# Metadata of the chapter that will be visualized in SpringerLink

| | |
|---|---|
| Book Title | Image Analysis and Recognition |
| Series Title | |
| Chapter Title | Human Action Recognition Using Fusion of Depth and Inertial Sensors |
| Copyright Year | 2018 |
| Copyright HolderName | Springer International Publishing AG, part of Springer Nature |

| | | |
|---|---|---|
| Author | Family Name | **Fuad** |
| | Particle | |
| | Given Name | **Zain** |
| | Prefix | |
| | Suffix | |
| | Role | |
| | Division | Faculty of Engineering and Natural Sciences |
| | Organization | Sabanci University |
| | Address | Istanbul, Turkey |
| | Email | zainfuad@sabanciuniv.edu |
| Corresponding Author | Family Name | **Unel** |
| | Particle | |
| | Given Name | **Mustafa** |
| | Prefix | |
| | Suffix | |
| | Role | |
| | Division | Faculty of Engineering and Natural Sciences |
| | Organization | Sabanci University |
| | Address | Istanbul, Turkey |
| | Email | munel@sabanciuniv.edu |

| | |
|---|---|
| Abstract | In this paper we present a human action recognition system that utilizes the fusion of depth and inertial sensor measurements. Robust depth and inertial signal features, that are subject-invariant, are used to train independent Neural Networks, and later decision level fusion is employed using a probabilistic framework in the form of Logarithmic Opinion Pool. The system is evaluated using UTD-Multimodal Human Action Dataset, and we achieve 95% accuracy in 8-fold cross-validation, which is not only higher than using each sensor separately, but is also better than the best accuracy obtained on the mentioned dataset by 3.5%. |

| | |
|---|---|
| Keywords<br>(separated by '-') | Human action recognition - Sensor fusion - Depth camera - Inertial sensor - Neural Network - Logarithmic opinion pool |

# Human Action Recognition Using Fusion of Depth and Inertial Sensors

Zain Fuad and Mustafa Unel(✉)

Faculty of Engineering and Natural Sciences, Sabanci University,
Istanbul, Turkey
{zainfuad,munel}@sabanciuniv.edu

AQ1

**Abstract.** In this paper we present a human action recognition system that utilizes the fusion of depth and inertial sensor measurements. Robust depth and inertial signal features, that are subject-invariant, are used to train independent Neural Networks, and later decision level fusion is employed using a probabilistic framework in the form of Logarithmic Opinion Pool. The system is evaluated using UTD-Multimodal Human Action Dataset, and we achieve 95% accuracy in 8-fold cross-validation, which is not only higher than using each sensor separately, but is also better than the best accuracy obtained on the mentioned dataset by 3.5%.

**Keywords:** Human action recognition · Sensor fusion
Depth camera · Inertial sensor · Neural Network
Logarithmic opinion pool

## 1 Introduction

Human action recognition consists of acquiring a person's gestures through various sensors, combining those gestures to form an action, and understanding those actions. It has found applications in domains such as surveillance, robotics, telemedicine, internet of things and human-machine interaction [1], and has extended to unorthodox areas, such as recognition of food preparation activities [2].

Recent advances such as the introduction of Microsoft Kinect and ASUS Xtion Pro Live, low-cost RGB-D cameras that acquire depth information in addition to RGB videos, have aided the capture of human motion, in contrast to the expensive detector based MoCap systems, or computationally-expensive 3-D reconstruction using stereo cameras [1]. RGB-D videos preserve discriminative information, such as shape and distance variations [3], and have reduced processing times as compared to traditional RGB cameras [4]. Thus, they have enabled researchers to use them in an action recognition structure. Han et al. [5] highlighted the utilization of Kinect for vision based algorithms, while Aggarwal et al. [6] discussed different approaches for feature extraction from 3D data and mentioned methodologies employed in the context of human activity recognition. Notable work in this area includes the proposition of a Hierarchical Recurrent

A. Campilho et al. (Eds.): ICIAR 2018, LNCS 10882, pp. 1–8, 2018.
https://doi.org/10.1007/978-3-319-93000-8_42

Neural Network framework which uses skeletal positions obtained from depth cameras, and understands the performed actions [7]. Similarly, Nie et al. [8] decomposed human actions into poses, and further decomposed these poses to mid-level spatio-temporal parts and used dynamic programming for classification purposes.

