Fusion of Multimodal Sensor Data for Effective Human Action Recognition in the Service of a Medicine-Oriented Platform

Panagiotis Giannakeris[1], Athina Tsanousa[1], Thanasis Mavropoulos[1], Georgios Meditskos[1], Konstantinos Ioannidis[1], Stefanos Vrochidis[1], and Ioannis Kompatsiaris[1]

Centre for Research & 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, we propose a framework that leverages multimodal data, namely inertial and depth sensor-originating data, is integrated in a health care oriented platform, and tackles the crucial issue of detecting patient actions, such as walking, standing and jogging, or even patient falls. To analyse person movement and consequently assess the patient's condition, an efficient methodology is presented that is two-fold: initially a sophisticated Kinect-based methodology is presented that exploits 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 a robust fusion strategy that provides an effective and reliable recognition methodology. The classification accuracy reported in a well-known benchmark dataset proves that the presented approach achieves competitive results and validates the applicability and efficiency of our human action recognition (HAR) methodology.

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. 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. Corresponding analysis results will be integrated into a unified multiuser-oriented platform, servicing both patients and caretakers (to avoid voiding the blind review, citation to be added upon paper acceptance). The contribution of the paper is two-fold, since it entails multimodal data analysis for action classification, coupled with a sophisticated fusion methodology. Specifically, we applied several classification algorithms on inertial and visual sensors separately in order to recognise 27 human actions of the UTD-MHAD public dataset [5]. Two different fusion methods were also utilised to combine the acquired information of heterogeneous sensors. The contribution of the paper could be summarised in the following:

- Efficient algorithms are presented for human action recognition based on inertial and depth sensors.
- Combination of depth and inertial sensors with two types of fusion (feature-level and decision-level) and their impact in performance is assessed.
- Extensive experimental evaluation is performed using numerous classifiers and evaluation protocols on a well-known multimodal benchmark dataset.

2 Related Work

Human action analysis and detection 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 electrocardiogram ones [21]. The multitude of existing sensor technologies is supplemented by respective analysis methodologies. Diverse studies elaborate on modern machine learning HAR approaches [31], such as the one found in [23] that focuses on deep learning, transfer learning, and active learning state-of-the-art techniques. Moreover, in [10] 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.

A common denominator when talking about inertial sensing, is the use of accelerometers and gyroscopes, and depending on the field of application [9, 1], 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 (focuses on the consequences of force and charge transduction methods, and mechanical system dynamics) governing typical gyroscope and accelerometer outputs is provided in [25].

Kinect revolutionised the field by providing an easily accessible and affordable tool, capable of skeleton, depth and RGB data provision. Since its introduction in the consumer market, researchers wholeheartedly embraced it and exploited its capabilities to present novel methods of tackling HAR [8], [30], [20], [18]. Despite the justified attention it gathered and the promising results that the respective approaches delivered, concerns were expressed regarding privacy issues (due to the RGB sensor), installation/setup complexity and computational efficiency [17]. As a consequence, many studies focus primarily on depth and skeleton information, with approaches that leverage RGB data being under-developed.

The importance and glaring popularity that the action recognition task has enjoyed has led to the existence of dedicated challenges [27, 24] and varying datasets [22, 5, 15] that have been created to promote it. Various methodologies have been tried and been evaluated, mostly focusing on individual depth camera or inertial sensor performance. 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 [32, 11, 4]. Three main fusion directions exist that apply to most HAR approaches and each is performed at a different workflow step. The first is called data-level fusion, the second feature-level fusion, and the third 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. Works that relate to various aspects of action recognition [6] via depth/skeleton and inertial sensor data fusion are being detailed next.

