Action Detection and Recognition in Continuous Action Streams by Deep Learning-Based Sensing Fusion

Neha Dawar, *Student Member, IEEE*, and Nasser Kehtarnavaz, *Fellow, IEEE*

***Abstract*— This paper presents a deep learning-based sensing fusion system to detect and recognize actions of interest from continuous action streams, which contain actions of interest occurring continuously and randomly among arbitrary actions of non-interest. The sensors used in the fusion system consist of a depth camera and a wearable inertial sensor. A convolutional neural network is utilized for depth images obtained from the depth sensor, and a combination of convolutional neural network and long short-term memory network is utilized for inertial signals obtained from the inertial sensor. Each sensing modality first performs segmentation of all actions and then detection of actions of interest for a particular application. A decision-level fusion of the two sensing modalities is carried out to achieve the recognition of the detected actions of interest. The developed fusion system is examined for two applications: one involving transition movements for home healthcare monitoring and the other involving smart TV hand gestures. The results obtained show the effectiveness of the developed fusion system in dealing with realistic continuous action streams.**

***Index Terms*— Deep learning-based continuous action detection and recognition, fusion of depth and inertial sensing, action detection and recognition in continuous action streams.**

I. Introduction

HUMAN action or gesture recognition has enabled natural interfacing between humans and computers, and has already found its way into consumer electronics products. Many applications have benefitted from human action or gesture recognition. For example, human action recognition has been increasingly used for activity monitoring of the elderly population in home environments to address the steady increase in healthcare costs [1].

Different sensing modalities including RGB cameras, e.g. [2], [3], depth cameras, e.g. [4], [5] and inertial sensors, e.g. [6], [7] have been mostly utilized individually for human action or gesture recognition. As discussed in our previous works [8]–[10], action or gesture recognition can be made more robust by fusing decision from two differing modality sensors as compared to a single modality sensor.

In the great majority of works reported in the literature on action or gesture recognition, actions or gestures of interest are already segmented from action streams. To operate a human computer interaction system in a real-world setting, it is required that the actions of interest are detected from unseen continuous action streams in which they occur randomly and continuously amongst arbitrary actions of non-interest or no actions. This real world setting is by far a more challenging scenario as compared to the scenario where action streams are segmented manually such that segments contain only one action of interest. Detection of actions of interest from continuous action streams requires first segmenting all possible actions, regardless whether they are actions of interest or actions of non-interest, followed by identifying and classifying the actions of interest for a particular application. In our previous works [11]–[13], several fusion approaches were developed to detect and recognize smart TV gestures from continuous action streams by using skeleton joint positions obtained from a depth camera and inertial signals obtained from an inertial sensor. In [14], a data flow synchronization technique was developed to enable the real-time implementation of our fusion approaches.

Most of the previously developed fusion systems use handcrafted features together with classifiers such as Hidden Markov Model (HMM), Collaborative Representation Classifier (CRC), and Maximum Entropy Markov Model (MEMM) [11]–[15]. With the growing popularity of deep learning neural networks due to their high performance in various recognition tasks, in particular Convolutional Neural Networks (CNN) [16] and Long Short Term Memory (LSTM) networks [17], a CNN+LSTM-based fusion system to automatically detect and recognize actions of interest from continuous action streams has been developed in this work. The developed fusion system is used to detect actions of interest from continuous action streams for two applications including human body transition movements monitoring and smart TV hand gesture recognition. The actions of interest in the transition movements monitoring application involve transitions between the body states of sitting, standing and lying down. Considering the importance of fall detection monitoring for elderly and patients [18], in addition to the transition movements, falls are also monitored and detected here.

The fusion system developed in this paper utilizes a depth camera and a wearable inertial sensor simultaneously to perform continuous action detection and recognition. Unlike video cameras, depth cameras do not provide identifying facial

Manuscript received September 9, 2018; accepted September 15, 2018. Date of publication September 28, 2018; date of current version November 13, 2018. The associate editor coordinating the review of this paper and approving it for publication was Prof. Danilo Demarchi. *(Corresponding author: Neha Dawar.)*

The authors are with the Department of Electrical and Computer Engineering, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: neha.dawar@utdallas.edu; kehtar@utdallas.edu).

Digital Object Identifier 10.1109/JSEN.2018.2872862

information thus avoiding any privacy concern. A continuous action dataset is also made available in this paper for public use. This dataset consists of synchronized depth images and inertial signals associated with body transition movements as well as falls that are performed in a continuous and random manner in between various actions of non-interest. In addition to this dataset, our continuous action dataset (named UTD-CAD) in [13] which consists of smart TV hand gestures performed continuously and randomly in between various actions of non-interest is also examined here. Noting that training a CNN or LSTM network often requires very large datasets, a data augmentation step is carried out to address the limited size of the above continuous datasets for CNN and LSTM training.

Basically, this work constitutes the first attempt at developing a deep learning-based fusion system based on a depth camera and an inertial sensor for the purpose of detecting and recognizing actions of interest of an application that are performed continuously and randomly in between arbitrary actions of non-interest.

The rest of the paper is organized as follows. An overview of related works appears in Section II. Section III covers a description of the transition movements dataset collected for this study which is provided for public use. Section IV covers the details of the developed deep learning-based fusion system. The experimental results and their discussion are then presented in Section V. Finally, the paper is concluded in Section VI.

II. Overview of Related Works

The bulk of research on action or gesture recognition involves the use of a single modality sensor. However, there are limitations associated with using a single modality sensor when performing action or gesture recognition in real-world settings [19] due to high intra-class variations and low inter-class variations in the actions performed for a particular application. No modality sensing can cope with such variations perfectly or flawlessly. Fusion is a way to address such limitations of using a single modality sensing. In [15], a fusion system using information from a depth camera and an inertial sensor was developed to achieve more robust gesture recognition. In [8], depth motion maps derived from depth images and statistical features derived from inertial signals were fused to achieve improved action recognition. The use of three data modalities of depth images, skeleton joint positions, and inertial signals was reported in [10].

