IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 31, 2019, accepted April 19, 2019, date of publication April 29, 2019, date of current version May 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913393

Robust Human Activity Recognition Using Multimodal Feature-Level Fusion

MUHAMMAD EHATISHAM-UL-HAQ[1], ALI JAVED[2], (Member, IEEE), MUHAMMAD AWAIS AZAM[1], HAFIZ M. A. MALIK[3], (Senior Member, IEEE), AUN IRTAZA[4], IK HYUN LEE[5], AND MUHAMMAD TARIQ MAHMOOD[6], (Senior Member, IEEE)

[1]Department of Computer Engineering, University of Engineering and Technology, Taxila 47080, Pakistan
[2]Department of Software Engineering, University of Engineering and Technology, Taxila 47080, Pakistan
[3]Electrical and Computer Engineering Department, University of Michigan–Dearborn, Dearborn, MI 48128, USA
[4]Department of Computer Science, University of Engineering and Technology, Taxila 47080, Pakistan
[5]Department of Mechatronics, Korea Polytechnic University, Gyeonggi-do 15073, South Korea
[6]School of Computer Science and Information Engineering, Korea University of Technology and Education, Cheonan 31253, South Korea

Corresponding author: Muhammad Tariq Mahmood (tariq@koreatech.ac.kr)

This work was supported in part by the Basic Science Research Program under Grant 2017R1D1A1B03033526 and Grant 2016R1D1A1B03933860, and in part by the Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2017R1A6A1A03015562.

ABSTRACT Automated recognition of human activities or actions has great significance as it incorporates wide-ranging applications, including surveillance, robotics, and personal health monitoring. Over the past few years, many computer vision-based methods have been developed for recognizing human actions from RGB and depth camera videos. These methods include space–time trajectory, motion encoding, key poses extraction, space–time occupancy patterns, depth motion maps, and skeleton joints. However, these camera-based approaches are affected by background clutter and illumination changes and applicable to a limited field of view only. Wearable inertial sensors provide a viable solution to these challenges but are subject to several limitations such as location and orientation sensitivity. Due to the complementary trait of the data obtained from the camera and inertial sensors, the utilization of multiple sensing modalities for accurate recognition of human actions is gradually increasing. This paper presents a viable multimodal feature-level fusion approach for robust human action recognition, which utilizes data from multiple sensors, including RGB camera, depth sensor, and wearable inertial sensors. We extracted the computationally efficient features from the data obtained from RGB-D video camera and inertial body sensors. These features include densely extracted histogram of oriented gradient (HOG) features from RGB/depth videos and statistical signal attributes from wearable sensors data. The proposed human action recognition (HAR) framework is tested on a publicly available multimodal human action dataset UTD-MHAD consisting of 27 different human actions. K-nearest neighbor and support vector machine classifiers are used for training and testing the proposed fusion model for HAR. The experimental results indicate that the proposed scheme achieves better recognition results as compared to the state of the art. The feature-level fusion of RGB and inertial sensors provides the overall best performance for the proposed system, with an accuracy rate of 97.6%.

INDEX TERMS Dense HOG, depth sensor, feature-level fusion, human action recognition, inertial sensor, RGB camera.

I. INTRODUCTION

Human action recognition (HAR) or activity recognition is an imperious area of research in signal and image processing. HAR mainly involves automatic detection, localization, recognition, and analysis of human actions from the data obtained from different types of sensors, including RGB camera, depth sensor, range sensor, or inertial sensor. Action detection involves determining the presence of the action of interest in a continuous data stream, whereas action localization estimates when and where an action of interest appears. The goal of action recognition or classification is

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaoxiang Zhang.

to determine which action appears in the data. In the past few years, the research on HAR has gained significant popularity and is becoming increasingly vital in a variety of disciplines. Detecting and recognizing human activities is the core of many human-computer interaction (HCI) applications, including visual surveillance, video analytics, assistive living, intelligent driving, robotics, telemedicine, sports annotation, and health monitoring [1]–[6]. Various sensor modalities have been utilized to monitor human beings and their activities. HAR approaches can generally be classified into two main categories depending upon the type of sensors used. These include vision-based HAR and inertial sensor-based HAR.

Earlier vision-based action recognition studies involved the use of RGB video sequences captured by conventional RGB cameras to recognize a human activity [7], [8]. These studies are mostly based on template-based or model-based approaches [9]–[11], space-time trajectory [12], motion encoding [13], and key poses extraction [14]. Numerous feature extraction methods have been proposed for HAR using RGB video data, which achieved successful recognition results. Particularly, these methods include 3D gradient-based spatiotemporal descriptor [15], spatiotemporal interest point (STIP) detector [16], motion-energy images (MEIs) and motion history images (MHIs) [17], [18]. The evolution of deep learning schemes, i.e., deep learning based convolutional neural networks (CNN) and Long Short-Term Memory (LSTM) networks, has motivated the researchers to explore its application for action recognition from RGB videos [19]–[22]. The increasing popularity of HAR using RGB camera has also been heavily investigated in recent years [23]–[26]. These papers have provided a comprehensive discussion on different features and algorithms used in the literature for efficient HAR. With all their benefits, there exist some limitations in utilizing RGB cameras for monitoring human activities. For example, conventional RGB images lack 3D action data, which ultimately affects the recognition performance.

The advancement in image acquisition technology has made it possible to capture 3D action data using depth sensors. The depth images obtained for these sensors are insensitive to changes in illumination compared to conventional RGB images. Moreover, these depth images also provide a way to obtain 3D information of a person's skeleton to recognize human actions in a better way. Therefore, many researchers have put their efforts in recognizing human actions based on depth imagery [27]–[31]. Several feature extraction, description, and representation techniques have been developed for depth sensor-based HAR. These include depth motion maps (DMMs) [32], bag of 3D points [33], projected depth maps [34], space-time occupancy patterns [35], spatiotemporal depth cuboid [36], surface normal [37], and skeleton joints [38]. Recently, a few research studies proposed deep learning based methods for HAR using depth camera and skeleton joints [39]–[42]. In [43], the authors utilized CNN and LSTM for skeleton-based activity recognition. The authors in [44] proposed a deep bilinear learning method for RGB-D action recognition. A comprehensive study about RGB-D based human motion recognition using deep learning approaches is presented in [45]. Although vision-based HAR is continuously progressing, it is exposed to many hindrances such as camera position, a limited angle of view, subject disparities in carrying out different actions, occlusion, and background clutter. Furthermore, camera-based HAR systems require an extensive amount of hardware resources to run computationally complex computer vision algorithms. These limitations are addressed by low-cost, computationally efficient, and miniaturized inertial sensors.

Wearable inertial sensors enable dealing with a much broader field of view and changing illumination conditions as compared to RGB and depth sensors. They are attached directly on the human body or entrenched into outfits, smartphones, footwear, and wrist watches to track human activities. They generate 3D acceleration and rotation signals conforming to human action. Hence, like depth sensors, the inertial sensors also track 3D action data entailing 3-axis acceleration in case of an accelerometer and 3-axis angular velocity in case of a gyroscope. Many researchers utilized smartphones, smart watches, and wearable inertial sensors, incorporating an accelerometer and gyroscope, for human activity recognition [46]–[48]. In [49], [50], the authors detected complex human activities by utilizing the built-in inertial sensors of the smartphone along-with wrist-worn motion sensors. With the growth of deep learning applications in vision-based action recognition systems, we witnessed the utilization of deep learning for sensor-based activity recognition. In [51], the authors used deep learning for smartphone-sensor based activity recognition, whereas the authors in [52] used body sensor data for recognizing human activities. These studies achieved successful results in detecting and recognizing human activities. However, with the continuous evolvement in pulling down the power consumption of wearable sensors, deep learning based approaches are becoming futile for unobtrusive human activity monitoring. Moreover, sensor-based activity recognition approaches have certain other limitations as well. For instance, sensor readings are sensitive to their orientation and location on the body. Also, wearing or placing these sensors on the bodies creates inconvenience for the users to carry out their tasks in a natural way. Table 1 provides the pros and cons regarding the use of different sensing modalities (i.e., RGB camera, depth camera, and inertial sensors) for HAR.

A conventional HAR system typically makes use of a single sensor modality, i.e., either a vision-based sensing modality or a wearable inertial sensor. However, under realistic operational settings, no sensor modality alone can handle varying conditions that may take place in real time. The RGB and depth images from an RGB-D camera and 3D inertial signals from a wearable sensor offer complementary information. For instance, vision-based sensors provide global motion features whereas inertial signals give 3D

TABLE 1. Pros and Cons of different sensing modalities for HAR.

