FUSION OF DEPTH, SKELETON, AND INERTIAL DATA FOR HUMAN ACTION RECOGNITION

Chen Chen[1], *Roozbeh Jafari*[2], *Nasser Kehtarnavaz*[1]

[1]Department of Electrical Engineering, University of Texas at Dallas
[2]Center for Remote Health Technologies and Systems, Texas A&M University

ABSTRACT

This paper presents a human action recognition approach by the simultaneous deployment of a second generation Kinect depth sensor and a wearable inertial sensor. Three data modalities consisting of depth images, skeleton joint positions, and inertial signals are fused by utilizing three collaborative representation classifiers. A database consisting of 10 actions performed by 6 subjects is put together to carry out two types of testing of the developed fusion approach: subject-generic and subject-specific. The overall recognition rates obtained from both types of testing indicate recognition improvements when fusing all the data modalities compared to the situations when data modalities are used individually.

Index Terms— Fusion of depth, skeleton and inertial data, human action recognition, second generation Kinect depth sensor, wearable inertial sensor

1. INTRODUCTION

Research on human action recognition has made significant progress in the last decade and is attracting growing attention in a number of application domains, e.g. [1-5]. Despite much progress made in human action recognition, achieving high recognition rates under realistic conditions still remains a challenge. In our previous works [6-10] and in [11], it has been shown that recognition rates can be improved by using the data simultaneously captured from a depth camera (e.g., Microsoft Kinect) and a wearable inertial sensor compared to situations when each sensor is used separately or individually. In these previous works, the first generation of Kinect or Kinect v1 sensor was used. Furthermore, the skeleton data was not utilized due to jitters in skeleton joint positions in the Kinect v1 sensor.

The second generation of Kinect or Kinect v2 sensor that has recently been released not only provides higher depth fidelity but also more stable skeleton tracking. The Kinect v2 sensor can track 25 joints per person (compared to 20 with the Kinect v1 sensor), and the tracked positions are more anatomically correct and stable. In this paper, a human action recognition solution is introduced by simultaneously using a Kinect v2 sensor and a wearable inertial sensor via fusing depth, skeleton, and inertial data. More specifically, in addition to the depth images from the Kinect v2 sensor and the acceleration and angular velocity signals from the inertial sensor, the skeleton positions from the Kinect v2 sensor are used to improve the recognition outcome. This work involves the utilization of the three data modalities of depth, skeleton, and inertial at the same time for human action recognition. Moreover, the dataset collected in this work are made available for public use.

The rest of the paper is organized as follows. Section 2 describes the sensors used in this work. Section 3 includes the collected multimodal dataset. The fusion framework for action recognition is then presented in Section 4. The experimental results and their discussion are stated in Section 5. Finally, the conclusion appears in Section 6.

2. SENSORS UTILIZED

The Microsoft Kinect v2 sensor comprises a color camera and an infrared depth camera. A picture of the Kinect sensor is shown in Fig. 1(a). This sensor has a depth image resolution of 512 × 424 pixels with a field of view of 70 × 60 degrees. An example depth image is depicted in Fig. 1(c). The effective range for depth is from 0.5m to 4.5m. The frame rate is approximately 30 frames per second. The Kinect for Windows SDK 2.0 [12] is a publicly available software package which allows tracking 25 skeleton body joints (see Fig. 1(d)) and their 3D spatial positions.

The wearable inertial sensor used in this work is a small size (1”×1.5”) wireless inertial sensor built in the Embedded Signal Processing (ESP) Laboratory at Texas A&M University [13]. This sensor captures 3-axis acceleration, 3-axis angular velocity and 3-axis magnetic strength, which are transmitted wirelessly via a Bluetooth link to a laptop/PC. Due to a lack of a controlled magnetic field in practice, only the signals associated with the 3-axis accelerometer and the 3-axis gyroscope are used here. This wearable inertial sensor is shown in Fig. 1(b). The sampling rate of the inertial sensor is 50Hz and its measuring range is ±8g for acceleration and ±1000 degrees/second for rotation. For practicality reasons or to avoid the intrusiveness associated with asking subjects to wear multiple inertial sensors, only one inertial sensor is utilized in this work bearing in mind that it is possible to utilize multiple inertial sensors if intrusiveness of wearing multiple sensors is not of concern. It is worth mentioning that both types of sensors, namely depth and inertial, are widely available commercially and are capable of generating 3D data of human actions.