Furthermore, in most action or gesture recognition approaches, actions or gestures are considered to be segmented actions or gestures with the start and end of the actions or gestures already known or manually identified. In [11], we reported an approach to detect and recognize actions of interest performed continuously and randomly amongst unknown actions of non-interest using skeleton joint positions obtained from a depth camera and inertial signals obtained from a wearable inertial sensor. Skeleton joint positions were used to perform detection and recognition while inertial signals were used to enhance the performance of recognition by removing false positives. In [12], we used skeleton joint positions to detect actions of interest from continuous action streams while recognition was achieved by fusing the outcomes of two CRC classifiers, one acting on skeleton joint positions and the other on depth images. In [13] and [14], we reported a fusion approach at both detection and recognition stages based on skeleton joint positions and inertial signals. In these works, the detection of actions of interest from continuous actions streams was achieved using one-class Support Vector Data Descriptor (SVDD) classifiers and the classification of the detected actions of interest was achieved using CRC classifiers.

Recently, deep learning neural networks, in particular CNN and LSTM, have been increasingly used for action and gesture recognition based on a single modality sensor. For example, weighted hierarchical depth motion maps (WHDMM) were used in a three channel CNN in [20] to improve the recognition performance. In [21], a 3D CNN was used to learn the spatio-temporal features from raw depth sequences and it was combined with the feature vectors obtained from skeleton joint positions. In [22], both CNN and LSTM networks were considered for depth image sequences to achieve recognition. In [23], inertial signals from a set of body worn sensors were used and fed as images into a CNN network to recognize human activity. In [24], CNN and LSTM layers were combined to achieve action recognition using information from multiple wearable inertial sensors. In [25], shallow features of inertial signals were used along with deep features extracted by a CNN network to achieve recognition.

The work reported in this paper differs from all of the previous works in the following manner. In comparison to single modality sensor solutions reported in the literature to perform action recognition, a fusion system is developed in this work by using CNN and LSTM networks to detect and recognize actions of interest from continuous action streams. Detection and recognition are performed for each of the two sensing modalities in parallel followed by a decision-level fusion. CNN is used to learn the spatio-temporal features from depth images, while both CNN and LSTM are used to learn the temporal features from inertial signals.

III. Continuous Datasets

Considering the unavailability of a public domain continuous dataset where depth images and inertial signals are captured simultaneously, a dataset is collected in this work for the transition movements application and is made available for public use. The depth images in the dataset are captured by a Microsoft Kinect v2 depth camera at a rate of approximately 30 frames per second and a resolution of 512 × 424. Examples of background subtracted depth images captured by this camera are provided in [13]. The camera is connected to a laptop computer via a USB port. The inertial signals are captured by the wearable inertial sensor reported in [26] at a rate of 200Hz. These signals consist of 3-axis acceleration and 3-axis angular velocity signals, which are transmitted via a Bluetooth link to the laptop computer running the fusion system software. The data from the two sensors are synchronized based on the time stamp scheme described in [27]. Basically, time stamps of depth image frames are used

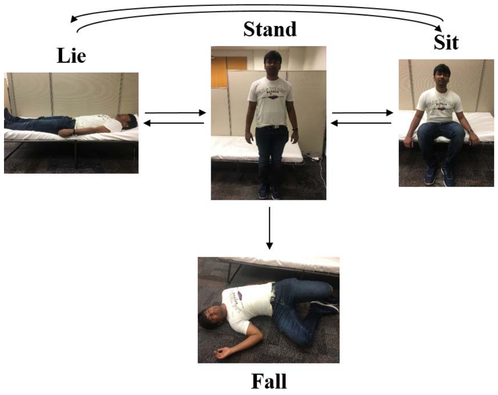


Fig. 1. Illustration of transition movements between the body states as well as fall in the continuous transition movements dataset: stand-to-lie, lie-to-stand, stand-to-sit, sit-to-stand, lie-to-sit, sit-to-lie, and fall.

as reference and inertial signals samples with the time stamp closest to a particular depth image frame are aligned with that depth image frame.

To collect data, a bed was placed in a room which had the depth camera installed at the room corner near the ceiling. The inertial sensor was worn on the waist while performing these actions. The dataset collected consists of 6 transition movements between the body states of sitting, standing and lying down, as well as falling down, thus forming these actions of interest ‘stand-to-sit’, ‘sit-to-stand’, ‘stand-to-lie’, ‘lie-to-stand’, ‘sit-to-lie’, ‘lie-to-sit’, and ‘fall’. Fig. 1 illustrates these transition movements between the body states. The continuous testing dataset was collected from 5 different subjects and a total of 5 continuous sets were collected from each subject resulting in a total of 25 continuous testing sets. Each continuous set contains the above 7 actions of interest performed in a continuous and random manner in between arbitrary actions of non-interest such as stretching, reading a book, drinking water, eating, combing, etc. The subjects were given complete freedom to perform any actions of non-interest as per their choice.

The collected continuous dataset was used only for testing or the operation phase. Training of the neural networks was performed using the segmented transition movement and fall actions provided in [28]. This dataset consists of segmented action data from 12 subjects. The continuous transition movement dataset collected in this work is made available for public use at this link: www.utdallas.edu/∼kehtar/UTD-Dataset-ContinuousTransitionMovements.htm.

For the smart TV application, the continuous dataset for the smart TV hand gestures in [13] is used here. The actions or gestures of interest in this dataset consist of ‘waving a hand’, ‘flip to left’, ‘flip to right’, ‘counterclockwise rotation’ and ‘clockwise rotation’. For these gestures, the inertial sensor was worn on the wrist. This dataset contains 5 continuous gesture streams each containing the above 5 gestures from 12 subjects.

