Multimed Tools Appl
DOI 10.1007/s11042-015-3177-1

A survey of depth and inertial sensor fusion for human action recognition

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

Received: 1 November 2015 / Revised: 9 December 2015 / Accepted: 15 December 2015

Abstract A number of review or survey articles have previously appeared on human action recognition where either vision sensors or inertial sensors are used individually. Considering that each sensor modality has its own limitations, in a number of previously published papers, it has been shown that the fusion of vision and inertial sensor data improves the accuracy of recognition. This survey article provides an overview of the recent investigations where both vision and inertial sensors are used together and simultaneously to perform human action recognition more effectively. The thrust of this survey is on the utilization of depth cameras and inertial sensors as these two types of sensors are cost-effective, commercially available, and more significantly they both provide 3D human action data. An overview of the components necessary to achieve fusion of data from depth and inertial sensors is provided. In addition, a review of the publicly available datasets that include depth and inertial data which are simultaneously captured via depth and inertial sensors is presented.

Keywords Human action recognition · Activity recognition · 3D action data · Depth sensor · Inertial sensor · Sensor fusion · Multimodal dataset

✉ Chen Chen
chenchen870713@gmail.com

Roozbeh Jafari
rjafari@tamu.edu

Nasser Kehtarnavaz
kehtar@utdallas.edu

[1] Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080, USA

[2] Biomedical Engineering, Computer Science and Engineering, and Electrical and Computer Engineering Departments, Texas A&M University, College Station, TX 77843, USA

Published online: 23 December 2015

1 Introduction

Human action recognition involves automatically detecting and analyzing human actions from the information acquired from sensors such as RGB cameras, depth cameras, range sensors, wearable inertial sensors, or other modality type sensors. Research on human action recognition has made significant progress in the last decade and is receiving growing attention in a wide variety of disciplines. Human action recognition has found its way into a wide range of applications including surveillance, video analytics, assistive living, robotics, telemedicine, and human computer interaction [16, 17, 60]. In a typical application, automated recognition of a number of actions is sought. In terms of sensor types that are used for human action recognition, there are two main approaches: vision-based action recognition and inertial-based action recognition.

In vision-based action recognition, many works have utilized conventional RGB cameras. The approaches developed based on video sequences can be classified into template-based approaches, where emphasis is placed on low- and mid-level features, and model-based approaches where emphasis is placed on high-level features [45]. A number of feature extraction methods, notably spatio-temporal interest point (STIP) detector [38], spatio-temporal descriptor based on 3D gradients [35], motion-energy images (MEI) and motion-history images (MHI) [7], have achieved successful outcomes for human action recognition using RGB video data. The popularity of human action recognition based on RGB cameras has led to several survey articles that have appeared in [1, 50, 51, 66]. These articles discuss various features and classifiers that have been used for human action recognition. As noted in [2], there are limitations associated with the utilization of RGB cameras. In practice, one requires to have a considerable amount of hardware resources in order to run computationally intensive image processing and computer vision algorithms and also one needs to deal with a lack of 3D action data in conventional images.

Recent emergence of cost-effective depth sensors has led to their widespread utilization for human action recognition considering that they provide 3D action data. There are basically three existing approaches towards obtaining 3D action data. The first approach uses relatively expensive marker-based motion capture systems such as MoCap.[1] Motion capture systems usually utilize optical sensing of markers placed in specific locations of a human body, and use triangulation from multiple cameras to estimate the 3D position data or the body skeleton. The second approach involves the use of stereo cameras. 3D data including depth are obtained via stereo matching and depth computation [4]. Stereo 3D reconstruction algorithms are computationally expensive and exhibit sensitivity to lighting changes and background clutter [2]. The third approach is based on range or depth sensors. More recently, depth sensors (in particular, Microsoft Kinect and Asus Xtion Pro) have provided cost-effective real-time 3D data for performing human action recognition. Compared to conventional RGB images captured by video cameras, depth images generated by depth cameras are shown to be insensitive to lighting changes and have led to gaining high performance in human action recognition. The human skeleton information can also be obtained from depth images [56].

Although vision-based human action recognition continues to advance, the recognition performance is subject to various challenges such as occlusion, camera position, subject variations in performing actions, background clutter, etc. In addition, vision-based approaches are applicable to a limited field of view or a constrained space defined by the camera position

[1] http://mocap.cs.cmu.edu/

and settings. To address such challenges, many researchers have utilized wearable inertial sensors incorporating accelerometers and gyroscopes, e.g., [34, 72]. This sensor technology has enabled coping with much wider field of views as well as changing lighting conditions. The continuous advancements in lowering the energy consumption and increasing the computational power of inertial sensors have enabled long-term recordings, computing, and continuous interaction. Furthermore, similar to depth sensors, wearable inertial sensors provide 3D action data consisting of 3-axis accelerations from their accelerometers and 3-axis angular velocities from their gyroscopes. However, wearable inertial sensors have their own limitations as well. For example, sensor drift may occur during long operation times and measurements are sensitive to sensor location on the body. In addition, for human action recognition, they require to be worn by subjects performing the actions, which creates the disadvantage of intrusiveness or inconvenience for subjects. A summary of the pros and cons associated with different modality sensors (i.e., RGB video cameras, depth cameras, and inertial sensors) for human action recognition is provided in Table 1.

A typical human action recognition system normally uses a single modality sensor, that is either a vision sensor or an inertial sensor alone. Under realistic operating conditions, it is known that no single sensor modality can cope with various situations that may occur in practice. One way to improve the performance of human action recognition systems is to combine data from these two differing modality sensors considering that depth images from a depth sensor and inertial signals from a wearable sensor provide complementary information. For example, depth images capture global (or full body) movement attributes while inertial signals capture local movement attributes. In [13, 14, 43], it was shown that fusing information from depth and inertial sensors leads to more robust recognition. Here, the emphasis has been placed on these two types of sensors due to the fact that commercially available depth cameras and wearable inertial sensors are both low-cost, widely available and more importantly they both provide 3D action data.

Table 1 A summary of pros and cons of different modality sensors for human action recognition

	RGB video cameras	Depth cameras	Inertial sensors
Pros	• Cost effective and widely available • Easy to operate • Provide rich texture information of the scene	• Cost effective and widely available • Insensitive to lighting conditions and illumination changes and can work in total darkness • Provide 3D structure information of the scene • Easy to operate • Not sensitive to color and texture change	• Cost effective and widely available • High sampling rate • Can work in total darkness • Can work in unconfined environment
Cons	• Require the subject to be in the field of view • Sensitive to lighting conditions, illumination changes, and background clutter • Sensitive to camera calibration • Algorithms can be computationally expensive	• Require the subject to be in the field of view • Different noise present in the images • Depth information sensitive to materials with different refection properties (e.g., transparent materials, light absorbing materials, etc.) • No color information	• Sensitive to sensor location on the body • Sensor drift • Power consumption for sensor onboard battery • Require multiple sensors for capturing full body movements • Intrusiveness of wearing single or multiple sensors

Although there exist several survey papers for human action recognition using depth sensors alone (e.g., [2, 19, 31, 76]) or inertial sensors alone (e.g., [5, 11, 30, 40]), there exists no survey article on the simultaneous utilization of these two differing modality sensors for human action recognition. After stating papers where an individual sensor modality is used for human action recognition, this paper reviews the approaches where both of these two sensor modalities are used simultaneously. In addition, a list of publicly available action/gesture recognition datasets that include both depth and inertial sensor data is provided. These datasets help researchers to evaluate recognition algorithms when fusing depth and inertial data. This survey paper is intended to inform researchers in the computer vision, pervasive computing, and multimodal fusion communities of the state-of-the-art works in fusing depth and inertial data for performing human action recognition. It is also worth mentioning that the fusion of depth and inertial sensing has been previously considered for other applications including skeleton estimation and tracking [22], human body tracking [32], and limb motion tracking [61]. However, the thrust of this survey is exclusively on the human action recognition application.

