

Article

# Fusion of Video and Inertial Sensing for Deep Learning–Based Human Action Recognition

Haoran Wei <sup>1,\*</sup> , Roozbeh Jafari <sup>2</sup> and Nasser Kehtarnavaz <sup>1</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, University of Texas at Dallas, Richardson, TX 75080, USA

<sup>2</sup> Center for Remote Health Technologies and System, Texas A&M University, College Station, TX 77843, USA

\* Correspondence: Haoran.Wei@utdallas.edu; Tel.: +01-972-883-6838

Received: 21 June 2019; Accepted: 22 August 2019; Published: 24 August 2019



**Abstract:** This paper presents the simultaneous utilization of video images and inertial signals that are captured at the same time via a video camera and a wearable inertial sensor within a fusion framework in order to achieve a more robust human action recognition compared to the situations when each sensing modality is used individually. The data captured by these sensors are turned into 3D video images and 2D inertial images that are then fed as inputs into a 3D convolutional neural network and a 2D convolutional neural network, respectively, for recognizing actions. Two types of fusion are considered—Decision-level fusion and feature-level fusion. Experiments are conducted using the publicly available dataset UTD-MHAD in which simultaneous video images and inertial signals are captured for a total of 27 actions. The results obtained indicate that both the decision-level and feature-level fusion approaches generate higher recognition accuracies compared to the approaches when each sensing modality is used individually. The highest accuracy of 95.6% is obtained for the decision-level fusion approach.

**Keywords:** fusion of video and inertial sensing for action recognition; deep learning-based action recognition; decision-level and feature-level fusion for action recognition

## 1. Introduction

Human action recognition has been extensively studied in the literature and has already been incorporated into commercial products. There are a wide range of applications of human action recognition including human–machine interface, e.g., [1–4], intelligent surveillance, e.g., [5–8], content-based data retrieval, e.g., [9,10], and gaming, e.g., [11,12].

Vision and inertial sensing modalities have been used individually to achieve human action recognition, e.g., [13–23]. Furthermore, the use of deep learning models or deep neural networks have proven to be more effective than conventional approaches for human action recognition. For example, in [15], it was shown that deep learning networks utilizing video images performed better than the previous conventional approaches. More recently, in [16], a three-dimensional (3D) convolutional neural network (CNN) was used by considering video data volumes. Video cameras are cost-effective and widely available. On the other hand, they have limitations in terms of their limited field of view and sensitivity to lighting and illumination changes. Depth cameras have also been utilized for human action recognition, e.g., [2,11]. The use of these cameras has been limited to indoor environments as they rely on infrared light for obtaining depth images.

Furthermore, wearable inertial sensors have been used to achieve human action recognition. Similarly, these sensors are cost-effective and widely available. The main advantage of these sensors include their wearability—Thus, they are not limited to a specific field of view. Often, 3-axis acceleration signals from their accelerometers and 3-axis angular velocity signals from their gyroscopes are used

for conducting human action recognition, e.g., [17–23]. These sensors have also limitations in terms of not capturing a complete representation of actions. The use of multiple inertial sensors, though possible, introduces its own challenge in terms of intrusiveness to wear multiple sensors on the body for capturing a complete representation of actions.

Basically, no sensing modality is perfect: no sensing modality provides the entire information associated with various actions. In previous works on fusing two sensing modalities [24–30], it was shown that the fusion of the sensing modalities of depth camera and inertial sensor generated more robust human action recognition compared to when using each sensing modality individually. In this paper, the fusion of the sensing modalities of video camera and inertial sensor is examined to achieve a more robust human action recognition in terms of recognition errors compared to the situations when each sensing modality is used individually. Furthermore, this paper considers the use of deep learning models for this fusion. A 3D convolutional neural network is used for video data captured by a video camera and a 2D (two dimensional) convolutional neural network is used for inertial signals captured by a wearable inertial sensor. Both feature-level and decision-level fusion are examined. This is the first time a simultaneous utilization of video images and inertial signals are considered to achieve human action recognition. No other works are reported in the literature in which both video images and inertial signals are captured and processed at the same time to conduct human action recognition.

The rest of the paper is organized as follows: Section 2 covers the dataset used. The architectures of the deep learning models used are presented in Section 3. The experimental results are reported in Section 4, followed by their discussion in Section 5. Finally, the paper is concluded in Section 6.

## 2. UTD-MHAD Dataset

This section gives a brief description of the dataset utilized and the way the video images and inertial signals are fed into the deep learning models described in Section 3. The dataset used is a public domain dataset called the University of Texas at Dallas Multimodal Human Action Dataset (UTD-MHAD). Interested readers are referred to [31] for the details of this dataset.

The UTD-MHAD dataset was collected by a Kinect camera and a wearable inertial sensor at the same time. The dataset consists of video images with a resolution of  $640 \times 480$  and depth images with a resolution of  $320 \times 240$  at 30 frames per second captured by the Kinect camera. It also consists of 3-axis acceleration and 3-axis angular velocity signals captured at the same time at a sampling rate of 50Hz by the inertial sensor. The details of the inertial sensor are provided in [32]. 27 different actions are included in this dataset. As listed in Table 1, actions numbered 1 through 21 are hands type movements. For these actions, the inertial sensor was worn on the right wrist. Actions numbered 22 to 27 are leg type movements. For these actions, the inertial sensor was worn on the right thigh.