IV. Developed Deep Learning-Based Continuous Detection and Recognition Fusion System

The developed fusion system carries out detection and recognition for each of the two differing sensing modalities of a depth camera and an inertial sensor, followed by a decision level fusion. The depth camera path uses a CNN network, while the inertial sensor path uses a CNN+LSTM network to perform detection and recognition. Segmentation is carried out along each modality path. The segmented actions are then passed through the CNN or CNN+LSTM networks of their respective paths to detect actions of interest or actions of non-interest and then to classify the detected actions of interest. A decision level fusion is conducted based on the output of the two paths. Fig. 2 illustrates a block diagram of the overall detection and recognition fusion system. Note that detection here means identifying the segmented actions as actions of interest or as actions of non-interest, while recognition means classifying the detected actions of interest.

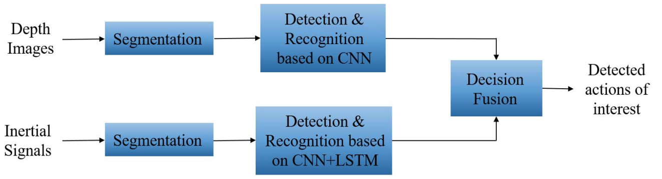

Depth Images
Segmentation
Detection & Recognition based on CNN
Decision Fusion
Detected actions of interest
Inertial Signals
Segmentation
Detection & Recognition based on CNN+LSTM

Fig. 2. Block diagram of the detection and recognition fusion system.

A. Segmentation

For segmentation of the transition movements in the depth camera path, the centroid (c_x, c_y) of the background subtracted depth images is obtained as follows:

$$c_x = \frac{\sum_{i=1}^{N} x_i m_i}{\sum_{i=1}^{N} m_i}, \quad c_y = \frac{\sum_{i=1}^{N} y_i m_i}{\sum_{i=1}^{N} m_i} \tag{1}$$

where (x_i, y_i) denotes a pixel location with m_i representing its intensity value. A sequence of centroids $C = (C^1, C^2, \ldots, C^t, \ldots)$ is then obtained where $C^t = (c_x, c_y)^t$ represents the centroid at frame t. A centroid difference C_d^t at t^{th} frame is then obtained as follows:

$$C_d^t = C^t - C^{t-1} \tag{2}$$

Noise related small fluctuations of centroid differences during no action are eliminated by setting centroid difference values below 5% level of the maximum centroid difference to zero. Frames with centroid differences above this level are used to denote the presence of movement or action. An example of centroid differences for a continuous action stream is shown in Fig. 3.

A similar segmentation process is carried out to segment actions using the inertial signals. If g_x^t, g_y^t, g_z^t denote the 3D angular velocities at a frame t, the angular velocity G^t at this frame is obtained as follows:

$$G^t = \sqrt{{g_x^t}^2 + {g_y^t}^2 + {g_z^t}^2} \tag{3}$$

Let $G = (G^1, G^2, \ldots, G^t, \ldots)$ represent a sequence of angular velocities. An angular velocity difference G_d^t at frame t is then obtained as follows:

$$G_d^t = G^t - G^{t-1} \tag{4}$$

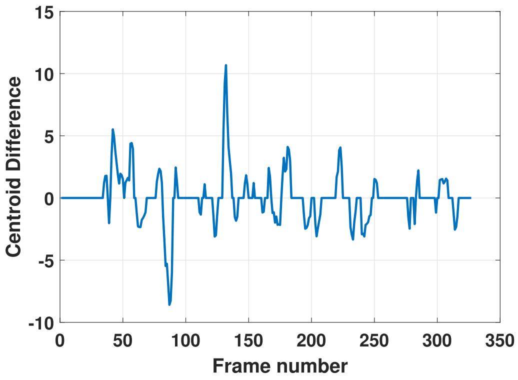


Fig. 3. An example of centroid differences of a continuous action stream.

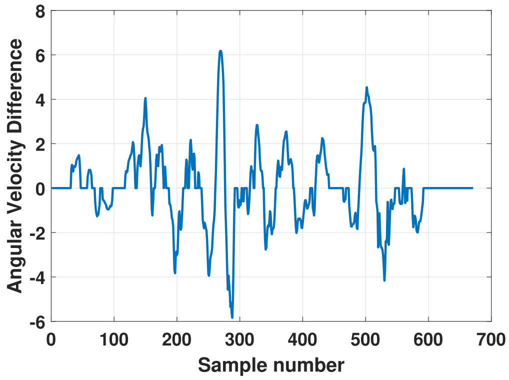


Fig. 4. An example of angular velocity differences of a continuous action stream.

Noise related small fluctuations of angular velocity differences during no action are eliminated by setting angular velocity differences below 5% level of the maximum angular velocity difference to zero. Frames with angular velocity differences above this level are used to denote the presence of movements or actions. An example of angular velocity differences for a continuous action stream is shown in Fig. 4.

For continuous smart TV hand gesture dataset, the same technique described in [13] is employed for segmentation. It is to be noted that the hand gestures involved in the continuous smart TV gesture dataset have the same starting and ending point. In other words, all the actions of interest start and end at more or less the same reference point. Hence, it is unnecessary to obtain the centroid or angular velocity differences for the purpose of capturing the peaks and dips of an action stream. Thresholding of the signals is sufficient to detect the start and end of the hand gestures. Furthermore, since the smart TV gestures are performed relatively close to the camera, skeleton joint positions are used as they provide more reliable information for segmentation instead of the depth image centroids considering that centroid positions do not change much when performing the hand gestures.

