Improving Human Action Recognition Using Decision Level Fusion of Classifiers Trained with Depth and Inertial Data

by

ZAIN FUAD

Submitted to
the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

SABANCI UNIVERSITY

July 2018

Improving Human Action Recognition Using Decision Level Fusion of Classifiers
Trained with Depth and Inertial Data

APPROVED BY:

Prof. Dr. Mustafa Ünel ....Mustafa Ünel..........
(Thesis Supervisor)

Asst. Prof. Dr. Hüseyin Özkan ....

Assoc. Prof. Dr. Şeref Naci Engin ....

DATE OF APPROVAL: ....30/7/2018..........

ABSTRACT

Improving Human Action Recognition Using Decision Level Fusion of Classifiers Trained with Depth and Inertial Data

Zain Fuad

Mechatronics Engineering M.Sc. Thesis, July 2018

Thesis Supervisor: Prof. Dr. Mustafa Ünel

Keywords: Human Action Recognition, Neural Networks, Classifier, Fusion, Logarithmic Opinion Pool, RGB-D Camera, Inertial Sensor

Improvement in sensor technology has aided research in the field of human action recognition (HAR), as acquiring data is easier and the obtained data is more accurate. However, each sensor has its own limitations and benefits, and a combination of these sensors can help improve the accuracy of recognition systems.

This thesis presents an in depth study of HAR using decision level fusion of classifiers that are trained using RGB-D camera and inertial sensor data. Extraction of robust and subject-invariant features is performed to train independent classifiers, i.e. neural networks, for action recognition purposes. This work employs decision level fusion on the outputs of the individual classifiers using a probabilistic approach in the form of Logarithmic Opinion Pool (LOP). The effect of varying the parameters of the proposed algorithm on the final 8-fold cross-validation accuracy is analyzed.

The proposed algorithm is tested on UTD-Multimodal Human Action Dataset that contains actions which are based upon the movement of different set of joints, and it achieves an average 8-fold cross-validation accuracy of 97.3%.

ÖZET

Derinlik ve Atalet Verileriyle Eğitilmiş Sınıflandırıcıların Karar Düzeyinde
Füzyonuyla İnsan Hareketi Tanımanın İyileştirilmesi

Zain Fuad

Mekatronik Mühendisliği Yüksek Lisans Tezi, Temmuz 2018

Tez Danışmanı: Prof. Dr. Mustafa Ünel

Anahtar Kelimeler: İnsan hareketi Tanıma, Sinir Ağları, Sınıflandırıcı, Füzyon, Logaritmik Düşünce Havuzu, RGB-D Kamera, Ataletsel Sensör

Sensör teknolojilerindeki ilerlemeler insan hareketi tanıma (İHT) alanındaki araştırmalara yardımcı oldu zira veri alımı kolaylaştı ve elde edilen verinin doğruluğu daha fazla. Bununla birlikte her sensörün kendine özgü sınırları ve yararları bulunmakta ve de bu sensörlerin füzyonu tanıma sistemlerinin doğruluğunu artırmada yardımcı olabilir.

Bu tezde RGB-D kamera ve ataletsel sensör verileri ile eğitilmiş bağımsız sınıflandırıcıların karar düzeyinde füzyonu kullanılarak İHT alanı derinlemesine irdelenmiştir. Gürbüz ve özneden bağımsız öznitelikler bağımsız hareket tanıma sınıflandırıcılarını (mesela sinir ağları) eğitmek için çıkarıldı. Bu çalışma Logaritmik Düşünce Havuzu (LDH) formunda olasılıksal yaklaşım kullanarak bireysel sınıflandırıcıların çıktıları üzerinde karar düzeyinde veri füzyonu uygulamıştır. Bu tez önerilen algoritmadaki parametreleri değiştirmenin son 8-katlı çapraz doğrulama üzerindeki etkisini incelemektedir.

Önerilen algoritma, içinde eklemlerin farklı hareketlerine göre sınıflandırılmış eylemler bulunan UTD-Multimodal Human Action Dataset üzerinde test edilmiş ve 8-katlı çapraz doğrulama sonucunda %97.3'lük bir doğruluk oranına ulaşılmıştır.