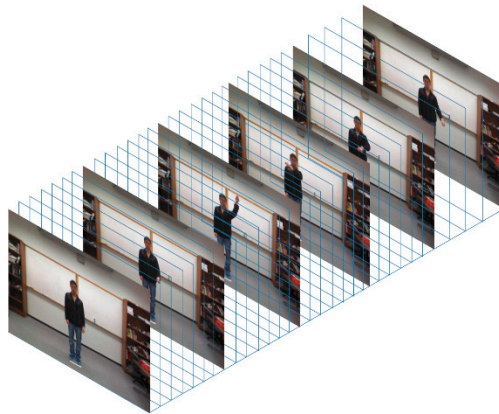
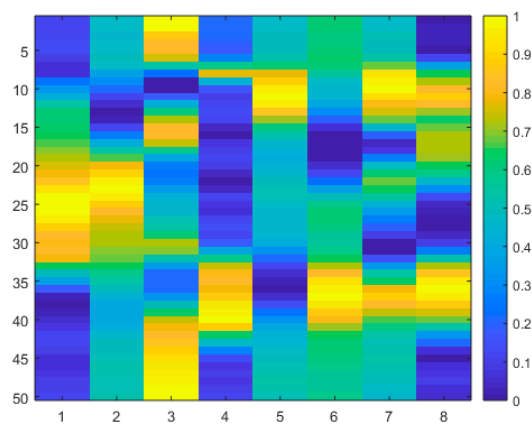
Each of the actions were performed by eight subjects (four females and four males), and each action was repeated four times. Hence, there are a total of 864 action clips, with three of the data corrupted action clips not included. Each action clip consists of videos, depth videos, skeleton joint positions, and inertial sensor signals. In this work, the videos and the inertial signals are used.

For the video data, since different actions have different video clip durations, 32 frames across each video clip are selected to form a consistent 3D input to the 3D convolutional neural network discussed in the next section. To gain computational efficiency, each frame is resized to  $320 \times 240$  pixels. As a result, each action clip has the same spatial and temporal size forming a  $320 \times 240 \times 32$  3D video data volume. An example 3D video data volume formed in this manner is shown in Figure 1.

For the inertial data, an eight-column input image is formed by combining the 3-axis acceleration signals, the 3-axis angular velocity signals, the overall acceleration signal, and the overall angular velocity signal at a sampling frequency of 50Hz. Thus, input images to the 2D convolutional neural network discussed in the next section are of size  $8 \times 50$ . An example inertial signals input image is shown in Figure 2.

**Table 1.** Actions in the University of Texas at Dallas Multimodal Human Action Dataset (UTD-MHAD) dataset.

Action Number	Hands Actions	Action Number	Legs Actions
1	right arm swipe to the left	22	jogging in place
2	right arm swipe to the right	23	walking in place
3	right hand wave	24	sit to stand
4	two hand front clap	25	stand to sit
5	right arm throw	26	forward lunge (left foot forward)
6	cross arms in the chest	27	squat (two arms stretch out)
7	basketball shoot		
8	right hand draw x		
9	right hand draw circle (clockwise)		
10	right hand draw circle (counter clockwise)		
11	draw triangle		
12	bowling (right hand)		
13	front boxing		
14	baseball swing from right		
15	tennis right hand forehand swing		
16	arm curl (two arms)		
17	tennis serve		
18	two hand push		
19	right hand knock on door		
20	right hand catch an object		
21	right hand pick up and throw		

**Figure 1.** Example video volume as input to the 3D convolution neural network.**Figure 2.** Example inertial signals image as input to the 2D convolution neural network.

### 3. Deep Learning Models

A 3D convolutional neural network is used to learn the video data for the 27 actions of the UTD-MHAD dataset. The architecture of this model is a simplified version of the model discussed in [16] and is shown on the left side of Figure 3. More specifically, the input is a  $320 \times 240 \times 32$  3D video volume. The first 3D convolutional layers have 16 filters, and their outputs are passed through 3D max pooling layers. The second 3D convolutional layers convolve the outputs of the pooling layers with 32 filters and pass them to another set of 3D max pooling layers. The third and fourth 3D convolutional layers have 64 and 128 filters, respectively, passing outputs to 3D max pooling layers. All of the 3D convolution filters are of size  $3 \times 3 \times 3$  with stride  $1 \times 1 \times 1$ . All the 3D pooling layers are of size  $2 \times 2 \times 2$  with stride  $2 \times 2 \times 2$ . Batch normalization layers and Rectified Linear Unit (ReLU) activation are used at each 3D convolutional layer. The output of the last 3D pooling layer is flattened and passed to a fully connected layer with 256 units based on the ReLU activation. The output of the fully connected layer is passed to a 50% dropout layer and then connected to a final fully connected layer using the softmax cost function, producing scores for the output classes denoting the 27 actions. The stochastic gradient descent with momentum (SGDM) optimization algorithm is used to train the networks. Table 2 provides the architecture and training parameters associated with the 3D convolutional neural network used.

For the inertial data, a 2D CNN is used. As inertial sensor signals are time-series signals, the 2D CNN model developed in [33] for speech processing is used here. This model is shown on the right side of Figure 3. It consists of 3 convolutional layers with 16, 32, and 64 filters, respectively, all having a size of  $3 \times 3$ . The 2D max pooling layers merge the filtering operations of the convolutional layers. The max pooling size used here is  $2 \times 2$ . At the last stage, there are two fully connected layers with 256 and 27 units, respectively. The scores of each class or action are then computed from the output layer using the softmax cost function. The ReLU activation is used at each convolutional layer and the first fully connected layer. The decision-level fusion is performed by multiplying the scores of the two sensing modalities and the class or action with the highest score is taken to be the recognized action. Table 3 provides the architecture and training parameters associated with the 2D convolutional neural network used.