B. CNN Architecture for Depth Images

A two stream CNN is used here to detect and recognize actions based on depth images, that is there are two separate streams that use depth images to detect and recognize actions of interest from continuous action streams. The first stream takes raw background subtracted depth images as its input and uses 3D convolutional layers to obtain deep features from the segmented actions as was reported in [21]. The raw images are resized to a common size of 32 × 32 and a fixed number of frames is evenly extracted from each action sequence. These resized depth images are used as the input to the first stream. For the transition movements dataset, 15 depth images per action sequence are used as the input noting that the average length of each action of interest is 15 frames. Similarly, 25 depth images are used for the continuous smart TV hand gesture dataset.

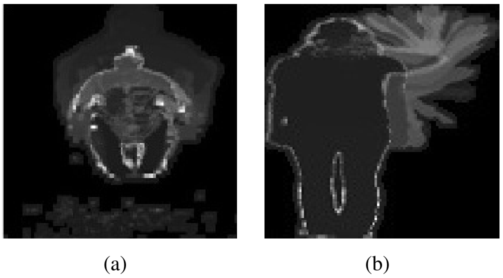


Fig. 5. Examples of weighted DMM images: (a) action ‘stand-to-sit’ from the transition movements dataset, (b) action ‘waving a hand’ from the smart TV gesture dataset.

As discussed in [25], when the dataset size is limited, one CNN stream alone is not able to capture the hierarchy of features in its entirety. Hence, another stream of CNN is used by considering handcrafted features of the segmented actions as its inputs. These features are the weighted depth motion map (DMM) images of the actions. To obtain the weighted DMM images, the depth images are projected onto three orthogonal planes corresponding to the front, side and top views. In order to keep the computational complexity low, only the projection onto the front view is utilized here. The projection map is weighted to obtain the DMM image. If map^n represents the front view projection map for n^{th} frame, for a sequence of N depth frames, the weighted DMM is computed as follows [29]:

$$DMM = \sum_{n=1}^{N-1} |map^{n+1} - map^n| \cdot weight(n+1) \qquad (5)$$

The motion areas are weighted linearly, that is $weight(n) = n/N$. The advantage of using weighted DMMs instead of traditional DMMs [8] is that motion areas around later frames appear brighter than earlier frames, thus making it easier to differentiate between the reversed transition movements such as ‘sit-to-stand’ and ‘stand-to-sit’, which would otherwise have a similar DMM. An example of a weighted DMM image from the transition movements dataset and an example from the smart TV gesture dataset are shown in Fig. 5. The weighted DMM images are resized to 50 × 50 and used as the input to the second CNN stream.

The overall architecture of the two CNN streams is shown in Fig. 6. The first CNN stream comprises two 3D convolutional layers. The first layer convolves the raw depth images with 16 convolution filters of size 5 × 5 × 5. The output is passed through a 3D subsampling layer employing max pooling. The second convolutional layer convolves the output

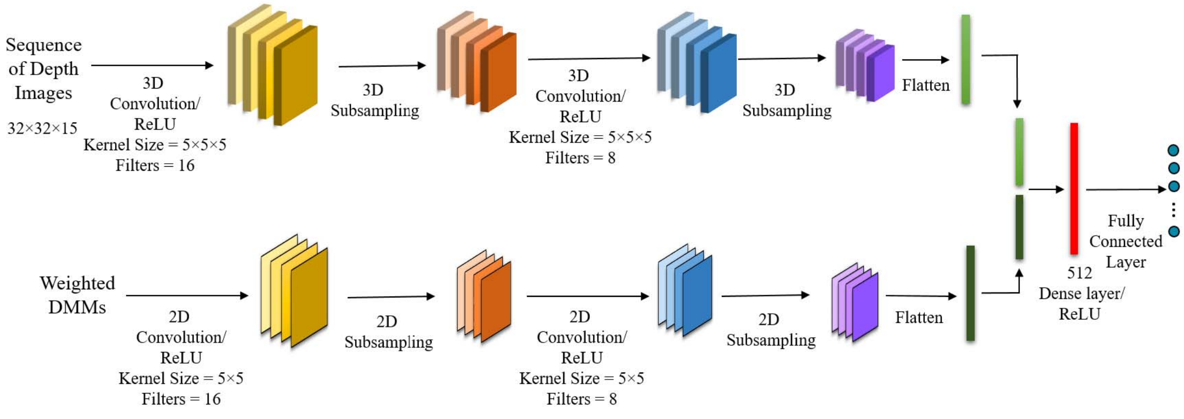


Fig. 6. CNN architecture used for continuous action detection and recognition based on depth images.

of the pooling layer with 8 filters of size 5 × 5 × 5 and passes it to another 3D subsampling layer. The second CNN stream convolves the input weighted DMM images with 16 2D convolution filters of size 5 × 5. The outputs of the convolution layer are subsampled and passed onto another 2D convolution layer comprising 8 filters of size 5 × 5, followed by another subsampling layer using max pooling. Rectified linear unit (ReLU) activation is used at each 2D and 3D convolution layer. The output of each CNN stream is flattened and concatenated before passing it to a dense layer, which maps it to a 512 × 1 vector based on the ReLU activation. The last layer is a fully connected layer using a sigmoid activation, which generates scores for the output classes.

C. CNN+LSTM Architecture for Inertial Signals

An architecture similar to the above two-stream CNN architecture is used for inertial signals, except that the second stream directly uses the handcrafted features with no further feature extraction. The first stream uses CNN and LSTM layers and the second stream directly uses the handcrafted inertial features. The input to the first stream is 8 time-series signals corresponding to the 3-axis acceleration signals, the 3-axis angular velocity signals, the overall acceleration signal, and the overall angular velocity signal. The overall angular velocity is obtained by computing the angular velocity at each frame via Eq. (3) and the overall acceleration is obtained using the acceleration A^t at each frame t of a sequence as follows:

$$A^t = \sqrt{{a_x^t}^2 + {a_y^t}^2 + {a_z^t}^2} \quad (6)$$

where a_x^t, a_y^t, a_z^t denote the 3D accelerations at frame t. These 8 time-series signals are sampled to obtain a total of 200 samples per sequence. These signals are normalized and sent to the CNN+LSTM stream as stacked inertial signal images of size 200 × 8.

