Real-Time Continuous Action Detection and Recognition Using Depth Images and Inertial Signals

Neha Dawar[1], Chen Chen[2], Roozbeh Jafari[3], Nasser Kehtarnavaz[1]
[1]Department of Electrical and Computer Engineering, University of Texas at Dallas, Texas, USA
Email: kehtar@utdallas.edu
[2]Center for Research in Computer Vision, University of Central Florida, Florida, USA
[3]Center for Remote Health Technologies and System, Texas A&M University, Texas, USA

Abstract— **This paper presents an approach to detect and recognize actions of interest in real-time from a continuous stream of data that are captured simultaneously from a Kinect depth camera and a wearable inertial sensor. Actions of interest are considered to appear continuously and in a random order among actions of non-interest. Skeleton depth images are first used to separate actions of interest from actions of non-interest based on pause and motion segments. Inertial signals from a wearable inertial sensor are then used to improve the recognition outcome. A dataset consisting of simultaneous depth and inertial data for the smart TV actions of interest occurring continuously and in a random order among actions of non-interest is studied and made publicly available. The results obtained indicate the effectiveness of the developed approach in coping with actions that are performed realistically in a continuous manner.**

Keywords— Real-time continuous action detection; action recognition from continuous data streams; simultaenous utilization of depth images and inertial signals for action recognition.

I. Introduction

Human action recognition is an extensively researched topic with a wide span of applications such as human-computer interaction, gaming, and rehabilitation. Different sensors or sensing modalities have been used for human action recognition including video cameras, depth cameras, and inertial sensors. For example, video camera images were used in [1] to perform human action recognition using 3D SIFT (Scale Invariant Feature Transform) based descriptors; depth camera images were used in [2] to conduct human action recognition using depth motion maps; and skeleton data from depth cameras were used in [3] to characterize different human actions. In [4-7], human action recognition was performed by using temporal and statistical features of inertial signals acquired from wearable inertial sensors. In [8-11], the data from both a depth camera and an inertial sensor were used simultaneously to achieve human action recognition with high accuracy.

It is important to note that the bulk of the research on action recognition has involved recognizing actions that are segmented into single actions. That is to say, data to process contain only a single action of interest. However, in practice, in applications such as smart TV and gaming, one needs to deal with recognizing an action of interest in real-time when there is a continuous stream of activities by a subject. In such cases, it becomes more challenging to accurately detect and recognize actions of interest. Our objective in this paper is to address this more challenging problem, that is to say, data to process are not segmented single actions but non-segmented actions of interest that appear continuously and in random order among actions of non-interest. Hence, in this work, both the detection and recognition processes are addressed within the same framework. First, actions of interest are separated from actions of non-interest and then the detected actions of interest are classified or recognized in real-time. In [12], a continuous recognition approach was discussed using only a depth camera, however, the continuous dataset used contained only actions of interest with no actions of non-interest randomly occurring between actions of interest.

This work presents an approach for real-time detection and recognition of actions of interest from a continuous data stream of activity by simultaneous utilization of a depth camera and an inertial sensor. Such data streams are considered to contain actions of interest randomly occurring among some arbitrary actions of non-interest. There are two main attributes that distinguish this work from the previous works on action recognition: (i) compared with the scenario of performing only actions of interest, a more realistic scenario of continuous activity is considered where actions of interest are performed continuously and in a random order among actions of noninterest, and (ii) a depth camera and an inertial sensor are used simultaneously for such a scenario.

The remainder of the paper is organized as follows: Section II provides a description of the two differing sensor modalities used. Section III describes the continuous dataset collected and examined for the experiments. The details of the approach developed are then provided in Section IV followed by the results and their discussion in Section V. Finally, the paper is concluded in Section VI.

