A Comparative Home Activity Monitoring Study using Visual and Inertial Sensors

L. Tao, T. Burghardt, S. Hannuna, M. Camplani, A. Paiement, D. Damen, M. Mirmehdi, I. Craddock
Visual Information Laboratory, Faculty of Engineering, University of Bristol, United Kingdom
Email: {lili.tao, tb2935, sh1670, massimo.camplani, csatmp, csxda, m.mirmehdi, ian.craddock}@bristol.ac.uk

***Abstract*—Monitoring actions at home can provide essential information for rehabilitation management. This paper presents a comparative study and a dataset for the fully automated, sample-accurate recognition of common home actions in the living room environment using commercial-grade, inexpensive inertial and visual sensors. We investigate the practical home-use of body-worn mobile phone inertial sensors together with an Asus Xmotion RGB-Depth camera to achieve monitoring of daily living scenarios. To test this setup against realistic data, we introduce the challenging *SPHERE-H130 action dataset* containing 130 sequences of 13 household actions recorded in a home environment. We report automatic recognition results at maximal temporal resolution, which indicate that a vision-based approach outperforms accelerometer provided by two phone-based inertial sensors by an average of 14.85% accuracy for home actions. Further, we report improved accuracy of a vision-based approach over accelerometry on particularly challenging actions as well as when generalising across subjects.**

I. Introduction

In this paper we focus on monitoring daily household activities in the home environment. It is here where patients, for whom activity monitoring is most challenging and necessary, spend most of their time after hospital discharge. Monitoring the level and type of patients' physical activity is of general interest to clinicians across a wide variety of subjects, including obesity, diabetes, and depression-related research, as well as regarding aftercare for orthopaedic, cardiac and other surgery [1]. Traditionally, physical activity levels have been monitored using questionnaires, occasional clinical check-ups, and more recently, wearable devices – with a focus on a coarse categorisation of activity levels by wrist-worn inertial sensors [2]. To date, the use of wearable accelerometers remains a popular choice as source for inferring human activity levels due to its low cost, low energy consumption and data simplicity. Among these, triaxial accelerometers are the most broadly used motion sensors to recognise ambulation activities [3].

Visual sensor based techniques have emerged over recent years for which there exists a significant body of literature describing the inference of activities from 2D colour intensity imagery [4]. However, the increasing availability of depth-measuring sensors, especially the introduction of the Microsoft Kinect, has generated an opportunity for utilizing depth in *conjunction* with traditional RGB camera data allowing for richer and more fine-grained analysis of human activity [5].

Recent work by Chen et al. [3] presents a comparative study of such RGB-D (colour and depth imaging) sensors versus accelerometer sensors. The work also introduces a fusion approach for both modalities of data. They show that RGB-D and accelerometer data can be used to generate comparable results when tested on the Berkeley MHAD dataset [6]. This dataset is, however, recorded in a laboratory environment where most actions in the dataset are related to body exercises (jump, punch etc.).

Developing a reliable home monitoring system has drawn much attention in recent years due to the growing demands for integrated health care. Existing approaches to current home monitoring systems often include custom-fit environmental, physiological and vision sensors, such as in the SPHERE project [1]. Such systems can enable several types of application, to increase personal safety for elderly patients and to facilitate clinicians to diagnose and monitor patients. This new patient-clinician interactive mode improves the reliability and effectiveness of diagnosis to some extent, significantly shortens the travel time and hospital stay for patients, and reduces the work load for clinicians [7]. Visual sensors in particular have the potential to address several limitations of current systems [8]: they are data-rich and able to capture multiple events simultaneously, and they are easy to integrate into already existing living environments.