A brief overview of the layers is described next. Interested readers are referred to many references that are available on deep neural networks, for example [34–36]. Three-dimensional convolutions are similar to 2D convolutions, but instead of using 2D filters, 3D filters are used. The 3D max pooling layers perform the down sampling operation by dividing 3D inputs to cuboidal pooling regions, and then by computing the maximum value of each such region. The ReLU activation function is a piecewise linear function, whose output is the same as the input when the input is positive and zero otherwise. Compared to other activation functions, ReLU exhibits more sensitivity to the input sum activation and avoids saturation and has become the default activation function in many computer vision and speech recognition applications. To speed up the training process and reduce sensitivity to network initialization, batch normalization layers are used between convolution layers and the ReLU layers. The batch normalization layers normalize each input layer across a batch by subtracting the batch mean and dividing by the batch standard deviation. Outputs of batch normalization layers maintain a mean value close to 0 and a standard deviation close to 1, allowing independency of the layers. Dropout layers reduce overfitting and improve the generalization capability of the networks by randomly dropping out nodes during training. The softmax cost function is used to transform a vector of input numbers to a probability distribution. After applying softmax, each output component lies in the interval between 0 and 1 with the sum adding up to 1.

Besides the above decision-level fusion approach of combining the decisions of two deep neural networks, one for video and one for inertial sensing, a feature-level fusion approach is also considered in this work. 256 features are extracted from the video network model and the inertial network model dropout layers as indicated in Figure 4. Then, these features are concatenated to form a 512-dimensional vector. This vector is used as the input to a backpropagation neural network with three fully connected

layers based on the ReLU activation function and the softmax cost function. The final output of the last layer is used as the score of the classes corresponding to the 27 actions. The class or action with the highest score is taken to be the recognized action.

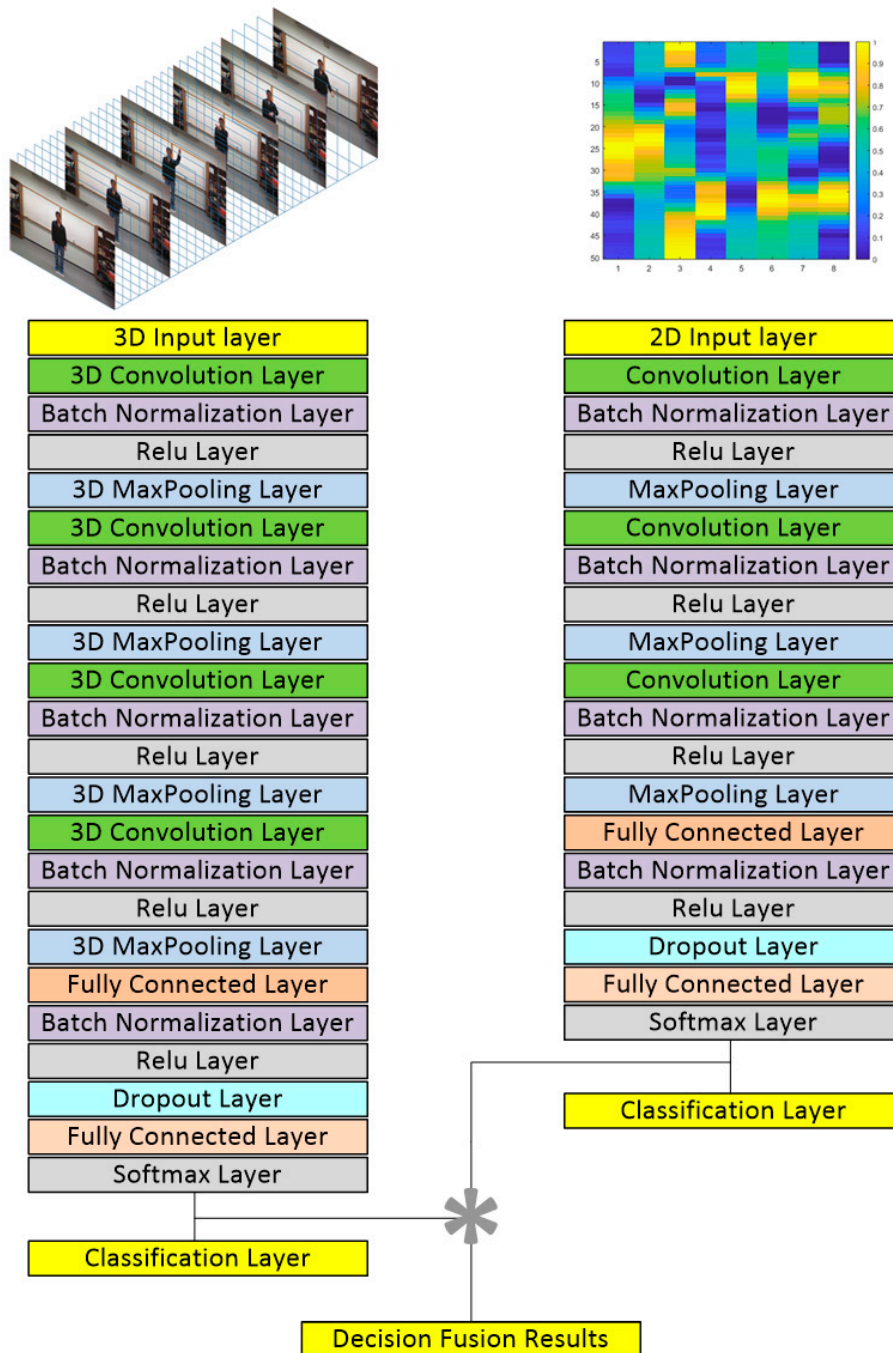


Figure 3. Network architecture used for the decision-level fusion of video and inertial sensing modalities.

