UTD-MHAD: A MULTIMODAL DATASET FOR HUMAN ACTION RECOGNITION UTILIZING A DEPTH CAMERA AND A WEARABLE INERTIAL SENSOR

Chen Chen, Roozbeh Jafari, Nasser Kehtarnavaz

Department of Electrical Engineering, University of Texas at Dallas, USA

ABSTRACT

Human action recognition has a wide range of applications including biometrics, surveillance, and human computer interaction. The use of multimodal sensors for human action recognition is steadily increasing. However, there are limited publicly available datasets where depth camera and inertial sensor data are captured at the same time. This paper describes a freely available dataset, named UTD-MHAD, which consists of four temporally synchronized data modalities. These modalities include RGB videos, depth videos, skeleton positions, and inertial signals from a Kinect camera and a wearable inertial sensor for a comprehensive set of 27 human actions. Experimental results are provided to show how this database can be used to study fusion approaches that involve using both depth camera data and inertial sensor data. This public domain dataset is of benefit to multimodality research activities being conducted for human action recognition by various research groups.

Index Terms— Multimodal human action dataset, human action recognition, fusion of depth and inertial data

1. INTRODUCTION

Human action recognition is an active research topic involving many applications such as biometrics, surveillance, human computer interaction, fitness monitoring, rehabilitation, etc. With the continuing advancements in sensor technology, human action recognition research is benefitting from the use of differing modality sensors such as RGB cameras, depth cameras, accelerometers, and gyroscopes.

In the past decade, recognizing human actions from video data captured by conventional RGB cameras has been extensively investigated. For example, space-time based methods (e.g., [1]) and motion history based methods (e.g., [2]) are popular methods developed for action recognition involving RGB cameras. There are RGB camera-based datasets such as KTH [1] and UCF50 [3] that have facilitated the comparison of different recognition approaches. Recent emergence of depth cameras (in particular, Microsoft Kinect) has made it possible to utilize depth images. Depth images have some advantages over conventional RGB images for human action recognition. Depth images are able to provide three-dimensional structure and motion information towards distinguishing different actions. They are also relatively insensitive to changes in lighting conditions. There exist human action datasets of depth images that have been created using the Kinect camera, e.g. MSR Action 3D dataset [4] and MSR Daily Activity 3D dataset [5].

With the advancement of Micro-Electro-Mechanical Systems (MEMS) and integrated circuit technologies, wearable inertial sensors such as accelerometers and gyroscopes are increasingly being utilized for human action recognition. For example, in [6] a wireless body area network composed of multiple wearable inertial sensors was developed to monitor daily activities for assisted physical rehabilitation. In [7], a wearable inertial sensor was employed to recognize the actions of hand-twist and hand-open for an intelligent medication adherence monitoring system. In [8], an action recognition framework based on a hidden Markov model (HMM) was specifically designed for the distributed architecture of body sensor networks. There are also publicly accessible datasets for human action recognition using wearable inertial sensors which have been used for comparison of recognition algorithms (e.g., UC Berkeley Wearable Action Recognition Database (WARD) [9], and USC Human Activity Dataset (HAD) [10]).

Depth cameras and wearable inertial sensors have been mostly used separately for human action recognition. In other words, the simultaneous utilization of both depth cameras and inertial sensors for action recognition has been limited in the literature. In [11, 12], data from a depth camera and an inertial sensor were fused within the framework of an HMM for robust hand gesture recognition. In [13], a fusion approach for improving human action recognition was developed based on depth and inertial sensors.

Currently, publicly available human action datasets that include both depth and inertial sensor data are the Berkeley Multimodal Human Action Database (MHAD) [14] and the University of Rzeszow (UR) fall detection dataset [15]. To facilitate research activities in multimodal sensor fusion for human action recognition, this paper also provides a multimodal human action dataset employing a Kinect depth camera and a wearable inertial sensor, named University of Texas at Dallas Multimodal Human Action Dataset (UTD-MHAD). Our dataset covers a more comprehensive set of human actions and is meant to be used for applications where the data from a depth camera and an inertial sensor are to be fused or used at the same time. Our dataset provides temporally synchronized RGB videos, depth videos, skeleton joint positions, and inertial signals for a comprehensive set of human actions. Such a dataset is of benefit to researchers working in different fields such as image processing, computer vision, wearable computing, and sensor fusion.