	RGB Cameras	Depth Sensors	Inertial Sensor
Pros	• Economical, easy-to-use and readily available • Offer color and texture information	• Inexpensive and widely available • Impermeable to variation in lighting and illumination settings • Insensitive to color and texture changes • Provide 3D action information • Insensitive to color and texture changes	• Very cheap and easily available • Impermeable to change in lighting and illumination settings • Provide a high sampling rate • Provide 3D action data • Can work in an unconstrained environment
Cons	• Limited viewing angle • Sensitive to variation in lighting and illumination • Affected by camera adjustment • Require computationally expensive processing algorithms	• Can work only in a constrained field of view • Sensitive to different types of noise • Depth information is sensitive to objects with varying refection properties • Lack of color an rich texture information	• Position and orientation sensitive • Sensor drift • Insufficient onboard power • Use multiple sensors for recording full body motion • Wearing these sensors create inconvenience for the users in performing their tasks

information about local body movement. Hence, by fusing data from two complementary sensing modalities, the performance of HAR systems can be improved. Few existing studies [53]–[56] utilized the fusion of depth and inertial sensors, aiming to increase the accuracy of action recognition and their results revealed significant improvement in recognition. Some authors also worked on using deep learning for multiple sensing modalities for robust action recognition [57]–[59]. In [60], the authors utilized deep learning based decision-level fusion for action recognition using depth camera and wearable inertial sensors. For depth cameras, CNN based features are extracted, whereas, for the inertial sensors, CNN and LSTM networks are used. Recently, in [61], the authors used skeleton-based LSTM and spatial CNN models to extract temporal and spatial features respectively for action recognition. The results of this study revealed that the fusion of multiple sensing modalities achieved a significant performance improvement compared to single modality based action recognition. Therefore, in this research work, we proposed a multimodal HAR framework that utilizes the combination of multiple sensing modalities (e.g., wearable inertial sensor, RGB camera sensor, and depth camera sensor) for action classification.

The fusion of multiple sensors can be performed at base-level (descriptor-level), feature-level (representation-level), or decision-level (score-level) [12]. Each fusion type has its own merits and demerits, and the selection of the fusion method is generally dependent on the type of features and descriptors. Existing studies for multimodal HAR mostly focus on the decision-level fusion due to its independence on the type, length, and numerical scale of different features extracted from multiple sensing modalities. Moreover, decision-level fusion does not require any post-processing of the extracted features and reduces the dimensions of the final feature vector for classification. The major drawback of the decision-level fusion is independent and stand-alone classification decisions relating to each sensing modality, which are then combined using some soft rule to make the final decision. Hence, for n different sensing modalities, the decision-level fusion requires n classifiers to be trained and tested independently on each sensing modality. For any multimodal HAR system, the acquisition of concurrent data from multiple sources is necessary to collect a sufficient amount of information for making improved decisions about human actions. However, with the decision-level fusion, it is not possible to combine multimodal data at an earlier stage to produce adequate information for recognizing human actions. In contrast, the feature-level fusion helps to collect concurrent features from multiple sensors and integrate them to generate sufficient information for making a strong decision. Moreover, it provides the best results in the case when the features extracted from different sensing modalities have the same dimensions and numerical scale. Therefore, in this study, we focused on the feature-level fusion of multiple sensing modalities for robust HAR. We extracted time domain features for inertial sensor data, whereas, to obtain the best results for feature-level fusion, we used densely extracted Histogram of Oriented Gradients (HOG) [62] as features for both RGB and depth video data. The features extracted from multiple sensors are then fused and used to train the machine learning algorithm for action classification.

The key contributions of this research work are as follows:

- A robust scheme is presented for HAR, which emphasized the feature-level fusion of RGB, depth, and inertial sensors to improve the accuracy of human action classification. Moreover, a detailed analysis is provided regarding the individual performance of these sensing modalities as well as their combination in HAR, using two common machine learning classifiers, i.e., *K-Nearest Neighbor* and *Support Vector Machine*.
- The existing approaches for RGB and depth sensor-based HAR use different types of features for both RGB

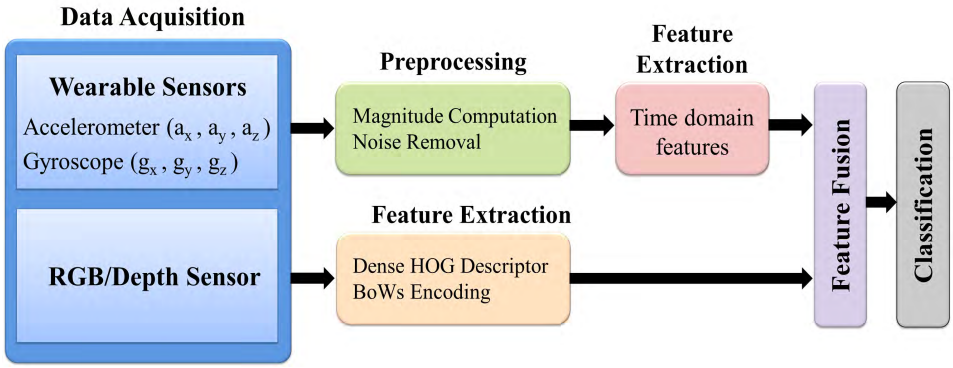

FIGURE 1. Block diagram of the proposed HAR method.

and depth videos, which becomes infeasible for the feature-level fusion. The proposed HAR method address this issue using RGB-D features based on densely extracted Histogram of Oriented Gradients (HOG). The obtained features are finally normalized to achieve the best recognition performance.

- The proposed HAR method is evaluated on publically available benchmark dataset *University of Texas at Dallas Multimodal Human Action Dataset (UTD-MHAD)* [53], which covers a wide-ranging set of 27 different human actions. The results achieved for the proposed scheme are better than state-of-the-art results. For demonstrating the effectiveness of the proposed feature-level fusion over decision-level fusion, the obtained results are also compared with the decision-level fusion results on *UTD-MHAD*.

The remaining part of the paper is organized as follows. Section II provides an in-depth discussion of the proposed method. Section III provides a discussion on the results of different experiments designed to measure the performance of the proposed HAR method. Also, we compared the performance of our method against different machine learning algorithms for HAR. Finally, Section IV concludes the outcomes of this research work and provide recommendations for future work.

II. METHODOLOGY OF RESEARCH

The proposed methodology for HAR is shown in Fig. 1, which consists of three main steps: feature extraction and description, feature fusion, and action classification. These steps are explained in detail in the following sub-sections.

A. FEATURE EXTRACTION AND DESCRIPTION

As this research work focuses on the feature-level fusion of multiple sensor modalities for robust HAR, hence we extracted different sets of features for inertial sensor data and RGB/depth videos. It is done because these features provide the best recognition rate when used for HAR with individual modality data. The following sections provide the detail of the feature extraction process for inertial sensor data and RGB/depth video sequences.

1) FEATURE EXTRACTION FOR INERTIAL SENSOR

The raw data obtained from wearable inertial sensors is orientation sensitive and often degraded by unwanted noise produced by either the instrument or unanticipated movement of the participant. Hence, it is crucial to preprocess the raw data obtained from wearable inertial sensors before any further processing. For this purpose, the magnitude s_{mag} of both acceleration and rotation signal is calculated, which is concatenated with existing three-dimensional data to make the form (s_x, s_y, s_z, s_{mag}), where s_x, s_y, and s_z represent the signal values along x, y, and z-axes respectively. The value of s_{mag} is calculated as : $s_{mag} = \sqrt{s_x^2 + s_y^2 + s_z^2}$.

For de-noising of the acquired signals, an average smoothing filter of size 1×3 is applied to the acquired data based on two nearest neighbors approach. After that, three time domain features are extracted from both acceleration and gyroscope signals obtained corresponding to each action trial. These features are presented in Eq. (1) to Eq. (3).

$$\mu = \frac{1}{N}\sum s\,(n) \tag{1}$$

$$\mu_\nabla = \frac{1}{N}\sum |s\,(n) - s(n-1)| \tag{2}$$

$$\mu_\Delta = \frac{1}{N}\sum |s\,(n+1) - 2s\,(n) + s(n-1)| \tag{3}$$

where, μ represents the mean of the signal $s(n)$, μ_∇ is the mean of absolute values of the first difference of the signal $s(n)$, μ_Δ is the mean of absolute values of the second difference of the signal $s(n)$, and N represents the count of total samples in the signal $s(n)$ at a sampling rate of 50 Hz. These features are extracted for all four channels, i.e., (s_x, s_y, s_z, s_{mag}), of the accelerometer and gyroscope and then concatenated for each sensor to form the resultant feature vector. Hence, for each data sequence, we obtained a feature vector of size [1 × (3 (# of features) × 4 (# of dimensions per sensor)] = [1×12] per sensor. As there are 861 data sequences in total, hence we get 861 different feature vectors per sensor with each feature vector having a length equal

to 12. These feature vectors are later used in the classification stage for HAR.

2) FEATURE EXTRACTION FOR RGB/DEPTH SENSOR

For RGB and depth video data, we employed the general Bag-of-Words (BoWs) pipeline for HAR, which is visualized in Fig. 2. The BoWs method [63] has been successfully adapted from static images to the motion clips and videos through local space-time descriptors. It has many successful applications in HAR [15], [64], [65]. For human action clips, BoWs may be specified as a bag of action patches that occur in the action frames for many times. We used the BoWs approach to transform locally extracted feature descriptors from an action clip into a fixed-sized vector needed for classification.

The proposed BoWs-based approach for HAR consists of the following steps:

1) *Local Feature Description:* For extracting features from RGB and depth videos, we utilized the dense sampling of local visual descriptors, since densely sampled descriptors are more accurate than keypoint-based sampling [66], [67]. As a type of local visual descriptors, we paid attention to densely extracted 3D volumes of HOG [68]. For calculating dense HOG, firstly the gradient magnitude response is computed in both horizontal and vertical directions, which resulted in a 2D vector field per frame. Haar features are used to calculate gradient magnitude response as these features are faster and obtain better results for HOG [62]. Next, we divided the input video into dense blocks of size 15×15 pixels $\times$ 20 frames. For every single block, the magnitude is quantized in O orientation bins (where $O = 8$), which is done by dividing each response magnitude linearly over two neighboring orientation bins. After that, we concatenated the responses of multiple adjacent blocks in both spatial and temporal directions. For this purpose, we concatenated the descriptors of 33 blocks in the spatial domain and two blocks in the temporal domain, resulting in a 144-dimensional HOG descriptor. The size of each HOG descriptor is then reduced to half using Principal Component Analysis (PCA), which lead to a 72-dimensional descriptor. Finally, L1-normalization is performed followed by the square root to obtain final descriptor representation.
2) *Visual Codebook Construction:* The number of significant interest points and densely extracted HOG features may change for different videos, which results in feature vectors having different size. However, to train a classifier, a fixed size feature vector is required for all data sequences. For this purpose, we clustered the features extracted from all training videos into '' clusters using k-means clustering. The center of each cluster is considered as a visual word. A group of these visual words together make a visual vocabulary or codebook.