The paper has two key contributions. Firstly, we introduce a dataset for fine-grained action recognition within a real home environment in the SPHERE project's house [9]. The dataset, exemplified in Figure 1, contains 13 household actions performed over 10 sessions - a total of 130 sequences. The setup consists of 1) an Asus Xmotion RGB-Depth camera mounted at the corner of a living room, and 2) two three-axis mobile phone accelerometer sensors worn at the waist and the dominant wrist. Secondly, we present a comparative study towards activity monitoring using these visual and accelerometer sensors in a living environment. We outline areas where a visual approach exceeds the performance of an accelerometer sensor, showing its merits in (a) detecting particularly challenging actions, and (b) in generalising across subjects.

II. Visual Data Collection and Processing

Visual Data Collection. We simultaneously collect RGB and depth images using the commercial product Asus Xmotion. Motion information can be recovered best from RGB data as it contains rich texture and colour information. Depth information, on the other hand, reveals details of the 3D configuration. We extract and encode both motion and depth features over the area of a bounding box as returned by the human detector and tracker provided by the OpenNI SDK [10]. Figure 2 gives an overview of the feature extraction process and illustrates the motion and depth information extracted. To

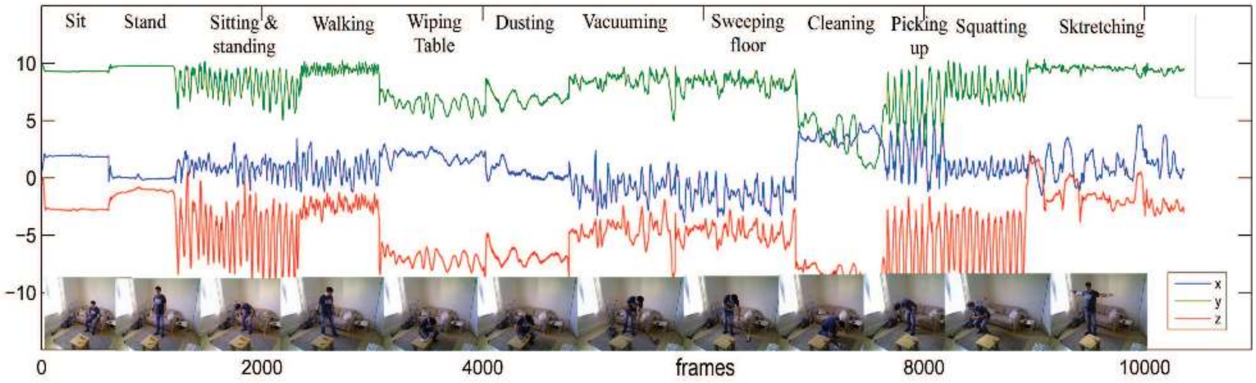


Fig. 1: Three-axis acceleration signals and sample colour images corresponding to the actions.

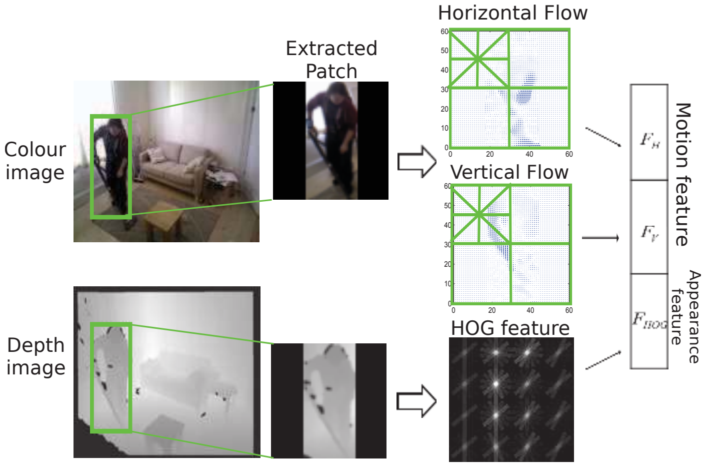


Fig. 2: Feature extraction. Feature descriptor is formed from motion and appearance information extracted from colour and depth images, respectively.

normalise the utilised image region due to varying heights of the subjects and their distance to the camera, the bounding box is scaled by fixing its longer side to $M = 60$ $[pixels]$ while maintaining aspect ratio. The scaled bounding box is then centred in a $M \times M$ square box and horizontally padded.

