Fusion of Multimodal Sensor Data for Effective Human Action Recognition in the Service of Medical Platforms

Panagiotis Giannakeris, Athina Tsanousa, Thanasis Mavropoulos, Georgios Meditskos, Konstantinos Ioannidis, Stefanos Vrochidis, and Ioannis Kompatsiaris

Centre for Research and Technology Hellas, Information Technologies Institute, Thessaloniki, Greece
{giannakeris,atsan,mavrathan,gmeditsk,kioannid,stefanos,ikom}@iti.gr

Abstract. In what has arguably been one of the most troubling periods of recent medical history, with a global pandemic emphasising the importance of staying healthy, innovative tools that shelter patient well-being gain momentum. In that view, a framework is proposed that leverages multimodal data, namely inertial and depth sensor-originating data, can be integrated in health care-oriented platforms, and tackles the crucial task of human action recognition (HAR). To analyse person movement and consequently assess the patient’s condition, an effective methodology is presented that is two-fold: initially, Kinect-based action representations are constructed from handcrafted 3DHOG depth features and the descriptive power of a Fisher encoding scheme. This is complemented by wearable sensor data analysis, using time domain features and then boosted by exploring fusion strategies of minimum expense. Finally, an extended experimental process reveals competitive results in a well-known benchmark dataset and indicates the applicability of our methodology for HAR.

Keywords: Action recognition · Sensor fusion · Depth sensors · Wearable sensors

1 Introduction

Considering the biological and psychological challenges that contemporary, urban mainly, settings pose for many people who are used to leading fast-paced but sedentary lives, it becomes apparent that maintaining a healthy lifestyle comprising mental and physical activities, as well as adequate rest is of paramount importance. Attaining the correct balance of activities is a task that greatly benefits from the latest advances in technologies such as pervasive sensors, artificial intelligence, human and health monitoring and assistive living [9]. They aid in the efficient logging of sleep/activity data [17] and thus the effective organisation of people’s routines via reminders/motivation actions and suggestions [24].
Particularly in unconventional circumstances, such as the present Covid-19 era, that people need to apply socially distancing criteria in all of their activities, often having to cope with the unavailability of experts, physical activity self-assessment via sensor-based methods is crucial. Specifically in the field of medicine, data analysis coming from small, low cost, high performance sensors has been providing researchers the tools to develop efficient and versatile methods of assisting patients, in order to improve their lifestyle. People in need of monitoring tend to be more autonomous and less attached to their caretakers, when having access to personalised activity information. Knowing that reliable mechanisms, such as automatic push notifications in case of patient fall, are in place to ensure timely intervention, it provides obvious benefits to both their physical state, and mental state and sense of self-sufficiency. Passive patient monitoring is an incontrovertible area of application of the abovementioned systems, where patients with mental diseases like dementia can be supervised to avoid or prevent potentially hazardous circumstances.

In the present work, focus is placed on monitoring certain well-defined actions/human movements, usually pertaining to a rehabilitation scenario, by fusing inertial and depth sensor data, since the technique has proven to provide excellent results, while the required training data are easily obtainable. To this end, manually crafted features are extracted first, from depth and sensor modalities and are adapted for our HAR framework. Then, multimodal data analysis is evaluated, with particular attention to fusion mechanisms of minimal expense. Specifically, several classification algorithms on inertial and visual sensors were applied, both separately and with two different fusion strategies, in order to recognise 27 human actions of the UTD-MHAD multimodal dataset \[5\]. The contribution of the paper could be summarised in the following:

- The methodology of Fisher encoding with 3DHOG depth features is adapted for HAR and evaluated. Time domain features based on inertial sensors are also evaluated.
- Two inexpensive fusion strategies (one feature-level and one decision-level) are deployed and performance comparisons are made between the two as well as the separate modalities.
- Extensive evaluation of several classifiers is performed with numerous evaluation protocols on a well-known multimodal HAR benchmark dataset.

Corresponding analysis results could be integrated into unified multi-user-oriented medical platforms, servicing both patients \[23\] and caretakers\[1\].