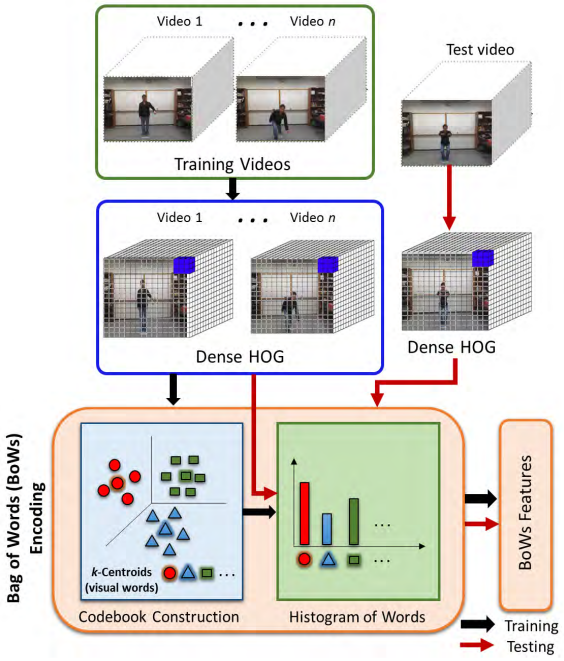

FIGURE 2. General pipeline for BoWs representation of dense HOG features extracted from RGB and depth video sequences.

3) *Histogram of Words Generation:* After constructing the visual vocabulary/codebook from the training videos, the next step is to quantize the HOG descriptors from each training/testing video into a fixed-sized vector known as a histogram of words. Histogram of words shows the frequency of each visual word that is present in a video sequence. So, for a given video, each of HOG descriptor is compared with all visual words and voting is performed for the best matching visual word, which resulted in a histogram of the visual words for that video. In this manner, all training and testing videos are quantized into k-dimensional vectors referred to as Bag-of-Words. After computing BoWs for training and testing video data, classifiers are applied for learning and recognition of human actions.

B. FEATURE FUSION

After extracting features from inertial sensors and RGB/depth videos, we performed their fusion for HAR. For this purpose, we independently computed feature vector for the data obtained from each sensing modality (i.e., RGB/depth sensor and inertial sensor) and concatenated the individual feature vectors obtained from the multimodal data related to the same action at the same time, which resulted in a new high dimensional feature vector. This resultant feature vector possessed more feature information to better recognize human actions compared to the feature vector obtained for single sensing modality.

For the feature-level fusion, it is necessary to balance different feature sets obtained corresponding to the data from different sensing modalities. Balancing different feature sets

means that the concatenated features must have the same numerical scale and similar length. Hence, we applied the *min-max* normalization technique [69] on the feature sets obtained for RGB/depth and inertial sensors before concatenating them to produce a single resultant vector. The purpose of employing feature normalization is to modify the numerical ranges and scaling parameters of the individual feature sets to transform these values into a new feature domain, having a similar numerical scale. The *min-max* normalization scheme preserves the original score distribution and maps the values into a standard range [0, 1] according to the formula given in Eq. (4).

$$x' = \frac{x - min\,(F_x)}{max(F_x) - min\,(F_x)} \tag{4}$$

where, x is the value to be normalized and x' is the normalized value, F_x represents the function that produces x, $min(F_x)$ and $max(F_x)$ donates the minimum and maximum values of F_x respectively for all possible values of x.

The size of the feature vector obtained in the case of inertial sensor data is fixed for each data sequence, i.e., $[1\times12]$. On the other hand, the feature vector extracted for RGB/depth video sequences is of size $[1\times k]$, where k is the number of visual words in BoWs representation of densely extracted HOG features. The variable k is introduced to balance the length of the fused feature vectors for RGB/depth and inertial sensor data, and to find out the effect of varying feature lengths on the feature-level fusion. The feature sets obtained are firstly normalized and then concatenated together for fusion. So, after feature-level fusion of a single inertial sensor and RGB/depth sensor, we obtained a final feature vector of size $[1 \times (12+k)]$. When performing the feature-level fusion of both accelerometer and gyroscope with RGB/depth sensor, we got a final feature vector of size $[1 \times (12 \times 2 + k)]$.

C. ACTION RECOGNITION

After feature extraction and fusion from multiple sensor modalities, the next process is choosing a suitable classifier for training the proposed framework for HAR and to test it. Two popular classifiers, i.e., K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM), are used for this purpose because of their efficient recognition performance in existing state-of-the-art studies [8], [70]–[72]. Moreover, we anticipated comparing their recognition performance when the fusion of different sensing modalities is used for HAR.

III. EXPERIMENTAL RESULTS

In this section, we first briefly describe the dataset used for experimentation along with experimental design and evaluation metrics. We then provide information regarding the implementation of our proposed framework. After that, we compare our algorithm with existing state-of-the-art HAR methods. Finally, we discuss the qualitative results to provide essential intuitions of the proposed method.

A. DATASET AND IMPLEMENTATION DETAILS

We evaluated the proposed method on a publicly accessible multimodal HAR dataset *UTD-MHAD*, which entails 27 human actions carried out by eight subjects (four females and four males). Fig. 3 provides a list of these actions with example images. Each subject repeated every action four times. Hence, there were overall 864 trimmed data sequences (8 (no. of subjects) $\times$ 4 (no. of trials per action per subject) $\times$ 27 (no. of action)). During data recording, three data sequences were corrupted; hence after removing the corrupted sequences, 861 data sequences were left in the dataset. Four sensing modalities including RGB, depth, skeleton joint positions, and the inertial sensors (3-axis acceleration and 3-axis rotation signals) were used for data recording purpose. The dataset was collected using a Microsoft Kinect sensor (at a rate of 30 frames per second) and a wearable inertial sensor (at a sampling rate of 50 Hz) in an indoor setting. A Bluetooth enabled hardware module was used as a wearable inertial sensor to record triaxial acceleration (using an accelerometer) and triaxial angular velocity (using a gyroscope). This sensing module was worn on the subject's *right wrist* for actions 1 to 21, whereas for actions 22 to 27, the sensor was placed on the subject's *right thigh*. For synchronizing data from different sensing modality, timestamp value was recorded for each data sample. The dataset is comprised of four data files for each segmented action trial, which correspond to four sensing modalities. A more detailed explanation regarding the dataset can be found in [53].

For implementing the proposed HAR method, K-NN and SVM classifiers are trained and tested on *UTD-MHAD*. For K-NN classifier, the parameter 'K' is set to 1, and an equal weight Euclidean distance metric is used for similarity measure. The Nearest neighbor parameter 'K' is different from 'k' as 'k' is the number of visual words in BoWs representation of RGB/depth video features. On the other hand, a *quadratic kernel* is applied for SVM classifier with a *one-vs-one* approach for multi-class classification. For ensuring any impartiality in results, an 8-fold stratified cross-validation method is used to assess the performance of these classifiers in action recognition. As a result, all action instances in the dataset are split randomly into eight sets and one set is used for testing while the remaining sets are used for training. This process is repeated eight times such that each set of instances participated in training and testing of the classifiers in different iterations. For all eight iterations, the classifiers are evaluated, and the average results of these iterations are computed, which are presented in this section. The performance metrics used for evaluating the classifier performance for the proposed HAR scheme are *accuracy*, *precision*, *recall*, and *f-measure*.

B. ACTION RECOGNITION RESULTS AND ANALYSIS

For feature-level fusion, we concatenated the individual feature sets extracted from inertial sensor data and the corresponding RGB and/or depth video sequence after

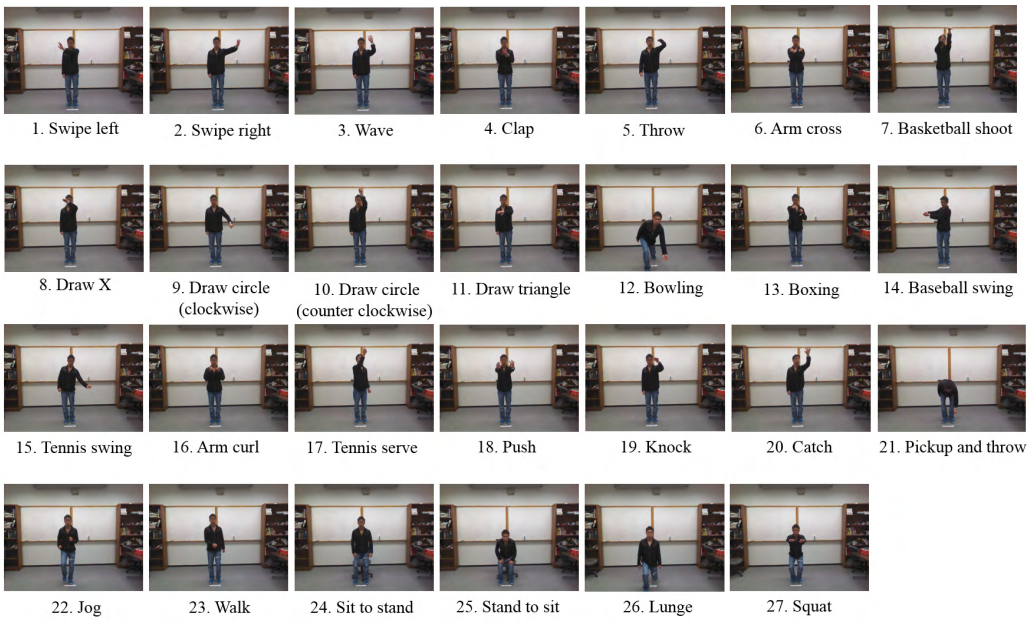

FIGURE 3. Set of 27 human actions in UTD-MHAD with sample image.