The remainder of the paper is organized as follows. In Sections 2 and 3, a brief overview of the human action recognition application based on depth sensor alone and inertial sensor alone is mentioned, respectively. In Section 4, the components of the fusion approaches for human action recognition involving a combination of depth and inertial sensors is covered. In Section 5, a list of human action/gesture recognition datasets that contain simultaneous data from depth and inertial sensors is mentioned. Finally, the paper is concluded in Section 6.

2 Human action recognition based on depth sensor alone

Recent advances in 3D depth cameras using structured light or time-of-flight strategies have led to a major breakthrough towards the problem of human action recognition. The release of affordable Microsoft's Kinect sensors has provided a commercially viable hardware platform to capture 3D data in real-time. The Kinect sensor comprises a color camera, an infrared (IR) emitter, an IR depth sensor, a tilt motor, a microphone array, and an LED light. The IR projector casts an IR speckle dot pattern into the 3-D scene while the IR camera captures the reflected IR speckles. Kinect is essentially a structured light depth sensor. A picture of the Kinect sensor or depth camera is shown in Fig. 1. This sensor can capture 16-bit depth images with a resolution of 320×240 pixels. It also offers color images at 640×480 pixels with 8-bit resolution per channel. A color image and the corresponding depth image of a scene generated from the Kinect sensor are depicted in Fig. 2. In a depth image, the value of each pixel indicates the distance between a 3D scene point and the sensor. The frame rate is approximately 30 frames per second. In addition, the Kinect software development kit (SDK) [33] is a publicly available software package which is widely used to track 20 body skeleton joints and their 3D spatial positions (see Fig. 3).

Fig. 1 Microsoft Kinect RGB-Depth sensor

Fig. 2 **a** An example color image (640 × 480) from the MSR Daily Activity 3D dataset [65], **b** corresponding depth image (320 × 240)

Various methods have been proposed for human action recognition from depth images in recent years. The developed major feature representation techniques for human action recognition based on depth sequences include a bag of 3D points [42], projected depth maps [15, 18, 75], space-time occupancy patterns [63, 64], spatio-temporal depth cuboid [69], surface normals [48, 74], and skeleton joints [25, 62, 73]. A summary of the existing survey papers on human action recognition using depth sensors is listed in Table 2. These papers include more details of different feature representation techniques based on depth images (or 3D skeletons).

3 Human action recognition based on inertial sensor alone

Over the past decade, low-power, low-cost, and miniaturized inertial sensors have provided yet another breakthrough towards the problem of human action recognition. Wearable inertial sensors are usually placed directly or indirectly on the human body. They have been embedded into clothes, shoes, wristwatches, mobile devices, etc. These sensors generate accelerometer and rotation signals corresponding to an action performed by a human. As an example, a

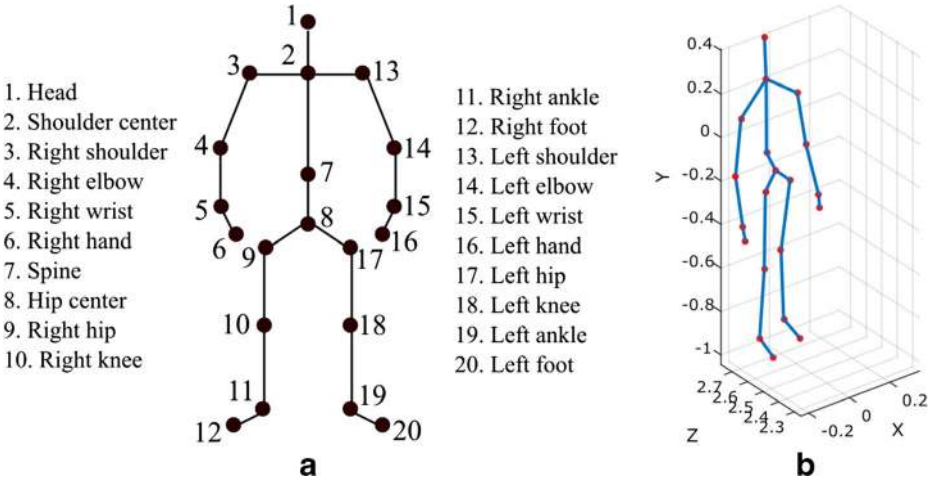

Fig. 3 **a** 20 skeleton joints tractable by the Kinect SDK, **b** an example skeleton frame in 3D

Table 2 Review or survey articles on human action recognition using depth sensors

References	Year	Title	Description
Chen et al. [19]	2013	A survey of human motion analysis using depth imagery	A review addressing articulated 3D body modelling for human pose estimation and human action recognition from depth images.
Han et al. [31]	2013	Enhanced computer vision with microsoft kinect sensor: A review	An overview of recent Kinect-based computer vision algorithms and applications including object tracking and recognition, human activity analysis, hand gesture analysis, and indoor 3D mapping.
Ye et al. [76]	2013	A survey on human motion analysis from depth data	An overview of recent approaches that perform human motion analysis which includes depth-based activity recognition, hand gesture recognition, facial feature detection, and head pose estimation.
Aggarwal et al. [2]	2014	Human activity recognition from 3d data: A review	Discussing the state-of-the-art methodologies on human-activity recognition using 3D data. Reviewing different features extracted from depth data in different scenarios for action recognition.

wearable inertial sensor in [12] that has been used for action recognition is shown in Fig. 4. This wearable sensor is a small size (1" × 1.5") wireless inertial sensor built in the Embedded Signal Processing (ESP) Laboratory at Texas A&M University [6]. 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. The sampling rate of the sensor is 50 Hz and its measuring range is ±8 g for acceleration and ±1000 degrees/second for rotation. Figure 5 shows the inertial sensor signals (3-axis accelerations and 3-axis angular velocities) generated by this sensor for the action *right hand high throw*.