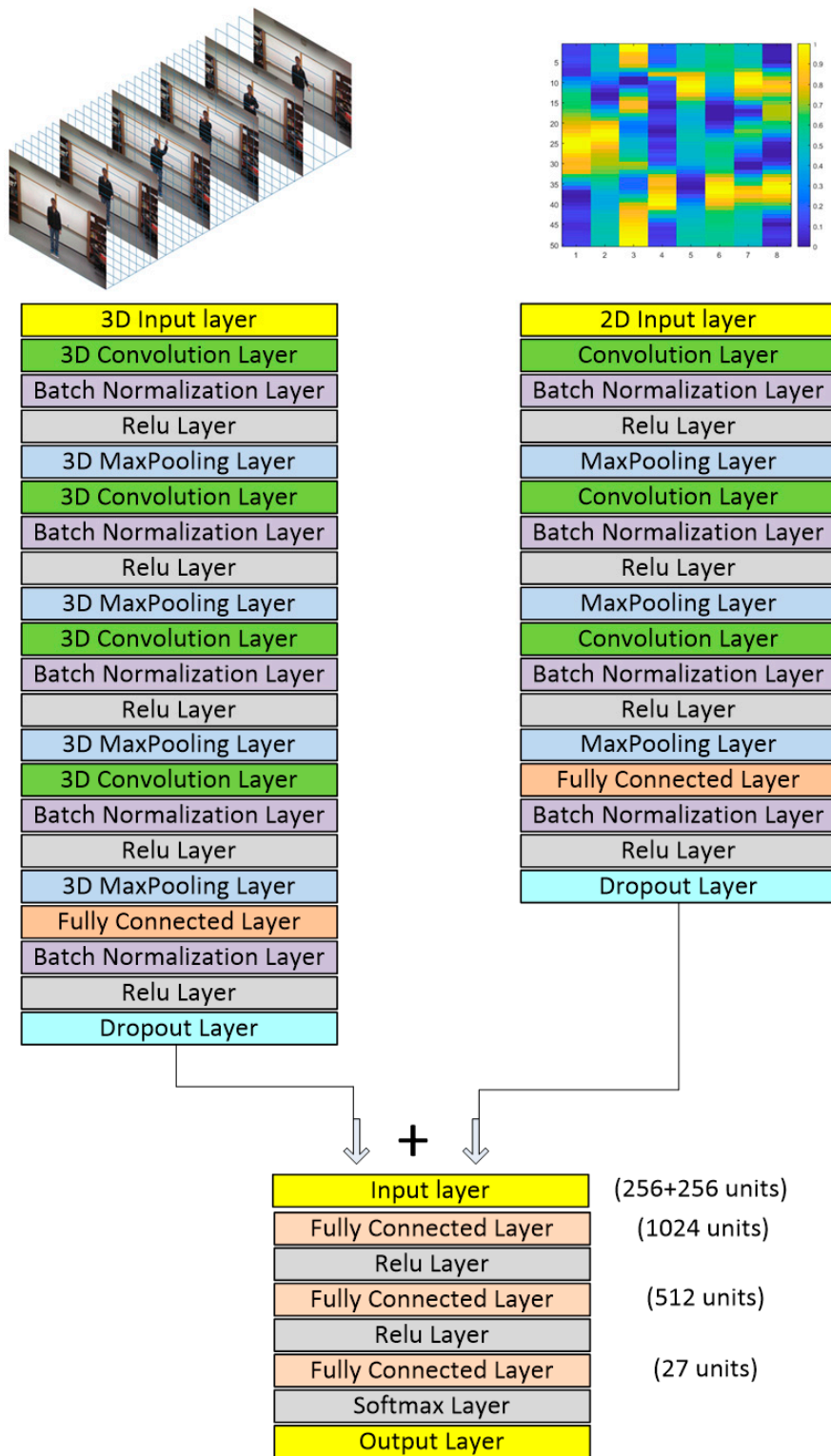


Figure 4. Network architecture used for the feature-level fusion of video and inertial sensing modalities.

**Table 2.** Architecture of the 3D convolutional neural network.

Layers and Training Parameters	Values
Input layer	$320 \times 240 \times 32$
1st 3D convolutional layer	16 filters, filter size $3 \times 3 \times 3$ , stride $1 \times 1 \times 1$
2nd 3D convolutional layer	32 filters, filter size $3 \times 3 \times 3$ , stride $1 \times 1 \times 1$
3rd 3D convolutional layer	64 filters, filter size $3 \times 3 \times 3$ , stride $1 \times 1 \times 1$
4th 3D convolutional layer	128 filters, filter size $3 \times 3 \times 3$ , stride $1 \times 1 \times 1$
All 3D max pooling layer	pooling size $2 \times 2 \times 2$ , stride $2 \times 2 \times 2$
1st fully connected layer	256 units
2nd fully connected layer	27 units
dropout layer	50%
Initial learn rate	0.0016
Learning rate drop factor	0.5
Learn rate drop period	4
Max epochs	20