TABLE 2. HAR results obtained using inertial sensors (accelerometer (Acc.), gyroscope (Gyro.), and their feature-level fusion).

Classifier	Sensor(s)	Precision	Recall	Accuracy	F-measure
K-NN	Acc.	0.790	0.785	0.785	0.788
	Gyro.	0.770	0.766	0.766	0.768
	Acc. + Gyro.	**0.918**	**0.916**	**0.916**	**0.917**
SVM	Acc.	0.782	0.767	0.766	0.774
	Gyro.	0.742	0.728	0.728	0.735
	Acc. + Gyro.	0.911	0.906	0.905	0.908

*The bold values in the table represent the best HAR performance obtained using inertial sensors.

min-max normalization. Although feature-level fusion seems to be simple and straightforward, it suffers from some several deficiencies. First, the increase in the dimensionality of the fused feature vector raises the computational complexity of classification. Second, the dimensionality of the RGB/depth features is typically much higher than the features extracted for inertial sensor data, which ultimately degrades the fusion purpose. We address these issues using a variable length feature vector for RGB and depth data sequences. The size of the feature vector obtained from inertial sensor data is equal to 1×12. On the other hand, the length of each feature vector extracted for RGB/depth video sequence is equal to the number of clusters k in BoWs representation of dense HOG features. We evaluated HAR results for varying values of k (starting from 10 to 30) to analyze the effect of varying feature vector length on recognition performance. Choosing k higher than 30 increases the difference between the lengths of the fused feature vectors obtained from RGB/depth and inertial sensor data. Hence, the feature sets become imbalanced and as a result, the feature-level fusion becomes ineffective. Also, higher values of k mean a higher number of clusters in BoWs feature representation and smaller distance between the cluster centroids or visual words. So, the chance of visual words misclassification enhances, which eventually decreases the recognition performance. The detailed results of HAR obtained using different sensor modalities individually as well as their combination are presented and discussed in the following sections.

1) PERFORMANCE ANALYSIS OF INERTIAL SENSOR-BASED HAR

This section discusses the results of HAR obtained using only the inertial sensors for recognition. Table 2 summarizes these results for the different combination of sensors. The results of HAR are provided individually for each inertial sensor as well as their feature-level fusion. It can be observed that K-NN classifier provides better performance than SVM classifier in recognizing human actions based on a single inertial sensor or their combination. The accuracy rate achieved for K-NN

TABLE 3. HAR results obtained using depth sensor, RGB sensor, and their feature-level fusion.

Classifier	Sensor(s)	k	Precision	Recall	Accuracy	F-measure
K-NN	Depth	10	0.695	0.683	0.684	0.689
		15	0.718	0.713	0.713	0.715
		20	0.774	0.768	0.768	0.771
		25	0.818	0.815	0.815	0.816
		30	0.815	0.818	0.815	0.816
	RGB	10	0.735	0.728	0.728	0.731
		15	0.806	0.803	0.803	0.805
		20	0.826	0.822	0.822	0.824
		25	0.855	0.852	0.852	0.854
		30	0.853	0.852	0.851	0.852
	Depth + RGB	10	0.826	0.823	0.823	0.825
		15	0.842	0.837	0.837	0.840
		20	0.87	0.873	0.873	0.874
		25	**0.898**	**0.893**	**0.893**	**0.896**
		30	0.896	0.891	0.893	0.896
SVM	Depth	10	0.491	0.473	0.473	0.482
		15	0.568	0.552	0.552	0.560
		20	0.667	0.649	0.649	0.658
		25	0.702	0.691	0.691	0.697
		30	0.736	0.720	0.720	0.728
	RGB	10	0.532	0.517	0.517	0.524
		15	0.691	0.676	0.676	0.683
		20	0.742	0.720	0.720	0.731
		25	0.734	0.720	0.720	0.727
		30	0.789	0.776	0.776	0.782
	Depth + RGB	10	0.711	0.698	0.698	0.705
		15	0.768	0.753	0.753	0.760
		20	0.816	0.809	0.810	0.813
		25	0.862	0.854	0.854	0.859
		30	0.861	0.854	0.854	0.858

*The parameter k is the length of the feature vector, which is equivalent to the number of visual words in BoWs representation of dense HOG features computed for an RGB/depth video sequence. The bold values in the table represent the best performance achieved for HAR using the feature-level fusion of RGB and depth sensors.

classifier in recognizing human actions using accelerometer and gyroscope individually is 78.5% and 76.6% respectively. These accuracy rates are 1.9% and 3.8% better than the accuracy values achieved for SVM classifier when using these sensors individually. The overall performance of an accelerometer in recognizing human actions is better than the gyroscope. Moreover, it can be observed that the fusion of these inertial sensors improves the overall recognition accuracy to 91.6% and 90.5% when classified using K-NN and SVM classifiers individually. Overall, K-NN classifier provides better results as compared to SVM classifier in classifying human actions based on the feature-level fusion of inertial sensors.

2) PERFORMANCE ANALYSIS OF RGB AND DEPTH SENSOR-BASED HAR

This section provides the detailed results obtained for HAR using depth and RGB sensors individually as well as their combination. These results are computed for different values of k, where k is the number of visual words in BoWs representation of dense HOG features extracted for each depth and RGB video sequence. This parameter k represents the length of the final feature vector obtained for depth and RGB video sequence. Varying the value of k affects the recognition results as depicted in Table 3. The lower value of k indicates less number of visual words in BoWs representation of dense HOG features, which provides lower action recognition performance. As we keep on increasing the value of k, the results become saturated. Hence, using a very high value of k might result in only a little performance improvement, but at the expense of increased computational cost. Hence, a moderate value of k leads to better recognition rate and lesser computational cost as well.

It can be observed from Table 3 that K-NN classifier achieves maximum accuracy rate for HAR using depth and RGB sensor individually, which is 81.5% and 85.2% respectively for $k = 25$. Also, the difference between the accuracy rate achieved for $k = 5$ and $k = 10$ is very high, which reduces as the value of k is increased. In the case of SVM classifier, the maximum accuracy rate achieved using depth and RGB sensor individually is 72% and 77.6% when k reaches 30. These results indicate that the individual performance of the RGB sensor in recognizing human actions, based on dense HOG features, is better than the performance of the depth sensor. It is because of the reason that RGB video provides rich texture information as compared to depth video, which is very useful for extracting dense HOG features. Moreover, the feature-level fusion of RGB and depth sensor improves HAR performance to 89.3% and 85.4% using K-NN and SVM

classifier respectively. However, it also increases the dimensionality of the fused feature vector, which raises the computational complexity of the classification process. Moreover, it might also degrade the overall recognition performance if the value of k is set too high.

3) PERFORMANCE ANALYSIS OF HAR BASED ON FEATURE-LEVEL FUSION OF RGB, DEPTH AND INERTIAL SENSORS

This section analyzes the performance of HAR when the feature-level fusion of RGB/depth and inertial sensors is performed. The statistical features computed from inertial sensor data are different from dense HOG-based features extracted for RGB/depth video data and have different dimensions. Feature-level fusion is practically possible when the dimensions of the fused feature vectors are not much different. In the case of inertial sensor data, the feature vector size is 1×12. Hence, the length of RGB/depth feature vector is kept from $k = 10$ to $k = 30$ for efficient recognition performance.

Table 4 presents the detailed results of HAR based on the feature-level fusion of RGB/depth and inertial sensors. It can be observed that K-NN classifier provides better results as compared to SVM classifier. When using only the accelerometer with a depth sensor, the maximum accuracy rate achieved for HAR using K-NN classifier is 94.8% (for $k = 25$). Whereas, SVM classifier provides a maximum accuracy rate of 90.6% (for $k = 25$) for the same combination of sensors. Adding gyroscope with a depth sensor for feature-level fusion achieves a maximum accuracy rate of 93.7% and 89.7% using K-NN and SVM classifier respectively when $k = 25$. It shows that adding accelerometer with a depth sensor provides better results for HAR as compared to the gyroscope. Adding both accelerometer and gyroscope with a depth sensor improves the recognition accuracy to 97% (for $k = 30$) using K-NN classifier. In the case of SVM classifier, the accuracy rate also improves to 95.1% when $k = 25$. These results indicate that KNN classifier performs better than SVM classifier in recognizing human actions.