As far as human action recognition based on inertial sensors is concerned, a number of solutions have appeared in the literature. For example, in [24], wearable inertial sensors were employed to recognize daily activities by using artificial neural networks within a tree

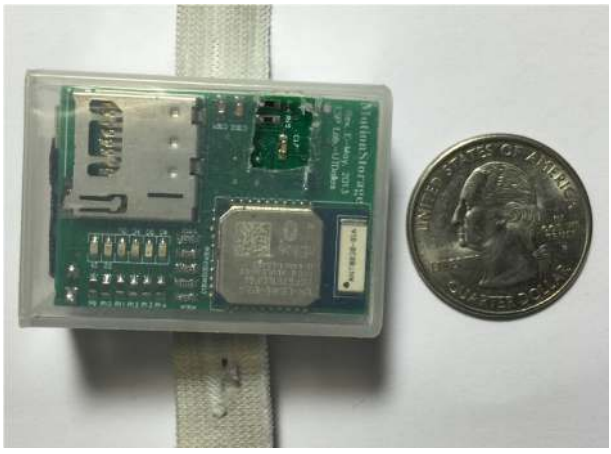

Fig. 4 Wearable inertial sensor developed in [6]

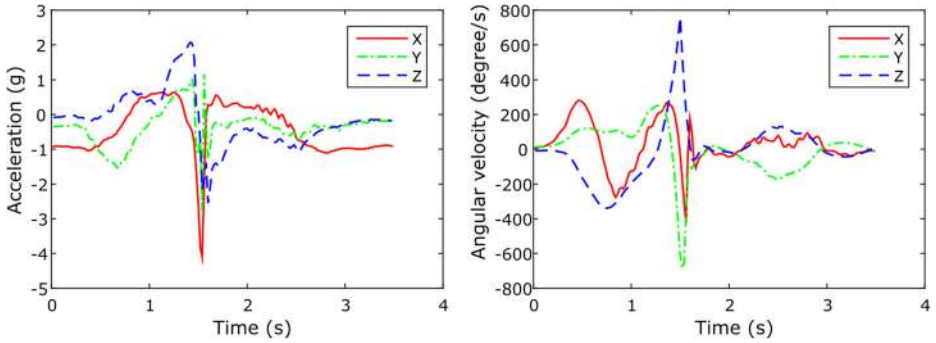

Fig. 5 Inertial sensor signals (3-axis accelerations and 3-axis angular velocities) for the action *right hand high throw*

structure. In [3], five inertial sensor units each comprising a triaxial gyroscope, a triaxial accelerometer, and a triaxial magnetometer were employed for classifying human activities and different classification techniques consisting of Bayesian decision, least-squares, and dynamic time warping, were implemented and compared. In [41], a fall detection system was presented based on wearable inertial sensors. In [16], a wearable inertial sensor was employed to recognize the hand-twist and hand-open actions for an intelligent medication adherence monitoring system.

A tutorial was presented in [8] covering a comprehensive discussion of designing and evaluating activity recognition systems using body-worn inertial sensors. A summary of the survey papers that have appeared on human action recognition based on inertial sensors is

Table 3 Review or survey articles on human action recognition using inertial sensors

References	Year	Title	Description
Chen et al. [11]	2012	Sensor-based activity recognition	An in-depth overview on the latest development of sensor-based activity recognition. Major data-driven and knowledge-driven approaches for activity recognition discussed and compared.
Lara et al. [40]	2013	A survey on human activity recognition using wearable sensors	A survey of the state of the art in human activity recognition based on wearable sensors, focusing on the main design issues for recognizing activities and the principal techniques applied in human activity recognition systems.
Avci et al. [5]	2010	Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey	Survey covering the current research directions of activity recognition using inertial sensors, with potential application in healthcare, wellbeing and sports. The five main steps involved in the activity recognition process discussed.
Guan et al. [30]	2011	Review of sensor-based activity recognition systems	Survey of wearable sensor based activity recognition systems. In addition, conventional video sensor (or camera) based activity recognition reviewed. For each type of activity recognition, main techniques, characteristics, strengths and limitations discussed and summarized.

listed in Table 3. These papers cover more details of the approaches and challenges associated with using inertial sensors for human action recognition.

4 Human action recognition using depth sensor and inertial sensor fusion

As stated above, depth sensors and wearable inertial sensors each have been used individually for the application of human action recognition. It has been established that there are recognition rate limitations when using a single modality sensor due to the fact that no single modality sensor can cope with various realistic situations that occur in real-world settings. Therefore, it is reasonable to expect that the utilization of both modalities simultaneously would improve the recognition performance due to the fact that each modality can complement the shortcomings of the other modality. In this section, the existing works on human action/gesture recognition by using depth and inertial sensors together are mentioned as well as the components required for the simultaneous utilization of these two differing modality sensors.

4.1 Data synchronization and preprocessing

Data fusion based on samples from differing modality sensors requires accurate time synchronization. In [43], a synchronization approach was developed by correlating the closest inertial sample to the depth frame according to the system time stamps. In [47], the temporal synchronization between the different modality sensors of video cameras, Kinect, MoCap, accelerometer was provided via the UNIX operating system time stamps which were included in the recordings of each modality. In these synchronizations, possible propagation delays [70] and variable intervals needed to generate samples were not taken into consideration. A more accurate time synchronization method was proposed in [20] by estimating the total delay occurring in the link between the camera and the PC. This method was later used in [26] for the depth and inertial sensors data synchronization for the fall detection application.

In addition to data synchronization, a signal filtering preprocessing component is often used. For example, a moving average window was used in [43] to reduce jitters in the raw sensor signals including skeleton joint positions from a Kinect depth camera and acceleration/angular velocity signals from an inertial sensor.

4.2 Action segmentation (detection of action start and end)

Most of the papers that have appeared in the literature on human action recognition have studied action signals (e.g., RGB videos, depth videos, inertial sensor signals) that have been segmented manually or by visual inspection providing the start and end of actions. However, to have an actual real-time working action recognition system, it is necessary to identify the start and end of actions. Action segmentation is a challenging task when actions are done in practice or in random time. A number of action segmentation methods have been developed when using vision sensors (e.g., [55, 66]) or inertial sensors (e.g., [57, 71]). When using depth and inertial sensors together within a fusion framework, depth and inertial sensor data can be utilized collaboratively to achieve improved action segmentation. Two examples of such action segmentation are presented next. In [14], the variances of the skeleton joint positions as well as the accelerations within a moving window were used to determine the start and the end of an action in real-time, with the requirement that each action began with a static posture

and ended with a static posture lasting for at least 1 s in an action sequence. In [77], a Gaussian model for the rest positions and a Gaussian model for non-rest positions were created using a training gesture dataset. Then, during recognition or testing of hand gestures, an observation (a combination of hand positions from a depth camera and inertial signals from an inertial sensor) was classified into a rest or a non-rest position based on the two Gaussian models. A sequence of continuous observations from non-rest positions longer than 0.25 s was then considered to denote a complete gesture.

4.3 Feature extraction

Different features extracted from depth and inertial data have been considered in the literature ranging from raw signals (e.g., raw accelerations) to high-level descriptors. In [77], a gesture spotting and recognition framework was used to fuse the data from a Kinect sensor and an inertial sensor. For hand feature extraction, linear acceleration signals of three directions (x, y, z), angular velocity signals of three directions (x, y, z) and Euler orientation signals of three directions (yaw, pitch, roll) from an inertial sensor on the hand were combined to form a 9-dimensinal feature vector for every time frame. The position of the gesturing hand in the 3D space (x, y, z) relative to the shoulder center joint was used to form a 3-dimensional feature vector for each frame from a Kinect sensor. The two sets of features were then concatenated. This approach can be viewed as using the raw data from the sensors (e.g., accelerations from the accelerometer and joint positions estimated using the Kinect SDK) without further processing. Similarly, in [43] and [44], the raw data from both a Kinect depth camera and an inertial body sensor (position data of the hand joint from a Kinect depth camera as well as acceleration data and angular velocity data from an inertial sensor) was utilized for hand gesture recognition. The advantage of using raw data is that there is no computational burden associated with feature extraction. However, raw signal data may not exhibit enough discriminatory power to achieve high accuracy for action recognition.