II. Sensors Utilized

The sensors used in the developed approach include a Kinect v2 depth camera and a wearable inertial sensor. The Kinect camera is a depth camera that is widely used for human action recognition. A picture of this camera is shown in Fig. 1(a). The Kinect SDK [13], which is a publicly available software package, provides 3D spatial positions of 25 skeleton body joints that are derived from depth images. Fig. 1(c) shows

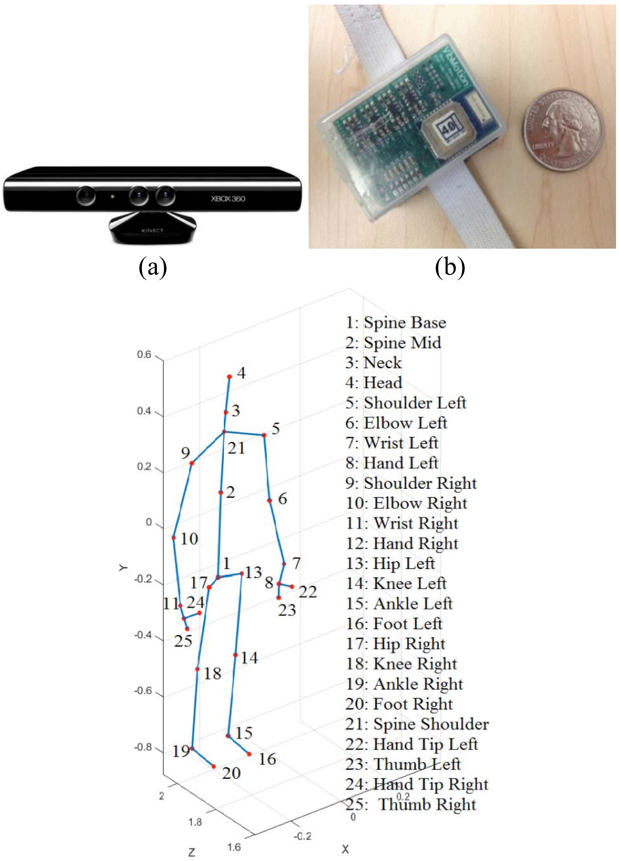
978-1-5090-1412-5/17/$31.00 ©2017 IEEE

the skeleton joints that Kinect v2 generates from captured depth images. As will be explained in Section IV, our approach uses these joint positions to segment a continuous stream of activity into pauses and motions as a prelude to action recognition.

The wearable inertial sensor used is a small wireless body sensor discussed in [14]. A picture of this inertial sensor is shown in Fig. 1(b). The sensor generates 3-axis acceleration and 3-axis angular velocity signals, which are wirelessly transmitted to a laptop via a Bluetooth link. It is worth mentioning that although wearing multiple inertial sensors on different parts of the body can increase the robustness of the system, due to the intrusiveness and thus the practicality aspect associated with wearing multiple inertial sensors, only one inertial sensor is used in this work.

III. Continuous Actions Dataset

Since the aim of this work is the recognition of some actions of interest from a continuous stream of activity, and given that there is no publicly available dataset that provides both a continuous and simultaneous stream of depth and inertial data, we have put together such a continuous dataset for the wrist actions involved in smart TV gestures. The dataset incorporates continuous streams of data that are simultaneously collected from the two differing modality sensors mentioned in Section II. The smart TV gestures include ‘Waving a Hand’, ‘Flip to Left’, ‘Flip to Right’, ‘Counterclockwise Rotation’ and ‘Clockwise Rotation’.


(a)

(b)

Fig. 1. (a) Kinect depth camera (b) wearable inertial sensor (c) human body skeleton joints obtained by Kinect v2

For data collection, the subjects performed the actions in front of a Kinect v2 camera while wearing the wearable inertial sensor on their right wrist. The data from the two sensors were synchronized by using the time stamp scheme described in [9]. For training, the subjects were asked to perform a single action of interest at a time and both depth and inertial data were recorded simultaneously. During actual operation or testing, the subjects were asked to perform the actions of interest in a continuous manner while randomly performing actions of non-interest in-between the actions of interest. Examples of actions of non-interest included picking up a water bottle, drinking water, wearing a pair of glasses, etc. As a result, a typical data stream consisted of both actions of interest and actions of non-interest appearing in a random order. Note that the subjects were given complete freedom to choose their own actions of non-interest while staying within the field of view of the Kinect camera.