The recognition results for the feature-level fusion of RGB and inertial sensors are also presented in Table 4. When adding accelerometer and gyroscope individually with RGB sensor, the maximum accuracy rate achieved for HAR using K-NN classifier is 96.1% (for $k = 25$) and 95.4% (for $k = 25$) respectively. In the case of SVM classifier, the addition of accelerometer with RGB sensor provides a maximum accuracy rate of 91.3% (for $k = 25$). Whereas, fusing gyroscope with RGB sensor gives a maximum accuracy of 90.1% (for $k = 30$). The best accuracy rate achieved for the proposed HAR framework is 97.6% (for $k = 25$) using K-NN classifier, which is achieved by the fusion of RGB and inertial sensors (both accelerometer and gyroscope). For the same combination of sensors, SVM classifier provides maximum accuracy of 95.5% when $k = 25$, which is lower as compared to the accuracy rate obtained for K-NN classifier. Adding depth sensor with RGB and inertial sensors provides an accuracy improvement of 0.7% (accuracy =98.3% for $k = 25$) and 0.6% (accuracy =96.1% for $k = 20$) when evaluated using K-NN and SVM classifier respectively as shown in Table 5. Hence, K-NN classifier provides the best accuracy rate of 98.3% for the proposed HAR system using the feature-level fusion of all four sensors (RGB, depth, accelerometer, and gyroscope). In general, for any combination of sensing modalities, the recognition rate achieved for the proposed HAR method using K-NN classifier is higher than the accuracy rate obtained for SVM classifier. Furthermore, K-NN classifier also provides lower computational complexity compared to SVM classifier. Therefore, K-NN classifier is concluded as the optimal choice for the proposed action recognition framework.

4) ANALYSIS OF FEATURE-LEVEL FUSION RESULTS FOR HAR USING K-NN CLASSIFIER

This section compares the best performance achieved for the proposed HAR method using K-NN classifier when different sensing modalities are used. Table 6 provides a comparison of the average accuracy attained using different sensors along with the final feature vector length and average processing time. It can be observed that the feature-level fusion of different sensors increases the length of the final feature vector, which in return increases the average computational time. The processing time for the proposed HAR method is computed using MATLAB on a laptop with a 2.3 GHz Intel Core-i5 CPU with 8 GB RAM. For each sensor or set of sensors, the average time taken for feature extraction and classification can be added to compute the overall average computational time. For RGB/depth sensor, the average time is calculated per frame, whereas, for inertial sensors, it is computed per sample.

From Table 6, it can be seen that the accuracy rate achieved for HAR with the accelerometer sensor only is 78.5%, whereas, for the gyroscope sensor, it is 76.6%. The fusion of accelerometer and gyroscope provides an accuracy of 91.6% at the expense of around 46% (53 microseconds (μs)) increase in average processing time per sample. The maximum accuracy rate achieved for HAR using depth and RGB sensor alone is 81.5% and 85.2% respectively with the feature vector length of 25. The fusion of depth and RGB features improved the recognition accuracy to 89.3%, which is 7.8% and 4.1% better than the individual accuracy rate achieved using depth and RGB sensor respectively. However, the fusion increased the average time for feature extraction to 7.34 milliseconds (ms) per frame, which is about 2.6 times (160%) and 1.6 times (60%) more than the average time taken for extracting depth and RGB features separately. The average classification time of the fused feature vector, in this case, is increased by 9.6% per frame. The fusion of inertial sensors (both accelerometer and gyroscope) with depth and RGB sensor separately achieved the maximum accuracy of 97% and 97.6% respectively. This accuracy rate is 12.4% and 15.5% more than the accuracy rate achieved for depth and RGB sensor individually, with an increase of 8.6% in average processing time.

TABLE 4. HAR results obtained using feature-level fusion of depth and inertial sensors (accelerometer (Acc.) and gyroscope (Gyro.)).

Classifier	Sensor(s)	k	Precision	Recall	Accuracy	F-measure
K-NN	Depth + Acc.	10	0.929	0.928	0.927	0.928
		15	0.938	0.937	0.937	0.938
		20	0.943	0.942	0.941	0.942
		25	0.948	0.946	0.948	0.948
		30	0.947	0.946	0.946	0.947
	Depth + Gyro.	10	0.929	0.926	0.926	0.928
		15	0.929	0.928	0.927	0.928
		20	0.929	0.928	0.927	0.928
		25	0.937	0.937	0.937	0.937
		30	0.933	0.931	0.931	0.932
	Depth + Acc. + Gyro.	10	0.966	0.964	0.963	0.965
		15	0.964	0.964	0.963	0.964
		20	0.969	0.960	0.968	0.969
		25	0.970	0.962	0.970	0.969
		30	0.971	0.969	0.969	0.970
	RGB + Acc.	10	0.933	0.931	0.931	0.932
		15	0.942	0.940	0.940	0.941
		20	0.949	0.948	0.948	0.949
		25	0.962	0.962	0.961	0.961
		30	0.958	0.957	0.957	0.957
	RGB + Gyro.	10	0.930	0.928	0.927	0.929
		15	0.949	0.947	0.947	0.948
		20	0.949	0.948	0.948	0.949
		25	0.956	0.954	0.954	0.955
		30	0.949	0.947	0.947	0.948
	RGB + Acc. +Gyro.	10	0.964	0.964	0.963	0.963
		15	0.964	0.964	0.963	0.964
		20	0.968	0.967	0.969	0.969
		25	**0.977**	**0.966**	**0.976**	**0.977**
		30	0.967	0.967	0.967	0.966
SVM	Depth + Acc.	10	0.890	0.878	0.878	0.884
		15	0.904	0.890	0.890	0.897
		20	0.898	0.890	0.890	0.894
		25	0.911	0.902	0.904	0.909
		30	0.915	0.906	0.906	0.911
	Depth + Gyro.	10	0.872	0.861	0.861	0.867
		15	0.887	0.877	0.877	0.882
		20	0.897	0.889	0.889	0.893
		25	0.900	0.891	0.891	0.896
		30	0.905	0.897	0.897	0.901
	Depth + Acc. +Gyro.	10	0.943	0.937	0.937	0.940
		15	0.942	0.936	0.936	0.939
		20	0.951	0.946	0.945	0.948
		25	0.955	0.951	0.951	0.953
		30	0.948	0.946	0.946	0.947
	RGB + Acc.	10	0.890	0.878	0.878	0.884
		15	0.907	0.900	0.900	0.904
		20	0.913	0.904	0.904	0.908
		25	0.918	0.913	0.913	0.914
		30	0.911	0.900	0.900	0.906
	RGB + Gyro.	10	0.900	0.890	0.890	0.895
		15	0.897	0.885	0.885	0.891
		20	0.908	0.892	0.892	0.900
		25	0.903	0.893	0.893	0.898
		30	0.911	0.902	0.901	0.906
	RGB + Acc. +Gyro.	10	0.952	0.950	0.950	0.951
		15	0.948	0.946	0.945	0.947
		20	0.950	0.948	0.948	0.949
		25	0.954	0.954	0.955	0.954
		30	0.950	0.946	0.946	0.947

* The parameter k is the length of the feature vector extracted for the RGB/depth video sequence. The bold values in the table represent the best HAR performance achieved using the feature-level fusion of RGB and inertial sensors.

TABLE 5. HAR results obtained using feature-level fusion of RGB, depth, and inertial sensors (accelerometer (Acc.) and gyroscope (Gyro.)).

Classifier	Sensor(s)	*k*	Precision	Recall	Accuracy	F-measure
KNN	Depth + RGB + Acc. +Gyro.	10	0.971	0.971	0.970	0.971
		15	0.980	0.970	0.979	0.979
		20	0.981	0.981	0.981	0.982
		25	**0.981**	**0.983**	**0.983**	**0.982**
		30	0.980	0.980	0.980	0.980
SVM	Depth + RGB + Acc. +Gyro.	10	0.954	0.951	0.951	0.953
		15	0.964	0.962	0.961	0.963
		20	0.964	0.962	0.961	0.963
		25	0.962	0.963	0.960	0.962
		30	0.954	0.949	0.949	0.951

*The parameter *k* is the length of the feature vector extracted for an RGB/depth video sequence. The bold values in the table represent the best HAR performance achieved using the feature-level fusion of RGB, depth, and inertial sensors.

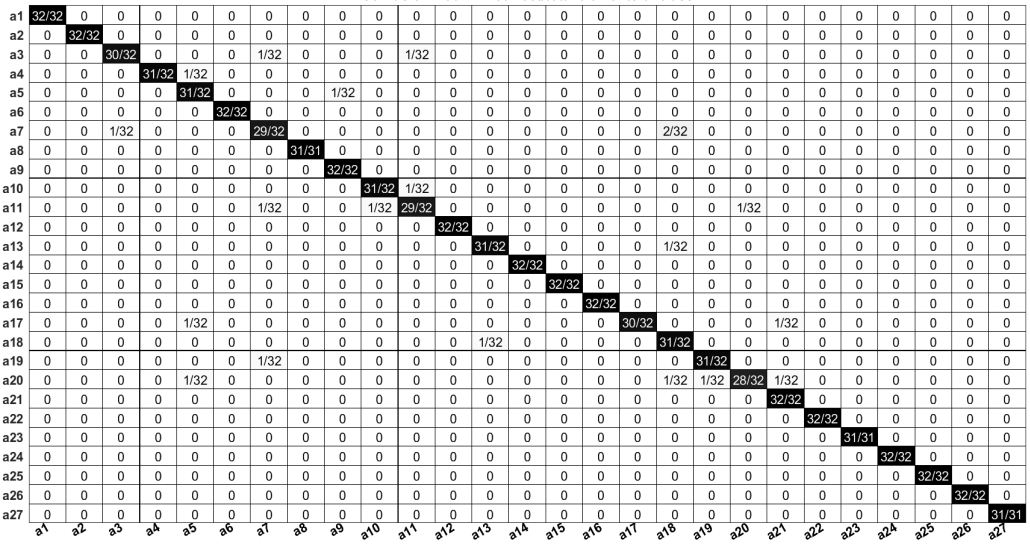

FIGURE 4. Confusion matrix for HAR results obtained for the feature-level fusion of RGB and inertial sensors (for $k = 25$).
∗ Each entry in the confusion matrix represents predicted/total elements for the given class (rows represent ground truth and columns represent predicted class).