In the next section, two existing datasets that include both depth and inertial sensor data are briefly reviewed. Our dataset is described in section 3. Section 4 provides an example utilization of this database for human action recognition. The conclusion is finally stated in section 5.

2. EXISTING DATASETS

2.1. Berkeley MHAD

The Berkeley MHAD dataset [14] contains temporally synchronized data from a motion capture system consisting of 12 RGB cameras, 2 Microsoft Kinect cameras, 6 wearable accelerometers, and 4 microphones. The dataset consists of 659 data sequences from 11 human actions performed by 12 subjects with 5 repetitions.

Although the dataset includes data from depth cameras and accelerometers, one issue to note here is the practicality or intrusiveness associated with wearing multiple inertial sensors or accelerometers in a real-world setting. More importantly, the dataset includes 11 human actions which are rather distinct and thus distinguishable from each other, whereas our dataset includes 27 human actions, and some of which are similar.

2.2. UR Fall Detection Dataset

The UR fall detection dataset [15] focuses on human fall detection. It contains 60 depth video and color video sequences recorded with 2 Microsoft Kinect cameras as well as the data from an accelerometer. The Kinect cameras were mounted on the ceiling and the accelerometer was worn near the spine on the lower back of a subject. Only two types of actions (falling and non-falling) are included in this dataset, limiting the applicability of this dataset for evaluating recognition algorithms for other actions.

3. UTD MULTIMODAL HUMAN ACTION DATASET

3.1. Sensors

For our multimodal action dataset, only one Microsoft Kinect camera and one wearable inertial sensor were used. This was intentional due to the practicality or relatively non-intrusiveness aspect of using these two differing modality sensors. Both of these sensors are widely available, low cost, easy to operate, and do not require much computational power for the real-time manipulation of data generated by them. A picture of the Kinect camera is shown in Fig. 1(a). It can capture a color image with a resolution of 640×480 pixels and a 16-bit depth image with a resolution of 320×240 pixels. The frame rate is approximately 30 frames per second. It is also important to note that the Kinect SDK [16] is a publicly available software package which can be used to track 20 skeleton joints and their 3D spatial positions.

The wearable inertial sensor used here is the low-cost wireless inertial sensor built in the ESSP Laboratory at the University of Texas at Dallas [17]. It consists of (i) a 9-axis MEMS sensor which captures 3-axis acceleration, 3-axis angular velocity and 3-axis magnetic strength, (ii) a 16-bit low power microcontroller, (iii) a dual mode Bluetooth low energy unit which streams data wirelessly to a laptop/PC, and (iv) a serial interface between the MEMS sensor and the microcontroller enabling control commands and data transmission. For the utilization of the magnetometer, a controlled magnetic field without any distortion is required. Due to a lack of such magnetic fields in practice, only the signals associated with the 3-axis accelerometer and the 3-axis gyroscope are considered here. The wearable inertial sensor is shown in Fig. 1(b). The sampling rate of the wearable inertial sensor is 50 Hz. The measuring range of the wearable inertial sensor is ±8g for acceleration and ±1000 degrees/second for rotation. Again, for practicality reasons or the intrusiveness associated with asking subjects to wear multiple inertial sensors, only one inertial sensor is considered here, either worn on the wrist (similar to a watch) or the thigh depending on the action of interest to be recognized in a particular application.

3.2. Dataset Description

Our dataset contains 27 actions performed by 8 subjects (4 females and 4 males). Each subject repeated each action 4 times. The protocol and consent were approved by the Institutional Review Board (IRB) at the University of Texas at Dallas. After removing three corrupted sequences, the dataset includes 861 data sequences. The 27 actions performed are listed in Table 1. As seen from this table, this list constitutes a comprehensive set of human actions covering sport actions (e.g., *bowling*, *tennis serve*, and *baseball swing*), hand gestures (e.g., *draw x*, *draw triangle*, and *draw circle*), daily activities (*knock on door*, *sit to stand*, and *stand to sit*), and training exercises (e.g., *arm curl*, *lunge*, and *squat*).

(a) (b)

Fig. 1. (a) Microsoft Kinect camera; (b) wearable inertial sensor.