The handcrafted features used in the second stream involve the statistical features of the inertial signals. The above 8 time-series signals are divided into 3 equal sized temporal segments and similar to [25], the statistical features of *mean*, *variance*, *standard deviation*, *root mean square*, *median*, *minimum*, and *maximum* of the segments of these signals, and *mean*, *variance*, *standard deviation*, and *root mean square* of the segments of their first derivatives are used as the handcrafted signals.

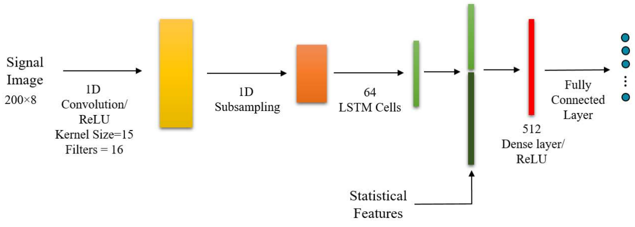


Fig. 7. CNN+LSTM architecture used for continuous action detection and recognition based on inertial signals.

The overall architecture of the CNN+LSTM network used for the inertial signals is shown in Fig. 7. The inertial signal images are first convolved with 16 1D filters of size 15 to obtain features from the time-series signals based on the ReLU activation. The output is subsampled using 1D max pooling and passed onto the LSTM layer with 64 cells. The output of the LSTM layer is then used along with the handcrafted features from the second stream to form a single vector. A dense layer maps this vector to a vector of size 512 × 1 based on the ReLU activation. The output of the dense layer is finally passed onto a fully connected layer which uses a sigmoid activation generating scores for the output classes.

For the continuous smart TV gesture dataset, another 1D convolution layer with 8 filters of size 9 and a subsampling layer is used before the LSTM layer to capture the entire dynamics of the hand gestures. Since the size of the time-series signals gets reduced further by adding another subsampling layer, only 16 LSTM cells are used here. The handcrafted features used for this dataset are the statistical features of *mean*, *variance*, *standard deviation*, *root mean square*, *median*, *minimum* and *maximum* of the three equal sized segments of the 8 time-series signals.

Here, it is worth stating that different architectures and parameters were examined to reach the architectures utilized here by using a subset of the training data as the validation data. One to three convolution layers with different numbers and sizes of filters were considered. Different numbers of LSTM cells were also examined. The architectures reported above for depth images and inertial signals were found to be the most effective ones. Apart from examining different architectures for decision-level fusion, a feature-level or data fusion was also considered by passing the data from the two sensors to a common network. It was found that the decision-level fusion by far was more effective than the feature-level or data fusion.

D. Continuous Detection and Recognition

A technique similar to the one reported in [30] is adopted here to perform detection and recognition from continuously segmented actions. Given a segment S^k with S^k_{start} representing its starting point and S^k_{stop} its stopping point, an examination action set $A^k = \{A^0_k, A^1_k, \ldots, A^l_k, \ldots, A^{K-1}_k\}$ is formed where A^l_k represents an action whose starting point is S^{k-l}_{start}. The stopping point of all the actions in A^k is S^k_{stop}. Hence, whenever a segment is obtained from a continuous action stream, it is examined along with $K - 1$ prior segments to detect the presence of an action of interest. Only the actions formed from these segments whose length lies within the range of actions of interest are examined further. Based on the number of segments that normally occur in the transitions movements or actions of interest, $K = 5$ was found to work best for depth image segments and $K = 10$ was found to work best for inertial signal segments. Similarly, depth images with a length falling in the frame range of $\{8, 40\}$ and the ones segmented from inertial signals having a length falling in the sample range of $\{70, 750\}$ were found to work best. The experimentations reported in the next section are based on these values.

The detection and recognition of the actions of interest are performed based on the output scores of the fully connected layer. Note that a softmax activation is not used here at the fully connected layers. The reason is that softmax activation would result in the output scores that add up to one. A sigmoid activation is used instead along with the mean squared error loss function. This ensures that all the classes are trained individually and this way the output scores at the fully connected layer do not need to add up to one. The main advantage of modifying the loss function and activation at the fully connected layer is that detection and recognition of actions of interest can be performed at the same time using the same network. This modification results in a low score throughout all the classes for most actions of non-interest. Also, an action with more than one high score class is indicative of the presence of actions of non-interest.

Based on the output scores at the fully connected layers in the two paths, an initial detection of actions of interest is performed. Based on the output scores of the depth image path, only the actions with scores > 0.9 are labeled as potential actions of interest. Similarly, for the output scores in the inertial signal path, the actions which have the first scores > 0.9 and the second scores < 0.1 are labeled as potential actions of interest. Only the actions which qualify as potential actions of interest from the two paths are passed onto the next stage. The output scores from the two paths are then multiplied to obtain the fusion scores. The actions which have fusion scores > 0.8 for exactly one class and fusion scores < 0.1 for the rest of the classes are considered indicative of actions of interest. Such actions are labeled as actions of interest and are classified or placed in the class with the highest score. Note that performing detection both before the fusion and during the fusion results in the rejection of the great majority of actions of non-interest.

E. Data Augmentation for Limited Datasets

Since training a CNN or LSTM network requires a very large amount of training data, a data augmentation step was performed in order to address the limited size of the dataset for training the CNN or LSTM networks. In case of depth images, the training samples were flipped, rotated and translated to generate multiple new training samples from a single sample. These operations were applied to both depth image sequences and weighted DMMs simultaneously to produce synchronized training samples. In case of inertial signals, white noise was added to random frames at the beginning or end or both at the beginning and end of the action streams. In addition to the data augmentation, a dropout ratio of 0.5 was used throughout the networks to control overfitting as noted in [31].