**Table 3.** Architecture and training parameters of the 2D convolutional neural network.

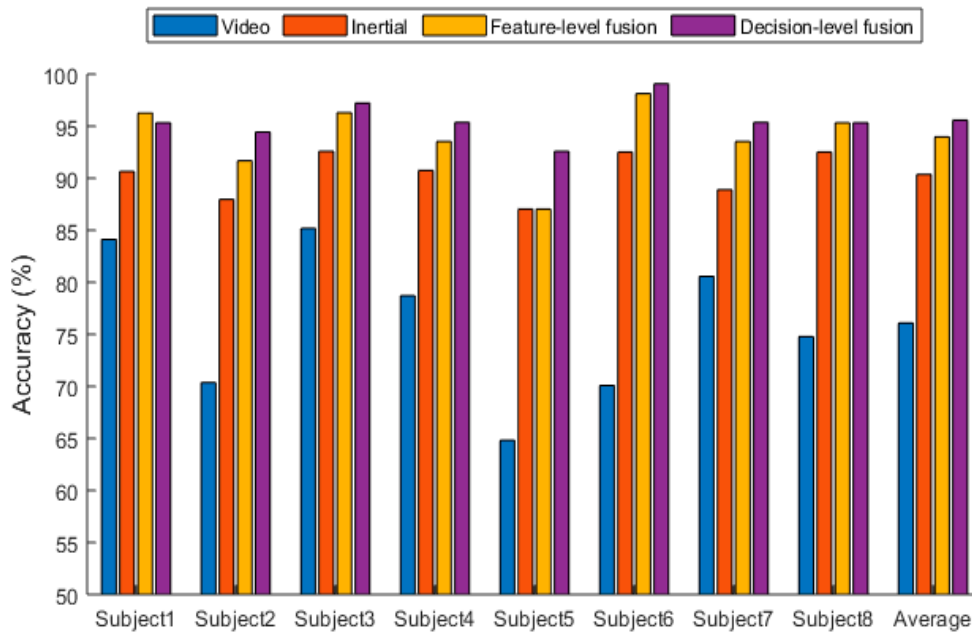
Layers and Training Parameters	Values
Input layer	$8 \times 50$
1st 2D convolutional layer	16 filters, filter size $3 \times 3$ , stride $1 \times 1$
2nd 2D convolutional layer	32 filters, filter size $3 \times 3$ , stride $1 \times 1$
3rd 2D convolutional layer	64 filters, filter size $3 \times 3$ , stride $1 \times 1$
All 2D max pooling layer	pooling size $2 \times 2$ , stride $2 \times 2$
1st fully connected layer	256 units
2nd fully connected layer	27 units
dropout layer	50%
Initial learn rate	0.0016
Learning rate drop factor	0.5
Learn rate drop period	4
Max epochs	20

#### 4. Experimental Results

This section presents the results of the experiments conducted to examine the effectiveness of the introduced fusion approach when the two sensing modalities of video images and inertial signals are used simultaneously. A leave-one-out cross validation was performed, meaning that the data from one subject was used for testing, and the data from the remaining seven subjects were used for training. This process was repeated for each of the subjects. Then, an average was taken across the subjects. The average accuracy of the following four approaches were obtained: (1) only using video sensing modality, (2) only using inertial sensing modality, (3) feature-level fusion of video and inertial sensing modalities, and (4) decision-level fusion of video and inertial sensing modalities. Table 4 provides the average recognition accuracy obtained for these four situations on the UTD-MHAD dataset, and Figure 5 illustrates the recognition performance for each of the eight subjects in the UTD-MHAD dataset.

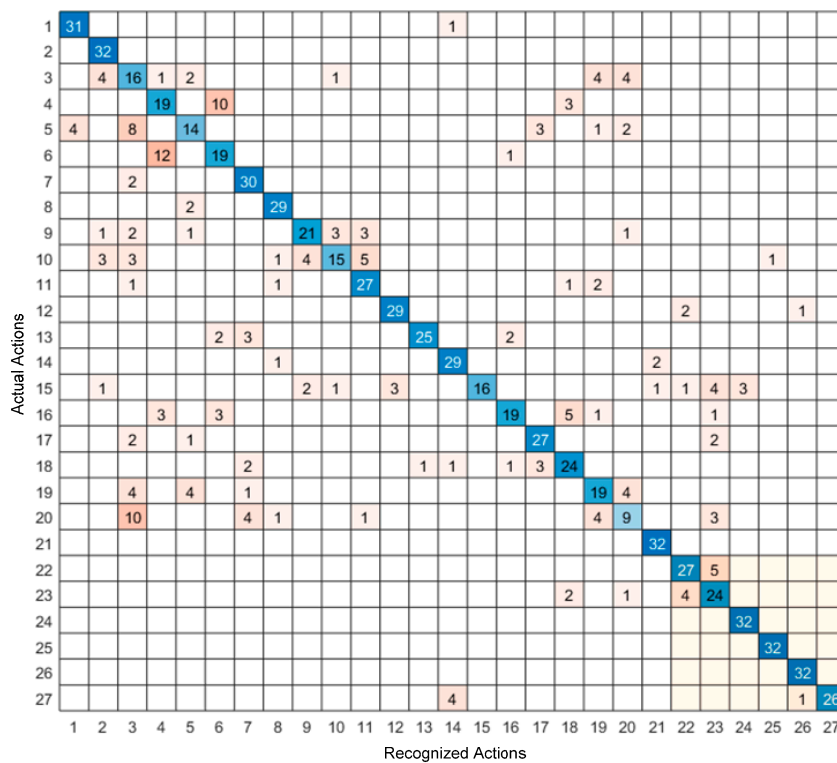
**Table 4.** Average accuracy of video sensing modality only, inertial sensing modality only, feature-level fusion of video and inertial sensing modalities, and decision-level fusion of video and inertial sensing modalities.

Approaches	Average Accuracy (%)
Video only	76.0
Inertial only	90.3
Feature-level fusion of video and inertial	94.1
Decision-level fusion of video and inertial	95.6



**Figure 5.** Recognition accuracy for the four situations of sensing modality across the eight subjects in UTD-MHAD dataset.

The confusion matrix of the video sensing modality only and the inertial sensing modality only are shown in Figures 6 and 7, respectively, and the confusion matrix of the feature-level fusion modality and the decision-level fusion modality are shown in Figures 8 and 9, respectively. The bottom gray parts of these confusion matrices correspond to leg actions (actions 22 through 27).