Table 1. Human Actions in UTD-MHAD

Wearable inertial sensor on right wrist		
1	*right arm swipe to the left*	*(swipe_left)*
2	*right arm swipe to the right*	*(swipe_right)*
3	*right hand wave*	*(wave)*
4	*two hand front clap*	*(clap)*
5	*right arm throw*	*(throw)*
6	*cross arms in the chest*	*(arm_cross)*
7	*basketball shoot*	*(basketball_shoot)*
8	*right hand draw x*	*(draw_x)*
9	*right hand draw circle (clockwise)*	*(draw_circle_CW)*
10	*right hand draw circle (counter clockwise)*	*(draw_circle_CCW)*
11	*draw triangle*	*(draw_triangle)*
12	*bowling (right hand)*	*(bowling)*
13	*front boxing*	*(boxing)*
14	*baseball swing from right*	*(baseball_swing)*
15	*tennis right hand forehand swing*	*(tennis_swing)*
16	*arm curl (two arms)*	*(arm_curl)*
17	*tennis serve*	*(tennis_serve)*
18	*two hand push*	*(push)*
19	*right hand knock on door*	*(knock)*
20	*right hand catch an object*	*(catch)*
21	*right hand pick up and throw*	*(pickup_throw)*
Wearable inertial sensor on right thigh		
22	*jogging in place*	*(jog)*
23	*walking in place*	*(walk)*
24	*sit to stand*	*(sit2stand)*
25	*stand to sit*	*(stand2sit)*
26	*forward lunge (left foot forward)*	*(lunge)*
27	*squat (two arms stretch out)*	*(squat)*

(a) (b)

Fig. 2. Placements of the wearable inertial sensor: (a) right wrist or (b) right thigh.

During data recording, the Kinect camera was placed on a tripod about 3 meters in front of the subjects to ensure that a subject's entire body appeared in the camera field of view. The wearable inertial sensor was worn on the subject's right wrist or the right thigh (see Fig. 2) depending on whether the action was mostly an arm or a leg type of action. Specifically, for actions 1 through 21, the wearable inertial sensor was placed on the subject's right wrist; for actions 22 through 27, the wearable inertial sensor was placed