Adding to this, low-cost, small and light-weight, wearable inertial sensors have made possible the use of these sensors for human-activity recognition, as they provide very little hindrance to the person performing these actions and can be used in real life scenarios [9]. Qaiser et al. [10] studied the classification of arm action in cricket using inertial sensors, while Ermes et al. [11] analyzed the use of inertial sensors in the detection of sports activities in controlled and natural environments. In the latter, a hybrid classifier was used, which was composed of a tree structure possessing a priori knowledge and artificial Neural Networks, and 3 reference classifiers.

Each of the sensors has their own advantages and short-comings, and a fusion of these sensors results in a higher action recognition performance [1]. This fusion can occur at the data-level, feature-level or decision-level and the literature suggests different approaches in this regard. For action recognition, Ofli et al. [12] used HOG and HOF features in a Bag-of-Features framework from the depth camera, and variance of acceleration for each temporal window from the inertial sensor. Chen et al. [13] performed a decision level fusion of depth motion maps from the depth sensor and statistical features obtained based on the temporal segments from inertial sensor. On the other hand, Stein and McKenna [2] proposed the use of statistical features from both Kinect and inertial sensor to gather visual displacement components and representations of acceleration signals respectively, to recognize food preparation activities.

The notion of human action recognition using a combination of different sensors and sensor fusion is still a maturing research area. Complex algorithms although give a good accuracy of recognition, however, their slow performance or huge training times limit their uses in a lot of scenarios. On the other hand, algorithms which can be employed in a real time framework have their use limited to controlled environments.

In this paper, we propose a human action recognition system that utilizes joint locations from the depth camera and acceleration and gyro measurements from the inertial sensor. After extracting subject invariant features, we implement separate Neural Network classifiers for each sensor, and perform a decision level fusion on the outputs of these networks in a probabilistic manner. The proposed algorithm can be employed in real time and performs well under noisy measurements. Furthermore, we test our algorithm on UTD-Multimodal Human Action Dataset [14], a publicly available dataset, that mimics the real world scenario due to the variety of actions available that incorporate the movement of different joints. The fusion accuracy is higher than when using each sensor alone as well as better than highest accuracy achieved [13] on the mentioned dataset.

The organization of this paper is as follows: Sect. 2 presents the proposed algorithm which incorporates features from depth and inertial sensors. Section 3 presents the results of the proposed algorithm, and finally, Sect. 4 concludes the paper discussing the future aspects of this research.

# 2 Proposed Method

Intra-class variations are a common observation in a human action recognition framework, due to the variations in the body shape (size) of the person performing the action coupled with the speed variations in the performed action. A robust classification algorithm is needed in this regard, as noise due to jitters makes the classification problem more challenging.

In the proposed algorithm (Fig. 1), depth information is acquired from Microsoft Kinect and 20 joints are tracked in 3-D Cartesian coordinates. Furthermore, linear accelerations and angular velocities are obtained from a wearable inertial sensor, i.e. IMU, located on different parts of the body which consists of a 3-axis accelerometer and a 3-axis gyroscope. All the signals are made the same size with respect to the entries from a particular sensor, using bicubic interpolation, to reduce the temporal variations. After performing normalization of the skeletal joint positions to achieve user independence and extraction of mean and standard deviation of the inertial data, the data obtained from each sensor is classified separately using Neural Networks and finally, decision level fusion is performed on the output of the Neural Networks using Logarithmic Opinion pool (LOGP) [15]. In this work we assume the use of one inertial and one depth sensor, however, the implemented algorithm can be scaled up to utilize information from more sensors.