Two scenarios were considered for data collection. The first scenario was done in a subject-specific manner in which the training and testing data were collected for the same subject. During training, a subject was asked to repeat each of the 5 actions of interest 15 times. For testing, continuous sets of actions were performed 6 times. In each set, the subject performed the 5 actions of interest with several actions of non-interest conducted randomly in-between the actions of interest in a continuous manner.

The second scenario was done in a subject-generic manner, i.e. training and testing data were collected from 5 different subjects. For training, the 5 subjects were asked to perform each of the 5 actions of interest 5 times, resulting in a total of 125 data streams consisting of simultaneous depth and inertial data. For testing, each subject performed the actions of interest with some actions of non-interest conducted randomly in-between the actions of interest in a continuous manner. This continuous dataset is made available for public use and can be downloaded from http://www.utdallas.edu/~kehtar/UTD-CAD-Both.htm. It is worth noting that this dataset is different than the one we previously reported in [9], which includes segmented single actions.

IV. Developed Continuous Action Detection And Recognition Approach

The approach developed in this paper relies on breaking down a continuous stream of skeleton activity data into pauses and motions, similar to the approach reported in [15-17]. A variable length Maximum Entropy Markov Model (MEMM) classifier is used in order to detect the presence of an action of interest in continuous data streams. This classifier operates similar to a Hidden Markov Model (HMM) classifier but is computationally more efficient to enable real-time operation. The acceleration and rotation signals from the wearable inertial sensor are used to remove false positive cases or to improve the recognition outcome by using a Collaborative Representation Classifier (CRC) as discussed in [18]. A block diagram of the components of the developed approach appears in Fig. 2. Note that the detection or segmentation task only uses the skeleton data, while both the skeleton and inertial data are used for the recognition task.

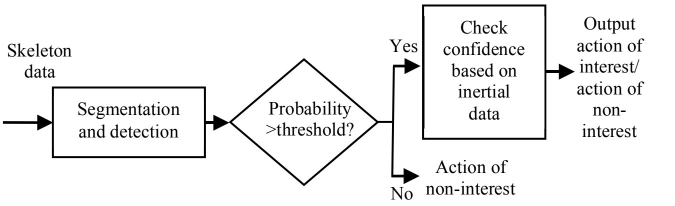

Skeleton data
Segmentation and detection
Probability >threshold?
Yes
No
Check confidence based on inertial data
Output action of interest/ action of non-interest
Action of non-interest

Fig. 2. Components of the developed continuous action detection and recognition approach

A. Training

Any continuous action is described as a sequence of pauses and motions. A training model similar to the one discussed in [17] is used to segment an activity sequence into pause and motion segments. In what follows, it is discussed how the model is modified in order to deal with a continuous data stream.

Pauses and motions from skeleton data are obtained by computing the length-invariant Normalized Relative Orientation (NRO) of the joints with respect to their rotating joints as follows:

$$F_{NRO}^{i} = \frac{L_i - L_j}{\|L_i - L_j\|} \tag{1}$$

where L_i and L_j denote 3D locations of the i^{th} and j^{th} joints, respectively, and j is the joint about which joint i rotates and $\|\cdot\|$ denotes the Euclidean distance.

Let $(F_1, F_2, \dots, F_t, \dots, F_n)$ represent a sequence of NRO features, where $F_t = (F_{NRO}^1, F_{NRO}^2, F_{NRO}^3, \dots)_t$ indicates the NRO features of the joints at frame t. Based on a reference NRO (F_r), a so called potential energy at frame t is computed as follows:

$$PE(t) = \|F_t - F_r\|^2 \tag{2}$$

Then, the following potential difference at frame t is obtained:

$$PD(t) = PE(t) - PE(t-1) \tag{3}$$

If the potential difference of data frames becomes less than a very low value close to zero (for example, for the dataset collected in this paper, 0.04 was found low enough to identify the start and end of all the motion segments), they are labeled as a pause segment, otherwise they are labeled as a motion segment. An example of pause and motion segments for the action 'Waving a Hand' is shown in Fig. 3. The horizontal portions represent pause segments and varying portions represent motion segments.