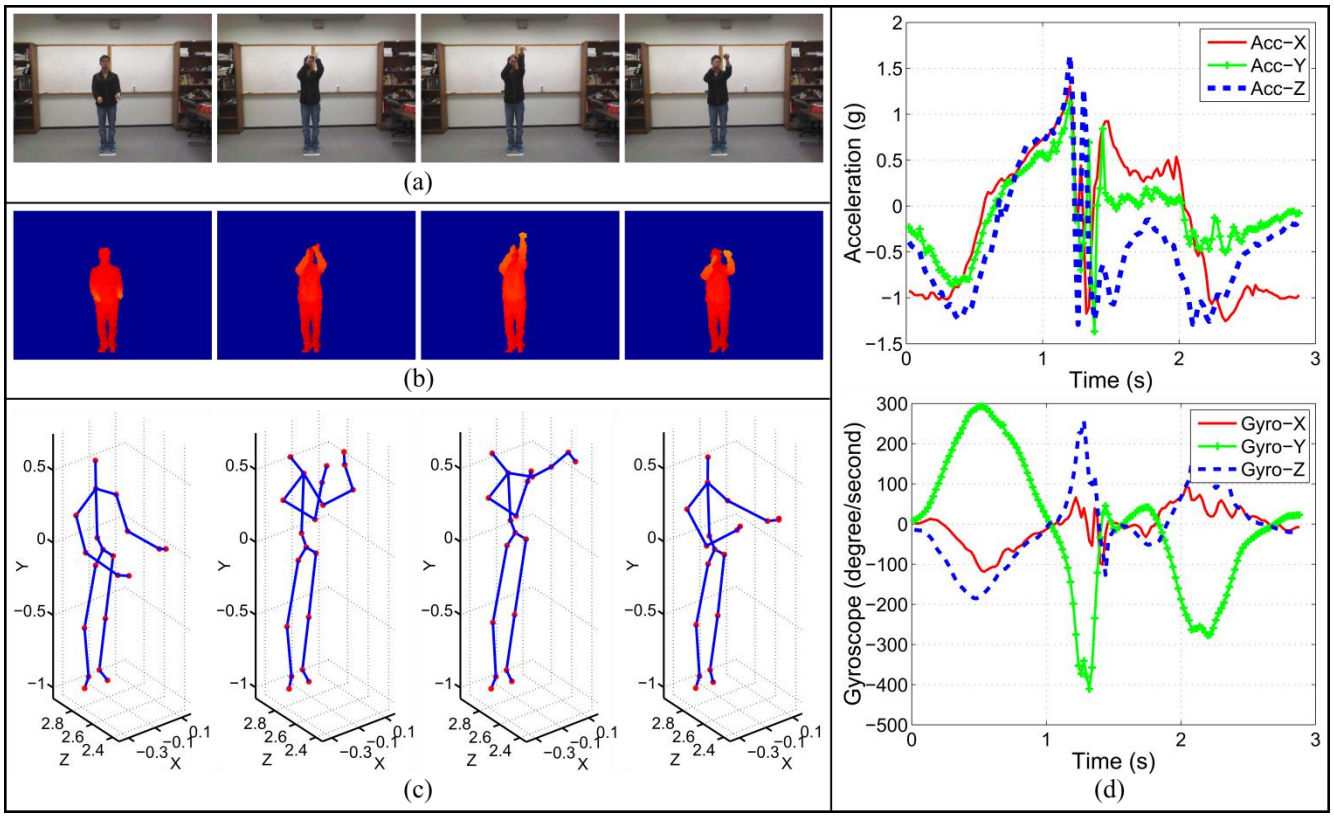

Fig. 3. An example of the multimodality data corresponding to the action *basketball-shoot*: (a) the color images, (b) the depth images (the background of each depth frame was removed), (c) the skeleton joint frames, and (d) the inertial sensor data (acceleration and gyroscope signals).

on the subject's right thigh. Each subject performed an action 4 times or 4 trials. The segmentation of each trial was conducted manually offline via visual inspection. This dataset possesses large intra-class variations due to the following reasons: (i) subjects performed the same action at different speeds in different trials, (ii) subjects had different heights, (iii) the same action was repeated in a natural way which made each trial slightly different. For example, the number of claps by a subject for the action *two-hand-front-clap* varied in different trials.

Four data modalities of RGB videos, depth videos, skeleton joint positions, and the inertial sensor signals were recorded in three channels or threads. One channel was used for simultaneous capture of depth videos and skeleton positions, one channel for RGB videos, and one channel for the inertial sensor signals (3-axis acceleration and 3-axis rotation signals). For data synchronization, a time stamp for each sample was recorded. Since the frame rate of the Kinect camera and the sampling rate of the wearable inertial sensor were different, the start and the end of an action were synchronized by using the time stamps of the depth images to serve as references.

More specifically, let the time stamp of the first depth frame (the starting frame) of an action sequence be t_D^s and the time stamp of the last depth frame (the ending frame) of an action sequence be t_D^e. Then, the two time stamps (denoted by t_I^s and t_I^e) of the inertial sensor data samples that were closest to t_D^s and t_D^e were found in order to identify the first and the last samples of the inertial sensor data. Note that the starting and ending depth frames of an action were annotated via visual inspection. An example of our multimodality data corresponding to the action *basketball-shoot* is illustrated in Fig. 3. For each segmented action trial, the color data was stored in video (*.avi*) files, and the depth, skeleton and inertial sensor data were stored using the MATLAB computing environment as three *.mat* files, respectively. As a result, four data files for an action trial are included in the dataset. The dataset can be downloaded from the link http://www.utdallas.edu/~kehtar/UTD-MHAD.html.

4. ACTION RECOGNITION USING DEPTH AND INERTIAL SENSOR FUSION

To demonstrate the utility of this multimodality dataset for human action recognition, this section provides the outcome of a data fusion approach for human action recognition when using our UTD-MHAD dataset. In the experiments conducted, the data from the subject numbers 1, 3, 5, 7 were used for training, and the data for the subject numbers 2, 4, 6, 8 were used for testing.

The existing feature extraction methods previously used to extract features from depth images and inertial sensors were considered here with the understanding that feature extraction or recognition algorithm is not the focus of this paper. Depth motion maps (DMMs) described in [18, 19] constituted the features which were extracted from the depth images. A brief explanation of these features is given here, more details appear in [18]. For a depth video sequence, all the depth frames were projected onto three orthogonal Cartesian planes to form the projected images corresponding to the three projection views - front (f), side (s), and top (t) views.