In [13], features extracted from depth sequences, named depth motion maps (DMMs) [18], were considered. Each depth frame in a depth video sequence was first projected onto three orthogonal Cartesian planes to form three projected maps. For each projection view, the absolute difference between two consecutive projected maps was accumulated through an entire depth video sequence forming a depth motion map. In this manner, three DMMs corresponding to three projection views (front view, side view and top view) were generated for a depth video sequence as the feature representation. For the inertial sensor, each acceleration and gyroscope signal sequence was partitioned into N temporal windows. Considering that statistical measures of mean, variance, standard deviation, and root mean square are computationally efficient and useful for capturing structural patterns in motion data, these four measures were computed along each direction in each temporal window. All the features from the temporal windows were concatenated to form a single combined feature vector. In [47], each depth video was first divided into 8 disjoint Depth-Layered Multi-Channel (DLMC) videos [46] by dividing the depth range into 8 equal depth layers and by keeping the pixels within the depth range of the corresponding depth layer. Then, so called Histogram of Gradients (HOG) and Histogram of Flow (HOF) features [39] were extracted from each DLMC video. This was followed by employing the Bag-of-Features (BoF) representation [53] to code DLMC videos into histograms to serve as the feature representation of depth sequences. As for inertial sensor data, $N_s = 30$ temporal windows from each accelerometer sequence were extracted and the variance of the acceleration along each direction in each temporal window was computed for each of the 6 accelerometers. By concatenating variance values from all the accelerometers, a local

temporal feature descriptor per temporal window was obtained. As a result, an action sequence was represented by a set of N_s feature descriptors. To further build a compact representation of these feature descriptors, the features were quantized into 20 codewords for classification purposes.

Table 4 lists a summary of the features that have been extracted simultaneously from a depth sensor and an inertial sensor for human action recognition. An important issue to keep in mind here is that these features have been designed to be computationally efficient allowing the developed action recognition systems to operate in real-time.

4.4 Classification and fusion approach

After feature extraction, classification and a fusion approach are needed for action recognition. Classical classifiers such as support vector machine (SVM) [9, 29] and hidden Markov model (HMM) [23] are often employed for action recognition, e.g., [43, 47]. Fusion of information from two sensors can be done in different ways. In general, three fusion approaches are applicable: (1) data-level fusion, (2) feature-level fusion, and (3) decision-level fusion. Data-level fusion occurs at the data level where incoming raw data from different sensors are combined. Feature-level fusion involves carrying out fusion of features after features are extracted from raw data. Decision-level fusion involves fusing the decisions made by individual classifiers or decision makers. Next, several representative classification and fusion approaches previously used for depth and inertial sensor fusion are mentioned.

In [13], both feature-level fusion and decision-level fusion were examined by using a collaborative representation classifier [78]. In the feature-level fusion approach, the features generated from the two differing modality sensors were merged before classification while in the decision-level fusion approach, the Dempster-Shafer theory [54] was used to combine the classification outcomes from two classifiers, each corresponding to one sensor. In [43], the depth and inertial sensor data were concatenated. An HMM classifier was employed for gesture recognition on the fused data. To further improve the gesture recognition performance, the data from the depth sensor and the inertial sensor were fed into multiple HMM classifiers [44]. Then, the probability outputs from the multiple HMM classifiers were combined to generate the final outcome. In [52] and [49], each gesture was considered to consist of three phases: pre-stroke, nucleus, and post-stroke. Each phase was modeled as an HMM. The gesture spotting and recognition was then performed based on concatenated HMMs trained for the three gesture phases.

Table 5 lists a summary of the classification and fusion approaches used for action/gesture recognition based on depth and inertial sensor fusion that have appeared in the literature.

As a general observation, it is seen that the simultaneous utilization of these two differing modality sensors allows one to achieve higher human action recognition performance compared to the situations when each sensor is used individually or on its own. For example, in [43], the gesture recognition accuracies of using a Kinect depth sensor alone and an inertial sensor alone were reported to be 84 and 88 %, respectively. However, when using the depth and inertial sensor fusion, the recognition accuracy was increased to 93 %. In [13], the fusion approach was evaluated based on the Berkeley multimodal human action database [47] and the results indicated that due to the complementary aspect of the data from the depth and inertial sensors, the fusion approach led to 2 to 23 % improvements in recognition accuracy depending on the type of action performed over the situations when each sensor was used individually. A 99 % correct walking pattern recognition rate was reported in [67] by combining an inertial sensor and a Kinect depth sensor as compared to the recognition rates of 89 and 65 % when using the inertial sensor alone and the depth sensor alone, respectively. For human fall detection in [36], a recognition accuracy of 98 % was achieved by using

Table 4 A summary of features extracted simultaneously from a depth sensor and an inertial sensor for action/gesture recognition involving depth and inertial sensor fusion

References	Features		Task
	Depth sensor	Inertial sensor	
[10, 43, 44, 77]	Raw data: skeleton joint positions	Raw data: accelerations, angular velocities, etc.	Gesture recognition
[12–14]	Depth motion maps [18]	Statistical features (e.g., mean, variance, etc.) computed from temporal segments of inertial sensor signals (e.g., accelerations and angular velocities)	Action recognition
[47]	Each depth video was into multiple Depth-Layered Multi-Channel (DLMC) videos [46]. HOG and HOF features were extracted from each DLMC video and coded into histograms using Bag-of-Features (BoF) model [53]	Variance of the acceleration in each direction in each temporal window was computed as feature descriptors. The feature descriptors were then quantized into a number of codewords.	Action recognition
[21]	Raw data: skeleton joint positions	Raw data: accelerations	Action recognition
[67]	Subject velocity, body azimuth angle, body inclination angle, leg separation distance, and leg separation angle calculated using 3D point cloud	Raw data: accelerations	Gait (different walking patterns) classification
[58]	Statistical features (mean, energy, standard deviation, entropy) for the visual displacement components extracted from a fixed temporal window of video frames	Statistical features (e.g., mean, variance, etc.) computed from temporal windows of acceleration signals.	Food preparation activities recognition
[68]	Skeleton joint positions were first transformed to be invariant to user's position, orientation, and body size. The final feature vector consists of four parts: 1. Absolute 3D position of joint points. 2. Relative 3D position of joint points, de fined on directly connected joint pairs. 3. First order difference in time of part 1 in the feature vector. 4. First order difference in time of part 2 in the feature vector.	Raw data: accelerations, angular velocities, etc.	Gesture recognition
[36, 37]	1. A ratio of width to height of the person's bounding box in the depth maps. 2. A proportion expressing the height of the person's surrounding box in the current frame to the physical height of the person, projected onto the depth map. 3. The distance of the person's centroid to the floor.	Raw data: accelerations and angular velocities	Fall detection

Table 4 (continued)

References	Features		Task
	Depth sensor	Inertial sensor	
	4. Standard deviation from the centroid for the abscissa and the applicate, respectively.		
[26]	Variation in the skeleton joint position	1. The acceleration magnitude of the wrist accelerometer 2. The angle between the X-axis and the gravity vector of the wrist accelerometer	Fall detection

a depth sensor and an accelerometer together, which was about 8 % higher than when using only the depth sensor and about 3 % when using only the accelerometer.