V. Experimental Results and Discussion

The effectiveness of the developed deep learning-based continuous action detection and recognition fusion system was examined on the two continuous datasets: continuous transition movements dataset and continuous smart TV hand gesture dataset. As mentioned earlier, the training of the CNN and CNN+LSTM networks was performed using the segmented datasets. Both the networks were trained individually from the two input layers to the fully connected output layer. The Adam optimizer was used to train the networks using the mean squared error loss function. For testing, the segmentation was carried out by using both of the sensing modalities and only the actions which qualified as potential actions of interest by both the modalities were passed onto the decision fusion stage to conduct the removal of false positives and to reach the final decision. The coding for both the training and testing of the developed continuous detection and recognition system was done in Python.

Here, it is worth mentioning that apart from the continuous datasets examined, all other existing datasets that provide simultaneous data from both a depth and an inertial sensor contain segmented or isolated actions, and thus it is not possible to test the performance of the detection and recognition system on these datasets. However, the developed fusion approach was compared with the existing fusion based recognition approaches in [13], [27], and [32] by using the UTD-MHAD dataset [27]. The fusion of CNN for depth images and CNN+LSTM for inertial signals was performed by multiplying the scores of their fully connected layers and the assigned class label was considered to be the one with the highest score. The UTD-MHAD dataset is a multimodal dataset comprising 27 actions performed by 8 subjects. To provide a fair comparison with the approach in [27], the data from the odd numbered subjects were used for training, while the data from the even numbered subjects were used for testing. The results obtained using different approaches are reported in Table I. As can be seen from this table, even with the limited

TABLE I

RECOGNITION ACCURACY FOR UTD-MHAD DATASET

Method	Accuracy (%)
ELC-KSVD [32]	76.2
Kinect and Inertial [27]	79.1
Skeleton Joints and Inertial [13]	86.3
Developed Deep Learning-based Fusion	92.8

TABLE II

RECOGNITION ACCURACIES WHEN USING SINGLE STREAMS VERSUS BOTH STREAMS

Dataset	Using first stream only	Using second stream only	Using both streams
UTD-MHAD	87.4%	87.4%	92.8%
Continuous Transition Movements	98.1%	95.2%	99.3%
Continuous Smart TV Gestures	80.3%	75.7%	86.3%

size of the dataset, a higher accuracy was achieved with the developed deep learning-based fusion system.

In order to see the effect of using two streams for recognition, the segmented data from the three datasets (UTD-MHAD, Continuous Transition Movements and Continuous Smart TV Gestures) were divided into a training and a validation set. The training sets were used to train the networks associated with single streams and the two networks associated with both streams. The recognition accuracies of the validation sets are reported in Table II. As can be seen from this table, the use of both streams led to higher accuracies compared to single streams.

To examine the performance of the overall system on the two continuous datasets, the ground truth actions were manually identified from the continuous action streams by visual inspection. The performance evaluation was based on the widely used measures of precision, recall and $F1$ score [33]. First, the detected actions were marked as either true positives or false positives. The actions of interest detected within a window of five frames from the ground truth and correctly classified were marked as true positives. The actions with no overlap with the ground truth, or the ones misclassified were marked as false positives. The ground truth actions which were not detected by the system, or the ones detected but not correctly classified were marked as false negatives. Based on the number of true positives N_{TP}, the number of false positives N_{FP} and the number of false negatives N_{FN} across all the continuous action streams, the measures of precision P, recall R and $F1$ scores were computed as follows [33]:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{7}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{8}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{9}$$

The results obtained for these measures are discussed further in the subsections that follow.

TABLE III

PRECISION, RECALL, AND $F1$ SCORE FOR THE CONTINUOUS TRANSITION MOVEMENTS DATASET

	Precision	Recall	$F1$ Score
Subject 1	91.3%	90.0%	90.3%
Subject 2	96.9%	88.6%	92.5%
Subject 3	94.2%	92.9%	93.5%
Subject 4	83.3%	100%	90.9%
Subject 5	86.1%	88.6%	87.3%
Average	90.4%	92.0%	90.9%

A. Continuous Transition Movements Dataset

The continuous action streams in the continuous transition movements dataset were segmented using centroid differences of the depth images and angular velocity differences of the inertial signals. The testing was repeated for each subject by not using the subject in training. A threshold falling in the range [1, 2] which corresponded to 5% of the maximum value was applied to remove negligible centroid differences in the continuous action streams. A threshold of 0.5 which corresponded to 5% of the maximum value was applied to remove angular velocity differences in the continuous action streams.

The measures of precision, recall and $F1$ score obtained by the developed deep learning-based fusion system for all the subjects are reported in Table III. This table also shows the overall or average precision, recall and $F1$ score obtained across all the subjects. This measure was also obtained by using a single modality of depth camera and inertial sensor and an improvement of more than 15% was achieved in $F1$ score when the fusion of the two modalities was used as compared to the cases where a single modality (either depth camera or inertial sensor) was used individually based on the same CNN or CNN+LSTM networks. This is due to the fact that the fusion system was able to reject most of the false positives that were detected by a single modality. An example showing the ground truth centroid difference signal and the detected actions is shown in Fig. 8. The confusion matrix indicating the recognition performance of the fusion system is reported in Table IV indicating an overall recognition accuracy of 98.8%. As indicated in this table, most misclassifications occurred due to the fact that the action ‘lie-to-stand’ can be regarded as a combination of the actions ‘lie-to-sit’ and ‘sit-to-stand’ performed in series, and the action ‘stand-to-lie’ can be regarded as a combination of the actions ‘stand-to-sit’ and ‘sit-to-lie’ performed in series. As a result, these actions were sometimes misdetected.