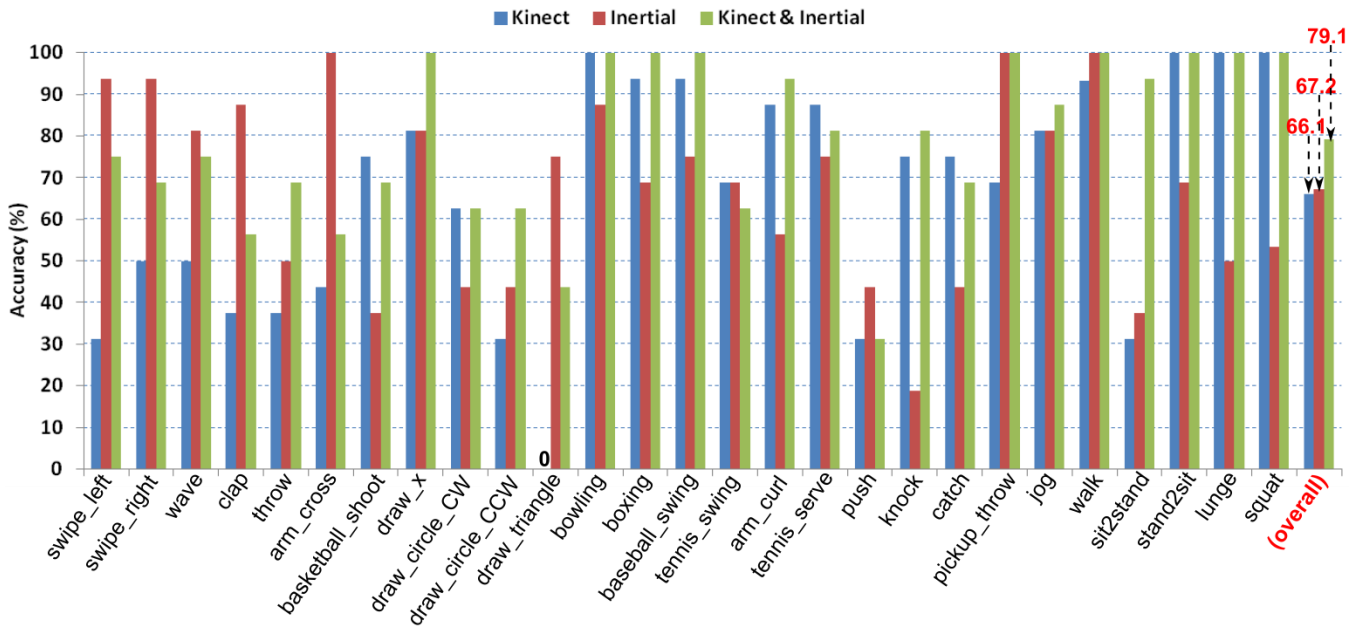

Fig. 4. Class specific accuracy and overall accuracy of the 27 UTD-MHAD human actions involving different sensor modalities when using a CRC classifier.

The absolute difference between two consecutive projected images was accumulated through the entire depth video creating three DMMs (DMM_f, DMM_s, and DMM_t) from the three projection views. In our experiments, the sizes of DMM_f, DMM_s, and DMM_t were set to 150×75, 150×100, and 100×75 as noted in [18]. The three DMMs of a depth sequence were stacked to form a feature vector. Principal component analysis (PCA) was applied to the concatenated feature vector to reduce the dimensionality. The principal components that accounted for 95% of the total variation of the training features were then used.

For the inertial sensor feature extraction, the method described in [13] was used. That is, each acceleration and gyroscope signal sequence was partitioned into N temporal windows. Statistical features including *mean*, *variance*, and *standard deviation*, were calculated for each direction per temporal window. All the features from N windows were concatenated to form a feature vector. N= 6 generated the best outcome and it was thus used in the subsequent experimentations. For the fusion approach, the feature sets from the depth camera and the inertial sensor were fused together as a single feature set before feeding it into a classifier.

For action recognition, the collaborative representation classifier (CRC) described in [20] was utilized to evaluate the effectiveness of the fusion approach. The regularization parameter λ of the CRC classifier was tuned based on a five-fold cross validation.

The recognition performance of the fusion was compared with the performance of each individual sensor modality. The results obtained are displayed in Fig. 4. As can be seem from this figure, by combining the features from the depth camera and the wearable inertial sensor, the overall recognition accuracy was improved by more than 11% over the situations when using the Kinect camera alone or the inertial sensor alone. This figure shows that the recognition accuracy of most of the actions was improved when using the fusion of depth and inertial sensor data. For example, the accuracies of the actions *right arm throw*, *draw x*, and *draw circle (counter clockwise)* were improved over 15% as compared to the situation when using the Kinect camera alone or the inertial sensor alone.