≪ *To my loving family* ≫

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisor Prof. Dr. Mustafa Ünel for his supervision, guidance, encouragement, and precious support. His perspective and passion for research has taught me a lot.

I would like to thank Asst. Prof. Dr. Hüseyin Özkan and Assoc. Prof. Dr. Şeref Naci Engin for spending their valuable time as my jurors.

I am thankful to my colleagues of the CVR group specially, Gökhan Alcan, Diyar Khalis Bilal, Hammad Zaki, Mehmet Emin Mumcuoğlu, Emre Yilmaz, Ekin Yağis and Mohammed Taleb Zamzam for their support. Moreover, I am thankful to Hande Karamahmutoğlu, Yağmur Yildizhan and Doğukan Kaygusuz for their friendship throughout my graduate life.

I would like to specially thank Fatih Emre Tosun, Ipeknaz Özden, Loreto Ferrer, Elisabeth Nikitidou, Serhat Emre Cebeci and Hassan Wasim Ursani for their endless support that has broadened my vision and perspective.

Lastly, I would like to thank my dear brother Daniyal Fuad, and my parents for their love, confidence and trust in me that has made me come so far.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human action recognition (HAR) is a multidiscipline research area and the goal can be simply put as acquiring a person's gestures through various sensors, merging these gestures to form an action, and lastly understanding or classifying those actions. In other words, the body movements acquired from different sensors are classified to understand the intended action. The applications consist of security or surveillance, robotics, telemedicine, internet of things and human-machine interaction [19], and have extended to unorthodox areas, such as recognition of food preparation activities [20]. Some examples of human actions can be seen in Figure 1.1.

There are different aspects that need to be looked into for solving the problem of HAR. One of these problems is to choose sensors that can acquire the significant human movements.

Due to the 3-dimensional nature of the world we live in, relying on RGB cameras result in the loss of the depth information, which in turn decreases the efficiency of the action recognition framework. Structure from motion [21] or stereo vision [22] although solve the issue of grasping the depth, however, they require high computational power, which restricts their use in many real world scenarios [23].

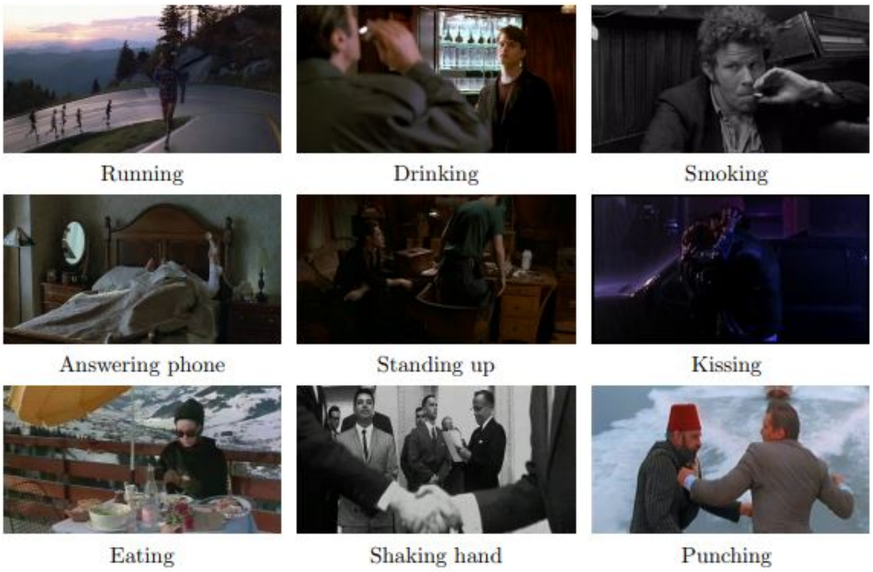
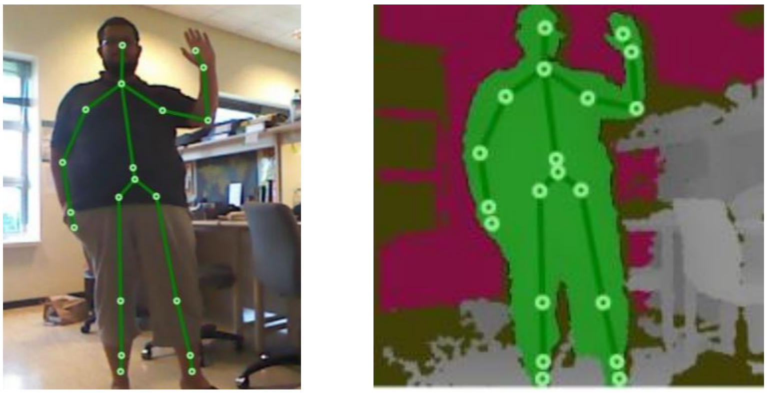