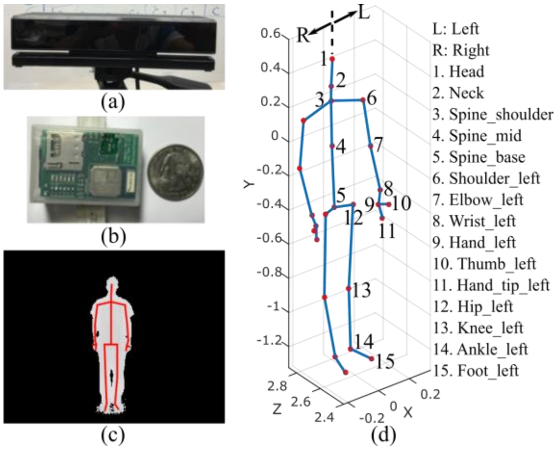

Fig. 1. (a) Microsoft Kinect v2 sensor. (b) Wearable inertial sensor. (c) An example depth image with the mapped 2D skeleton. (d) An example 3D skeleton with 25 tracked skeleton joints.

3. DATASET COLLECTION

Since there is no publicly available dataset that contains data from a Kinect v2 sensor and a wearable inertial sensor, the first task in this work involved collecting a multimodal dataset using the simultaneous utilization of these sensors. The actions considered correspond to the 10 similar actions in the Microsoft Research (MSR) Action3D dataset [14]. These actions are listed in Table 1. Note that the similarity of these actions makes the recognition task more challenging. Six subjects (3 female and 3 male subjects) were asked to perform these 10 actions. Each subject repeated an action 5 times, which resulted in a total of 300 action samples. The collected dataset incorporate intra-class variations due to different subject heights and subjects performing the same action differently.

During the data collection, subjects were standing in front of the Kinect v2 sensor with the wearable sensor worn on their right wrists noting that the 10 actions considered were hand type of movements. The experimental setup used for the data collection is illustrated in Fig. 2.

Three data modalities of depth images, skeleton joint positions, and inertial sensor signals (3-axis acceleration and angular velocity signals) were recorded in two channels or threads. One channel was used for capturing of depth images and skeleton positions, and one channel for the simultaneous capturing of inertial sensor signals. Each action sample was generated in one recording. The background of the depth images was removed during recording using the body tracking functionality provided in the Kinect SDK 2.0.

As reported in [7], for data synchronization, a time stamp for each action sample was utilized. Since the frame rate of the Kinect sensor and the sampling rate of the wearable inertial sensor were different, the start and end of an action were synchronized by using the time stamps of the depth images to serve as references.

Table 1. 10 actions used in the experiments.

1. Right hand high wave	6. Right hand draw circle
2. Right hand catch	7. Right hand horizontal wave
3. Right hand high throw	8. Right hand forward punch
4. Right hand draw X	9. Right hand hammer
5. Right hand draw tick	10. Hand clap (two hands)

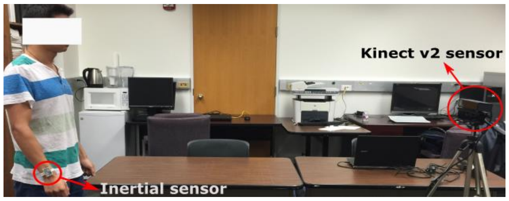

Fig. 2. Experimental setup for the dataset collection.

The dataset collected is made available for public use and can be downloaded from the link: http://www.utdallas.edu/~kehtar/UTD-MHAD.html.

4. ACTION RECOGNITION VIA SENSOR FUSION

4.1. Feature extraction

To extract features from depth images, depth motion maps (DMMs) discussed in [15] are used due to their computational efficiency. Each 3D depth image in a depth sequence is first projected onto three orthogonal Cartesian planes to generate three 2D projected maps corresponding to front, side, and top views, denoted by map_f, map_s, and map_t, respectively. For a depth sequence with N frames, the DMMs are obtained as follows:

$$DMM_{\{f,s,t\}} = \sum_{i=1}^{N-1} \left| map_{\{f,s,t\}}^{i+1} - map_{\{f,s,t\}}^{i} \right| , \qquad (1)$$

where i represents frame index. A bounding box is considered to extract the foreground (non-zero region) in each DMM. Since foreground DMMs of different video sequences may have different sizes, a bicubic interpolation is applied to resize all such DMMs to a fixed size and thus to reduce the intra-class variability.