5 Public-domain multimodal human action datasets

A number of public-domain human action datasets have been created based on depth sensor only (e.g., MSR Action3D dataset [42]) and inertial sensor only (e.g., Wearable Action Recognition Database (WARD) [72]). This section includes a review of the publicly available multimodal human action datasets that contain simultaneously captured data from depth and inertial sensors. The reader is referred to the publications listed for the details of the datasets.

5.1 Berkeley multimodal human action database (MHAD)

The Berkeley MHAD dataset[2] [47] contains temporally synchronized data from a motion capture system, 12 RGB cameras, 2 Microsoft Kinect depth cameras, 6 wearable accelerometers, and 4 microphones. The dataset consists of 659 data sequences from 11 human actions performed by 7 male and 5 female subjects of 23–30 years age except for one elderly subject with 5 repetitions of each action. These 11 actions include: *jumping in place*, *jumping jacks*, *bending-hands up all the way down*, *punching*, *waving two hands*, *waving right hand*, *clapping hands*, *throwing a ball*, *sit down and stand up*, *sit down*, and *stand up*. The 6 accelerometers were placed at the wrists, ankles and hips. The two Microsoft Kinect cameras were placed approximately in opposite directions to prevent interference between the two active projection patterns. Each Kinect camera captured a color image with a resolution of 640 × 480 pixels and a 16-bit depth image, both with an acquisition rate of 30 Hz.

5.2 University of Rzeszow fall detection (URFD) dataset

The UR fall detection dataset[3] [36] focuses on the human fall detection application. It contains 70 (30 falls and 40 activities of daily living) sequences from 5 subjects. The 30 fall sequences were recorded with 2 Kinect cameras as well as a 3-axis accelerometer. One of the Kinect

[2] http://tele-immersion.citris-uc.org/berkeley_mhad
[3] http://fenix.univ.rzeszow.pl/~mkepski/ds/uf.html

Table 5 A summary of classification and fusion approaches used for action/gesture recognition

References	Classification and fusion method	Task
[13]	1. Feature-level fusion: features from depth and inertial sensor data were concatenated. The concatenated features were used as input to a collaborative representation classifier [78]. 2. Decision-level fusion: features extracted from depth sensor data and inertial sensor data were used individually as input to two collaborative representation classifiers. Dempster-Shafer theory [54] was utilized to combine the combine the classification outcomes from two classifiers.	Action recognition
[43]	Depth and inertial sensor data were concatenated. HMM was employed for classification on the fused sensor data. (Data-level fusion)	Gesture recognition
[44]	Data from the depth sensor and the inertial sensor were fed into multiple HMM classifiers. The probability outputs from the multiple HMM classifiers were combined to generate the final outcome. (Decision-level fusion)	Gesture recognition
[68]	A Bayesian co-boosting framework was proposed to combine features from depth and inertial sensor modalities. HMM was used as the week classifier in the boosting framework. (Feature-level fusion)	Gesture recognition
[47]	Multiple kernel learning (MKL) strategy [28] was used for combining various modalities for action recognition. Different weights computed by MKL were assigned to different modalities. SVM was used as the classifier. (Feature-level fusion)	Action recognition
[10]	A coupled hidden Markov model (CHMM) was employed to discover the correlation and complementary information across different modalities. (Data-level fusion)	Gesture recognition
[36, 37]	Initial fall detection was based on acceleration data. The acceleration data was also employed to trigger the processing of the depth images for fall detection based on K nearest neighbor classifier or SVM classifier to further reduce the false alarm. (Decision-level fusion)	Fall detection

cameras was mounted on the ceiling and the other was placed in front of the subjects. The accelerometer was worn near the spine on the lower back of a subject. Depth data and RGB data from the two Kinect cameras were recorded for the fall sequences. This dataset also consists of data corresponding to the normal activities of walking, sitting down, crouching down and lying in order to evaluate the performance of the fall detection algorithm. The 40 sequences of the activities provided in this dataset only contain the depth and RGB data from one Kinect camera.

5.3 University of Texas at Dallas multimodal human action dataset (UTD-MHAD)

The UTD-MHAD dataset[4] [12] is a comprehensive multimodal human action dataset that consists of data from a Kinect sensor and a wearable inertial sensor capturing 3-axis acceleration and 3-axis angular velocity signals [17]. This dataset consists of four temporally synchronized data modalities, which include RGB videos, depth videos, skeleton positions from a Kinect camera sensor, and inertial signals (acceleration and angular velocity) from a

[4] http://www.utdallas.edu/~kehtar/UTD-MHAD.html

wearable inertial sensor for a comprehensive set of 27 human actions encountered in the literature on human action recognition. The 27 actions include: (1) *right arm swipe to the left*, (2) *right arm swipe to the right*, (3) *right hand wave*, (4) *two hand front clap*, (5) *right arm throw*, (6) *cross arms in the chest*, (7) *basketball shoot*, (8) *right hand draw X*, (9) *right hand draw circle* (*clockwise*), (10) *right hand draw circle* (*counter clockwise*), (11) *draw triangle*, (12) *bowling* (*right hand*), (13) *front boxing*, (14) *baseball swing from right*, (15) *tennis right hand forehand swing*, (16) *arm curl* (*two arms*), (17) *tennis serve*, (18) *two hand push*, (19) *right hand knock on door*, (20) *right hand catch an object*, (21) *right hand pick up and throw*, (22) *jogging in place*, (23) *walking in place*, (24) *sit to stand*, (25) *stand to sit*, (26) *forward lunge* (*left foot forward*), (27) *squat* (*two arms stretch out*). These actions were performed by 8 subjects (4 females and 4 males). Each subject repeated each action 4 times. After removing three corrupted sequences, the dataset included 861 data sequences. The wearable inertial sensor was placed on the subjects' right wrists for actions (1) through (21) which were hand type movements, and on the subjects' right thigh for actions (22) through (27) which were leg type movements.

5.4 50 salads dataset

The 50 salads dataset[5] [58] is a publicly available dataset of complex activities that involve manipulative gestures. It captures 25 people preparing 2 mixed salads and contains over 4 h of annotated accelerometer and RGB-D video data (i.e., acceleration data, RGB video data and depth data). A Kinect camera was mounted on the wall to have a top-down view onto the work space. Accelerometers were embedded in the handles of a knife, a mixing spoon and a peeler. Additional accelerometers were attached to a small spoon, a glass, an oil bottle, and a pepper dispenser. 27 subjects prepared a mixed salad two times. Two subjects were excluded from the final dataset due to data loss. Preparing the mixed salad involved preparing a dressing with salt, pepper, olive oil and balsamic vinegar, cutting ingredients (cucumber, tomato, feta cheese and lettuce) into pieces, mixing ingredients, adding the dressing to the salad and serving the salad onto a plate. The following activities were annotated: *add oil*, *add vinegar*, *add salt*, *add pepper*, *mix dressing*, *peel cucumber*, *cut cucumber*, *place cucumber into bowl*, *cut cheese*, *place cheese into bowl*, *cut lettuce*, *place lettuce into bowl*, *cut tomato*, *place tomato into bowl*, *mix ingredients*, *serve salad onto plate*, and *add dressing*. Each activity was split into three phases which were annotated individually: pre-, core- and post-phase. Each activity was associated with one of three stages in the recipe which were also annotated: prepare dressing, cut and mix ingredients and serve salad. In total, 966 activity instances were annotated. The 50 salads dataset is useful for carrying out research in activity recognition, activity spotting, sequence analysis, and sensor fusion.