B. Continuous Smart TV Gesture Dataset

As mentioned earlier, since the continuous smart TV gesture dataset consists of hand gestures, it was easier to segment these hand gestures using the skeleton joint positions and inertial signals via the technique described in [13]. Once the segmentation was done, the deep learning-based fusion system was used to identify the actions of interest in order to provide a comparison with the subject-specific results reported in [13]. The subject-specific scenario means the system is

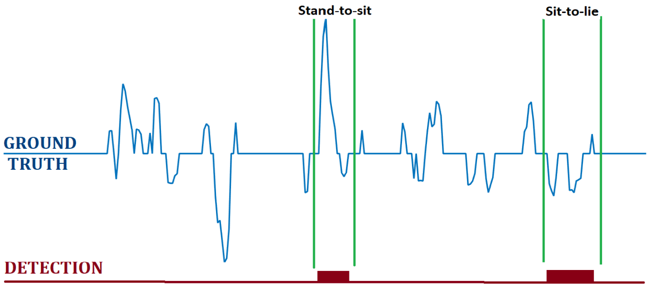


Fig. 8. An example of detected actions of interest versus the ground truth for the centroid difference signal in a continuous action stream.

TABLE IV

CONFUSION MATRIX FOR THE CONTINUOUS TRANSITION MOVEMENTS DATASET (IN %)

	St-S	St-L	S-St	S-L	L-S	L-St	F
St-S	100	-	-	-	-	-	-
St-L	-	96	-	4	-	-	-
S-St	-	-	100	-	-	-	-
S-L	-	4	-	96	-	-	-
L-S	-	-	-	-	100	-	-
L-St	-	-	-	-	-	100	-
F	-	-	-	-	-	-	100

St-S: stand-to-sit, St-L: stand-to-lie, S-St: sit-to-stand, S-L: sit-to-lie, L-S: lie-to-sit, L-St: lie-to-stand, F: fall

TABLE V

PRECISION, RECALL, AND $F1$ SCORE FOR THE CONTINUOUS SMART TV GESTURE DATASET

	Precision	Recall	*F*1 Score
[13]	96.6%	95.7%	96.2%
Deep Learning Fusion	97.5%	96.5%	97.0%

trained using the segmented data of the subject for whom testing is performed. The measures of precision, recall and $F1$ score were obtained and averaged for the 12 subjects in the dataset and compared to the results obtained by the SVDD and CRC-based continuous detection and recognition approach reported in [13]. Table V provides the comparison of the measures between the fusion approach in [13] and the one developed here. It should be noted that the approach developed in [13] utilizes skeleton joint positions while the approach developed here utilizes depth images. Although skeleton joint positions are more informative, the use of depth images is more general purpose in terms of applicability to different action recognition applications since in practice skeleton joint positions appear overlapping in many action recognition applications. To allow proper tracking of the skeleton joints, the joints should be visible at all times with no overlap. In practice, however, overlapping occurs in many action recognition applications. The confusion matrix of the recognition performance for the continuous smart TV dataset is provided in Table VI indicating an overall recognition accuracy of 97.6%.

C. System Operation Processing Time

The times to process segments and obtain their handcrafted features were measured on a laptop computer running the fusion system with the depth camera connected to it via a USB port and the wearable inertial sensor connected to it via a Bluetooth link. This laptop was equipped with a 4.2GHz processor and 64GB RAM. It was found that the computation of the weighted DMMs from the actions obtained from the depth segments and the final scores of the CNN network took 94ms on average. Similarly, the formation of the handcrafted statistical features from the inertial segments and the final score computation using the CNN+LSTM networks took 3ms on average. As a result, the detection and recognition of actions of interest from the continuous action streams was made 100ms after the completion of an action. It is worth mentioning here that this time represents the algorithmic complexity of the system to perform continuous detection and recognition via a modern laptop without the need to use any additional dedicated processing hardware. Two video clips of the operation of the fusion system running in real-time on continuous action streams corresponding to the two applications considered can be viewed at these links: www.utdallas.edu/~kehtar/DeepLearningFusionSystem-TranistionMovements.avi and www.utdallas.edu/~kehtar/DeepLearningFusionSystem-SmartTV.avi.

TABLE VI

CONFUSION MATRIX FOR THE CONTINUOUS SMART TV GESTURE DATASET (IN %)

	WH	FL	FR	CCR	CR
WH	100	-	-	-	-
FL	-	98.3	1.7	-	-
FR	-	1.7	96.6	1.7	-
CCR	-	1.7	-	93.3	5
CR	-	-	-	-	100

WH: waving a hand, FL: flip to left, FR: flip to right, CCR: counterclockwise rotation, CR: clockwise rotation

VI. CONCLUSION

In this paper, a deep learning-based fusion system to detect and recognize actions of interest from continuous action streams has been developed. Continuous action streams reflect the way actions are performed in real-world situations, that is when actions of interest are performed continuously and randomly among arbitrary and unknown actions of non-interest. The system uses depth images from a depth camera and inertial signals from a wearable inertial sensor. Decision-level fusion is applied to the actions of interest that are detected by both of the modalities in order to reject actions of non-interest and classify the detected actions of interest. The developed fusion system has been examined for two applications: one involving transition movements for home healthcare monitoring and the other for smart TV hand gestures. The results obtained indicate the effectiveness of the developed fusion system in the detection and recognition of actions of interest in realistic continuous action streams.

REFERENCES

[1] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *Proc. 23rd Int. Conf. Archit. Comput. Syst. (ARCS)*, Feb. 2010, pp. 1–10.