## 2 Related Work

Human action recognition in the context of Ambient Assisted Living (AAL) is facilitated by a variety of sensors, which may include inertial, range and magnetic sensors, depth and RGB cameras and even atypical modality type sensors, such as:

\[1\] https://www.gatekeeper-project.eu/.
as electrocardiogram ones [22]. The multitude of existing sensor technologies is supplemented by respective analysis methodologies. Diverse studies elaborate on modern machine learning HAR approaches, such as the one found in [25] or [31] that focuses on deep learning, state-of-the-art techniques. Moreover, in [11] distinct neural networks are exploited for depth and inertial sensing before decision-level fusion is performed. However, to leverage the performance improvement of deep learning, large amounts of training data and computational resources are often required.

Kinect revolutionised the field by providing an easily accessible and affordable tool, capable of synchronised skeleton, depth and RGB data provision without the need for additional post-processing. Since its introduction in the consumer market, researchers wholeheartedly embraced it and exploited its capabilities to present novel methods of tackling HAR [8, 19, 21, 30]. Despite the justified attention it gathered and the promising results, concerns are expressed regarding installation/setup complexity and computational efficiency [18]. In addition, privacy issues due to the RGB data are raised nowadays more often than before. As a consequence, many studies focus primarily on depth and skeleton information to deal with data anonymisation requirements.

A common denominator when talking about inertial sensing, is the use of accelerometers and gyroscopes, and depending on the field of application [1, 10], they may be complemented by more specialised sensors, such as magnetometers or barometric altimeters. Applications and trends favourable to inertial sensing are illustrated in [2], which also includes details on the history of devices and predictions on future directions. An in-depth view of the most important features and technologies, coupled with significant drawbacks governing typical gyroscope and accelerometer outputs is provided in [26].

Since certain real life challenges are impossible to be tackled just by one modality, approaches that combine the two have also been tested with promising results and helped overcome certain otherwise insurmountable issues [4, 12, 32]. Three main fusion directions exist that apply to most HAR approaches and each is performed at a different workflow step [6]: (a) data-level fusion, (b) feature-level fusion, and (c) decision-level fusion. Data-level fusion corresponds to the concatenation of raw data as they are directly collected from the respective sensors. Feature-level fusion (early fusion) is performed after features have been extracted from raw data and entails fusion of retrieved feature sets. Lastly, decision-level fusion (late fusion) combines the results of individual sensors after the classification has been completed.

Depending on the problem, different fusion mechanisms and theories have been attempted, such as exploitation of Hidden Markov Models (HMM) for hand gesture recognition [20] to tackle different modality synchronisation issues or the Dempster-Shafer theory for late (decision-level) fusion for action recognition in [3]. The former methodology [20] reported individual recognition sensing accuracy of 84% (Kinect) and 88% (inertial), while the concatenated model achieved accuracy of 93%. In the latter [3], early (feature-level) fusion is achieved by merging each sensor’s individually extracted feature sets (first represented as vectors
and then normalised) before the classification process is activated. Reported scores varied between 2–23% compared to the individual ones. Similar improvements are exhibited in [33] when the authors combine ear-worn sensors and RGB-D (Red, Green, Blue and Depth) to perform walking analysis. Moreover, an ensemble of binary one-vs-all neural network classifiers is explored in [13] to improve indoor human action recognition robustness, which once trained, is able to be effortlessly embedded on portable devices. Furthermore, a task that benefits greatly (2–8% improvement) from sensor data fusion is identified in [16], which describes an approach that leverages an SVM classifier and combines depth maps with accelerometer data to perform fall detection.

3 Methodology

3.1 Inertial Sensors

One wearable inertial sensor was used to record human actions in UTD-MHAD dataset [5], which is used in this work. The sensor provided recordings of acceleration, angular velocity and magnetic strength. To perform the analysis on the inertial sensor signals, the features suggested in [14], a paper that conducts experiments on the same dataset, were extracted. Firstly the magnitude of the raw signals of accelerometers and gyroscopes was calculated, using the formula in Eq. 1, where \( a \) stands for the signal values of each axis. For the preprocessing stage, a moving window average for each 3 rows of data was applied. Following, three features were extracted from the filtered signal vectors of each axis and of the calculated magnitude. More specifically, the mean of each vector (Eq. 2), the average of the absolute first difference of each signal vector \( a \) (Eq. 3), as well as the average of the corresponding second difference of the signal vectors \( a \) (Eq. 4) were calculated. Analysis was performed on accelerometer and gyroscope signals, as well as on their concatenated features.