Figure 1.1: Sample human actions [1]



(a) Skeletal joint positions shwon on an RGB image of a person

(b) Skeletal joint positions shown on depth image of a person

Figure 1.2: Microsoft Kinect skeletal joint positions illustration

Nonetheless, with the advent of RGB-D cameras such as Microsoft Kinect, depth information can be acquired with less computational effort. An example of the skeletal joints obtained from Microsoft Kinect is shown in Figure 1.2.

The advent of wearable inertial sensors (Figure 1.3) made their application possible in everyday usage as they provide little or no hindrance to the person performing

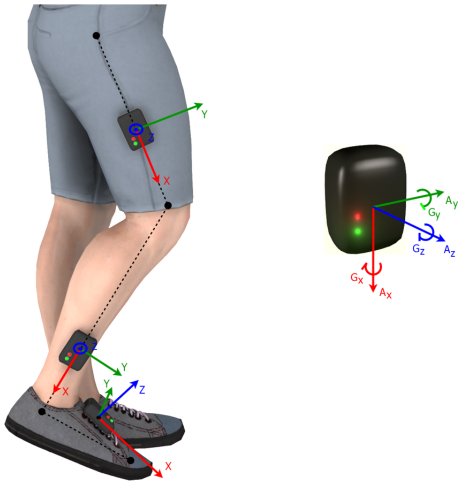
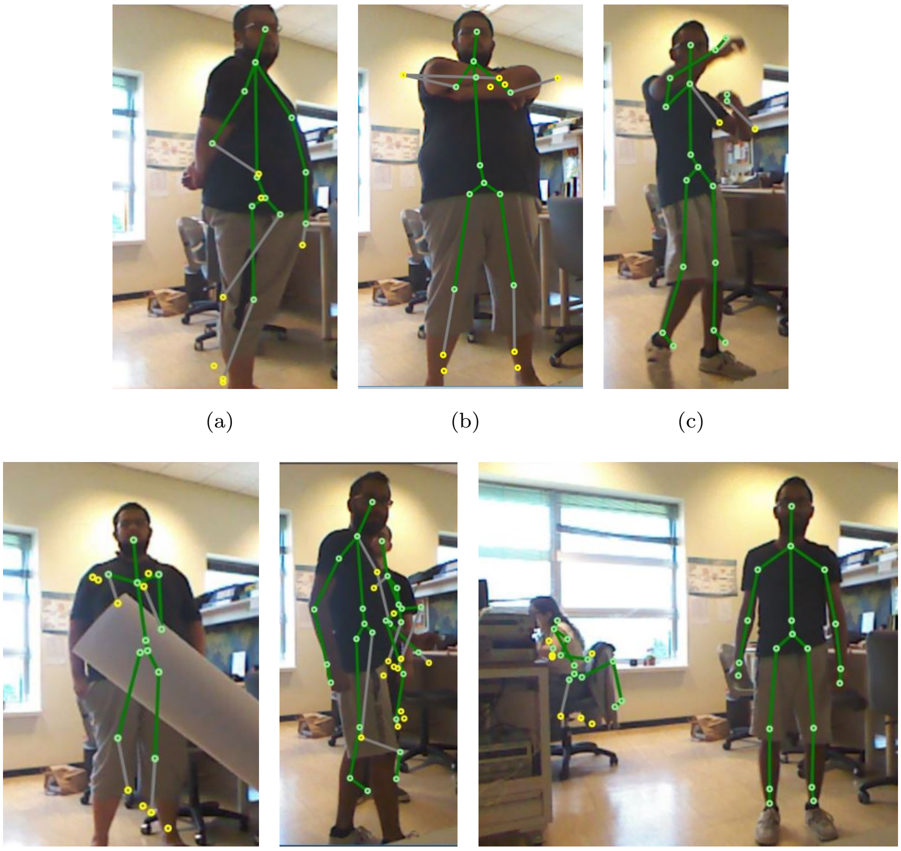