Motion Feature Encoding. Motion information can generally be readily extracted from this box independent of varying human appearance, Inspired by [11], optical flow measurements are taken and split into horizontal and vertical components. These are re-sampled to fit the normalised box and a median filter with kernel size 5×5 is applied to smooth the data in each direction. The normalised bounding box is divided into an $N \times N$ non-overlapping grid and the orientations of each grid cell are quantised into n_b bins. The parameters for our experiments are empirically determined as (optimal) values of $N = 2$, $n_b = 9$. The second and third rows in Figure 3 show optical flow patterns and its motion features for different actions. Here, we only show the magnitude of horizontal flow F_H and vertical flow F_V in one figure to save space.

These patterns in hand, a local motion feature descriptor $F = F_H \sqcup F_V$ is constructed by concatenating the histogram of optical flow features in each block from both orientations. To encode motion spatio-temporally, we choose a temporal window of 15 frames suggested in [6] which is approximately half a second around the current frame to be concatenated for establishing the final descriptor.

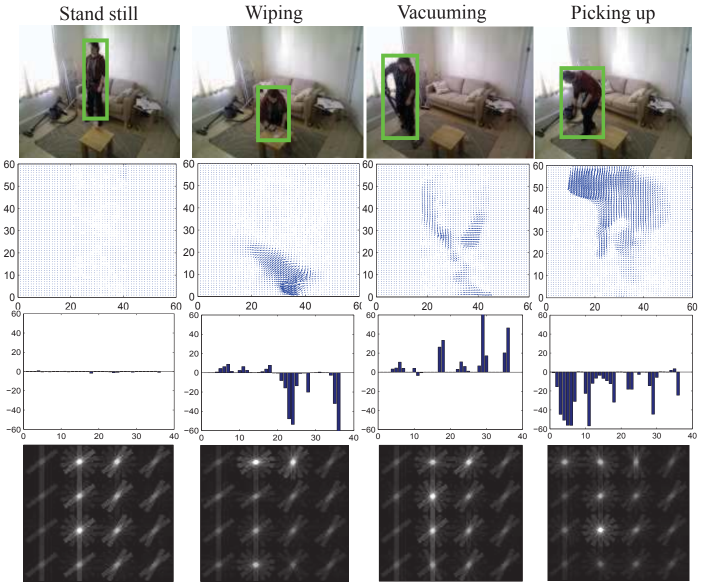


Fig. 3: First row: colour images with detected human from different action sequence. Second row: optical flows patterns. Third row: motion features. Last row: visual representation of histogram of gradients.

Depth Feature Encoding. Information computed from structured light alleviates the effect of appearance and lighting variations, allowing for independent depth recovery. For home environments with partial occlusions and unconstrained object interaction, however, we found that the performance of available skeleton trackers [12] is unreliable – specially when the subject is not facing the camera. Instead, we opt to extract features directly from depth imagery employing the histogram of gradients (HOG) feature on raw depth images [13] applied to the normalised box. The last row in Figure 3 shows the visual representation of the HOG feature for different actions. Essentially, these descriptors are able to encode a person's silhouette, its contours and the edges and depth gradients within its area.

The complete visual feature descriptor consists of appearance features extracted from the depth image of the current frame and the 15-frame motion context from colour images.

III. Inertial Data Collection and Processing

As shown in [3], accelerometers placed on wrist and waist were found to be the most effective for human action recognition. Having more sensors [6] may improve performance, but it is