\[
\begin{align*}
\text{a}_\text{mag} &= \sqrt{a_x^2 + a_y^2 + a_z^2} \\
\text{mean} &= \frac{1}{N} \sum \text{a}(n) \\
\text{mean}_{fd} &= \frac{1}{N} \sum |\text{a}(n) - \text{a}(n - 1)| \\
\text{mean}_{sd} &= \frac{1}{N} \sum |\text{a}(n + 1) - 2\text{a}(n) + \text{a}(n - 1)|
\end{align*}
\]

3.2 Depth Sensors

Local Features. In order to extract features from depth videos, the well-established efficiency of the HOG descriptor (Histograms of Oriented Gradients) was leveraged. The process was performed on 3D volumes, as in [14], to
capture spatio-temporal features that encode the actor’s body shape and limp movements that happen when an action is performed. The 3DHOG descriptors are calculated based on the gradient magnitude responses in the horizontal and vertical directions of frames. Next, the responses are aggregated over spatio-temporal blocks of pixels. A histogram of gradient responses, quantised into 8 bins (8 orientations), is constructed for each block and the responses of all pixels in that block are assigned linearly into neighboring bins. Finally, the histograms of a neighborhood of blocks are concatenated together to form a local 3DHOG descriptor. Our method is different in that aspect from the approach of [28] or [14], and does not result in 3D chunks of perfectly neighboring blocks. Instead, in order to speed up the calculations, strided sampling was applied, where a fixed number of pixels are skipped before the next block is taken. The blocks were chosen to have a size of $15 \times 15$ pixels in space, and 20 frames in time, as in [14]. The 3D chunks are created with the concatenation of $3 \times 3$ blocks in space and 2 blocks in time, and the stride parameter is set to 5 pixels on all directions. Therefore, each chunk is compiled by 18 histograms ($3 \times 3 \times 2$ blocks), resulting in a 144-dimensional 3DHOG descriptor. Finally, the local 3DHOGs are L1-normalised and reduced to half their size (70 components) using PCA.

**Action Representation.** The local 3DHOG descriptor’s dimensionality depends on the choices for the spatial and temporal dimensions of the concatenation chunks and is fixed in a given setting (144 reduced to 70 after PCA). However, the number of local 3DHOG descriptors extracted in a sequence can be arbitrary and is determined by the duration of each video, which is not the same for every sequence in the dataset. Thus, we ought to apply a method that will allow us to aggregate the set of collected local 3DHOGs to a final fixed size meaningful representation for each sequence.

In order to build the final descriptors, Fisher encoding is applied, which is proven to be a more efficient and powerful method to synthesise action representations compared to other bag-of-words techniques [7, 28, 29]. First, a visual vocabulary based on the most prominent visual clues of the whole depth sequence is built. The computation of the most discriminating samples is performed by applying unsupervised clustering (Gaussian Mixture Model (GMM)) in the shallow representation hyperspace, as formed by the feature collection of each depth sequence.

Let $\{\mu_j, \Sigma_j, \pi_j; j \in \mathbb{R}^L\}$ be the set of parameters for $L$ Gaussian models, with $\mu_j, \Sigma_j$ and $\pi_j$ standing respectively for the mean, the covariance and the prior probability weights of the $j$th Gaussian. Assuming that the $D$-dimensional 3DHOG descriptor is represented as $x_i \in \mathbb{R}^D; i = \{1, \ldots, N\}$, with $N$ denoting the total number of descriptors, Fisher encoding is then built upon the first and second order statistics:
\[ f_{1j} = \frac{1}{\sqrt{\pi_j}} \frac{1}{N} \sum_{i=1}^{N} q_{ij} \sigma_j^{-1} (x_i - \mu_j) \]

\[ f_{2j} = \frac{1}{\sqrt{2\pi_j}} \frac{1}{N} \sum_{i=1}^{N} q_{ij} \left[ \frac{(x_i - \mu_j)^2}{\sigma_j^2} - 1 \right] \]  

where \( q_{ij} \) is the Gaussian soft assignment of descriptor \( x_i \) to the \( j^{th} \) Gaussian and is given by:

\[ q_{ij} = \frac{\exp \left[ -\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right]}{\sum_{t=1}^{L} \exp \left[ -\frac{1}{2} (x_i - \mu_t)^T \Sigma_j^{-1} (x_i - \mu_t) \right]} \]

Distances, as calculated by Eq. 5, are next concatenated to form the resulting Fisher vector, \( F_X = [f_{11}, f_{21}, \ldots, f_{1L}, f_{2L}] \). Finally, L2 and power normalisation is applied to all Fisher vectors.