Figure 1.3: Example of a wearable inertial sensor [2]

the action. These sensors can be placed on any part of the human body and they can capture the motion to a great accuracy.

Despite the advances, there are still a lot of challenges in this regard, which basically arise from the way an action is performed and can be influenced by environmental, cultural, personal and emotional factors [24]. These factors may include view point occlusions or signal distortions of a particular sensor, view point differences or the presence of different type of clothing for vision-based sensors and unwillingness of a person using a particular type of a sensor [25]. In regards to the wearable sensors of any sort, wearing them loosely affects their performance, as there is relative movement between the sensor and the body, in contrast to the sensor being firmly fixed to capture solely the body movements. Moreover, there is a limit to the number of sensors being worn, as they can cause physical and/or psychological discomfort to the person wearing them.

Some of the problems associated with getting skeletal joint positions from Microsoft Kinect are shown in Figure 1.4. The problem mainly arises when the camera looses the track of the human body parts. This causes broken joints, joints being at unrealistic locations and/or unreasonable skeletal form.



(a) (b) (c)

(d) (e) (f)

FIGURE 1.4: Faulty skeletal joint positions obtained from Kinect in a real world environment

For this reason, the idea of sensor fusion comes into practice, as the deficiencies and limitations of one sensor can be compensated by other sensor(s). The purpose of this thesis is to recognize human actions using data acquired from depth and inertial sensors. Adding to this, this work makes use of neural networks as the main classifiers due to their robustness and high classification accuracy, and applies Logarithmic Opinion Pool (LOP) as the decision level fusion method to merge the outputs of the individual classifiers.

1.1 Contributions of the Thesis

The goal of this work is to design a framework that is able to recognize human actions to a high degree of accuracy. For this purpose, existing work in the literature is investigated and a new method has been developed that is based upon the idea of sensor/data fusion.

This thesis has the following main contributions:

- It provides instances where the sensors that are typically used for HAR fail in one way or the other, and proposes to use a fusion of joint locations from depth sensor and linear acceleration and angular velocities from an inertial sensor, to acquire the body movements to a significant degree of accuracy.
- A new algorithm is developed that is subject-invariant, performs well under noisy measurements, and can be employed in real time. The proposed algorithm consists of neural network classifiers to classify the data from each sensor. Decision level fusion is performed on the outputs of these classifiers in a probabilistic manner. Moreover, a discussion about the free parameters of the proposed algorithm has been presented that can be used to tune the algorithm.
- The algorithm has been tested and benchmarked on UTD-MHAD dataset [17]. This dataset contains a variety of actions that are performed by the movement of different joints, and depict real world scenarios. The achieved accuracy on this dataset is 97.3%.

1.2 Outline of the Thesis

The organization of this thesis is as follows:

Chapter 2 reviews works that address the issue of HAR. The works are divided into different categories based on the modality of the sensor that they make use of.

Chapter 3 provides a general overview of the characteristics of the sensors that are used in the framework of HAR. Moreover, a list of the publicly available datasets used in this framework are also presented.

In Chapter 4 the proposed algorithm is highlighted and explanations are provided for each step of the algorithm.

Chapter 5 presents the results of the algorithm and Chapter 6 concludes the thesis and indicates possible future directions.