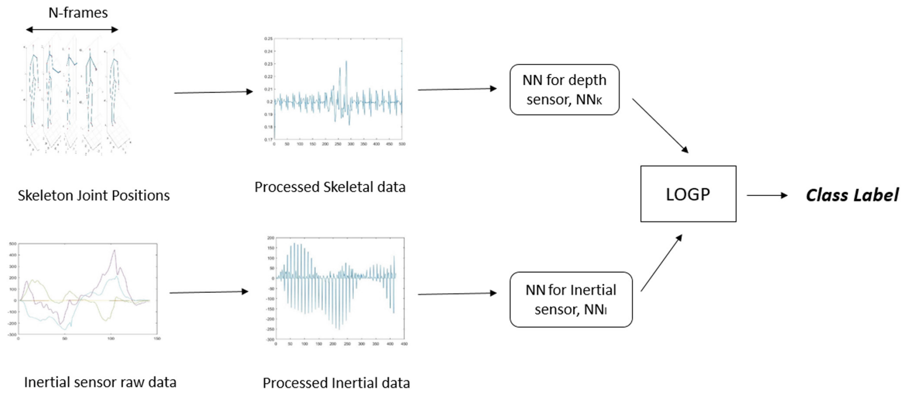

**Fig. 1.** Overview of the proposed method

## 2.1 Feature Extraction from Depth Sensor

The depth sensor outputs the 3D world coordinates of each tracked joint. Let $[x_{i,j}\ y_{i,j}\ z_{i,j}]$ be the spatial coordinates of each joint, where $i$ is the joint number and $j$ is the frame number. Then the output $I_K$ of the depth sensor can be represented as

$$I_K = \begin{bmatrix} x_{1,1} & y_{1,1} & z_{1,1} & x_{1,2} & y_{1,2} & z_{1,2} & \cdots & x_{1,N} & y_{1,N} & z_{1,N} \\ x_{2,1} & y_{2,1} & z_{2,1} & x_{2,2} & y_{2,2} & z_{2,2} & \cdots & x_{2,N} & y_{2,N} & z_{2,N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{M,1} & y_{M,1} & z_{M,1} & x_{M,2} & y_{M,2} & z_{M,2} & \cdots & x_{M,N} & y_{M,N} & z_{M,N} \end{bmatrix} \quad (1)$$

where each row of $I_k$ is the 3D spatial coordinates of each joint and $N$ is the total number of frames. Since the number of joints tracked by the sensor is 20, $M = 20$.

Due to the variations in speed in performing actions, the total number of frames for each action may differ. To eliminate this, the proposed algorithm utilizes bicubic interpolation. After the interpolation operation, the number of columns in $I_K$ reduces to

$$\hat{N} = \lambda N_{min} \quad (2)$$

where $N_{min}$, a data dependent parameter, is the least number of frames corresponding to the entry in the training dataset, and $\lambda$ is a scaling constant that helps in dimensionality reduction. In our implementation, we set $\lambda$ as 0.6.

Each row of $I_K$ is divided by its norm to produce a unit vector of length one, which not only gets rid of dependence on any specific person performing the task, however, it also makes sure that the individual joint movements does not affect other joints.

The rows of the reduced matrix are stacked column-wise to produce a $20\hat{N} \times 1$ input vector to the neural network, labeled as $\text{NN}_K$ in Fig. 1. However, there is noise present in the form of spikes and for that Savitzky-Golay [16] filter is applied to reduce the spikes.

### 2.2 Feature Extraction from Inertial Sensor

A wearable inertial sensor can be placed at any part of the body, and outputs 3-axis acceleration and gyro measurements. The output of the inertial sensor for each frame is $[a_x \; a_y \; a_z \; \omega_x \; \omega_y \; \omega_z]$, where $a_i$ represent linear acceleration, $\omega_i$ is the angular velocity and $i$ depicts the respective axis. Then the data obtained from the inertial sensor can be described as

$$I_I = \begin{bmatrix} a_{x,1} & a_{y,1} & a_{z,1} & \omega_{x,1} & \omega_{y,1} & \omega_{z,1} \\ a_{x,2} & a_{y,2} & a_{z,2} & \omega_{x,2} & \omega_{y,2} & \omega_{z,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{x,N} & a_{y,N} & a_{z,N} & \omega_{x,N} & \omega_{y,N} & \omega_{z,N} \end{bmatrix} \quad (3)$$

However, if there is more than one inertial sensors utilized, the structure of $I_I$ can be changed to incorporate them in a similar manner as skeleton joints in $I_K$.

As with the skeleton data, the inertial sensor data has different signal sizes. To reduce this variation, all the signals are resized using bicubic interpolation.

The size of $I_I$ is reduced to $N_{min} \times 6$, where $N_{min}$ is chosen from the inertial sensor training data in the same manner as in the case of the depth sensor. Furthermore, the inertial sensor measurements are partitioned into windows, the statistical features of mean and standard deviation are calculated for each window per direction and are used as features.

Finally, the features are stacked column-wise to produce a feature vector which is the input to the neural network, labeled as $\mathrm{NN}_I$ in Fig. 1.

### 2.3 Feature Classification

Data from each sensor is classified using a separate Neural Network, and each Neural network contains one hidden layer. The hidden layer of $\mathrm{NN}_K$ contains 86 neurons, while that of $\mathrm{NN}_I$ contains 90 neurons. Moreover, the networks are trained using Conjugate gradient backpropagation with Polak-Ribiére updates [17].