### 3.3 Sensor Fusion

For the fusion of depth and inertial sensors, both early and late fusion schemes were deployed. Accelerometer and gyroscope features were combined with the features extracted from the depth videos. In order to combine the heterogeneous sources at feature level (early fusion), the sensor data were first L2-normalised and then concatenated with the Fisher vectors. To perform late fusion, the probability vectors of the predicted classes were combined by averaging: using the same classifier, the probabilities obtained from inertial and depth modalities were averaged and the class with the highest averaged probability was assigned to each test case. The amount of actions included in the dataset would not favour other forms of late fusion, like weighted late fusion, that compute weights based on the classification metrics of each class. The additional cost of fusing the modalities is low, given that concatenation and averaging calculations are simple as well as highly parallelizable.

### 4 Experiments and Results

#### 4.1 Dataset and Evaluation Description

The evaluation of our methods was performed on a well-known public multimodal dataset for action recognition, \textbf{UTD-MHAD} [5]. This dataset provides captured data for 27 different types of actions, carried out by 8 subjects (4 female, 4 male), performing 1 to 4 trials for each action. The set contains in total 861 samples. Please refer to [5] for a detailed description and the full class list. This is a challenging dataset because it contains a high number of classes with substantial variability. Specifically, only about 30 samples correspond to each class on average.
Table 1. Performance of inertial sensors.

|                | Sbj Generic | Sbj Specific | Cross Sbj |
|----------------|-------------|--------------|-----------|
|                | Acc+Gyro    | Acc          | Acc+Gyro  | Acc          | Acc+Gyro    | Acc        | Gyro      | Gyro      |
| LDA            | 0.787       | 0.665        | 0.806     | 0.806       | 0.890       | 0.786      | 0.609     | 0.637     |
| kNN            | 0.476       | 0.515        | 0.476     | 0.806       | 0.876       | 0.437      | 0.500     | 0.437     |
| NB             | 0.597       | 0.499        | 0.506     | 0.527       | 0.424       | 0.384      | 0.586     | 0.474     | 0.523     |
| RF             | 0.693       | 0.569        | 0.592     | 0.913       | 0.813       | 0.828      | 0.656     | 0.527     | 0.567     |
| LSVM           | 0.618       | 0.469        | 0.602     | 0.802       | 0.846       | 0.802      | 0.574     | 0.460     | 0.574     |
| KSVM           | 0.318       | 0.418        | 0.336     | 0.602       | 0.792       | 0.599      | 0.346     | 0.451     | 0.351     |

In our effort to comply with all the evaluation scenarios that have been previously proposed for this dataset, we conduct our experiments based on three different evaluation protocols: a) subject-generic protocol, where each subject was used once as a test set. b) The subject-specific protocol, where each subject was examined separately. For each subject, two of the trials constitute the training set and the other two trials form the test set. c) The cross-subject protocol, where the models are trained on half of the subjects (1, 3, 5, 7) and tested on the other half (2, 4, 6, 8). The respective results refer to the average accuracy of all rounds of experiments. The classification algorithms evaluated in this work are: Linear Discriminant Analysis (LDA), k-Nearest Neighbours with 1 neighbour (k-NN), Naive Bayes (NB), Random Forests (RF), Linear Support Vector Machine (LSVM) and Kernel SVM (KSVM) with quadratic kernel. We also experimented with a higher number of neighbours for the k-NN classifier, but, the accuracy dropped significantly, mainly because the training set is small relative to the high number of classes.