For the skeleton feature extraction, the method described in [16] is used due to its low computational complexity. Each skeleton sequence is partitioned into K temporal windows. Four statistical features of *mean*, *variance*, *standard deviation*, and *root mean square* are computed for the skeleton joint positions along three axes per temporal window. Since the 10 actions in our dataset involve hand type movements, only the position data from the upper body joints are used. More specifically, these 13 skeleton joints are used: left elbow, right elbow, left hand, right hand, left hip, right hip, left shoulder, right shoulder, base of the spine, middle of the spine, spine at the shoulder, left wrist, and right wrist. All the features from the temporal windows are then concatenated to form a single skeleton feature vector of dimensionality $4\times3\times13\times K=156K$.

For the inertial sensor, the same feature extraction method used for the skeleton data is utilized. Each acceleration and gyroscope signal sequence is partitioned into M temporal windows as reported in [7, 16]. The four statistical features are computed for each direction per temporal window. All the features from the temporal windows are concatenated to form a single inertial feature vector of dimensionality $4\times3\times2\times M=24M$.

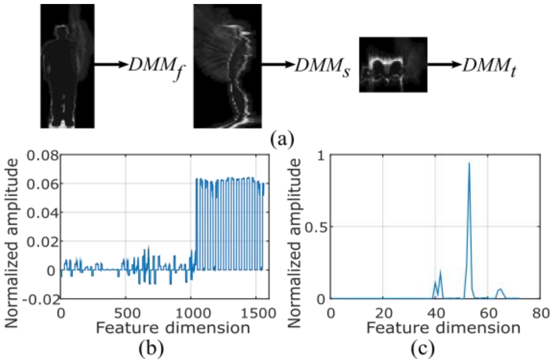

Fig. 3. Example data for the action *right hand high throw*: (a) DMM features, (b) skeleton joint features, and (c) inertial signal features.

An example of the three different types of features for the action *right hand high throw* is shown in Fig. 3. To gain computational efficiency, principal component analysis (PCA) is performed to reduce the dimensionality of features by retaining 95% of the total variation of the data.

4.2. Recognition using decision-level fusion

For action recognition, the collaborative representation classifier (CRC) [17], previously reported in [15], was utilized due to its computational efficiency and good classification performance. This classifier is briefly described here. Let $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_C]\in\mathbb{R}^{D\times n}$ denote n training samples from C classes and $\mathbf{X}_j\in\mathbb{R}^{D\times n_j}$ denote n_j training samples that are associated with class j. In CRC, a test sample $\mathbf{y}\in\mathbb{R}^D$ is encoded on $\mathbf{X}$ via l_2-minimization

$$\hat{\boldsymbol{\alpha}}=\arg\min_{\boldsymbol{\alpha}}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}\right\|_2^2+\lambda\left\|\boldsymbol{\alpha}\right\|_2^2, \tag{2}$$

where $\boldsymbol{\alpha}\in\mathbb{R}^n$ is a coefficient vector corresponding to all the training samples and λ is a regularization parameter. The solution of this minimization problem is given by

$$\hat{\boldsymbol{\alpha}}=\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}\,. \tag{3}$$

The classification decision is made according to the class which minimizes the reconstruction error, i.e.

$$\text{class}\left(\mathbf{y}\right)=\arg\min_{j\in\{1,2,\ldots,C\}}\left\{e_j\left(\mathbf{y}\right)=\left\|\mathbf{y}-\mathbf{X}_j\hat{\boldsymbol{\alpha}}_j\right\|\right\} \tag{4}$$

where $\hat{\boldsymbol{\alpha}}_j$ is the coefficient vector related to $\mathbf{X}_j$ and $e_j\left(\mathbf{y}\right)$ is the corresponding reconstruction error.