The best accuracy rate obtained for the proposed HAR approach is 98.3%, which is achieved as a result of the feature-level fusion of four different sensors including RGB, depth, accelerometer, and gyroscope sensor. However, in this case, the average time for classification is increased by 13.5% when compared with the average classification time taken for the fusion of inertial sensors with RGB or depth sensor. The average time required for feature extraction is also increased with a maximum factor of approximately 2.6, which can be observed from Table 6. On the other hand, adding inertial sensors with RGB or depth sensor only results in a slight increase in the average processing time and provides an accuracy rate comparable to the maximum accuracy rate. The accuracy rate achieved by the fusion of RGB and inertial sensors is 97.6%, which is 8.3% more than that obtained by feature-level fusion of RGB and depth sensor using K-NN classifier. Hence, it is evident that the overall performance of the proposed HAR method (considering the accuracy rate and the computational time as a trade-off) is better for the feature level fusion of inertial sensors with only RGB or depth sensor. In particular, as the RGB sensor provides rich texture information and inertial sensor tracks 3D motion information, it is concluded that their feature-level fusion provides the overall best performance for the proposed HAR framework.

Fig. 4 provides the confusion matrix of the best overall results achieved for the feature-level fusion (using RGB and inertial sensors) to demonstrate per class recognition accuracy of all 27 actions in *UTD-MHAD*. It can be observed from the figure that most of the actions are recognized with a very high individual accuracy. The lowest individual recognition accuracy achieved is 87.5% for action 20, i.e., *right-hand catch an object*.

5) COMPARISON OF FEATURE-LEVEL FUSION AND DECISION-LEVEL FUSION RESULTS FOR PROPOSED HAR METHOD

Our proposed method for HAR relies on the feature-level fusion of multiple sensors for robust action recognition.

TABLE 6. Comparison of HAR results obtained for the proposed scheme using K-NN classifier with single and multiple sensing modalities.

Sensor(s)	Accuracy Rate	Feature Length	Average Computational Time	
			Average Time for Feature Extraction	Average Time for Classification
A	78.5%	12	0.31 µs	0.82 µs
G	76.6%	12	0.31µs	0.82 µs
A + G	91.6%	12+12=24	0.62 µs	1.04 µs
D	81.5% (for k=25)	25	2.80 ms	1.97 µs
RGB	85.2% (for k=25)	25	4.54 ms	1.97 µs
D + RGB	89.3% (for k=25)	25+25=50	7.34 ms	2.16 µs
D + A	94.8% (for k=25)	25+12=37	2.80 ms	2.04 µs
D + G	93.7% (for k=25)	25+12=37	2.80 ms	2.04 µs
D + A + G	97.0% (for k=25)	25+12+12=49	2.80 ms	2.14 µs
RGB + A	95.7% (for k=25)	25+12=37	4.54 ms	2.04 µs
RGB + G	95.4% (for k=25)	25+12=37	4.54 ms	2.04 µs
RGB + A + G	97.6% (for k=25)	25+12+12=49	4.54 ms	2.14 µs
D + RGB + A + G	98.3% (for k=25)	25+25+12+12=74	7.34 ms	2.43 µs

* The parameter k is the length of the feature vector extracted for the RGB/depth video sequence. Here, 'A' represents the accelerometer sensor, 'G' represents the gyroscope, and 'D' is the depth sensor. For the RGB-D sensor, the average time is measured per frame, whereas, for inertial sensor, it is computed per sample. The feature-level fusion of RGB and inertial sensor acquired the overall best performance considering the accuracy rate and computational time as the performance measures.

However, most of the existing studies for multimodal action recognition [53], [54] focused on decision-level fusion to achieve effective recognition results as the features being extracted from different sensors are independent. Instead, the feature-level fusion requires the numerical scale and dimensions of the fused feature vectors to be similar, which is not possible with the type of features extracted for RGB and depth video sequences in the existing studies. Also, the dimensions of the RGB and depth features are often quite higher as compared to the inertial sensor features, which is infeasible for the feature-level fusion. Consequently, the results obtained for the feature-level fusion are not much consistent and accurate as compared to the decision-level fusion results in the literature. In our proposed study, we first balanced the dimensions and numerical scale of the RGB-D features (densely extracted HOG) and the statistical signal attributes computed from the inertial sensors. After that, we performed multimodal feature-level fusion to achieve the desired HAR results.

To validate the effectiveness of our feature-level fusion approach, we also computed the decision-level fusion results for the proposed scheme and compared both results. For the decision-level fusion, we followed the same approach as proposed by the authors in [53], [54]. For the fusion of n different sensors, we trained n K-NN classifiers separately by passing the corresponding set of features as an input to each classifier. During testing, we merged the decision of each classifier using a logarithmic opinion pool (LOGP) [73] at the posterior-probability level. For calculating the posterior probability of each classifier, we used Euclidean distance to compute the error vector. The final class label for each testing instance is assigned to the action class with the smallest error. Fig. 5 shows the comparison of the accuracy rate achieved for the proposed HAR with the feature-level and decision-level fusion. For any set of sensing modalities, the percentage accuracy achieved for the multimodal feature-level fusion is higher than that obtained for the decision-level fusion. The proposed HAR framework with the feature-level fusion of RGB and inertial sensors provides a 19.3% increase in the accuracy rate as compared to the decision-level fusion of the same set of sensors. It substantiates the efficacy of the proposed feature-level fusion over the decision-level fusion.

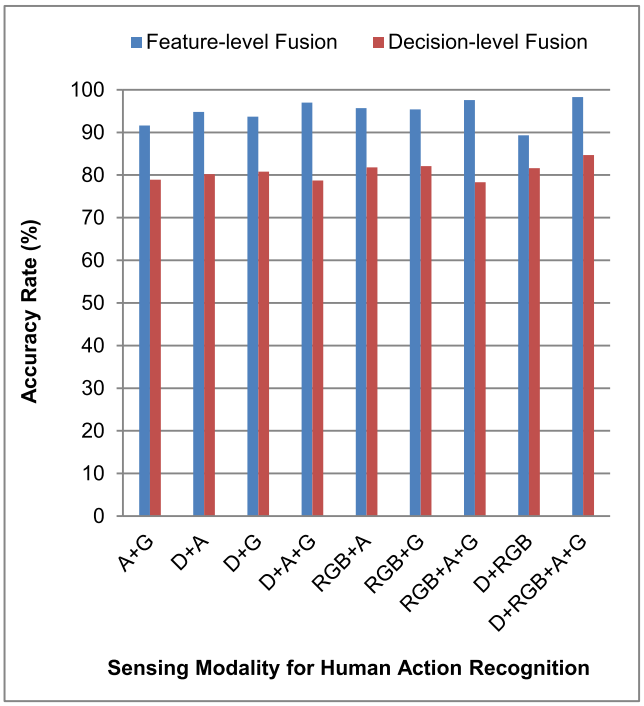

FIGURE 5. Comparison of the maximum accuracy rate achieved for the proposed HAR framework with the feature-level and decision-level fusion of different sensors using K-NN classifier. For any combination of sensors, the feature-level fusion outperforms the decision-level fusion.
*** Here, 'A' represents the accelerometer sensor, 'G' represents the gyroscope, 'D' is the depth sensor, and 'RGB' represents the RGB sensor.**

TABLE 7. Comparison of proposed HAR method results with existing studies.

Study	Sensor(s)	Accuracy
Chen et al. (2015) [53]	Depth, Accelerometer, Gyroscope	79.1%
Chen et al. (2016) [54]	Depth, Accelerometer, Gyroscope	91.5% (subject-generic) 97.2% (subject-specific)
Ben et al. (2017) [8]	RGB	70.3%
Wang et al. (2018) [40]	Skeleton	87.9%
Kamel et al. (2018) [41]	Depth, Skeleton	88.1%
Hou et. al (2018) [39]	Skeleton	86.9%
Neha et al. (2019) [60]	Depth, Accelerometer, Gyroscope	89.2%
Cui et al. (2019) [61]	Skeleton	87.0%
Proposed Method	**RGB, Accelerometer, Gyroscope**	**97.6%** (subject-generic) **98.2%** (subject-specific)

*The proposed scheme provides better results on *UTD-MHAD* dataset in recognizing human actions, as highlighted in bold.