Depending on the problem, different fusion mechanisms and theories have been attempted, such as exploitation of Hidden Markov Models (HMM) for hand gesture recognition [19] to tackle different modality synchronisation issues or the Dempster-Shafer theory for late (decision-level) fusion for action recognition in [3]. The former methodology [19] 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 [12] 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 we use in this work. The sensor provided recordings of acceleration, angular velocity and magnetic strength. To perform the analysis on the inertial sensor signals, we utilised the feature set suggested in [13], a paper that conducts experiments on the same dataset. Firstly we calculated the magnitude of the raw signals of accelerometers and gyroscopes, using the formula in Eq. 1, where a stands for the signal values of each axis. For the preprocessing stage, the authors in [13] proposed a moving window average for each 3 rows of data. Following, three features were extracted from the filtered signal vectors of each axis and of the calculated magnitude. More specifically, in accordance with [13], we calculated 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). Analysis was performed on accelerometer and gyroscope signals, as well as on their concatenated features.

$$a_{mag} = \sqrt{a_x^2 + a_y^2 + a_z^2} \tag{1}$$

$$mean = \frac{1}{N} \sum a(n) \tag{2}$$

$$mean_{fd} = \frac{1}{N} \sum |a(n) - a(n-1)| \tag{3}$$

$$mean_{sd} = \frac{1}{N} \sum |a(n+1) - 2a(n) + a(n-1)| \tag{4}$$

3.2 Depth Sensors

Local Features In order to extract features from depth videos, we leverage the well-established efficiency of the HOG descriptor (Histograms of Oriented Gradients). We apply the process on 3D volumes so as 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 horizontal and vertical directions of a given set 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 [13], and does not result in 3D chunks of perfectly neighboring blocks. Instead, in order to maximise efficiency and speed up the calculations, we apply strided sampling which skips a fixed number of pixels before taking the next block. We chose to construct blocks with a size of 15x15 pixels and 20 frames as in [13]. The 3D chunks are created with the concatenation of 3x3 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 (3x3x2 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, we apply a Fisher encoding scheme, which is proven to be a more efficient and powerful method to synthesise action representations compared to other bag-of-words techniques [29, 28, 7]. 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 $\{\mu_j, \Sigma_j, \pi_j; j \in R^L\}$ be the set of parameters for L Gaussian models, with μ_j, Σ_j and π_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 $\overline{x}_i \in R^D; i = \{1, \ldots, N\}$, with N denoting the total number of descriptors, Fisher encoding is then built upon the first and second order statistics:

$$
\begin{aligned}
f_{1j} &= \frac{1}{N\sqrt{\pi_j}} \sum_{i=1}^{N} q_{ij}\sigma_j^{-1}(\overline{x}_i - \overline{\mu}_j) \\
f_{2j} &= \frac{1}{N\sqrt{2\pi_j}} \sum_{i=1}^{N} q_{ij}[\frac{(\overline{x}_i - \overline{\mu}_j)^2}{\sigma_j^2} - 1]
\end{aligned}
\tag{5}
$$

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

$$
q_{ij} = \frac{exp[-\frac{1}{2}(x_i - \mu_j)^T \varSigma_j^{-1}(x_i - \mu_j)]}{\sum_{t=1}^{L} exp[-\frac{1}{2}(x_i - \mu_t)^T \varSigma_j^{-1}(x_i - \mu_t)]}
\tag{6}
$$

Distances, as calculated by Eq. 5, are next concatenated to form the resulting Fisher vector, $F_X = [f_{11}, f_{21}, \dots, f_{1L}, f_{2L}]$. 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, we combined the probability vectors of the predicted classes 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.

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, **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.

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 values 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 and quadratic Support Vector Machines.

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 [13] did not yield the same results, probably because of a misconception in the description of the evaluation or feature extraction steps.

Table 1: Inertial sensor performance.

	Sbj Generic			Sbj Specific			Cross Sbj		
	Acc+Gyro	Acc	Gyro	Acc+Gyro	Acc	Gyro	Acc+Gyro	Acc	Gyro
LDA	0.7875	0.6654	0.6597	0.8066	0.8063	0.8907	0.7860	0.6093	0.6372
kNN	0.4767	0.5150	0.4763	0.8068	0.8764	0.8067	0.4370	0.5000	0.4372
NB	0.5970	0.4990	0.5064	0.5270	0.4243	0.3849	0.5860	0.4744	0.5233
RF	0.6934	0.5690	0.5924	0.9139	0.8138	0.8282	0.6560	0.5279	0.5674
Linear SVM	0.6180	0.4690	0.6029	0.8022	0.8463	0.8023	0.5740	0.4605	0.5744
Kernel SVM	0.3181	0.4189	0.3366	0.6020	0.7928	0.5990	0.3465	0.4512	0.3511

4.3 Depth Sensor Performance Analysis

Specifically for the depth sensor methodology, it was first imperative to infer the optimal value for the number of Gaussians of the GMM clustering procedure.