Sequence ID		Sit still	Stand still	Sitting	Standing	Walking	Wiping	Dusting	Vacuuming	Sweeping	Cleaning	Picking	Squatting	Stretching	**total**
1		586	580	318	382	802	741	615	807	823	642	574	764	678	**8312**
2		834	287	332	353	952	623	664	735	759	789	536	747	621	**8232**
3		525	665	210	268	876	691	918	744	752	742	493	835	639	**8358**
4		905	872	548	672	840	1005	713	742	755	627	580	914	918	**10091**
5		609	604	517	619	714	962	756	963	1091	834	529	730	1430	**10358**
6		621	568	423	542	1014	751	750	1062	1002	724	814	799	1010	**10080**
7		557	1076	426	397	1110	359	499	841	880	615	1011	422	673	**8866**
8		682	1466	371	378	1257	827	437	1180	1294	888	417	640	1034	**10871**
9		399	442	344	432	673	759	553	892	754	933	445	849	852	**8327**
10		517	408	267	422	618	587	602	807	720	679	432	676	716	**7451**
Sum		6235	6968	3756	4465	8856	7305	6507	8773	8830	7473	5831	7376	8571	**90946**
ACC		63.02	80.18	35.36	43.56	76.95	66.04	22.59	62.16	51.57	59.60	49.73	42.34	60.21	**56.67**
Visual	HOG	24.76	35.19	42.81	58.01	79.05	54.41	15.98	64.98	57.71	90.15	58.70	25.33	51.77	**52.20**
	FLOW	80.42	73.82	82.22	83.61	83.69	63.55	41.40	65.95	61.14	79.90	70.86	65.43	52.70	**68.57**
	FLOW+HOG	76.65	77.43	83.57	85.98	85.83	67.00	29.32	75.09	68.89	88.75	77.24	60.52	60.87	**71.52**
Visual+ACC		76.71	82.46	83.65	86.67	86.91	67.32	34.01	79.92	69.43	89.01	78.70	64.05	67.89	**73.99**

TABLE I: Number of frames per sequence and action, together with recognition rate(%) of visual(HOG, FLOW feature) and accelerometer (ACC feature) data.

not convenient for participants to wear and charge many on-body sensors, especially in a daily living scenario. We opt for subjects to wear two accelerometers mounted at the centre of the waist and the dominant wrist only.

Inertial Data. Raw time series data from accelerometers is measured as $[X, Y, Z]$ vectors, where each column corresponds to acceleration in orthogonal spatial dimensions. Figure 1 illustrates readings of the accelerometer for various actions. Raw data cannot be directly utilised for action recognition, and instead three commonly used features [14] are extracted from each of the three axes for each device, giving a total of 18 attributes. These features include the first- and second-order statistics, namely the mean and the variance; we also use correlation measures between each axes pair. The features are generated from a temporal window of 30 samples taken over approximately 1 second. Mean acceleration is calculated by the average value of the signal over the window for each axis and Standard deviation captures the range of acceleration values over the same window. Correlation is calculated between each pair of axes, such as the correlation between the X and Y axis is $corr(X, Y) = \text{cov}(X, Y)/\sigma_X \sigma_Y$.

IV. Experimental Data

We introduce the *SPHERE-H130 action dataset* for human action recognition from RGB-Depth and inertial sensor data captured in a real living environment. The dataset is generated over 10 sessions by 5 subjects containing 13 action categories per session: *sit still, stand still, sitting down, standing up, walking, wiping table, dusting, vacuuming, sweeping floor, cleaning stain, picking up, squatting, upper body stretching*. Overall, recordings correspond to about 70 minutes of total time captured. Actions were simultaneously captured by the Asus Xmotion RGB-depth camera and the two wireless accelerometers. Colour and depth images were acquired at a rate of 30Hz. The accelerometer data was captured at about 100Hz sampled down to 30Hz, a frequency recognised as optimal for human action recognition [15]. Figure 1 shows an example of accelerometer data reading (waist) and sample colour images for the actions in the dataset. Note, that activities of daily living have low inter-class variability, but high intra-class variability due to diverse subjects and living habits. Figure 4 shows snapshots from "*stretching*" as an example that actions may be performed differently and can vary significantly across different subjects.