It is important to note that the accuracies of the fusion approach for some actions did not improve compared to when using the inertial sensor alone or when using the depth camera alone. This demonstrated that a fusion approach in general is helpful for those actions that generate depth and inertial data that are complementary. In other words, for those actions that a single modality sensor provides adequate discriminatory power, fusion may not provide any improvement.

5. CONCLUSION

This paper has provided a public domain dataset, named the UTD Multimodal Human Action Dataset (UTD-MHAD), for the examination and comparison of different human action recognition methods, in particular those involving fusion or using both a depth camera and an inertial sensor. The dataset includes four data modalities including RGB videos, depth videos, skeleton positions from a Kinect camera and inertial signals (acceleration and rotation signals) from a wearable inertial sensor. It incorporates 861 data sequences by 8 subjects for a comprehensive set of 27 human actions. This public domain dataset is of benefit to multi-modality research activities being conducted for human action recognition by various research groups.

6. ACKNOLEDGMENT

This work was supported in part by the National Science Foundation, under grants CNS-1150079. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations.

7. REFERENCES

[1] C. Schuldt, I, Laptev, and B. Caputo, "Recognition human actions: a local SVM approach," in *Proceedings of IEEE International Conference on Pattern Recognition*, Cambridge, UK, pp. 32-36, August 2004.

[2] J. Davis, "Hierarchical motion history images for recognizing human motion," *Proceedings of IEEE Workshop on Detection and Recognition of Events in Video*, Vancouver, BC, pp.39-46, July 2001.

[3] K. Reddy, and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971-981, 2012.

[4] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, San Francisco, CA, pp. 9-14, June 2010.

[5] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, pp. 1290-1297, June 2012.

[6] E. Jovanov, A. Milenkovic, C. Otto, and P. C. de Groen, "A wireless bodyarea network of intelligent motion sensors for computer assisted physical rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 2, no. 6, pp. 1-10, 2005.

[7] C. Chen, N. Kehtarnavaz, and R. Jafari, "A medication adherence monitoring system for pill bottles based on a wearable inertial sensor," in *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, pp. 4983-4986, August 2014.

[8] E. Guenterberg, H. Ghasemzadeh, V. Loseu, and R. Jafari, "Distributed continuous action recognition using a hidden markov model in body sensor networks," *Distributed Computing in Sensor Systems*, Springer Berlin Heidelberg, pp. 145-158, 2009.

[9] A. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy, "Distributed recognition of human actions using wearable motion sensor networks," *Journal of Ambient Intelligence and Smart Environments*, vol. 1, no. 2, pp. 103-115, 2009.

[10] M. Zhang, and A. A. Sawchuk, "USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proceedings of ACM International Conference on Ubiquitous Computing Workshop on Situation, Activity and Goal Awareness*, Pittsburgh, PA, pp. 1036-1043, September 2012.

[11] K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of inertial and depth sensor data for robust hand gesture recognition," *IEEE Sensors Journal*, vol. 14, no. 6, pp. 1898-1903, June 2014.

[12] K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz, "Multi-HMM classification for hand gesture recognition using two differing modality sensors," in *Proceedings of the 10th IEEE Dallas Circuits and Systems Conference*, Richardson, TX, pp. 1-4, October 2014.

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

[14] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley MHAD: A comprehensive multimodal human action database," In *Proceedings of IEEE Workshop on Applications of Computer Vision*, Tampa, FL, pp. 53-60, January 2013.

[15] M. Kepski, and B. Kwolek, "Fall detection using ceiling-mounted 3D depth camera," In *Proceedings of International Conference on Computer Vision Theory and Applications*, Lisbon, Portugal, pp. 640-647, January 2014.

[16] http://www.microsoft.com/en-us/kinectforwindows/

[17] C. Chen, K. Liu, R. Jafari, and N. Kehtarnavaz, "Home-based senior fitness test measurement system using collaborative inertial and depth sensors," in *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, pp. 4135-4138, August 2014.

[18] C. Chen, K. Liu and N. Kehtarnavaz, "Real-time human action recognition based depth motion maps," *Journal of Real-Time Image Processing*, August 2013, doi: 10.1007/s11554-013-0370-1.

[19] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, Waikoloa Beach, HI, pp. 1092-1099, January 2015.

[20] L. Zhang, M. Yang and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proceedings of IEEE International Conference on Computer Vision*, Barcelona, Spain, pp. 471-478, November 2011.