6) PERFORMANCE COMPARISON OF PROPOSED HAR SCHEME WITH STATE-OF-THE-ARTS

This section provides a performance comparison of the proposed scheme for HAR with the existing techniques. The proposed HAR scheme, based on the feature-level fusion of RGB and inertial sensors, provides superior recognition performance on *UTD-MHAD* compared to existing methods as shown in Table 7. Chen *et al.* [53] presented *UTD-MHAD* in their study and utilized the decision-level fusion of depth and inertial sensors (accelerometer and gyroscope) for HAR. They computed three statistical features for inertial sensor data and extracted DMMs for depth video sequences. The authors partitioned the dataset into two equal splits for training and testing. The data corresponding to four different users was utilized for training whereas the data from the rest of the users was used for testing, which resulted in an average accuracy rate of 79.1%. The authors modified their existing methodology in [54] to incorporate real-time HAR, which achieved recognition accuracy of 91.5% using an 8-fold cross-validation scheme for subject-generic experiments. The authors also conducted experiments using subject-specific training and testing, which achieved an average accuracy rate of 97.2%. Ben Mahjoub and Atri [8] proposed an RGB sensor-based scheme that utilized the STIP for detecting significant changes in an action clip. Moreover, they used the HOG and Histogram of Optical Flow (HOF) as feature descriptors and achieved an accuracy rate of 70.37% using SVM classifier. Wang *et al.* [40] used CNN for HAR and utilized the skeleton information from the Kinect sensor to achieve an overall recognition accuracy of 88.1% on *UTD-MHAD*. Kamel *et al.* [41] applied deep CNN for HAR using depth maps and skeleton information and achieved an accuracy rate of 87.9% on *UTD-MHAD* dataset. The research work in [39] proposed the skeleton optical spectra (SOS) method based on CNNs to recognize human actions. The authors encoded the skeleton sequence information into color texture images for HAR and achieved an accuracy rate of 86.9% on *UTD-MHAD*. The authors in [60] utilized the decision-level fusion for HAR using depth camera and wearable inertial sensors. They extracted CNN based features for depth sensor and used CNN and LSTM networks for inertial sensors. Their study achieved an accuracy of 89.2% on *UTD-MHAD*. Recently, Cui *et al.* [61] used the skeletal data to extract the temporal and spatial features for action recognition using LSTM and spatial CNN models respectively. They achieved a maximum accuracy rate of 87.0% on *UTD-MHAD*.

Our proposed scheme combines the color and rich texture information from the RGB sensor with 3D motion information obtained from inertial sensors for robust HAR. The proposed scheme for HAR, based on the feature-level fusion of RGB and inertial sensors, obtained the maximum recognition accuracy of 97.6% using 8-fold cross-validation, which is better than the reported results of existing techniques. Furthermore, the proposed scheme is computationally efficient as the overall length of the fused feature vector is very small, i.e., 49 ($25 + 2 \times 12$), for the case when performance is achieved for K-NN classifier using the fusion of RGB and inertial sensors. On the other hand, in existing techniques, generally, the dimensions of the feature vector obtained for RGB/depth video sequence are very high, which makes the HAR system computationally expensive. Moreover, the application of CNN for HAR also increases the computational cost of the system. Moreover, in the case of RGB and depth sensor fusion, the computational complexity and the dimensions of the fused feature vector increases significantly. However, in our proposed method, we quantized the dense HOG features computed on RGB or depth video sequences to have a maximum length of 30. Then, we concatenated these features with those obtained from inertial sensor data for the

feature-level fusion. In this way, we increased the accuracy of HAR without making the proposed framework computationally expensive. Finally, to have a fair comparison with the results reported for subject-specific experiments in [54], we also evaluated the proposed HAR scheme using the same protocols. Using the feature-level fusion of RGB and inertial sensors, the subject-specific experiments for our proposed scheme obtained an accuracy rate of 98.2%, which is better than previously reported results in [54]. Hence, it is concluded that the proposed scheme provides better recognition results than state-of-the-art.

IV. CONCLUSION

In this paper, a feature-level fusion method has been proposed for human action recognition, which utilizes data from two differing sensing modalities: vision and inertial. The proposed system merges the features extracted from individual sensing modalities to recognize an action using a supervised machine learning approach. The detailed experimental results indicate the robustness of our proposed method regarding classifying human actions as compared to the settings where each sensor modality is used individually. Also, the feature-level fusion of time domain features computed from inertial sensors and densely extracted HOG features from depth/RGB videos reduces the computational complexity and improves the recognition accuracy of the system as compared to state-of-the-art deep CNN methods. Regarding classifier performance, K-NN classifier provides better results for the proposed HAR system as compared to SVM classifier.

The proposed HAR methods also have some limitations. For example, it works with pre-segmented actions, which do not exist in practice. Moreover, it does not incorporate multi-view HAR, and the orientation of the person whose action is being recognized remains the same with respect to the camera. In the future, we plan to extend the proposed HAR method to address these limitations. Furthermore, we aim to investigate the specific applications of the proposed fusion framework using RGB-D camera and wearable inertial sensors.

REFERENCES

[1] W. Chi, J. Wang, and M. Q.-H. Meng, "A gait recognition method for human following in service robots," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 48, no. 9, pp. 1429–1440, Sep. 2018.

[2] G. Liang, X. Lan, J. Wang, J. Wang, and N. Zheng, "A limb-based graphical model for human pose estimation," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 48, no. 7, pp. 1080–1092, Jul. 2018.

[3] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in human action recognition: A survey," *New J. Phys.*, vol. 17, no. 8, pp. 1–30, 2015.

[4] L. Fademrecht, I. Bülthoff, and S. de la Rosa, "Action recognition in the visual periphery," *J. Vis.*, vol. 16, no. 33, pp. 1–14, 2016.

[5] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Advances in Computer Vision and Pattern Recognition*, vol. 71. Cham, Switzerland: Springer, 2014, pp. 181–208.

[6] X. Li, R. Chen, and T. Chu, "A crowdsourcing solution for road surface roughness detection using smartphones," in *Proc. 27th Int. Tech. Meeting Satellite Division Inst. Navigat. (ION GNSS)*, vol. 1, 2014, pp. 498–502.

[7] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[8] A. B. Mahjoub and M. Atri, "Human action recognition using RGB data," in *Proc. 11th Int. Design Test Workshop*, Dec. 2016, pp. 83–87.

[9] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[10] F. Lv, R. Nevatia, and M. W. Lee, "3D human action recognition using spatio-temporal motion templates," in *Computer Vision in Human-Computer Interaction*, vol. 3766. Berlin, Germany: Springer, 2005, pp. 120–130.

[11] N. Noorit, N. Suvonvorn, and M. Karnchanadecha, "Model-based human action recognition," in *Proc. 2nd Int. Conf. Digit. Image Process.*, vol. 7546, 2010, Art. no. 75460P. doi: 10.1117/12.853223.

[12] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Comput. Vis. Image Understand.*, vol. 150, pp. 109–125, Sep. 2016.

[13] B. Fernando, E. Gavves, J. Oramas M., A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 773–787, Apr. 2017.

[14] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1860–1870, Dec. 2013.

[15] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, p. 99.1.

[16] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.

[17] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description," *IET Comput. Vis.*, vol. 10, no. 7, pp. 758–767, Oct. 2016.

[18] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "PMHI: Proposals from motion history images for temporal segmentation of long uncut videos," *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 179–183, Feb. 2018.

[19] L. Wang *et al.*, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[20] H. Chen, J. Chen, C. Chen, and R. Hu, "Action recognition with gradient boundary convolutional network," in *Proc. Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1047–1051.

[21] B. Meng, X. J. Liu, and X. Wang, "Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos," *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 26901–26918, 2018.

[22] D. Wu, N. Sharma, and M. Blumenstein, "Recent advances in video-based human action recognition using deep learning: A review," in *Proc. Int. Joint Conf. Neural Netw.*, May 2017, pp. 2865–2872.

[23] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human action recognition with video data: Research and evaluation challenges," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 5, pp. 650–663, Oct. 2014.

[24] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 2, pp. 224–241, Feb. 2011.

[25] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[26] M. Koohzadi and N. M. Charkari, "Survey on deep learning methods in human action recognition," *IET Comput. Vis.*, vol. 11, no. 8, pp. 623–632, Dec. 2017.

[27] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 20–27.

[28] V. Megavannan, B. Agarwal, and R. V. Babu, "Human action recognition using depth maps," in *Proc. Int. Conf. Signal Process. Commun. (SPCOM)*, Jul. 2012, pp. 1–5.

[29] X. Yang, C. Zhang, and Y. L. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. 20th ACM Int. Conf. Multimedia (MM)*, 2012, pp. 1057–1060.

[30] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. M. Campos, "STOP: Space-time occupancy patterns for 3D action recognition from depth map sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7441. Berlin, Germany: Springer, 2012, pp. 252–259.

[31] X. Yang and Y. L. Tian, "EigenJoints-based action recognition using Naïve–Bayes-nearest-neighbor," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 14–19.