5.5 ChAirGest multimodal dataset

The ChAirGest multimodal dataset[6] [52] was designed to encourage researchers to take advantage of data recorded from multiple sensors to optimize and evaluate

[5] http://cvip.computing.dundee.ac.uk/datasets/foodpreparation/50salads/
[6] https://project.eia-fr.ch/chairgest/Pages/Overview.aspx

methods for gesture spotting and recognition. The dataset contains 6 h of continuous multimodal recordings. 10 subjects were asked to mimick gestures seen on a computer screen. The data were acquired from a Kinect camera and 4 inertial motion units (IMUs) attached to the right arm and the neck of the subjects. The dataset contains 10 different gestures, started from 3 different resting postures and recorded in two different lighting conditions by 10 different subjects. The gestures considered include: *swipe left*, *swipe right*, *push to screen*, *take from screen*, *palm-up rotation*, *palm-down rotation*, *draw a circle I* (*longitudinally*), *draw a circle II*, *wave hello*, and *shake hand*. The total dataset contains 1200 annotated gestures split in continuous video sequences. The RGB stream, the depth stream and the 3D position of the upper-body skeleton joints from the Kinect camera were recorded. Each IMU provided linear acceleration, angular acceleration, magnetometer, Euler orientation and orientation quaternion at a frame rate of 50Hz.

5.6 Telecommunication systems team (TST) fall detection database

The TST fall detection database[7] [27] is another multimodal database which focuses on fall detection. The dataset was collected using a Microsoft Kinect v2 sensor and two IMUs. It is composed of daily living activities and the fall action simulated by 11 subjects. The actions performed by a single subject were separated in two main groups: daily living activity and fall. Each activity was repeated three times by each subject. The daily living activities include: *sit on a chair*, *walk and grasp an object from the floor*, *walk back and forth*, and *lie down on the mattress*. Four types of fall were performed including *fall from the front and ends up lying*, *fall backward and ends up lying*, *fall to the side and ends up lying*, and *fall backward and ends up sitting*. The data include two raw acceleration streams provided by the two IMUs placed on the waist and the right wrist, and the depth frames and skeleton joints captured by the Kinect sensor. This database contains a total of 264 action sequences.

5.7 Huawei/3DLife dataset

The Huawei/3DLife dataset[8] [59] is a multimodal dataset designed for 3D human reconstruction and action recognition from multiple active and passive sensors. The dataset consists of RGB and depth video streams from 5 Kinect sensors at different viewpoints covering the entire body, audio streams captured by the 5 Kinect sensors, and inertial sensor data captured from 8 IMUs on a subject's body. The 8 IMUs were attached to the following locations on the body: left wrist, right wrist, chest, hips, left ankle, right ankle, left foot and right foot. These devices captured 3D acceleration (using accelerometers), 3D magnetic flux (using magnetometers) and 3D angular rate (using gyroscopes). The data were captured in two sessions with different spatial arrangements of the sensors. 17 subjects participated in the data collection, all performing at least 5 instances of 22 different types of gestures/movements. Thus, the dataset consists of approximately 3740 instances. The actions performed are of three types and include: (i) simple actions that involve mainly the upper human body

[7] http://www.tlc.dii.univpm.it/blog/databases4kinect#IDFall
[8] http://mmv.eecs.qmul.ac.uk/mmgc2013/

(*hand waving*, *knocking on the door*, *clapping*, *throwing*, *punching*, and *push away with both hands*), (ii) training exercises (*jumping jacks*, *lunges*, *squats*, *punching and hen kicking*, and *weight lifting*), (iii) sports related activities (*golf drive*, *golf chip*, *golf putt*, *tennis forehand*, *tennis backhand*, and *walking on the treadmill*), and (iv) static gestures (*arms folded*, *T-pose*, *hand on the hips*, *T-pose with bent arms*, and *forward arms raise*).

Table 6 lists a summary of the number of sensors, number of subjects, number of actions, and number of action sequences for publicly available human action datasets involving differing modality sensors.

6 Conclusion and future directions

In this survey paper, a review of different approaches for human action recognition that involve the simultaneous utilization of both depth and inertial sensors has been presented. After covering the main challenges or limitations associated with using each sensor modality individually, the existing literature on fusing depth and inertial sensor data for human action recognition is outlined. In addition, the components in the existing fusion approaches are reviewed. A review of publicly available human action datasets where both depth and inertial sensor data are collected simultaneously has been provided. As a general observation, it is noted that the simultaneous utilization of these two differing modality sensors allows one to achieve higher human action recognition performance compared to situations when each sensor is used individually. Considering that both of these two sensors are cost-effective and able to provide 3D data, it is anticipated that the research on action recognition based on depth and inertial sensor fusion will receive growing attention. However, there still remain challenges when using fusion of depth and inertial information for action recognition. A few possible research directions to address these challenges are noted below.

1) Developing view-invariant features for depth images. Under realistic operating conditions, a subject may perform an action at an arbitrary orientation with respect to the camera. Noting that inertial sensor signals (accelerations and angular velocities) are view invariant,

Table 6 Publicly available multimodal human action datasets involving different modalities: (M)otion capture, RGB (V)ideo, (D)epth, (A)udio, and (I)nertial

Dataset	Modality					# Sub	# Act	# Seq
	M	V	D	A	I			
Berkeley MHAD [47]	1	12	2	4	6	12	11	660
URFD [36]	–	2	2	–	1	5	>5	70
UTD-MHAD [12]	–	1	1	–	1	8	27	861
50 salads [58]	–	1	1	–	7	25	17	966
ChAirGest [52]	–	1	1	–	4	10	10	1200
TST Fall detection database [27]	–	–	1	–	2	11	8	264
Huawei/3DLife dataset [59]	–	5	5	5	8	17	22	3740

a future research direction would be developing view invariant features for depth images towards achieving view invariant action recognition.

2) Developing intelligent fusion approaches. Realistic limitations that may occur in practice such as occlusion in case of depth cameras or loss of signal in case of inertial sensors can be studied further. In other words, when data from one modality sensor are not reliable, a future research direction would be developing techniques for the system to intelligently switch to the sensor that is providing reliable data.
3) Examining human and object interactions. Many activities involve human object interactions. Such activities normally consist of multiple sub-actions that involve interactions between a human and objects. A future research direction would be developing fusion schemes for situations when inertial sensors are attached to the objects that are being interacted with.

Acknowledgments This work was supported in part by the National Science Foundation, under grant 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.