Based on pause and motion segments, a codebook for pauses and motions is then set up which is used for action recognition via a variable-length MEMM classifier. Basically, an action is characterized by its sequence of motion segments.

Similar motion segments occur in some actions of interest, for example the actions 'Waving a Hand' and 'Flip to Left' begin similarly, i.e. their first motion segments are similar. Hence, unlike [17], clustering is first applied to motion segments using a Gaussian Mixture Model (GMM) to group similar motion segments. The cluster representatives are used for action detection. The clustering is carried out by dividing each motion segment into three equal portions. Then, the averaged features from each of these three portions of motion segments are computed and used to cluster motion segments of an action into M clusters.

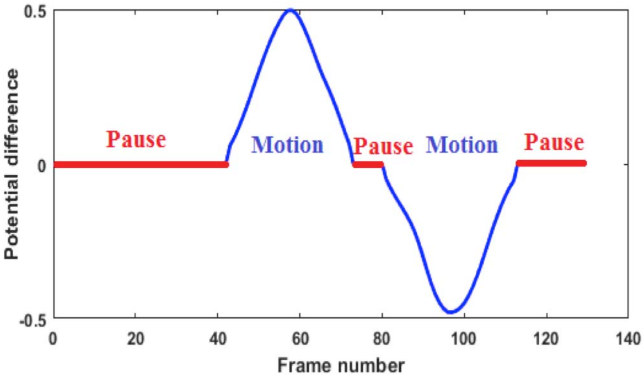


Fig 3. Pause and motion segments in the action 'Waving a Hand'

Next, the transition probabilities amongst the clusters are obtained. Since a pause segment is present between two motion segments, for every pair of motion clusters $MC1$ and $MC2$, the mean feature of the pause segments is obtained and stored in a so-called 'dynamic cluster' $DC(MC1, MC2)$. This way, a codebook of motion and pause clusters is generated from the training data.

B. Detection and Recognition

The task of continuous action detection and classification or recognition is carried out using likelihood probabilities. As the skeleton data is generated frame by frame, the corresponding NROs are generated and potential differences are used to classify the segments as pause or motion ones. Whenever a motion segment ends, the likelihood probabilities of the motion segments are obtained for each motion cluster. The likelihood probability of a motion segment for a motion cluster m of an action n is obtained as follows:

$$D(n,m) = \|FM - MC_{m,n}\|^2 \tag{4}$$

$$P_{motion}(n,m) = \frac{{}^{1}/_{D(n,m)}}{\sum_n \sum_m {}^{1}/_{D(n,m)}} \tag{5}$$

where FM denotes the feature vector corresponding to the three portions of a motion segment and $MC_{m,n}$ denotes the feature vector of the m^{th} motion cluster of the n^{th} action.

Similarly, the likelihood of a pause segment to lie between the motion clusters $m1$ and $m2$ of an action n is obtained as follows:

$$C(n, m1, m2) = \|FP - PC_n(m1, m2)\|^2 \tag{6}$$

$$P_{pause}(n, m1, m2) = \frac{{}^{1}/_{C(n,m1,m2)}}{\sum_n \sum_{m1} \sum_{m2} {}^{1}/_{C(n,m1,m2)}} \tag{7}$$

where FP denotes the mean feature vector of a pause segment and $PC_n(m1, m2)$ represents the feature vector of the pause cluster associated with the motion clusters $m1$ and $m2$ of the n^{th} action.

Once the likelihood probabilities are obtained, the variable-length MEMM classifier is used to assign the probability of a motion segment for each of the action classes. If the probability of a motion segment is greater than a set threshold, to be discussed later, the presence of an action of interest is indicated and the segment is assigned to the action class with the likelihood probability greater than the threshold.