4.2 Inertial Sensor Performance Analysis

The recordings of the wearable inertial sensor were tested for their performance together and separately. As seen in Table 1, which presents the accuracy levels of all experiments of the three evaluation scenarios, we cannot draw conclusions on which scheme performs best, as it seems that this varies depending on the classifier. In case of the subject specific evaluation scenario, the combination of accelerometer and gyroscope performs better. This is not the case though in the other two evaluation scenarios, where there are classifiers that produce better results using the readings of the one sensor only. Such observations are usually reported in relevant studies, where there is always present heterogeneity caused by different subjects, different sampling frequencies or even different placement of sensors. Another reason would be the number of actions recorded in the current dataset. Regarding the performance of the classification algorithms, LDA and RF produced the best accuracy levels. The experiments reproduced from the baseline paper [14] did not yield the same results, probably because of a misconception in the description of the evaluation or feature extraction steps.
Table 2. 8-fold cross validation using various GMM vocabulary sizes.

|          | 4 words | 8 words | 16 words | 32 words | 64 words |
|----------|---------|---------|----------|----------|----------|
| LDA      | 0.886   | 0.959   | 0.973    | **0.979**| 0.962    |
| kNN      | 0.926   | 0.957   | 0.968    | **0.980**| 0.979    |
| NB       | 0.792   | 0.828   | **0.853**| 0.851    | 0.838    |
| RF       | 0.921   | 0.938   | 0.956    | **0.965**| 0.954    |
| L SVM    | 0.902   | 0.956   | 0.976    | **0.990**| 0.986    |
| KSVM     | 0.011   | 0.008   | 0.008    | **0.015**| 0.005    |

Table 3. Performance of depth sensor.

|          | Sbj Generic | Sbj Specific | Cross Sbj |
|----------|-------------|--------------|-----------|
| LDA      | **0.856**   | 0.860        | 0.781     |
| kNN      | 0.572       | 0.993        | 0.458     |
| NB       | 0.796       | 0.670        | 0.681     |
| RF       | 0.826       | 0.984        | **0.809** |
| L SVM    | 0.779       | **0.998**    | 0.747     |
| KSVM     | 0.502       | 0.970        | 0.433     |

4.3 Depth Sensor Performance Analysis

To infer the optimal number of Gaussians of the GMM clustering, that is, the number of visual words of the vocabulary, an initial experiment was conducted, using 8-fold cross validation on the entire dataset with random splits. The values for the size of the codebook that were tested are: 4, 8, 16, 32 and 64 words. Table 2 shows the results. Nearly all the classifiers achieve their peak performance with 32 GMM words, therefore, the sweet spot is roughly around this value and is used in all further experiments. Table 3 shows the performance of the depth sensor for every classifier in every evaluation protocol. It can be seen that in general, LDA, Random Forests and Linear SVM perform consistently better than the others in all the tests. Moreover, the method performs better in the subject specific protocol, where there are no unseen subjects in the test set.

4.4 Sensor Fusion Performance Analysis

Figure 1 shows a comparison of the fusion approaches with the individual modalities for each evaluation protocol. In most cases the early fusion scheme performs better, or at least equal, to both the inertial and depth modalities and the late fusion scheme. This conclusion holds true for the majority of the classifiers in all tests. On the contrary, there are cases where late fusion performs worse than the separate modalities. In general, we can safely conclude that early fusion is the most appropriate technique, irrespective of the classifier.
Fig. 1. Performance comparison of individual modalities with early and late fusion.

Table 4. Comparison with state-of-the-art. I = Inertial, D = Depth, I+D = Fusion.

| Work                | Modality | Sbj Generic | Sbj Specific | Cross Sbj | 8-fold cv |
|---------------------|----------|-------------|--------------|-----------|-----------|
| Chen et al. 2015 [5]| I        | –           | –            | 0.661     | –         |
|                     | D        | –           | –            | 0.672     | –         |
|                     | I+D      | –           | –            | 0.791     | –         |
| Elmadany et al. 2015 [15]| D         | –           | –            | 0.734     | –         |
| Chen et al. 2015 [4]| I        | 0.764       | 0.883        | –         | –         |
|                     | D        | 0.747       | 0.851        | –         | –         |
|                     | I+D      | 0.915       | 0.972        | –         | –         |
| Zhang et al. 2017 [34]| D         | –           | –            | 0.844     | –         |
| Ehatisham et al. 2019 [14]| I       | –           | –            | –         | 0.916     |
|                     | D        | –           | –            | –         | 0.815     |
|                     | I+D      | –           | –            | –         | 0.970     |
| Dawar et al. 2019 [12]| I       | –           | –            | 0.815     | –         |
|                     | D        | –           | –            | 0.759     | –         |
|                     | I+D      | –           | –            | **0.892** | –         |
| Weiyao et al. 2019 [32]| D         | –           | –            | 0.887     | –         |
| Sidor et al. 2020 [27]| D        | 0.886       | 0.993        | –         | –         |
| Ours                | I        | 0.787       | 0.913        | 0.786     | 0.904     |
|                     | D        | 0.856       | 0.998        | 0.809     | 0.990     |
|                     | I+D      | 0.873       | **0.998**    | 0.853     | **0.997** |
4.5 Comparison with State-of-the-Art