The implemented algorithm utilizes a probabilistic framework, which is achieved using a soft-max output layer. The output vector (4) is a $C \times 1$ vector and each entry represents the probability of the respective label being assigned to the input sample $y$.

$$Output = [p_q(1|y)\ p_q(2|y)\ .\ .\ .\ p_q(C|y)]^T \tag{4}$$

where $C$ is the total number of classes, and $q \in [1, 2]$ represents each sensor.

Logarithmic opinion pool (LOGP) [15] is employed to merge the individual posterior probabilities of the classifiers and estimate the global membership function given in (5). A uniform distribution is assumed when fusing the two sensors, similar to [13].

$$P(c|y) = \prod_{q=1}^{2} p_q(c|y)^{1/2} \tag{5}$$

where $c \in [1, 2, ..., C]$ represents a class label.

The final label, to any sample, is assigned to the class label that has the highest probability according to

$$Label = \underset{c=1...C}{\mathrm{argmax}}\ P(c|y) \tag{6}$$

## 3 Experiments

We assessed the performance of the implemented algorithm on University of Texas at Dallas Multimodal Human Action Dataset [14]. The Dataset is publicly available and comprises of data synchronized from RGB videos, skeleton joint positions and depth positions obtained from Kinect, and inertial signals obtained from a wearable inertial sensor. There are a total of 27 actions registered, performed by 8 subjects (4 male and 4 female). Each action is performed 4 times by each subject. Moreover, due to 3 corrupt sequences being removed, the size of the dataset is 861 entries. For our purpose, we use the skeletal and inertial signal information.

This dataset has been chosen because it mimics the real world scenario. The dataset comprises of actions that utilize the movement of different parts of the body. Moreover, the position for the inertial sensor is also changed for different actions (the sensor is placed on the subject's right wrist for 21 actions and placed on the subject's right thigh for 6 actions).

We perform 8-fold cross validation, as in [13], by training on 7 subjects and testing on the left out 1 subject. This procedure is repeated for every subject in turn. The results are shown and compared in Table 1.

**Table 1.** Recognition accuracies for subject-generic experiment

| Subject-Generic Test | Skeletal Accuracy (%) | Inertial Acc. (%) | Fusion Acc. (%) |
|---|---|---|---|
| Chen et al. [13] | 74.7 | 76.4 | 91.5 |
| Proposed Algorithm | 74.8 | 81.2 | 95.0 |

Figure 2 illustrates the performance of the algorithm for each subject. It can be seen that the fusion always results in a better accuracy than using either of the sensors alone. Moreover, the depth sensor performed better than the inertial sensor for person 1 and person 4, while during the other trials the inertial sensor was more accurate than the depth sensor. Lastly, the accuracy for person 8 (85% fusion accuracy) was fairly lower than that of the other subjects and this reduced the cross-validation accuracy.

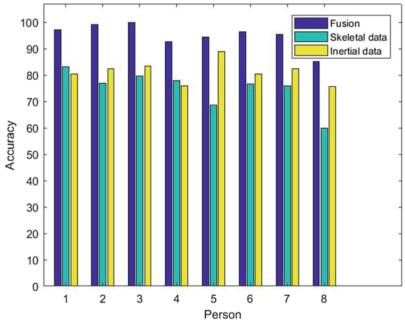

**Fig. 2.** Classification performance for each subject tested

Figure 3 displays the confusion matrix of the classification results when using depth sensor only, Inertial sensor only and a fusion of both. It can be seen that actions such as drawing circle clockwise (action 9) and drawing circle counter-clockwise (action 10) had misclassifications amongst them, and jogging in place (action 22) was misclassified as walking in place (action 23) a number of times, which is because of these actions being of very similar nature.