C. Inertial Data

The probability threshold impacts the detection and recognition outcome for a continuous data stream. If the threshold is set too low, the rate of incorrect detections will be high, i.e. many actions of non-interest will be classified as actions of interest. On the other hand, if the threshold is set too high, actions of interest will be missed. It is therefore important to set the threshold such that the highest number of actions of interest are detected while minimizing the number of wrong detections.

The data from the inertial sensor are used to improve the detection accuracy by ruling out false detections. A CRC classifier is used here for this purpose. Whenever the likelihood probability of a particular motion segment using the skeleton data is greater than the threshold, the inertial data are used to verify the detection. Based on the CRC classifier, a residual error for each class is obtained and the segment is assigned to the class corresponding to the minimum error. If the detected action for a particular segment using the inertial data is the same as the one obtained using the skeleton data, that particular segment is considered to belong to that action, otherwise it is considered to be an action of non-interest.

V. Results And Discussion

In this section, the results of the developed approach on continuous data streams are reported. Since one cannot match the sequence predicted in a continuous data stream to serve as the ground truth, the evaluation framework proposed in [19] is used here. That is, a true positive was flagged whenever an action was detected within a window of 4 frames from the ground truth, while a false positive was flagged when the predicted action lied outside the window of 4 frames or when the action classified did not match the ground truth. The actions in the dataset examined contained a maximum of 4 motion segments, hence the number of motion clusters per action was set to 4, or $M = 4$, for the experimentations reported below.

For each continuous data stream, the number of true positives, false positives and false negatives were found and the performance was evaluated based on F1 score discussed in [20-21]. This score is derived from the precision and recall indices that are defined as follows:

$$P = \frac{\#TP}{\#TP + \#FP} \qquad (8)$$

$$R = \frac{\#TP}{\#TP + \#FN} \qquad (9)$$

$$F1 = 2\frac{P \cdot R}{(P + R)} \qquad (10)$$

where P denotes the precision index, R the recall index, $\#TP$ the number of true positives, $\#FP$ the number of false positives, and $\#FN$ the number of false negatives.

For the subject-specific scenario, the precision values, recall values and F1 scores that were obtained for different values of the probability threshold p with and without using the inertial data are listed in Tables I, II and III, respectively. As can be seen from these tables, for high threshold values, there were very few false positive detections resulting in a high value of precision. For such cases, some of the true positives were wrongly rejected when using the inertial data, which led to a drop in the recall value. Furthermore, many of the true positives could not be identified, hence the F1 score became low. As the threshold probability p was decreased, the F1 score and the recall value improved since more and more true positives were detected. However, upon further decreasing the threshold probability p, the number of true positives did not increase much but the number of false positives grew, which was reflected in the decrease in the F1 score.

As can be observed from the tables, the precision values, recall values and F1 scores improved when both the skeleton and inertial data were used. Note that the improvement when using the inertial data was more for the precision values as compared to the recall values. This is because the inertial data was used for the purpose of rejecting false positives.

TABLE I. Precision values for subject specific scenario

p values	Without Inertial	With Inertial
p=0.60	85.2%	94.1%
p=0.55	82.4%	89.9%
p=0.50	80.5%	89.1%
p=0.45	74.7%	87.1%
p=0.40	64.0%	76.4%
p=0.35	51.6%	67.0%

TABLE II. Recall values for subject specific scenario

p values	Without Inertial	With Inertial
p=0.60	61.6%	61.2%
p=0.55	77.4%	75.1%
p=0.50	86.7%	84.5%
p=0.45	92.5%	93.5%
p=0.40	92.9%	96.1%
p=0.35	93.8%	97.1%

TABLE III. F1 scores for subject specific scenario

p values	Without Inertial	With Inertial
p=0.60	71.5%	74.2%
p=0.55	79.8%	81.9%
p=0.50	83.5%	86.7%
p=0.45	82.7%	90.2%
p=0.40	75.7%	85.1%
p=0.35	66.5%	79.3%

The precision values, recall values and F1 scores for the subject-generic scenario with different values of the threshold probability *p* are listed in Tables IV, V and VI, respectively. These tables show the results with and without using the inertial data.