To fuse the three sets of features (e.g., depth features, skeleton features, and inertial features), a decision-level fusion scheme is adopted here by passing each set of features through a CRC classifier. Hence, for a test action sample $\mathbf{y}$, three CRC classifiers are used, each handling one type of features. Then, a logarithmic opinion pool (LOGP) [18] is applied to achieve fusion at the posterior-probability level. LOGP uses the individual posterior probability $p_q\left(\omega|\mathbf{y}\right)$ of each classifier to estimate this global membership function

$$P\left(\omega|\mathbf{y}\right)=\prod_{q=1}^{m}p_q\left(\omega|\mathbf{y}\right)^{\beta_q}, \tag{5}$$

where m is the number of classifiers and β_q is a uniformly distributed classifier weight ($\beta_q=1/m$). Here, the Gaussian mass function denoted below is considered

$$p_q\left(\omega|\mathbf{y}\right)=\exp\left(-e_j\left(\mathbf{y}\right)\right) \tag{6}$$

This function indicates that a smaller reconstruction error $e_j\left(\mathbf{y}\right)$ yields a higher probability $p_q\left(\omega|\mathbf{y}\right)$. The final class label for $\mathbf{y}$ is then assigned to be the one with the largest probability $P\left(\omega|\mathbf{y}\right)$.

5. EXPERIMENTAL RESULTS AND DISCUSSION

To examine the effectiveness of sensor fusion for action recognition, the following experiments involving two different settings were conducted. The first experimental setting is named subject-generic. This setting involved the leave-one-out subject test, i.e., each time a subject was assigned as the testing subject (the action samples associated with this subject were regarded as testing samples) and the remaining five subjects were assigned as the training subjects (the action samples associated with these five subjects were regarded as training samples). The second experimental setting is named subject-specific. This setting involved dividing the samples from only one specific subject into a training and a testing set. Since each subject had performed an action 5 times, the first two repetitions of an action were used to form the training set and the remaining repetitions to form the testing set. In this case, all the training and testing samples were associated with the same subject.

5.1. Parameter setting

In the experiments, the sizes of DMM_f , DMM_s and DMM_t were set to 170×86, 170×72 and 72×86, respectively, which were the average sizes of DMM_f ,

DMM_s and DMM_t from all the action samples as discussed in [15]. As shown in Fig. 4, the number of windows generating the best recognition outcome was used for the subsequent experimentations. The number of windows being 10 generated the best outcome when using the skeleton features (i.e., $K = 10$) and the number of windows being 3 generated the best outcome when using the inertial features (i.e., $M = 3$). The parameter λ in the CRC was set to the value that maximized the training accuracy via a five-fold cross-validation.

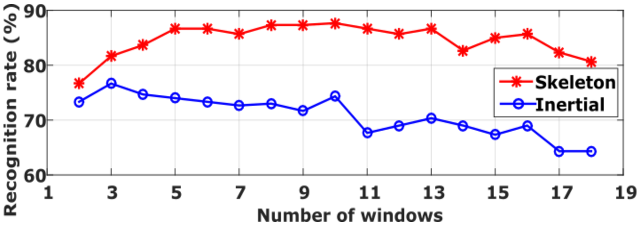

Fig. 4. Recognition rates (%) with different number of windows for skeleton and inertial features in the leave-one-subject-out test.

5.2. Recognition rates and processing time

In this sub-section, the recognition rates of the developed fusion approach are reported. Under the two experimental settings, i.e. subject-generic and subject-specific, the recognition rates were obtained using only the depth features (D), only the skeleton features (S), only the inertial features (I), the combination of the depth and inertial features (D+I), the combination of the skeleton and inertial features (S+I), and the combination of the depth, skeleton and inertial features (D+S+I). Table 2 lists the average recognition rates per class and the average overall recognition rate over the 6 subjects in the subject-generic test. As can be seen from this table, the action recognition performance was noticeably improved by the simultaneous utilization or fusion of the data from the Kinect sensor and the wearable inertial sensor (e.g., D *vs*. D+I and I *vs*. S+I). By fusing the three modalities of depth, skeleton and inertial data, the overall recognition rate reached a relatively high accuracy of 93.7% for the subject-generic test.

We also conducted experiments reflecting different combinations of features for the subject-specific test. The outcome of these experiments is provided in Table 3. As evident from this table, the fusion approach led to higher recognition rates compared to situations when each set of features was used individually. In general, the recognition rates in subject-specific experiments would be higher than those in subject-generic experiments because of the fact there exist smaller intra-class variations in subject-specific experiments.