**Figure 6.** Confusion matrix of the video only sensing modality.



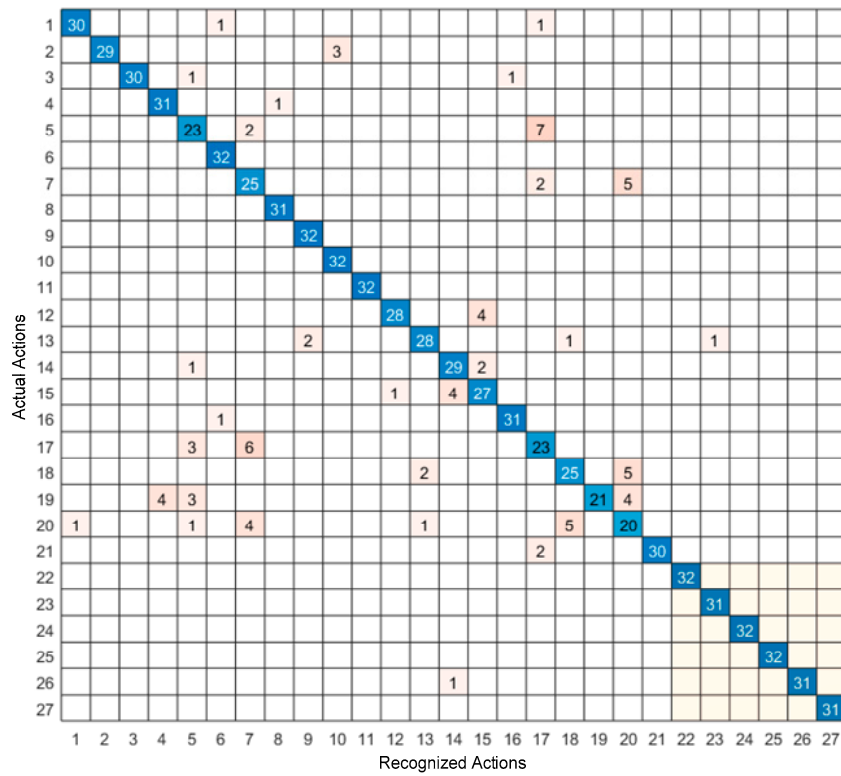


Figure 7. Confusion matrix of the inertial only sensing modality.

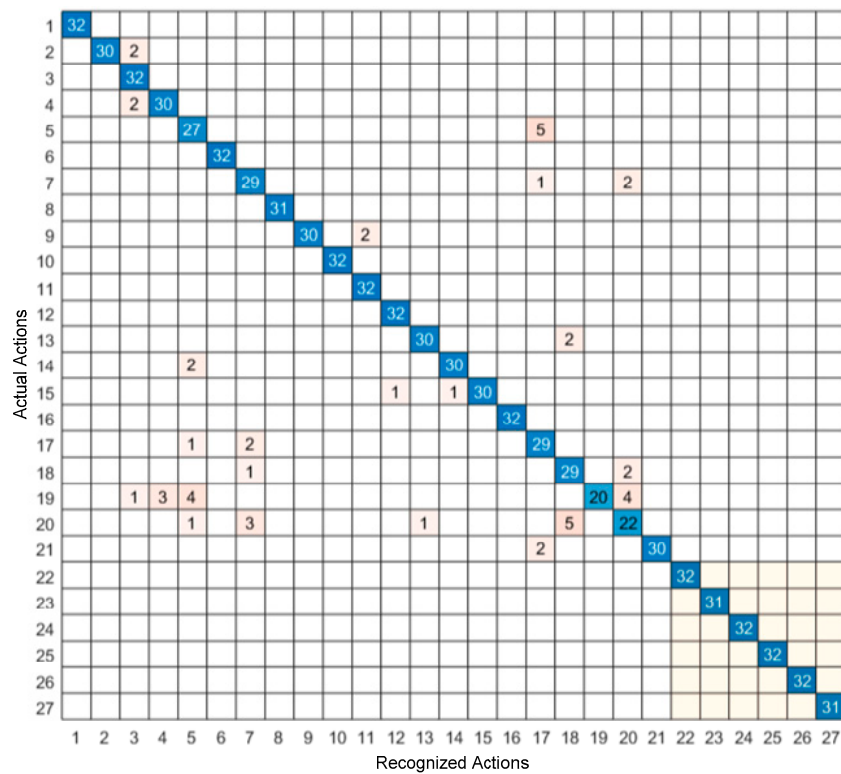


Figure 8. Confusion matrix of the feature-level fusion of video and inertial sensing modalities.