Table 4 shows a detailed comparison with the state-of-the-art works in the same dataset. Our method’s results are taken from the best performing classifier on the corresponding evaluation protocol and for each one of the inertial, depth and early fusion approaches. For other works, the reported results are presented on the corresponding field, depending on what protocols have been followed. It can be seen, that our method outperforms all other works on the subject specific and 8-fold cross validation protocols. Regarding the subject generic evaluation, our early fusion technique is surpassed by the decision-level fusion of [4], despite the fact that the separate modalities in our methodology perform better. This is an indication that more sophisticated fusion may boost our results in the case of unseen subjects. Regarding the cross subject evaluation, which is the most popular protocol, our fusion technique is surpassed by the deep learning-based fusion of [12], but our depth modality scores higher. Still, our method’s early fusion scheme achieves competitive accuracy (down by a factor of 0.04) without data augmentation which is required in [12] to train deep CNNs.

5 Conclusions

In this work we have presented an effective methodology for human action recognition, based on fusion of inertial and depth data. Regarding the depth sensors, the 3DHOG and Fisher encoding methodology can produce discriminative features of actions and even compete with deep learning approaches, particularly for actions of previously seen subjects. The dataset used for this work consisted of many subjects and recorded actions. This heterogeneity seems to have affected the results of the inertial sensors’ analysis. However, any discriminative information in the features can be exploited with a simple and inexpensive early fusion, as the results suggest. LDA, Random Forests and Linear SVMs are the best choices for HAR classification using these features. Overall, there is still room for improvement regarding actions of unseen subjects, which would require robustness to arbitrary physical dimensions or specific movement patterns of subjects.

Acknowledgment. This research has been financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (T1EDK-00686) and the EC funded project GATEKEEPER (H2020-857223).