The best performance in both the scenarios was observed at the threshold probability of *p*=0.45 and at this threshold, the improvement in the F1 score by using the inertial data was about 8% for the subject specific scenario and 2% for the subject generic scenario. Due to large variations of the same actions associated with different subjects, in general, the subject specific scenario is recommended for any practical deployment.

TABLE IV. PRECISION VALUES FOR SUBJECT GENERIC SCENARIO

***p* values**	**Without Inertial**	**With Inertial**
p=0.55	95.9%	100.0%
p=0.50	93.1%	97.7%
p=0.45	85.1%	92.0%
p=0.40	68.6%	78.5%
p=0.35	55.4%	68.5%
p=0.30	43.2%	53.2%

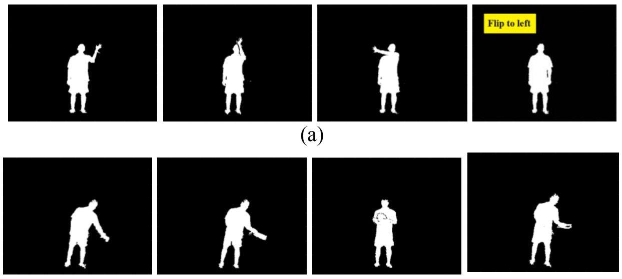
TABLE V. RECALL VALUES FOR SUBJECT GENERIC SCENARIO

***p* values**	**Without Inertial**	**With Inertial**
p=0.55	56.8%	56.0%
p=0.50	70.8%	69.6%
p=0.45	82.4%	79.2%
p=0.40	89.2%	90.8%
p=0.35	93.2%	96.0%
p=0.30	95.6%	97.2%

TABLE VI. F1 SCORES FOR SUBJECT GENERIC SCENARIO

***p* values**	**Without Inertial**	**With Inertial**
p=0.55	71.3%	71.7%
p=0.50	80.4%	81.3%
p=0.45	83.7%	85.1%
p=0.40	77.5%	84.2%
p=0.35	69.5%	80.0%
p=0.30	59.6%	68.8%

An important point to note here is that action detection and recognition were performed continuously in real-time (30 frames per second). For each incoming depth image frame and inertial signals, the features were extracted and pause and motion segments were obtained in real-time. The classification was performed when a motion segment was completed. An example of the depth and inertial data for an action of interest and an action of non-interest in a test sequence is shown in Figs. 4 and 5, respectively. The vertical lines in Fig. 5 exhibit the segments associated with the actions of interest and actions of non-interest in the potential difference function. The other two graphs in this figure show the acceleration along the z-direction for an action of interest and an action of non-interest. A videoclip of the developed continuous action detection and recognition approach running in real-time can be viewed at www.utdallas.edu/~kehtar/ContinuousAction.avi.



(a)

(b)

Fig. 4. (a) Snapshots of depth images from an action of interest 'Flip to Left' (b) an action of non-interest 'picking up and reading a book'

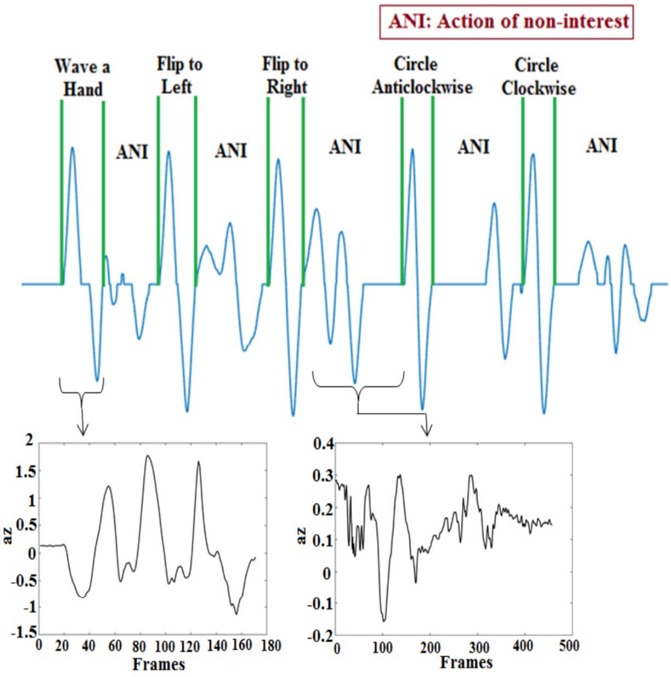


Fig. 5. Potential difference (top curve) and acceleration along z-axis (bottom curves) in an activity sequence consisting of actions of interest and actions of non-interest