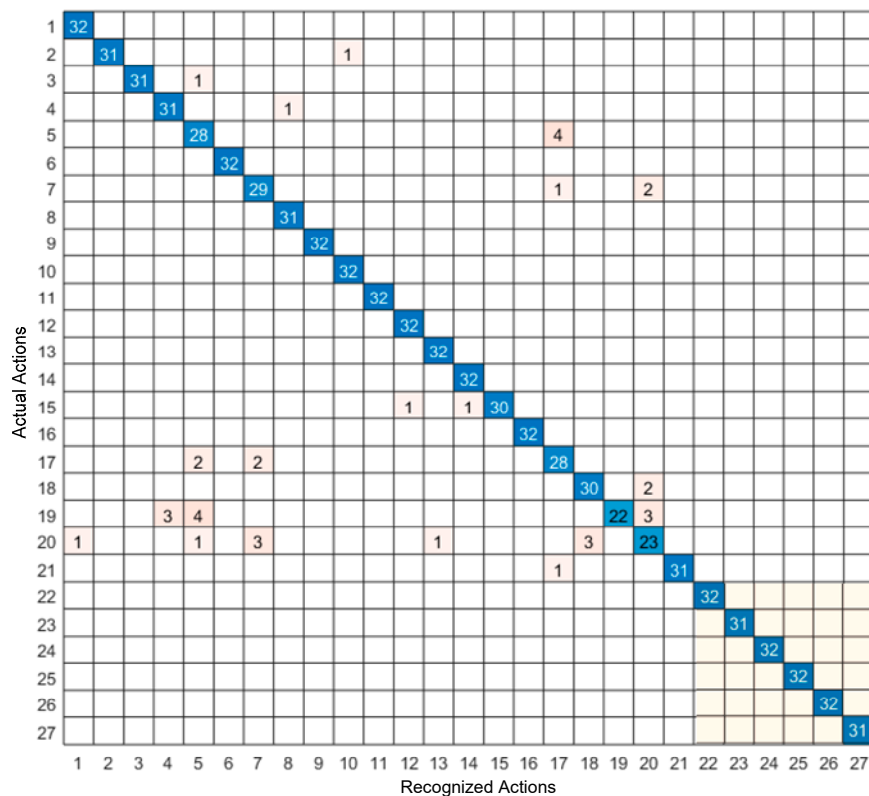


Figure 9. Confusion matrix of the decision-level fusion of video and inertial sensing modalities.

## 5. Discussion of Results

From Table 4, it can be seen that both the feature-level fusion and decision-level fusion generated higher accuracies than the individual sensing modalities. In addition, the decision-level fusion generated fewer errors compared to the feature-level fusion. More specifically, as can be seen from Figure 6, most of the errors occurred between actions 4 and 6, between actions 3 and 20, between actions 3 and 5, and between actions 22 and 23. These errors were caused due to the similarities of the volume data between these actions. For example, the images associated with action 4 corresponding to two hand clap, and action 6 corresponding to crossing arms appeared close. From Figure 7, it can be seen that most of the errors occurred between actions 18 and 20, between actions 5 and 17, between actions 7 and 20, and between actions 7 and 17. These errors were caused by the similarities of the inertial signals between these actions. For example, the inertial signals associated with action 7 corresponding to shooting a basketball, action 17 corresponding to a tennis serve, and action 20 corresponding to a right-hand catch appeared close. As can be seen from Figures 8 and 9, in both the feature-level fusion and decision-level fusion, most of the errors occurred between actions 18 and 20 and actions 5 and 17, with fewer number of errors compared to the single sensing modalities, indicating the positive impact made by fusing the individual sensing modalities. From the experiments conducted, both the feature-level fusion and the decision-level fusion exhibited more robustness to errors compared to individual sensing modalities.

## 6. Conclusions

In this paper, for the first time, the simultaneous utilization of video and inertial sensing modalities were considered within a fusion framework to achieve human action recognition based on deep learning models. The following four approaches were compared in terms of recognition accuracies: using only video data as input to a deep neural network, using only inertial data as input to a deep neural network, using both video data and inertial data as inputs to two deep neural networks within

a decision-level fusion framework, and using both video data and inertial data as inputs to two deep neural networks within a feature-level fusion framework. The experiments conducted based on the publicly available UTD-MHAD dataset have shown that the decision-level fusion approach provided the highest recognition accuracy of 95.6%—The fusion of the two sensing modalities exhibited more robust recognition outcome in terms of misclassification errors compared to the situations when each sensing modality was used individually.

**Author Contributions:** The first and third authors contributed equally to this work. The second author participated in the discussions and provided partial internal funding for this work.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Presti, L.L.; Cascia, M.L. 3D Skeleton-based Human Action Classification: A Survey. *Pattern Recognit.* **2016**, *53*, 130–147. [[CrossRef](#)]
2. Dawar, N.; Kehtarnavaz, N. Continuous detection and recognition of actions of interest among actions of non-interest using a depth camera. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 4227–4231.
3. Eum, H.; Yoon, C.; Lee, H.; Park, M. Continuous human action recognition using depth-MHI-HOG and a spotter model. *Sensors* **2015**, *15*, 5197–5227. [[CrossRef](#)] [[PubMed](#)]
4. Chu, X.; Ouyang, W.; Li, H.; Wang, X. Structured feature learning for pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4715–4723.
5. Chaaoui, A.A.; Padilla-Lopez, J.R.; Ferrandez-Pastor, F.J.; Nieto-Hidalgo, M.; Florez-Revuelta, F. A vision-based system for intelligent monitoring: Human behaviour analysis and privacy by context. *Sensors* **2014**, *14*, 8895–8925. [[CrossRef](#)] [[PubMed](#)]
6. Wei, H.; Laszewski, M.; Kehtarnavaz, N. Deep Learning-Based Person Detection and Classification for Far Field Video Surveillance. In Proceedings of the 13th IEEE Dallas Circuits and Systems Conference, Dallas, TX, USA, 2–12 November 2018; pp. 1–4.
7. Ziaefard, M.; Bergevin, R. Semantic human activity recognition: A literature review. *Pattern Recognit.* **2015**, *48*, 2329–2345. [[CrossRef](#)]
8. Wei, H.; Kehtarnavaz, N. Semi-Supervised Faster RCNN-Based Person Detection and Load Classification for Far Field Video Surveillance. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 756–767. [[CrossRef](#)]
9. Van Gemert, J.C.; Jain, M.; Gati, E.; Snoek, C.G. APT: Action localization proposals from dense trajectories. In Proceedings of the British Machine Vision Conference 2015: BMVC 2015, Swansea, UK, 7–10 September 2015; p. 4.
10. Zhu, H.; Vial, R.; Lu, S. Tornado: A spatio-temporal convolutional regression network for video action proposal. In Proceedings of the CVPR, Venice, Italy, 22–29 October 2017; pp. 5813–5821.
11. Bloom, V.; Makris, D.; Argyriou, V. G3D: A gaming action dataset and real time action recognition evaluation framework. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 7–12.
12. Wang, Y.; Yu, T.; Shi, L.; Li, Z. Using human body gestures as inputs for gaming via depth analysis. In Proceedings of the IEEE International Conference on Multimedia and Expo, Hannover, Germany, 26 April–23 June 2008; pp. 993–996.
13. Wang, L.; Zang, J.; Zhang, Q.; Niu, Z.; Hua, G.; Zheng, N. Action Recognition by an Attention-Aware Temporal Weighted Convolutional Neural Network. *Sensors* **2018**, *7*, 1979. [[CrossRef](#)] [[PubMed](#)]
14. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 3–6 December 2013; pp. 3551–3558.
15. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 8–13 December 2014; pp. 568–576.

16. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 4489–4497.
17. Avilés-Cruz, C.; Ferreyra-Ramírez, A.; Zúñiga-López, A.; Villegas-Cortéz, J. Coarse-Fine Convolutional Deep-Learning Strategy for Human Activity Recognition. *Sensors* **2019**, *19*, 1556. [[CrossRef](#)] [[PubMed](#)]
18. Chen, C.; Kehtarnavaz, N.; Jafari, R. A medication adherence monitoring system for pill bottles based on a wearable inertial sensor. In Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 4983–4986.
19. Yang, A.Y.; Jafari, R.; Sastry, S.S.; Bajcsy, R. Distributed recognition of human actions using wearable motion sensor networks. *J. Ambient Intell. Smart Environ.* **2009**, *1*, 103–115.
20. Nathan, V.; Paul, S.; Prioleau, T.; Niu, L.; Mortazavi, B.J.; Cambone, S.A.; Veeraghavan, A.; Sabharwal, A.; Jafari, R. A Survey on Smart Homes for Aging in Place: Toward Solutions to the Specific Needs of the Elderly. *IEEE Signal Process. Mag.* **2018**, *35*, 111–119. [[CrossRef](#)]
21. Wu, J.; Jafari, R. Orientation independent activity/gesture recognition using wearable motion sensors. *IEEE Internet Things J.* **2018**, *6*, 1427–1437. [[CrossRef](#)]
22. Liu, J.; Wang, Z.; Zhong, L.; Wickramasuriya, J.; Vasudevan, V. uWave: Accelerometer-Based Personalized Gesture Recognition and Its Applications. In Proceedings of the Seventh Annual IEEE International Conference on Pervasive Computing and Communications (PerCom 2009), Galveston, TX, USA, 9–13 March 2009.
23. Alves, J.; Silva, J.; Grifo, E.; Resende, C.; Sousa, I. Wearable Embedded Intelligence for Detection of Falls Independently of on-Body Location. *Sensors* **2019**, *19*, 2426. [[CrossRef](#)] [[PubMed](#)]
24. Chen, C.; Jafari, R.; Kehtarnavaz, N. Improving human action recognition using fusion of depth camera and inertial sensors. *IEEE Trans. Hum. Mach. Syst.* **2015**, *45*, 51–61. [[CrossRef](#)]
25. Chen, C.; Jafari, R.; Kehtarnavaz, N. A real-time human action recognition system using depth and inertial sensor fusion. *IEEE Sens. J.* **2016**, *16*, 773–781. [[CrossRef](#)]
26. Chen, C.; Jafari, R.; Kehtarnavaz, N. A survey of depth and inertial sensor fusion for human action recognition. *Multimed. Tools Appl.* **2017**, *76*, 4405–4425. [[CrossRef](#)]
27. Dawar, N.; Kehtarnavaz, N. A convolutional neural network-based sensor fusion system for monitoring transition movements in healthcare applications. In Proceedings of the IEEE 14th International Conference on Control and Automation, Anchorage, AK, USA, 12–15 June 2018; pp. 482–485.
28. Dawar, N.; Kehtarnavaz, N. Action detection and recognition in continuous action streams by deep learning-based sensing fusion. *IEEE Sens. J.* **2018**, *18*, 9660–9668. [[CrossRef](#)]
29. Rwigema, J.; Choi, H.R.; Kim, T. A Differential Evolution Approach to Optimize Weights of Dynamic Time Warping for Multi-Sensor Based Gesture Recognition. *Sensors* **2019**, *19*, 1007. [[CrossRef](#)] [[PubMed](#)]
30. Dawar, N.; Kehtarnavaz, N. Real-time continuous detection and recognition of subject-specific smart tv gestures via fusion of depth and inertial sensing. *IEEE Access* **2018**, *6*, 7019–7028. [[CrossRef](#)]
31. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Proceedings of the 2015 IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015; pp. 168–172.
32. Chen, C.; Liu, K.; Jafari, R.; Kehtarnavaz, N. Home-based senior fitness test measurement system using collaborative inertial and depth sensors. In Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 4135–4138.
33. Wei, H.; Kehtarnavaz, N. Determining Number of Speakers from Single Microphone Speech Signals by Multi-Label Convolutional Neural Network. In Proceedings of the 44th Annual Conference of the IEEE Industrial Electronics Society, Washington, DC, USA, 21–23 October 2018; pp. 2706–2710.
34. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [[CrossRef](#)] [[PubMed](#)]
35. Tao, F.; Busso, C. Aligning audiovisual features for audiovisual speech recognition. In Proceedings of the IEEE International Conference on Multimedia and Expo, San Diego, CA, USA, 23–27 July 2018; pp. 1–6.

36. Wang, Z.; Kong, Z.; Changra, S.; Tao, H.; Khan, L. Robust High Dimensional Stream Classification with Novel Class Detection. In Proceedings of the IEEE 35th International Conference on Data Engineering, Macao, China, 8–11 April 2019; pp. 1418–1429.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).