/up/international
23. Jiang M, Shang H, Wang Z, Li H, Wang Y (2011) A method to deal with installation errors of wearable accelerometers for human activity recognition. Physiol Measure 32(3):347
24. Kailas A (2012) Capturing basic movements for mobile platforms embedded with motion sensors. In: International conference of the IEEE engineering in medicine and biology society, pp 2480–2483
25. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on artificial intelligence, pp 1137–1143. San Francisco, CA, USA
26. Kusserow M, Amft O, Gubelmann H, Troester G (2010) Arousal pattern analysis of an olympic champion in ski jumping. Sports Technol 3(3):192–203
27. Kwapisz JR, Weiss GM, Moore SA (2011) Activity recognition using cell phone accelerometers. ACM SigKDD Explor 12(2):74–82
28. Lara O, Labrador M (2012) A survey on human activity recognition using wearable sensors. IEEE Commun Surv Tutor 99:1–18
29. Lester J, Choudhury T, Kern N, Borriello G, Hannaford B (2005) A hybrid discriminative/generative approach for modeling human activities. In: Proceedings of the 19th international joint conference on artificial intelligence, pp 766–772. San Francisco, CA, USA
30. Mannini A, Intille SS, Rosenberger M, Sabatini AM, Haskell W (2013) Activity recognition using a single accelerometer placed at the wrist or ankle. Med Sci Sports Exer 45(11):2193–2203

31. Mäntyjärvi J, Himberg J, Seppanen T (2001) Recognizing human motion with multiple acceleration sensors. In: Proceedings of the international IEEE conference on systems, man and cybernetics, pp 747–752
32. Mathie MJ, Coster ACF, Lovell NH, Celler BG (2004) Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement. Physiol Measure 25(2):1–20
33. Maurer U, Smailagic A, Siewiorek D, Deisher M (2006) Activity recognition and monitoring using multiple sensors on different body positions. In: International workshop on wearable and implantable body sensor networks, pp 113–116
34. Maurtua I, Kirisci PT, Stiefmeier T, Sbodio ML, Witt H (2007) A wearable computing prototype for supporting training activities in automative production. In: 4th international forum on applied wearable computing
35. Najafi B, Aminian K, Paraschiv-Ionescu A, Loew F, Bula CJ, Robert P (2003) Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. IEEE Trans Biomed Eng 50(6):711–723
36. Nike® (2013) Nike + running. http://nikeplus.nike.com/plus/products/gps_app/
37. Nike® (2013) Nike + sportwatch. http://nikeplus.nike.com/plus/products/sport_watch/
38. Parkka J, Ermes M, Korpipaa P, Mantyjarvi J, Peltola J, Korhonen I (2006) Activity classification using realistic data from wearable sensors. IEEE Trans Inf Technol Biomed 10(1):119–128
39. Pirttikangas S, Fujinami K, Seppanen T (2006) Feature selection and activity recognition from wearable sensors. In: Third international symposium ubiquitous computing systems. LNCS vol 4239, pp 516–527. Springer, Berlin
40. Preece SJ, Goulermas JY, Kenney LPJ, Howard D, Meijer K, Crompton R (2009) Activity identification using body-mounted sensors—a review of classification techniques. Physiol Measure 30(4):1–33
41. Ravi N, Mysore P, Littman ML (2005) Activity recognition from accelerometer data. In: Proceedings of the seventeenth conference on innovative applications of artificial intelligence, pp 1541–1546
42. Roggen D, Calatroni A, Rossi M, Holleczek T, Förster K, Tröster G, Lukowicz P, Bannach D, Pirkl G, Ferscha A, Doppler J, Holzmann C, Kurz M, Holl G, Chavarriaga R, Creatura M, del Millán JR (2010) Collecting complex activity data sets in highly rich networked sensor environments. In: 7th international conference on networked sensing systems, pp 233–240
43. Roggen D, Magnenat S, Waibel M, Tröster G (2011) Wearable computing: designing and sharing activity-recognition systems across platforms. IEEE Robot Automat Mag 18(2):83–95
44. Sagha H, Bayati H, del Millan JR, Chavarriaga R (2013) On-line anomaly detection and resilience in classifier ensembles. Pattern Recognit Lett 34(15):1916–1927
45. Sazonov E, Fulk G, Sazonova N, Schuckers S (2009) Automatic recognition of postures and activities in stroke patients. In: International conference of the IEEE engineering in medicine and biology society, pp 2200–2203
46. Selles R, Formanoy M, Bussmann J, Janssens P, Stam H (2005) Automated estimation of initial and terminal contact timing using accelerometers; development and validation in transtibial amputees and controls. IEEE Trans Neural Syst Rehab Eng 13(1):81–88
47. Stiefmeier T, Roggen D, Ogris G, Lukowicz P, Tröster G (2008) Wearable activity tracking in car manufacturing. IEEE Perv Comput Mag 7(2):42–50
48. Stone M (1977) Asymptotics for and against cross-validation. Biometrika 64(1):29–35
49. Sun L, Zhang D, Li B, Guo B, Li S (2010) Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations. In: Yu Z, Liscano R, Chen G, Zhang D, Zhou X (eds) Ubiquitous intelligence and computing. Proceedings of the 7th international conference, UIC 2010, Xi'an, China, October 2010. LNCS pp. 548–562. Springer, Berlin
50. Theodoridis S, Koutroumbas K (2008) Pattern recognition, 4th edn. Academic, San Diego
51. UnderArmour® (2014) Armour39. http://www.underarmour.com/shop/us/en/armour39
52. Ward JA, Lukowicz P, Gellersen HW (2011) Performance metrics for activity recognition. ACM Trans Intell Syst Technol 2(1):6:1–6:23
53. Zwartjes D, Heida T, van Vugt J, Geelen J, Veltink P (2010) Ambulatory monitoring of activities and motor symptoms in Parkinson's disease. IEEE Trans Biomed Eng 57(11):2778–2786