[32] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *J. Real-Time Image Process.*, vol. 12, no. 1, pp. 155–163, Aug. 2013.
[33] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2010, pp. 9–14.
[34] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2015, pp. 1092–1099.
[35] Z. Shi, "P-SNV: Pyramid-super normal vector descriptor for human action recognition based on depth sequences," *J. Inf. Comput. Sci.*, vol. 12, no. 18, pp. 7061–7070, 2015.
[36] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2834–2841.
[37] X. Yang and Y. Tian, "Super normal vector for human activity recognition with depth cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1028–1039, May 2017.
[38] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2014, pp. 4513–4518.
[39] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 807–811, Mar. 2018.
[40] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowledge-Based Systems*, vol. 158, pp. 43–53, Oct. 2018.
[41] A. Kamel, B. Sheng, P. Yang, T. Li, R. Shen, and D. D. Feng, "Deep convolutional neural networks for human action recognition using depth maps and postures," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published,
[42] H.-H. Pham, L. Khoudoura, P. Zegersc, A. Crouzil, and S. A. Velastin, "Exploiting deep residual networks for human action recognition from skeletal data," *Comput. Vis. Image Understand.*, vol. 170, pp. 51–66, May 2018.
[43] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognit.*, vol. 76, pp. 80–94, Apr. 2018.
[44] J.-F. Hu, W.-S. Zheng, J. Pan, J. Lai, and J. Zhang, "Deep bilinear learning for RGB-D action recognition," in *Computer Vision—ECCV* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11211. Cham, Switzerland: Springer, 2018, pp. 346–362.
[45] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Comput. Vis. Image Understand.*, vol. 171, pp. 118–139, Jun. 2018.
[46] M. Ehatisham-ul-Haq, M. A. Azam, U. Naeem, Y. Amin, and J. Loo, "Continuous authentication of smartphone users based on activity pattern recognition using passive mobile sensing," *J. Netw. Comput. Appl.*, vol. 109, pp. 24–35, May 2018.
[47] J. Wannenburg and R. Malekian, "Physical activity recognition from smartphone accelerometer data for user context awareness sensing," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 12, pp. 3142–3149, Dec. 2017.
[48] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, 3rd Quart., 2013.
[49] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga, "Complex human activity recognition using smartphone and wrist-worn motion sensors," *Sensors*, vol. 16, no. 4, p. 426, 2016.
[50] M. Shoaib, S. Bosch, H. Scholten, P. J. M. Havinga, and O. D. Incel, "Towards detection of bad habits by fusing smartphone and smartwatch sensors," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2015, pp. 591–596.
[51] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Future Gener. Comput. Syst.*, vol. 81, pp. 307–313, Apr. 2018.
[52] M. M. Hassan, S. Huda, M. Z. Uddin, A. Almogren, and M. Alrubaian, "Human activity recognition from body sensor data using deep learning," *J. Med. Syst.*, vol. 42, no. 6, p. 99, 2018.
[53] 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," in *Proc. Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 168–172.
[54] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors J.*, vol. 16, no. 3, pp. 773–781, Feb. 2016.
[55] C. Chen, R. Jafari, and N. Kehtarnavaz, "Improving human action recognition using fusion of depth camera and inertial sensors," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 1, pp. 51–61, Feb. 2015.
[56] C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of depth, Skeleton, and inertial data for human action recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2712–2716.
[57] I. Hwang, G. Cha, and S. Oh, "Multi-modal human action recognition using deep neural networks fusing image and inertial sensor data," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst.*, Nov. 2017, pp. 278–283.
[58] N. Dawar and N. Kehtarnavaz, "Action detection and recognition in continuous action streams by deep learning-based sensing fusion," *IEEE Sensors J.*, vol. 18, no. 23, pp. 9660–9668, Dec. 2018.
[59] Z.-Z. Wu, S.-H. Wan, L. Yan, and L.-H. Yue, "Autoencoder-based feature learning from a 2D depth map and 3D skeleton for action recognition," *J. Comput.*, vol. 29, no. 4, pp. 82–95, 2018.
[60] N. Dawar, S. Ostadabbas, and N. Kehtarnavaz, "Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition," *IEEE Sensors Lett.*, vol. 3, no. 1, Jan. 2019, Art. no. 7101004.
[61] R. Cui, G. Hua, A. Zhu, J. Wu, and H. Liu, "Hard sample mining and learning for skeleton-based human action recognition and identification," *IEEE Access*, vol. 7, pp. 8245–8257, 2017.
[62] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[63] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Int. Workshop Stat. Eur. Conf. Comput. Vis. (ECCV)*, 2004, pp. 59–74.
[64] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[65] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of Web videos," *Mach. Vis. Appl.*, vol. 24, no. 5, pp. 971–981, 2013.
[66] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1, Oct. 2005, pp. 604–610.
[67] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 124.1–124.11.
[68] J. R. R. Uijlings, I. C. Duta, N. Rostamzadeh, and N. Sebe, "Realtime video classification using dense HOF/HOG," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*, 2014, pp. 145–152.
[69] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, Dec. 2005.
[70] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors*, vol. 14, no. 6, pp. 10146–10176, 2014.
[71] A. Anjum and M. U. Ilyas, "Activity recognition using smartphone sensors," in *Proc. IEEE 10th Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2013, pp. 914–919.
[72] X. Su, H. Tong, and P. Ji, "Activity recognition with smartphone sensors," *Tsinghua Sci. Technol.*, vol. 19, no. 3, pp. 235–249, Jun. 2014.
[73] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.

MUHAMMAD EHATISHAM-UL-HAQ received the B.Sc. degree in computer engineering from the University of Engineering and Technology (UET), Taxila, Pakistan, in 2014, winning the Gold Medal, and the M.Sc. degree in computer engineering from UET Taxila, in 2017, winning the Chancellor's Gold Medal, where he is currently pursuing the Ph.D. degree in computer engineering.

His field of specialization is pervasive and ubiquitous computing. His research interests are within the areas of signal, image, and video processing, biomedical signal processing, mobile sensing, machine learning, human activity and emotion recognition, and human behavior analysis.

ALI JAVED (M'16) received the B.Sc. degree (Hons.) in software engineering from the University of Engineering and Technology (UET), Taxila, Pakistan, in 2007, and the M.S. and Ph.D. degrees in computer engineering from UET Taxila, Pakistan, in 2010 and 2016, respectively.

He was a visiting Ph.D. Research Scholar with the ISSF Lab, University of Michigan, MI, USA, in 2015. He has also served as the HOD of the Software Engineering Department, UET Taxila, in 2014, where he is currently an Assistant Professor. He is also serving as a Postdoctoral Scholar with the Computer Science and Engineering Department, Oakland University, Rochester, MI, USA. He was awarded the HEC Scholarship to pursue his Ph.D. research work at the University of Michigan. His areas of interests include digital image processing, computer vision, video content analysis, machine learning, multimedia signal processing, medical image processing, and digital forensics.

Dr. Javed got selected as an Ambassador of the Asian Council of Science Editors from Pakistan, in 2016. He has been a member of the Pakistan Engineering Council, since 2007. He has secured 3rd position in Software Batch–2003F, and received the Chancellor's Gold Medal for his M.S. degree in computer engineering. He is a recipient of various research grants from HEC Pakistan, National ICT R n D Fund Pakistan, and UET Taxila Pakistan.

MUHAMMAD AWAIS AZAM received the B.Sc. degree in computer engineering from the University of Engineering and Technology (UET), Taxila, Pakistan, in 2006, the M.Sc. degree (Hons.) in wireless networks from Queen Mary University, London, U.K., in 2008, and the Ph.D. degree in pervasive and ubiquitous computing from London, in 2012. From 2006 to 2007, he was a Lecturer with UET. From 2012 to 2013, he was the Head of Academics in the Cromwell College of IT & Management, London. Since 2013, he has been an Assistant Professor with the Department of Computer Engineering, UET. He leads a research team of M.Sc. and Ph.D. students in the area of pervasive and ubiquitous computing. His research interests include network architecture, the IoT, network security, embedded systems, ambient intelligence, wireless communications, opportunistic networks, and recommender systems. He received the Gold Medal for his B.Sc. degree.

HAFIZ M. A. MALIK (S'02–M'06–SM'10) received the B.E. degree (Hons.) in electronics and communications engineering from the University of Engineering and Technology Lahore, Pakistan, in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Illinois, Chicago, in 2006. After the Ph.D. degree, he joined the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA, where he was a Postdoctoral Research Fellow. He is currently serving as an Associate Professor with the ECE Department, University of Michigan–Dearborn, MI. He has published more than 15 technical papers and book chapters in refereed conferences and journals in the area of multimedia security, steganography, steganalysis, multimedia processing, audio analysis/synthesis, and statistical signal processing. His research interests include the general areas of digital content protection and digital signal processing, and the focus of current research includes information security, steganography, steganalysis, statistical signal processing, audio analysis/synthesis, and digital forensic analysis. He has served as an Organizing Committee Member of the special track on Doctoral Dissertation in the IEEE International Symposium on Multimedia (ISM) 2006. He was a member of the technical program committees of several conferences.

AUN IRTAZA received the Ph.D. degree from FAST-National University of Computer and Emerging Sciences, in 2016. During his Ph.D., he was a Research Scientist with the Signal and Image Processing Laboratory, Gwangju Institute of Science and Technology (GIST), South Korea. His research interests include computer vision, pattern analysis, and big data analytics.

IK HYUN LEE received the B.S. degree in control and instrument engineering from Korea University, South Korea, in 2004, and the M.S. and Ph.D. degrees from the School of Information and Mechatronics, Gwangju Institute of Science and Technology, South Korea, in 2008 and 2013, respectively. He was a Postdoctoral Researcher with the Media Laboratory, Massachusetts Institute of Technology, and a Senior Researcher with the Korea Aerospace Institute of Research. He is currently an Assistant Professor with the Department of Mechatronics Engineering, Korea Polytechnic University, South Korea. His research interests include image registration, image fusion, depth estimation, and medical image processing.

MUHAMMAD TARIQ MAHMOOD (M'12–SM'16) received the M.S. degree in computer science from the Blekinge Institute of Technology, Sweden, in 2006, and the Ph.D. degree in informatics and mechatronics from the Gwangju University of Science and Technology, South Korea, in 2011. He is currently an Assistant Professor with the School of Computer Science and Engineering, Korea University of Technology and Education, South Korea. His research interests include image processing, machine learning, and pattern recognition.