VI. CONCLUSION

In this paper, a real-time action detection and recognition approach has been introduced which is capable of processing a continuous stream of depth images and inertial signals, a more challenging scenario than the conventional single action scenario normally reported in the literature. Continuous data implies actions of interest occur in a continuous manner while having actions of non-interest randomly located in-between them. The developed approach was applied to a continuous dataset for the smart TV application by simultaneously using a Kinect depth camera and a wearable inertial sensor. The results obtained show the effectiveness of the developed approach when activities are done continuously. In our future work, we plan to apply this approach to other applications or other sets of actions of interest.

REFERENCES

[1] S. Paul, S. Ali, and M. Shah, “A 3-dimensional sift descriptor and its application to action recognition,” *Proceedings of the 15th ACM International Conference on Multimedia*, pp. 357–360, September 2007.

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

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

[4] K. Altun, and B. Barshan, "Human activity recognition using inertial/magnetic sensor units," *Proceedings of International Workshop on Human Behavior Understanding*, Springer Berlin Heidelberg, pp. 38–51, August 2010.

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

[6] M. Ermes, J. Parkka, J. Mantyjarvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 1, pp. 20–26, January 2008.

[7] E. Guenterberg, H. Ghasemzadeh, and R. Jafari, "Automatic segmentation and recognition in body sensor networks using a hidden Markov model," *ACM Transactions on Embedded Computing Systems*, vol. 11, no. S2, pp. 46:1–46:19, August 2012.

[8] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, vol. 76, pp. 4405–4425, Feb 2017.

[9] C. Chen, R. Jafari, and N. Kehtarnavaz, “UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor,” *Proceedings of the IEEE International Conference on Image Processing*, pp. 168–172, Quebec City, Canada, September 2015.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz, “A real-time human action recognition system using depth and inertial sensor fusion,” *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773–781, February 2016.

[11] B. Delachaux, J. Rebetez, A. Perez-Uribe, and H. Mejia, “Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors,” *Proceedings of the 12th International Work-Conference on Artificial Neural Networks*, pp 216–223, June 2013.

[12] D. Huang, S. Yao, Y. Wang, and F. De La Torre, "Sequential max-margin event detectors," *Proceedings of European Conference on Computer Vision*, Springer International Publishing, pp. 410–424, September 2014.

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

[14] A. Y. Yang, R. Jafari, 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, January 2009.

[15] J. Shan, and S. Akella, "3D human action segmentation and recognition using pose kinetic energy," *Proceedings of the IEEE Workshop on Advanced Robotics and its Social Impacts*, pp. 69–75, September 2014.

[16] G. Zhu, L. Zhang, P. Shen, J. Song, L. Zhi and K. Yi, "Human action recognition using key poses and atomic motions," *Proceedings of the IEEE International Conference on Robotics and Biomimetics*, pp. 1209–1214, December 2015.

[17] G. Zhu, L. Zhang, P. Shen, and J. Song, "An Online Continuous Human Action Recognition Algorithm Based on the Kinect Sensor," *IEEE Sensors Journal*, vol. 16, no. 2, pp. 161, January 2016.

[18] 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.

[19] V. Bloom, D. Makris, and V. Argyriou, "G3d: A gaming action dataset and real time action recognition evaluation framework," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7–12, June 2012.

[20] J. Davis, and M. Goadrich, "The relationship between Precision-Recall and ROC curves," *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, June 2006.

[21] C. Goutte, and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," *Proceedings of European Conference on Information Retrieval*, Springer Berlin Heidelberg, pp. 345–359, March 2005.