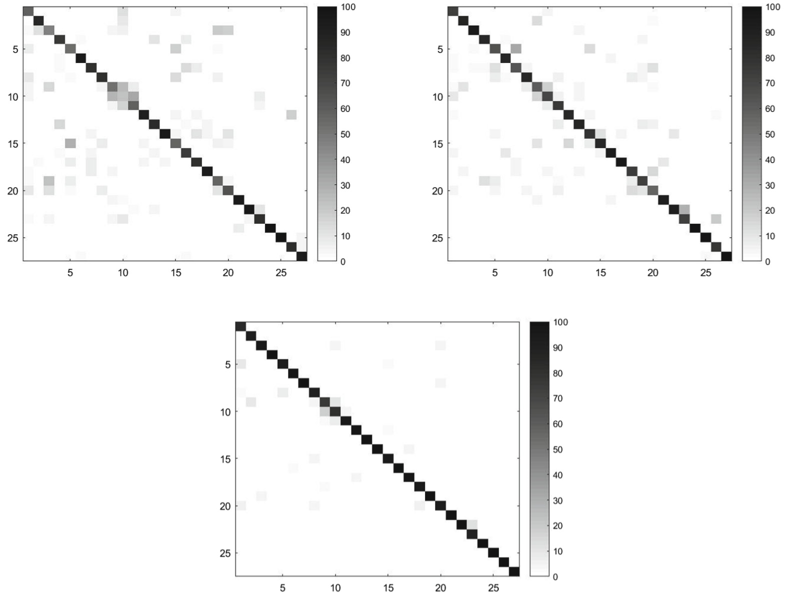

**Fig. 3.** Confusion Matrix of Skeletal (top left), Inertial (top right) and Fusion (bottom)

## 4 Conclusion

In this paper, we have presented a new approach to human action recognition that incorporates skeletal information from a depth camera, and acceleration and gyro measurements from an inertial sensor. We successfully extracted subject-invariant features that were classified using Neural Network classifiers. To improve the performance of our system we performed decision level fusion. Moreover, we performed extensive experimentation on a publicly available dataset and obtained good results. In particular, we have achieved 95% accuracy in 8-fold cross-validation, which is not only higher than using each sensor separately, but is also better than the best accuracy obtained on the mentioned dataset by 3.5%. As a future work, we plan on scaling up the system and testing it on datasets having more than one sensor of each modularity.

## References

1. Chen, C., Jafari, R., Kehtarnavaz, N.: A survey of depth and inertial sensor fusion for human action recognition. Multimed. Tools Appl. **76**(3), 4405–4425 (2017)
2. Stein, S., McKenna, S.J.: Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM (2013)

3. Ming, Y., Wang, G., Fan, C.: Uniform local binary pattern based texture-edge feature for 3D human behavior recognition. PloS one **10**(5), e0124640 (2015)
4. Ustundag, B.C., Unel, M.: Human action recognition using histograms of oriented optical flows from depth. In: Bebis, G., et al. (eds.) ISVC 2014. LNCS, vol. 8887, pp. 629–638. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-14249-4_60
5. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans. Cybernet. **43**(5), 1318–1334 (2013)
6. Aggarwal, J.K., Xia, L.: Human activity recognition from 3d data: a review. Pattern Recogn. Lett. **48**, 70–80 (2014)
7. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1110–1118 (2015)
8. Nie, B.X., Xiong, C., Zhu, S.C.: Joint action recognition and pose estimation from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1293–1301 (2015)
9. Altun, K., Barshan, B., Tunçel, O.: Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recogn. **43**(10), 3605–3620 (2010)
10. Qaisar, S., et al.: A hidden markov model for detection & classification of arm action in cricket using wearable sensors. J. Mob. Multimed. **9**(1&2), 128–144 (2013)
11. Ermes, M., Parkka, J., Mantyjarvi, J., Korhonen, I.: Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. IEEE Trans. Inf. Technol. Biomed. **12**(1), 20–26 (2008)
12. Ofli, F., et al.: Berkeley MHAD: a comprehensive multimodal human action database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp. 53–60. IEEE (2013)
13. Chen, C., Jafari, R., Kehtarnavaz, N.: A real-time human action recognition system using depth and inertial sensor fusion. IEEE Sens. J. **16**(3), 773–781 (2016)
14. 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 IEEE International Conference on Image Processing (2015)
15. Li, W., Chen, C., Su, H., Du, Q.: Local binary patterns for spatial-spectral classification of hyperspectral imagery. IEEE Trans. Geosci. Remote Sens. **53**(7), 3681–3693 (2015)
16. Orfanidis, S.J.: Introduction to Signal Processing. Prentice-Hall, Englewood Cliffs (1996)
17. Scales, L.E.: Introduction to Non-Linear Optimization. Springer, New York (1985)

# Author Queries

**Chapter 42**

| Query Refs. | Details Required | Author's response |
|---|---|---|
| AQ1 | This is to inform you that corresponding author has been identified as per the information available in the Copyright form. | |