All the coding was done in MATLAB on a laptop equipped with a 2.6 GHz Intel quad-core i7 CPU with 8 GB RAM. The processing time of the major components of the recognition program is listed in Table 4. As evident from these processing times, the recognition program runs in real-time by being able to process 30 frames per second.

Table 2. Action recognition rates (%) using different feature combinations for the subject-generic test.

Action	D	S	I	D+I	S+I	D+S+I
1	96.7	96.7	73.3	96.7	100	100
2	50.0	56.7	80.0	76.7	80.0	90.0
3	50.0	86.7	73.3	76.7	80.0	80.0
4	63.3	80.0	66.7	80.0	80.0	80.0
5	86.7	100	76.7	100	96.7	100
6	50.0	86.7	90.0	83.3	93.3	93.3
7	53.3	83.3	53.3	76.7	96.7	93.3
8	93.3	96.7	73.3	90.0	100	100
9	86.7	90.0	96.7	96.7	100	100
10	100	100	83.3	96.7	100	100
Overall	73.0	87.7	76.7	87.3	92.7	93.7

Table 3. Action recognition rates (%) using different feature combinations for the subject-specific test.

Action	D	S	I	D+I	S+I	D+S+I
1	83.3	100	94.4	100	100	100
2	50.0	94.4	94.4	100	100	100
3	100	100	100	100	100	100
4	83.3	88.9	88.9	88.9	88.9	94.4
5	72.2	100	100	100	100	100
6	77.8	94.4	100	100	100	100
7	72.2	100	100	94.4	100	100
8	83.3	100	88.9	100	100	100
9	77.8	94.4	100	100	100	100
10	94.4	100	100	100	100	100
Overall	79.4	97.2	96.7	98.3	98.9	99.4

Table 4. Processing times (in ms) of the major components of the fusion recognition program.

Code components	Average processing time (ms)
DMMs computation	11.2
Skeleton feature computation	8.4
Inertial feature computation	0.6
PCA dimensionality reduction	6.4
Fusion classification	7.9

6. CONCLUSION

In this paper, a data fusion approach for human action recognition has been developed by using a second generation Kinect depth sensor and a wearable inertial sensor. Depth images and skeleton joint positions from the Kinect sensor are collected together with the acceleration and angular velocity signals from the inertial sensor. These multimodality data are then used to carry out a decision-level fusion via collaborative representation classifiers. The extensive experimentations performed have indicated the effectiveness of the fusion approach for action recognition compared to the situations when using each data modality individually. Possible extensions of this work include examining adaptive weights for combining features instead of using equal weights as done in this work and utilizing multiple inertial sensors for more complicated actions.

6. REFERENCES

[1] 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 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, pp. 4983-4986, August 2014.

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

[3] W. Lin, M. Sun, R. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," *Proceedings of the IEEE International Symposium on Circuit and Systems*, Seattle, WA, pp. 2737-2740, May 2008.

[4] J. Lockhart, T. Pulickal, and G. Weiss, "Applications of mobile activity recognition," *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, Pittsburgh, PA, pp. 1054-1058, September 2012.

[5] C. Chen, Z. Hou, B. Zhang, J. Jiang, and Y. Yang, "Gradient local auto-correlations and extreme learning machine for depth-based activity recognition," *Proceedings of the 11th International Symposium on Visual Computing*, Las Vegas, NV, pp. 613-623, December 2015.

[6] 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.

[7] 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 IEEE International Conference on Image Processing*, Quebec city, Canada, pp. 168-172, September 2015.

[8] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors Journal*, doi: 10.1109/JSEN.2015.2487358, available online, Oct. 2015.

[9] 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.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, doi: 10.1007/s11042-015-3177-1, available online, Dec 2015.

[11] B. Delachaux, J. Rebetez, A. Perez-Uribe, and H. F. S. Mejia, "Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors," *Proceedings of the 12th International Conference on Artificial Neural Networks: Advances in Computational Intelligence*, pp. 216–223, Puerto de la Cruz, Spain, June 2013.

[12] https://www.microsoft.com/en-us/kinectforwindows/develop/

[13] A. 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, 2009.

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

[15] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based depth motion maps," *Journal of Real-Time Image Processing*, pp. 1-9, August 2013.

[16] 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*, pp. 53-60, Tampa, FL, January 2013.

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

[18] 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 Beach, HI, 2015.