1.3 Publications

The following papers are produced during my MS thesis work

- Fuad Z., Unel M. (2018) Human Action Recognition Using Fusion of Depth and Inertial Sensors. In: Campilho A., Karray F., ter Haar Romeny B. (eds) Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science, vol 10882. Springer, Cham.

- Fuad, Z. and Unel, M. Improving Human Action Recognition Based on Decision Level Fusion of Classifiers Trained with Depth and Inertial Data **(under preparation)**

Chapter 2

Literature Survey

The literature contains a lot of techniques and solutions to the problem of HAR. Hachaj et al. [26] proposed a method for template generation, matching, comparing and visualization which they applied on MoCap recordings of highly-skilled karate athletes. On the other hand, Chaaraoui et al. [27] propose a multi-view setup approach to recognize human behavior for health purposes and they extend this approach to maintain the privacy of the users of the system.

Nazir et al. [3] proposed a Bag of Expression framework which is based on the bag of words approach, and formed a codebook of visual expressions based on the training videos. Later a non-linear SVM was used as the action classification algorithm.

On the other hand, Nie et al. [4] decomposed human actions into poses, and further decomposed these poses to mid-level spatio-temporal parts and used dynamic programming for classification purposes. They claim that this way they are able to capture the geometric and appearance variations of the poses at each frame. The results they obtain are shown in Figure 2.1.

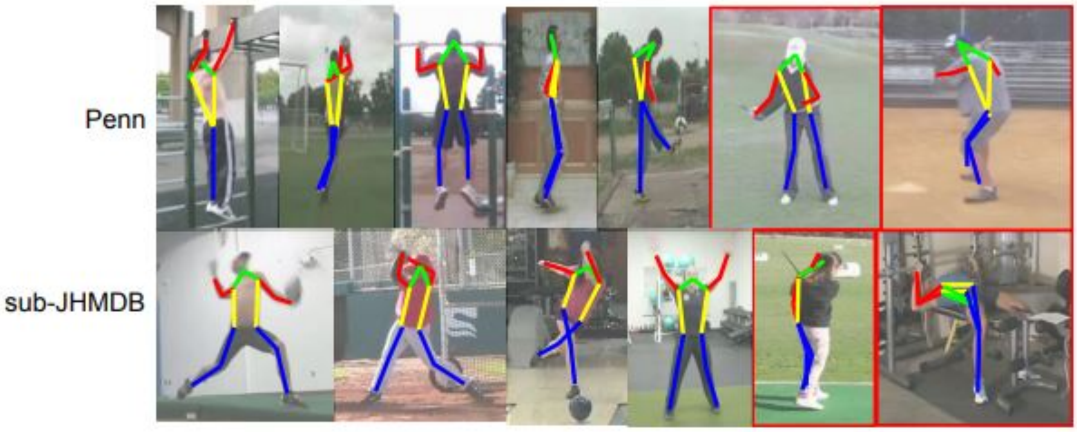


FIGURE 2.1: Pose estimation from videos [4]

Due to the nature of this thesis, the literature has been classified according to the sensor modalities they employ.

2.1 HAR Based on Depth Sensor

The idea of HAR from depth sensors is a well-established idea [28]. Advances such as Microsoft Kinect and ASUS Xtion Pro Live, low-cost RGB-D cameras that can acquire depth information in addition to RGB videos, have aided the encapsulation of human motion, in contrast to the expensive detector based MoCap systems, or computationally-expensive 3-D reconstruction using stereo cameras [19]. In other words, RGB-D videos preserve discriminative information, such as shape and distance variations [29], and have reduced processing times as compared to traditional RGB cameras [23]. 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, and covered the topics regarding preprocessing, object tracking and recognition, human activity analysis, hand gesture analysis and indoor 3-D mapping (Figure 2.2).