References

1. Aggarwal JK, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv (CSUR) 43(3):16
2. Aggarwal JK, Xia L (2014) Human activity recognition from 3d data: a review. Pattern Recogn Lett 48:70–80
3. Altun K, Barshan B (2010) Human activity recognition using inertial/magnetic sensor units. In: Human behavior understanding, pp 38–51
4. Argyriou V, Petrou M, Barsky S (2010) Photometric stereo with an arbitrary number of illuminants. Comput Vis Image Underst 114(8):887–900
5. Avci A, Bosch S, Marin-Perianu M, Marin-Perianu R, Havinga P (2010) Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: a survey. In: Architecture of Computing Systems (ARCS), 2010 23rd International Conference on, pp 1–10
6. Bidmeshki MM, Jafari R (2013) Low power programmable architecture for periodic activity monitoring. In: Proceedings of the ACM/IEEE 4th International Conference on Cyber-Physical Systems, pp 81–88
7. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267
8. Bulling A, Blanke U, Schiele B (2014) A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput Surv (CSUR) 46(3):33
9. Burges CJ (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2): 121–167
10. Cao C, Zhang Y, Lu H (2015) Multi-modal learning for gesture recognition. In: Multimedia and Expo (ICME), 2015 I.E. International Conference on, pp 1–6
11. Chen L, Hoey J, Nugent CD, Cook DJ, Yu Z (2012) Sensor-based activity recognition. IEEE Trans Syst Man Cybern Part C Appl Rev 42(6):790–808
12. Chen C, Jafari R, Kehtarnavaz N (2015) UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: Proceedings of the IEEE International Conference on Image Processing. Canada
13. Chen C, Jafari R, Kehtarnavaz N (2015) Improving human action recognition using fusion of depth camera and inertial sensors. IEEE Trans Human-Machine Syst 45(1):51–61
14. Chen C, Jafari R, Kehtarnavaz N (2015) A real-time human action recognition system using depth and inertial sensor fusion. IEEE Sensors J 2015

15. Chen C, Jafari R, Kehtarnavaz N (2015) Action recognition from depth sequences using depth motion maps-based local binary patterns. In: Applications of Computer Vision (WACV), 2015 I.E. Winter Conference on, pp 1092–1099
16. Chen C, Kehtarnavaz N, Jafari R (2014) A medication adherence monitoring system for pill bottles based on a wearable inertial sensor. In: Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, pp 4983–4986
17. Chen C, Liu K, Jafari R, Kehtarnavaz N (2014) Home-based senior fitness test measurement system using collaborative inertial and depth sensors. In: Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, pp 4135–4138
18. Chen C, Liu K, Kehtarnavaz N (2013) Real-time human action recognition based on depth motion maps. J Real-Time Image Proc 1–9
19. Chen L, Wei H, Ferryman J (2013) A survey of human motion analysis using depth imagery. Pattern Recogn Lett 34(15):1995–2006
20. Cippitelli E, Gasparrini S, Gambi E, Spinsante S, Wahsleny J, Orhany I, Lindhy T (2015) Time synchronization and data fusion for RGB-depth cameras and inertial sensors in AAL applications. In: Communication Workshop (ICCW), 2015 I.E. International Conference on, pp 265–270
21. Delachaux B, Rebetez J, Perez-Uribe A, Mejia HFS (2013) Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors. In: Advances in Computational Intelligence, pp 216–223
22. Destelle F, Ahmadi A, O'Connor NE, Moran K, Chatzitofis A, Zarpalas D, Daras P (2014) Low-cost accurate skeleton tracking based on fusion of kinect and wearable inertial sensors. In: Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, pp 371–375
23. Eddy SR (2004) What is a hidden Markov model? Nat Biotechnol 22(10):1315–1316
24. Ermes M, Parkka J, Mantyjarvi J, Korhonen I (2008) Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. IEEE Trans Inf Technol Biomed 12(1): 20–26
25. Evangelidis G, Singh G, Horaud R (2014) Skeletal quads: human action recognition using joint quadruples. In: Pattern Recognition (ICPR), 2014 22nd International Conference on, pp 4513–4518
26. Gasparrini S, Cippitelli E, Gambi E, Spinsante S, Wåhslén J, Orhan I, Lindh T (2016) Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion. In: ICT Innovations 2015, pp 99–108
27. Gasparrini S, Cippitelli E, Spinsante S, Gambi E (2014) A depth-based fall detection system using a Kinect® sensor. Sensors 14(2):2756–2775
28. Gehler P, Nowozin S (2009) On feature combination for multiclass object classification. In: Computer Vision, 2009 I.E. 12th International Conference on, pp 221–228
29. Gu B, Sheng VS, Tay KY, Romano W, Li S (2015) Incremental support vector learning for ordinal regression. IEEE Trans Neural Netw Learn Syst 26(7):1403–1416
30. Guan D, Ma T, Yuan W, Lee YK, Jehad Sarkar AM (2011) Review of sensor-based activity recognition systems. IETE Tech Rev 28(5):418–433
31. Han J, Shao L, Xu D, Shotton J (2013) Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans Cybernet 43(5):1318–1334
32. Helten T, Muller M, Seidel HP, Theobalt C (2013) Real-time body tracking with one depth camera and inertial sensors. In: Computer Vision (ICCV), 2013 I.E. International Conference on, pp 1105–1112
33. http://www.microsoft.com/en-us/kinectforwindows/
34. Jovanov E, Milenkovic A, Otto C, De Groen PC (2005) A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. J NeuroEng Rehabil 2(1):6
35. Klaser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients. In: BMVC 2008-19th British Machine Vision Conference, pp 275–1. British Machine Vision Association
36. Kwolek B, Kepski M (2014) Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput Methods Prog Biomed 117(3):489–501
37. Kwolek B, Kepski M (2015) Improving fall detection by the use of depth sensor and accelerometer. Neurocomputing 168:637–645
38. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2–3):107–123
39. Laptev I, Marszałek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. IEEE Conference on, pp 1–8
40. Lara OD, Labrador MA (2013) A survey on human activity recognition using wearable sensors. IEEE Commun Surv Tutorials 15(3):1192–1209