V. Experimental Results and Evaluation

For evaluation, we perform leave-one-subject-out cross validation (CV1) where final recognition results reported are averaged over all subjects to remove any bias. We utilise a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel to classify the data[1]. Table I lists the number of frames for each action and sequence. Associated with this, we show the percentage of the total number of frames for each action. The recognition rates reported in the table include appearance features only (HOG), motion features only (FLOW), appearance and motion feature fusion (FLOW+HOG), accelerometer data (ACC), and the fusion of both visual[2] and accelerometer data (Visual+ACC).

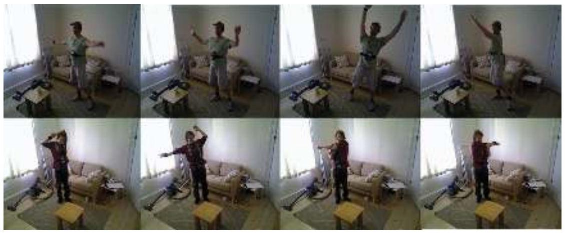


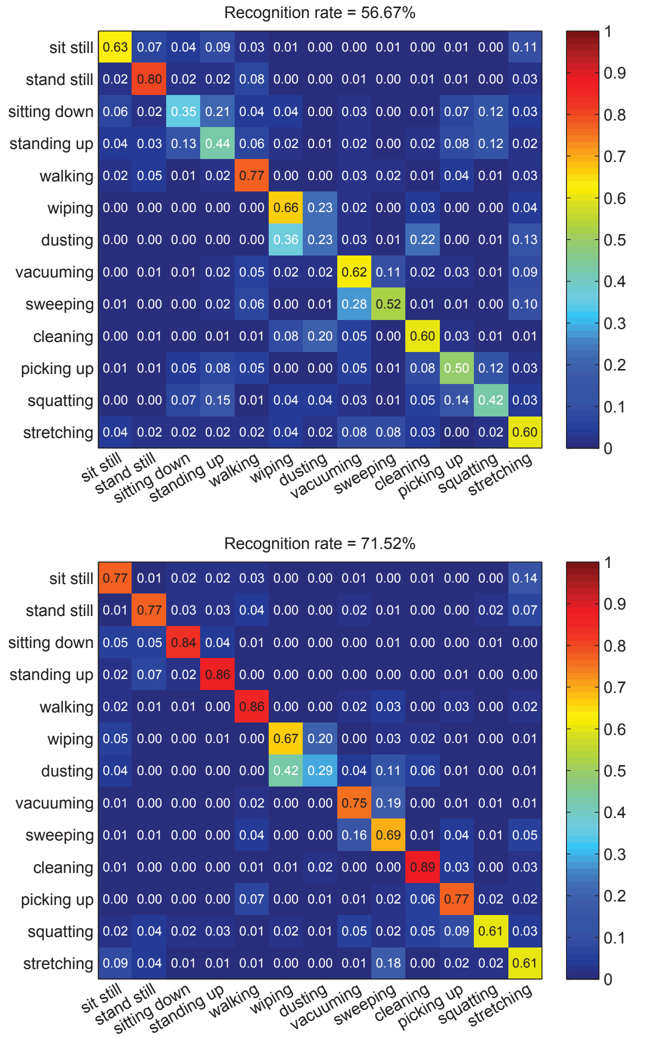
Fig. 4: Snapshots from two "Stretching" sequences. Actions may be performed differently and can vary across subjects.

In general, the overall recognition rate of visual sensors is found to be 14.85% higher than that of accelerometers. Notice that a substantial recognition improvement can be attributed to actions, for which part of the containing movements are similar. For example, "sitting down" and "standing up" are misclassified as "picking up" by accelerometers, but can be recognised by cameras due to different body poses. Some actions, e.g. wiping the table and dusting, are confused by both sensors, as these actions are performed in a very similar way. It can also be observed that a combined visual-inertial approach does not lead to a significant recognition improvement than when using only visual data. Figure 5 depicts the recognition confusion matrices corresponding to the use of inertial and visual sensors, respectively.