That is, the number of visual words of the vocabulary. To this end, we conducted an initial experiment performing 8-fold cross validation on the entire dataset using random splits, with various values for the size of the codebook: 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 we hypothesise that the sweat spot is roughly around this value and we use it in all further experiments. Table 3 shows the performance of the depth sensor for every classifier for 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, our method seems to perform better in the subject specific protocol, where there are no unseen subjects in the test set.

Table 2: 8-fold cross validation with various GMM vocabulary sizes.

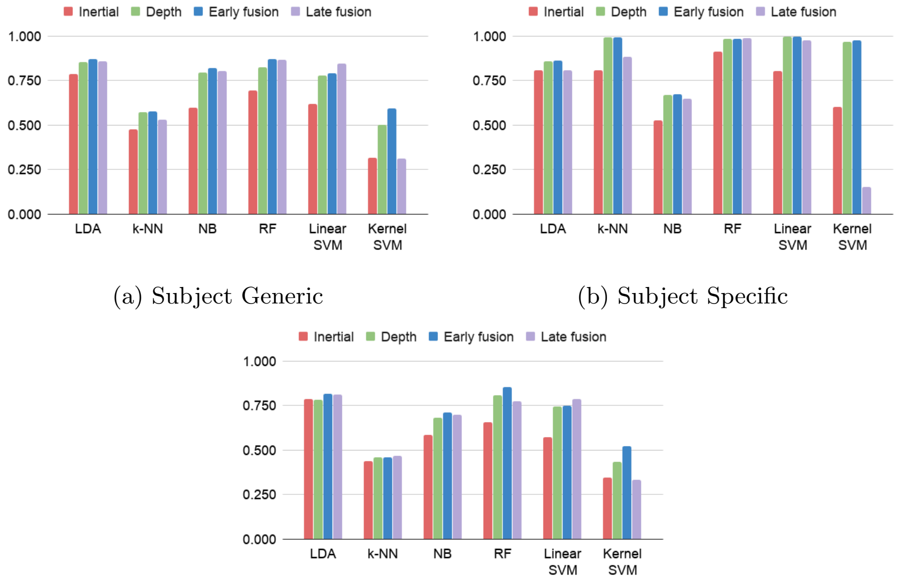
	GMM vocabulary				
	4 words	8 words	16 words	32 words	64 words
LDA	0.886	0.959	0.973	**0.979**	0.962
kNN	0.926	0.957	0.968	**0.980**	0.979
NB	0.792	0.828	**0.853**	0.851	0.838
RF	0.921	0.938	0.956	**0.965**	0.954
Linear SVM	0.902	0.956	0.976	**0.990**	0.986
Kernel SVM	0.011	0.008	0.008	**0.015**	0.005

Table 3: Depth sensor performance.

	Sbj Generic	Sbj Specific	Cross Subject
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**
Linear SVM	0.779	**0.998**	0.747
Kernel SVM	0.502	0.970	0.433

4.4 Sensor Fusion Performance Analysis

In this section we present the results of the combination of depth and inertial sensors. Accelerometers and gyroscopes were combined with the depth sensors. Figure 1 shows a comparison of the two fusion approaches with the individual modalities for each one of the three evaluation protocols. As we can see in most cases the early fusion scheme outperforms, or is equal, compared 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. Moreover, there are cases where late fusion performs worse than either one of the two sensors. In general, we can safely conclude that our early fusion technique is the best choice irrespective of the classifier used.



(a) Subject Generic

(b) Subject Specific

(c) Cross Subject evaluation

Fig. 1: Performance comparison of individual modalities with early and late fusion approaches.

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 modalities. For other works, we present the reported results on the appropriate field depending on what protocols have been followed. As we can see, 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], although the individual modalities in our methodology perform better. This indicates that a more sophisticated fusion technique is possibly needed to maximize the capabilities of our methodology. Regarding the cross subject evaluation, which is the most popular protocol, our fusion technique is surpassed by the deep learning-based fusion of [11], but our depth modality scores higher. Still, our method's early fusion scheme achieves competitive results without the data augmentation step of [11] which is needed to train deep CNNs.