41. Li Q, Stankovic J, Hanson M, Barth AT, Lach J, Zhou G (2009) Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information. In: Wearable and Implantable Body Sensor Networks, 2009. BSN 2009. Sixth International Workshop on, pp 138–143
42. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3d points. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 I.E. Computer Society Conference on, pp 9–14
43. Liu K, Chen C, Jafari R, Kehtarnavaz N (2014) Fusion of inertial and depth sensor data for robust hand gesture recognition. IEEE Sensors J 14(6):1898–1903
44. Liu K, Chen C, Jafari R, Kehtarnavaz N (2014) Multi-HMM classification for hand gesture recognition using two differing modality sensors. In: Circuits and Systems Conference (DCAS), 2014 I.E. Dallas, pp 1–4
45. Mukherjee S, Biswas SK, Mukherjee DP (2011) Recognizing human action at a distance in video by key poses. IEEE Trans Circuits Syst Video Technol 21(9):1228–1241
46. Ni B, Wang G, Moulin P (2013) Rgbd-hudaact: a color-depth video database for human daily activity recognition. In: Consumer Depth Cameras for Computer Vision, pp 193–208
47. Ofli F, Chaudhry R, Kurillo G, Vidal R, Bajcsy R (2013) Berkeley mhad: a comprehensive multimodal human action database. In: Applications of Computer Vision (WACV), 2013 I.E. Workshop on, pp 53–60
48. Oreifej O, Liu Z (2013) Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences. In: Computer Vision and Pattern Recognition (CVPR), 2013 I.E. Conference on, pp 716–723
49. Pavlovic V, Sharma R, Huang TS (1997) Visual interpretation of hand gestures for human-computer interaction: a review. IEEE Trans Pattern Anal Mach Intell 19(7):677–695
50. Poppe R (2010) A survey on vision-based human action recognition. Image Vis Comput 28(6): 976–990
51. Ramanathan M, Yau WY, Teoh EK (2014) Human action recognition with video data: research and evaluation challenges. IEEE Trans Human-Machine Syst 44(5):650–663
52. Ruffieux S, Lalanne D, Mugellini E (2013) ChAirGest: a challenge for multimodal mid-air gesture recognition for close HCI. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp 483–488
53. Schüldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3, pp 32–36
54. Shafer G (1976) A mathematical theory of evidence, vol 1. Princeton University Press, Princeton
55. Shan J, Akella S (2014) 3D human action segmentation and recognition using pose kinetic energy. In: Advanced Robotics and its Social Impacts (ARSO), 2014 I.E. Workshop on, pp 69–75
56. Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Moore R (2013) Real-time human pose recognition in parts from single depth images. Commun ACM 56(1):116–124
57. Spriggs EH, De La Torre F, Hebert M (2009) Temporal segmentation and activity classification from first-person sensing. In: Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on, pp 17–24
58. Stein S, McKenna SJ (2013) Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp 729–738
59. Sun L, Aizawa K (2013) Action recognition using invariant features under unexampled viewing conditions. In: Proceedings of the 21st ACM International Conference on Multimedia, pp 389–392
60. Theodoridis T, Agapitos A, Hu H, Lucas SM (2008) Ubiquitous robotics in physical human action recognition: a comparison between dynamic anns and gp. In: Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on, pp 3064–3069
61. Tian Y, Meng X, Tao D, Liu D, Feng C (2015) Upper limb motion tracking with the integration of IMU and Kinect. Neurocomputing 159:207–218
62. Vemulapalli R, Arrate F, Chellappa R (2014) Human action recognition by representing 3d skeletons as points in a lie group. In: Computer Vision and Pattern Recognition (CVPR), 2014 I.E. Conference on, pp 588–595
63. Vieira AW, Nascimento ER, Oliveira GL, Liu Z, Campos MF (2012) Stop: space-time occupancy patterns for 3d action recognition from depth map sequences. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp 252–259
64. Wang J, Liu Z, Chorowski J, Chen Z, Wu Y (2012) Robust 3d Action Recognition with Random Occupancy Patterns. In: Computer Vision–ECCV 2012, pp 872–885
65. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: Computer Vision and Pattern Recognition (CVPR), 2012 I.E. Conference on, pp 1290–1297
66. Weinland D, Ronfard R, Boyer E (2011) A survey of vision-based methods for action representation, segmentation and recognition. Comput Vis Image Underst 115(2):224–241

67. Wong C, McKeague S, Correa J, Liu J, Yang G Z (2012) Enhanced classification of abnormal gait using BSN and depth. In: Wearable and Implantable Body Sensor Networks (BSN), 2012 Ninth International Conference on, pp 166–171
68. Wu J, Cheng J (2014) Bayesian co-boosting for multi-modal gesture recognition. J Mach Learn Res 15(1): 3013–3036
69. Xia L, Aggarwal JK (2013) Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: Computer Vision and Pattern Recognition (CVPR), 2013 I.E. Conference on, pp 2834–2841
70. Xie S, Wang Y (2014) Construction of tree network with limited delivery latency in homogeneous wireless sensor networks. Wirel Pers Commun 78(1):231–246
71. Yang AY, Iyengar S, Sastry S, Bajcsy R, Kuryloski P, Jafari R (2008) Distributed segmentation and classification of human actions using a wearable motion sensor network. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pp 1–8
72. Yang AY, Jafari R, Sastry SS, Bajcsy R (2009) Distributed recognition of human actions using wearable motion sensor networks. J Ambient Intell Smart Environ 1(2):103–115
73. Yang X, Tian Y (2012) Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 I.E. Computer Society Conference on, pp 14–19
74. Yang X, Tian Y (2014) Super normal vector for activity recognition using depth sequences. In: Computer Vision and Pattern Recognition (CVPR), 2014 I.E. Conference on, pp 804–811
75. Yang X, Zhang C, Tian Y (2012) Recognizing actions using depth motion maps-based histograms of oriented gradients. In: Proceedings of the 20th ACM International Conference on Multimedia, pp 1057–1060
76. Ye M, Zhang Q, Wang L, Zhu J, Yang R, Gall J (2013) A survey on human motion analysis from depth data. In: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications. Springer Berlin Heidelberg, pp 149–187
77. Yin Y, Davis R (2013) Gesture spotting and recognition using salience detection and concatenated hidden markov models. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp 489–494
78. Zhang L, Yang M, Feng X (2011) Sparse representation or collaborative representation: which helps face recognition?. In: Computer Vision (ICCV), 2011 I.E. International Conference on, pp 471–478

Chen Chen received the BE degree in automation from Beijing Forestry University, Beijing, China, in 2009 and the MS degree in electrical engineering from Mississippi State University, Starkville, in 2012. He is a PhD candidate in the Department of Electrical Engineering at the University of Texas at Dallas, Richardson, TX. His research interests include compressed sensing, signal and image processing, pattern recognition and computer vision.

Roozbeh Jafari is an associate professor in Biomedical Engineering, Computer Science and Engineering and Electrical and Computer Engineering at Texas A&M University. He received his PhD in Computer Science from UCLA and completed a postdoctoral fellowship at UC-Berkeley. His research interest lies in the area of wearable computer design and signal processing. His research has been funded by the NSF, NIH, DoD (TATRC), AFRL, AFOSR, DARPA, SRC and industry (Texas Instruments, Tektronix, Samsung & Telecom Italia). He has published over 100 papers in refereed journals and conferences. He is the recipient of the NSF CAREER award in 2012, IEEE Real-Time & Embedded Technology & Applications Symposium (RTAS) best paper award in 2011 and Andrew P. Sage best transactions paper award from IEEE Systems, Man and Cybernetics Society in 2014. He is an associate editor for the IEEE Sensors Journal, IEEE Internet of Things Journal and IEEE Journal of Biomedical and Health Informatics.

Nasser Kehtarnavaz received the Ph.D. degree in electrical and computer engineering from Rice University in 1987. He is a Professor of Electrical Engineering and Director of the Signal and Image Processing Laboratory at the University of Texas at Dallas. His research areas include signal and image processing, real-time processing on embedded processors, biomedical image analysis, and pattern recognition. He has authored or co-authored 9 books and more than 300 publications in these areas. He has had industrial experience in various capacities at Texas Instruments, AT&T Bell Labs, US Army TACOM Research Lab, and Houston Health Science Center. He is currently Editor-in-Chief of Journal of Real-Time Image Processing, and Chair of the SPIE Conference on Real-Time Image and Video Processing. Dr. Kehtarnavaz is a Fellow of IEEE, a Fellow of SPIE, and a licensed Professional Engineer.