[2] N. Zerrouki, F. Harrou, Y. Sun, and A. Houacine, "Vision-based human action classification using adaptive boosting algorithm," *IEEE Sensors J.*, vol. 18, no. 12, pp. 5115–5121, Jun. 2018.

[3] W. Lao, J. Han, and P. H. N. De With, "Automatic video-based human motion analyzer for consumer surveillance system," *IEEE Trans. Consum. Electron.*, vol. 55, no. 2, pp. 591–598, May 2009.

[4] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1290–1297.

[5] B. Ghojogh, H. Mohammadzade, and M. Mokari, "Fisherposes for human action recognition using Kinect sensor data," *IEEE Sensors J.*, vol. 18, no. 4, pp. 1612–1627, Feb. 2018.

[6] R. Xu, S. Zhou, and W. Li, "MEMS accelerometer based nonspecific-user hand gesture recognition," *IEEE Sensors J.*, vol. 12, no. 5, pp. 1166–1173, May 2012.

[7] A. Wang, G. Chen, J. Yang, S. Zhao, and C.-Y. Chang, "A comparative study on human activity recognition using inertial sensors in a smartphone," *IEEE Sensors J.*, vol. 16, no. 11, pp. 4566–4578, Jun. 2016.

[8] 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.

[9] 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.

[10] C. Chen, R. Jafari, and N. Kehtarnavaz, "Fusion of depth, skeleton, and inertial data for human action recognition," in *Proc. ICASSP*, Mar. 2016, pp. 2712–2716.

[11] N. Dawar, C. Chen, R. Jafari, and N. Kehtarnavaz, "Real-time continuous action detection and recognition using depth images and inertial signals," in *Proc. IEEE Int. Symp. Ind. Electron.*, Jun. 2017, pp. 1342–1347.

[12] N. Dawar and N. Kehtarnavaz, "Continuous detection and recognition of actions of interest among actions of non-interest using a depth camera," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4227–4231.

[13] N. Dawar and N. Kehtarnavaz, "Real-time continuous detection and recognition of subject-specific smart tv gestures via fusion of depth and inertial sensing," *IEEE Access*, vol. 6, pp. 7019–7028, 2018.

[14] N. Dawar and N. Kehtarnavaz, "Data flow synchronization of a real-time fusion system to detect and recognize smart TV gestures," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2018, pp. 1–4.

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

[16] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[17] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, Jun. 2015, pp. 2625–2634.

[18] J. Chen, K. Kwong, D. Chang, J. Luk, and R. Bajcsy, "Wearable sensors for reliable fall detection," in *Proc. IEEE-EMBS 27th Annu. Int. Conf. Eng. Med. Biol. Soc.*, Jan. 2005, pp. 3551–3554.

[19] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools Appl.*, vol. 76, no. 3, pp. 4405–4425, 2017.

[20] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Trans. Human Mach. Syst.*, vol. 46, no. 4, pp. 498–509, Aug. 2016.

[21] Z. Liu, C. Zhang, and Y. Tian, "3D-based deep convolutional neural network for action recognition with depth sequences," *Image Vis. Comput.*, vol. 55, pp. 93–100, Nov. 2016.

[22] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based action recognition using LSTM and CNN," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 585–590.

[23] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. 24th Int. Conf. Artif. Intell. (IJCAI)*, Jul. 2015, pp. 3995–4001.

[24] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016.

[25] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, "A deep learning approach to on-node sensor data analytics for mobile or wearable devices," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 56–64, Jan. 2017.

[26] A. Y. Yang, R. Jafari, S. S. Sastry, and R. Bajcsy, "Distributed recognition of human actions using wearable motion sensor networks," *J. Ambient Intell. Smart Environ.*, vol. 1, no. 2, pp. 103–115, Jan. 2009.

[27] 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. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 168–172.

[28] N. Dawar and N. Kehtarnavaz, "A convolutional neural network-based sensor fusion system for monitoring transition movements in healthcare applications," in *Proc. IEEE Int. Conf. Control Autom. (ICCA)*, Jun. 2018, pp. 482–485.

[29] C. Chen, M. Liu, H. Liu, B. Zhang, J. Han, and N. Kehtarnavaz, "Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition," *IEEE Access*, vol. 5, pp. 22590–22604, 2017.

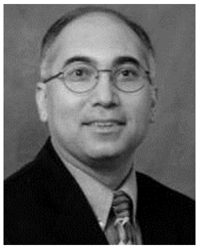
[30] G. Zhu, L. Zhang, P. Shen, and J. Song, "An online continuous human action recognition algorithm based on the Kinect sensor," *Sensors*, vol. 16, no. 2, p. 161, Jan. 2016.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, and H. Zhang, "Discriminative key pose extraction using extended LC-KSVD for action recognition," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2014, pp. 1–8.

[33] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. Eur. Conf. Inf. Retr.*, Mar. 2005, pp. 345–359.

Neha Dawar (S'15) received the B.Tech. degree in communication and computer engineering from The LNM Institute of Information Technology, Jaipur, India, in 2011, and the M.S. degree in electrical engineering from the University of Calgary, Canada, in 2014. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Texas at Dallas, Richardson, TX, USA. Her research interests include signal and image processing, computer vision, and machine learning.



Nasser Kehtarnavaz (S'82–M'86–SM'92–F'12) is currently an Erik Jonsson Distinguished Professor with the Department of Electrical and Computer Engineering and the Director of the Signal and Image Processing Laboratory, University of Texas at Dallas, Richardson, TX, USA. His research interests include signal and image processing, machine learning, and real-time implementation on embedded processors. He has authored or co-authored 10 books and more than 370 journal papers, conference papers, patents, manuals, and editorials in these areas. He is a Fellow of SPIE, a licensed Professional Engineer, and an Editor-in-Chief of the *Journal of Real-Time Image Processing*.