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

		Evaluation protocol			
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 [14]	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 [13]	I				0.916
	D				0.815
	I+D				0.970
Dawar et al., 2019 [11]	I			0.815	
	D			0.759	
	I+D			**0.892**	
Weiyao et al., 2019 [32]	D			0.887	
Sidor et al., 2020 [26]	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**

5 Conclusions and Future Work

In this work we have presented an efficient methodology for human action recognition, based on inertial and depth data and their fusion. We have compared the early and late fusion schemes for an array of classifiers and confirmed the superiority of early fusion. Our method yields competitive results to other works without the need for deep learning or elaborate fusion schemes. However, we intend to explore these alternatives in our future work so as to push towards more accurate action classification.

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).

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 Transactions on Human-Machine Systems **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 Sensors Journal **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 and Applications **76**(3), 4405–4425 (2017)
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 Recognition Letters **34**(15), 1995–2006 (2013)
9. Collin, J., Davidson, P., Kirkko-Jaakkola, M., Leppäkoski, H.: Inertial sensors and their applications. In: Handbook of Signal Processing Systems, pp. 51–85. Springer (2019)
10. Dawar, N., Ostadabbas, S., Kehtarnavaz, N.: Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. IEEE Sensors Letters **3**(1), 1–4 (2019)
11. Dawar, N., Ostadabbas, S., Kehtarnavaz, N.: Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. IEEE Sensors Letters **3**(1), 1–4 (2018)
12. Delachaux, B., Rebetez, J., Perez-Uribe, A., Mejia, H.F.S.: Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors. In: International Work-Conference on Artificial Neural Networks. pp. 216–223. Springer (2013)
13. Ehatisham-Ul-Haq, M., Javed, A., Azam, M.A., Malik, H.M., Irtaza, A., Lee, I.H., Mahmood, M.T.: Robust human activity recognition using multimodal feature-level fusion. IEEE Access **7**, 60736–60751 (2019)
14. 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)
15. Kepski, M., Kwolek, B.: Embedded system for fall detection using body-worn accelerometer and depth sensor (09 2015). https://doi.org/10.1109/IDAACS.2015.7341404
16. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Computer methods and programs in biomedicine **117**(3), 489–501 (2014)

17. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE communications surveys & tutorials **15**(3), 1192–1209 (2012)
18. 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)
19. Liu, K., Chen, C., Jafari, R., Kehtarnavaz, N.: Fusion of inertial and depth sensor data for robust hand gesture recognition. IEEE Sensors Journal **14**(6), 1898–1903 (2014)
20. Liu, L., Shao, L.: Learning discriminative representations from rgb-d video data. In: Twenty-Third International Joint Conference on Artificial Intelligence (2013)
21. 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)
22. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley mhad: A comprehensive multimodal human action database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV). pp. 53–60. IEEE (2013)
23. Ramasamy Ramamurthy, S., Roy, N.: Recent trends in machine learning for human activity recognition—a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), e1254 (2018)
24. Ruffieux, S., Lalanne, D., Mugellini, E.: Chairgest: a challenge for multimodal mid-air gesture recognition for close hci. In: Proceedings of the 15th ACM on International conference on multimodal interaction. pp. 483–488 (2013)
25. Shaeffer, D.K.: Mems inertial sensors: A tutorial overview. IEEE Communications Magazine **51**(4), 100–109 (2013)
26. Sidor, K., Wysocki, M.: Recognition of human activities using depth maps and the viewpoint feature histogram descriptor. Sensors **20**(10), 2940 (2020)
27. Twomey, N., Diethe, T., Kull, M., Song, H., Camplani, M., Hannuna, S., Fafoutis, X., Zhu, N., Woznowski, P., Flach, P., et al.: The sphere challenge: Activity recognition with multimodal sensor data. arXiv preprint arXiv:1603.00797 (2016)
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 Recognition Letters **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 Transactions on Image processing **26**(10), 4648–4660 (2017)