Activities performed by different subjects may vary significantly. To investigate the transfer of learned action descriptions across different subjects, we conduct an experiment by using one sequence for training and another sequence from the same subject for testing (CV2). The results listed in Table II show the overall recognition rate over all actions for CV1 and CV2. It is noticeable that in CV2 there is a significant improvement of accelerometer performance compared to the CV1 test, while similar results are produced by visual data and the fusion of visual and accelerometer data. In practice, it is unrealistic to have all the users' data available for system training. This demonstrates one advantage of using *visual* sensors for action

[1] The libsvm [16] implementation was used in the experiments

[2] FLOW+HOG is referred to as visual data in the following sections



Recognition rate = 56.67%

	sit still	stand still	sitting down	standing up	walking	wiping	dusting	vacuuming	sweeping	cleaning	picking up	squatting	stretching
sit still	0.63	0.07	0.04	0.09	0.03	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.11
stand still	0.02	0.80	0.02	0.02	0.08	0.00	0.00	0.01	0.00	0.01	0.01	0.00	0.03
sitting down	0.06	0.02	0.35	0.21	0.04	0.04	0.00	0.03	0.00	0.01	0.07	0.12	0.03
standing up	0.04	0.03	0.13	0.44	0.06	0.02	0.01	0.02	0.00	0.02	0.08	0.12	0.02
walking	0.02	0.05	0.01	0.02	0.77	0.00	0.00	0.03	0.02	0.01	0.04	0.01	0.03
wiping	0.00	0.00	0.00	0.00	0.00	0.66	0.23	0.02	0.00	0.03	0.00	0.00	0.04
dusting	0.00	0.00	0.00	0.00	0.00	0.36	0.23	0.03	0.01	0.22	0.00	0.01	0.13
vacuuming	0.00	0.01	0.01	0.02	0.05	0.02	0.02	0.62	0.11	0.02	0.03	0.01	0.09
sweeping	0.01	0.00	0.00	0.02	0.06	0.00	0.01	0.28	0.52	0.01	0.01	0.00	0.10
cleaning	0.00	0.01	0.00	0.01	0.01	0.08	0.20	0.05	0.00	0.60	0.03	0.01	0.01
picking up	0.01	0.01	0.05	0.08	0.05	0.00	0.00	0.05	0.01	0.08	0.50	0.12	0.03
squatting	0.00	0.00	0.07	0.15	0.01	0.04	0.04	0.03	0.01	0.05	0.14	0.42	0.03
stretching	0.04	0.02	0.02	0.02	0.02	0.04	0.02	0.08	0.08	0.03	0.00	0.02	0.60

Recognition rate = 71.52%

	sit still	stand still	sitting down	standing up	walking	wiping	dusting	vacuuming	sweeping	cleaning	picking up	squatting	stretching
sit still	0.77	0.01	0.02	0.02	0.03	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.14
stand still	0.01	0.77	0.03	0.03	0.04	0.00	0.00	0.02	0.01	0.00	0.00	0.02	0.07
sitting down	0.05	0.05	0.84	0.04	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00
standing up	0.02	0.07	0.02	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
walking	0.02	0.01	0.01	0.00	0.86	0.00	0.00	0.02	0.03	0.00	0.03	0.00	0.02
wiping	0.05	0.00	0.00	0.01	0.00	0.67	0.20	0.00	0.03	0.02	0.01	0.00	0.01
dusting	0.04	0.00	0.00	0.00	0.00	0.42	0.29	0.04	0.11	0.06	0.01	0.00	0.01
vacuuming	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.75	0.19	0.00	0.01	0.01	0.01
sweeping	0.01	0.01	0.00	0.00	0.04	0.00	0.00	0.16	0.69	0.01	0.04	0.01	0.05
cleaning	0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.00	0.00	0.89	0.03	0.00	0.03
picking up	0.00	0.00	0.00	0.00	0.07	0.00	0.01	0.01	0.02	0.06	0.77	0.02	0.02
squatting	0.02	0.04	0.02	0.03	0.01	0.02	0.01	0.05	0.02	0.05	0.09	0.61	0.03
stretching	0.09	0.04	0.01	0.01	0.01	0.00	0.00	0.01	0.18	0.00	0.02	0.02	0.61

Fig. 5: The confusion matrices by (a) accelerometer data and (b) visual data.

recognition for the home dataset at hand: visual information learned for action recognition can be more readily transferred across subjects than inertial information.