References

1. Avci, A., Bosch, S., Marin-Perianu, M., Marin-Perianu, R., Havinga, P.: Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: a survey. In: 23th International Conference on Architecture of Computing Systems 2010, pp. 1–10. VDE (2010)
2. Benser, E.T.: Trends in inertial sensors and applications. In: 2015 IEEE International Symposium on Inertial Sensors and Systems (ISISS) Proceedings, pp. 1–4 (2015)
3. Chen, C., Jafari, R., Kehtarnavaz, N.: Improving human action recognition using fusion of depth camera and inertial sensors. IEEE Trans. Hum.-Mach. Syst. 45(1), 51–61 (2015)
4. Chen, C., Jafari, R., Kehtarnavaz, N.: A real-time human action recognition system using depth and inertial sensor fusion. IEEE Sens. J. 16(3), 773–781 (2015)
5. Chen, C., Jafari, R., Kehtarnavaz, N.: UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 168–172. IEEE (2015)
6. Chen, C., Jafari, R., Kehtarnavaz, N.: A survey of depth and inertial sensor fusion for human action recognition. Multimedia Tools Appl. 76(3), 4405–4425 (2015). https://doi.org/10.1007/s11042-015-3177-1
7. Chen, C., Liu, M., Zhang, B., Han, J., Jiang, J., Liu, H.: 3D action recognition using multi-temporal depth motion maps and fisher vector. In: IJCAI, pp. 3331–3337 (2016)
8. Chen, L., Wei, H., Ferryman, J.: A survey of human motion analysis using depth imagery. Pattern Recogn. Lett. 34(15), 1995–2006 (2013)
9. Chen, Y., Le, D., Yumak, Z., Pu, P.: EHR: a sensing technology readiness model for lifestyle changes. Mob. Netw. Appl. 22(3), 478–492 (2017)
10. Collin, J., Davidson, P., Kirkko-Jaakkola, M., Leppäkoski, H.: Inertial sensors and their applications. In: Bhattacharyya, S.S., Deprettere, E.F., Leupers, R., Takala, J. (eds.) Handbook of Signal Processing Systems, pp. 51–85. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-91734-4_2
11. Dawar, N., Ostadabbas, S., Kehtarnavaz, N.: Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. IEEE Sens. Lett. 3(1), 1–4 (2019)
12. Dawar, N., Ostadabbas, S., Kehtarnavaz, N.: Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. IEEE Sens. Lett. 3(1), 1–4 (2018)
13. Delachaux, B., Rebetez, J., Perez-Uribe, A., Satizábal Mejia, H.F.: Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors. In: Rojas, I., Joya, G., Cabestany, J. (eds.) IWANN 2013. LNCS, vol. 7903, pp. 216–223. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38682-4_25
14. Ehatisham-Ul-Haq, M., et al.: Robust human activity recognition using multimodal feature-level fusion. IEEE Access 7, 60736–60751 (2019)
15. Elmadany, N.E.D., He, Y., Guan, L.: Human action recognition using hybrid centroid canonical correlation analysis. In: 2015 IEEE International Symposium on Multimedia (ISM), pp. 205–210. IEEE (2015)
16. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Programs Biomed. 117(3), 489–501 (2014)
17. Lane, N.D., et al.: Bewell: sensing sleep, physical activities and social interactions to promote wellbeing. Mob. Netw. Appl. 19(3), 345–359 (2014)
18. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutor. 15(3), 1192–1209 (2012)
19. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 9–14. IEEE (2010)
20. Liu, K., Chen, C., Jafari, R., Kehtarnavaz, N.: Fusion of inertial and depth sensor data for robust hand gesture recognition. IEEE Sens. J. 14(6), 1898–1903 (2014)
21. Liu, L., Shao, L.: Learning discriminative representations from RGB-D video data. In: Twenty-Third International Joint Conference on Artificial Intelligence (2013)
22. Masum, A.K.M., Bahadur, E.H., Shan-A-Alahi, A., Uz Zaman Chowdhury, M.A., Uddin, M.R., Al Noman, A.: Human activity recognition using accelerometer, gyroscope and magnetometer sensors: deep neural network approaches. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–6 (2019)
23. Mavropoulos, T., et al.: A smart dialogue-competent monitoring framework supporting people in rehabilitation. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp. 499–508 (2019)
24. Munson, S.A., Consolvo, S.: Exploring goal-setting, rewards, self-monitoring, and sharing to motivate physical activity. In: 2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, pp. 25–32. IEEE (2012)
25. Ramasamy Ramanurthy, S., Roy, N.: Recent trends in machine learning for human activity recognition—a survey. Wiley Interdisc. Rev. Data Mining Knowl. Discov. 8(4), e1254 (2018)
26. Shaeffer, D.K.: Mems inertial sensors: a tutorial overview. IEEE Commun. Mag. 51(4), 100–109 (2013)
27. Sidor, K., Wysocki, M.: Recognition of human activities using depth maps and the viewpoint feature histogram descriptor. Sensors 20(10), 2940 (2020)
28. Uijlings, J.R., Duta, I.C., Rostamzadeh, N., Sebe, N.: Realtime video classification using dense HOF/HOG. In: Proceedings of International Conference on Multimedia Retrieval, pp. 145–152 (2014)
29. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2013
30. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1290–1297. IEEE (2012)
31. Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L.: Deep learning for sensor-based activity recognition: a survey. Pattern Recogn. Lett. 119, 3–11 (2019)
32. Weiyao, X., Muqing, W., Min, Z., Yifeng, L., Bo, L., Ting, X.: Human action recognition using multilevel depth motion maps. IEEE Access 7, 41811–41822 (2019)
33. Wong, C., McKeague, S., Correa, J., Liu, J., Yang, G.Z.: Enhanced classification of abnormal gait using BSN and depth. In: 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks, pp. 166–171. IEEE (2012)
34. Zhang, B., Yang, Y., Chen, C., Yang, L., Han, J., Shao, L.: Action recognition using 3D histograms of texture and a multi-class boosting classifier. IEEE Trans. Image Process. 26(10), 4648–4660 (2017)