	ACC	Visual	Visual+ACC
CV1	56.67	71.52	73.99
CV2	70.16	72.12	75.01

TABLE II: Overall recognition rate (%) for CV1 vs CV2 tests.

VI. Conclusion and Future Work

This paper presented a comparative study and relevant dataset for the fully automated recognition of common home activities via inertial(accelerometer) and visual sensors. We introduced the challenging *SPHERE-H130 action dataset*[3] that covers home-typical human activities in 130 sequences of 70 minutes of multi-modal recordings. Both vision and inertial sensors are low cost, easy to operate, and suitable for deployment among different applications, residents, and households. Comparing vision and inertial sensors for actions in the home environment, results indicated that a vision-based approach outperforms body-worn accelerometer sensors by an average of 14.85% accuracy for the dataset. A combined descriptor only marginally outperformed vision descriptors. More importantly, we found that vision provides better generalisation across subjects and is able to differentiate some complex actions that accelerometry fails to decouple. We conclude that visual approaches should play a role in future monitoring systems for the home. Future work will include comparisons between different modalities for particular target variables including energy expenditure and related monitoring tasks.

Acknowledgment

This work was performed under the SPHERE IRC funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1.

References

[1] N. Zhu, T. Diethe, M. Camplani, L. Tao, A. Burrows, N. Twomey, D. Kaleshi, M. Mirmehdi, P. Flach, and I. Craddock, "Bridging ehealth and the internet of things: The SPHERE project," *IEEE Intelligent Systems*, 2015.

[2] M. Altini, J. Penders, R. Vullers, and O. Amft, "Estimating energy expenditure using body-worn accelerometers: a comparison of methods, sensors number and positioning," *Biomedical and Health Informatics, IEEE Journal of*, vol. 19, no. 1, pp. 219–226, 2015.

[3] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *Human-Machine Systems, IEEE Transactions on*, vol. 45, no. 1, pp. 51–61, 2015.

[4] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, p. 16, 2011.

[5] J. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[6] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *IEEE Workshop on Applications of Computer Vision*, 2013, pp. 53–60.

[7] B. R. Bloem, Y. Grimbergen, M. Cramer, M. Willemsen, and A. H. Zwinderman, "Prospective assessment of falls in parkinson's disease," *Journal of Neurology*, vol. 248, no. 11, pp. 950–958, 2001.

[8] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 873–10 888, 2012.

[9] P. Woznowski *et al.*, "A multi-modal sensor infrastructure for healthcare in a residential environment," in *IEEE International Conference on Communications, Workshop on ICT-enabled services and technologies for eHealth and Ambient Assisted Living*, 2015.

[10] *OpenNI User Guide*, OpenNI organization, November 2010. [Online]. Available: http://www.openni.org/documentation

[11] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *European Conference on Computer Vision*, 2008, pp. 548–561.

[12] A. Paiement, L. Tao, S. Hannuna, M. Camplani, D. Damen, and M. Mirmehdi, "Online quality assessment of human movement from skeleton data," in *British Machine Vision Conference*, 2014.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005*, vol. 1, 2005, pp. 886–893.

[14] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *AAAI*, vol. 5, 2005, pp. 1541–1546.

[15] C. V. Bouten, K. Koekkoek, M. Verduin, R. Kodde, and J. D. Janssen, "A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity," *Biomedical Engineering, IEEE Transactions on*, vol. 44, no. 3, pp. 136–147, 1997.

[16] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[3] The dataset is released on SPHERE website http://www.irc-sphere.ac.uk/work-package-2/ar