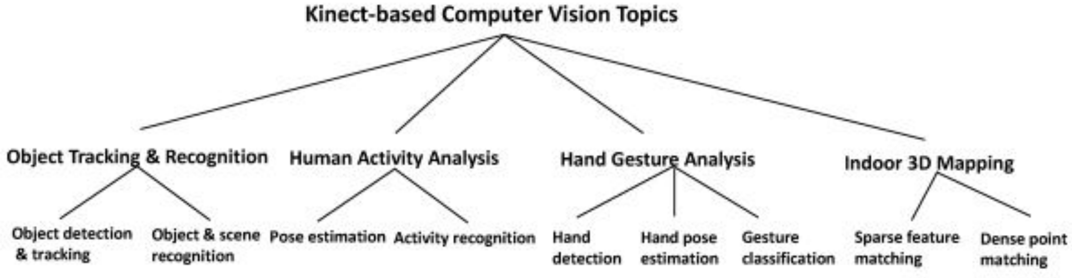


FIGURE 2.2: Applications of Kinect [5]

Aggarwal et al. [6] discussed different approaches for feature extraction from depth data and mentioned methodologies employed in the context of human activity recognition. They further highlight the pros and cons of each algorithm they analyzed. The taxonomy of their review can be seen in Figure 2.3.

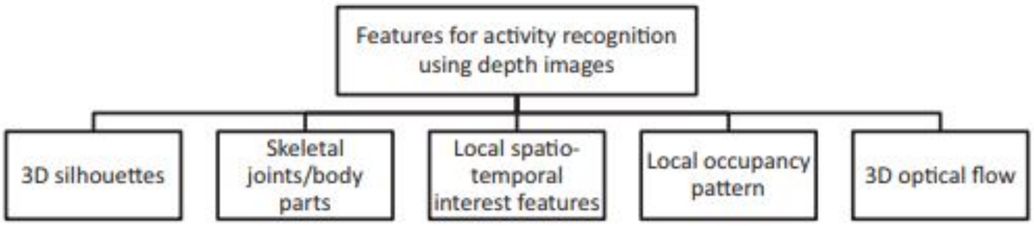

Features for activity recognition using depth images
3D silhouettes
Skeletal joints/body parts
Local spatio-temporal interest features
Local occupancy pattern
3D optical flow

FIGURE 2.3: Features used in the context of for HAR from depth images [6]

Notable work in this area includes the proposition of a Hierarchical Recurrent Neural Network framework which uses skeletal positions obtained from depth cameras, and understands the performed actions [7]. The authors divide the skeleton into 5 parts and feed them into 5 subnets, as opposed to taking the whole skeleton as the input. A sketch is shown in Figure 2.4 where the skeleton is divided into 5 parts and fed into the proposed framework.

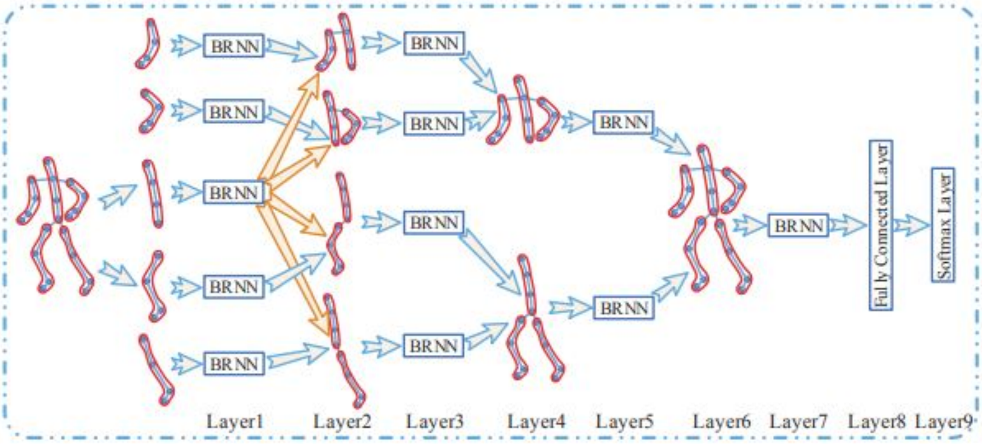


FIGURE 2.4: Sketch of a RNN framework for HAR [7]

On the other hand, Luzivon et al. [8] extracted sets of spatial and temporal features form subgroups of joints, which were later combined and k-NN was used to classify the actions. An overview of their proposed algorithm is shown in 2.5.

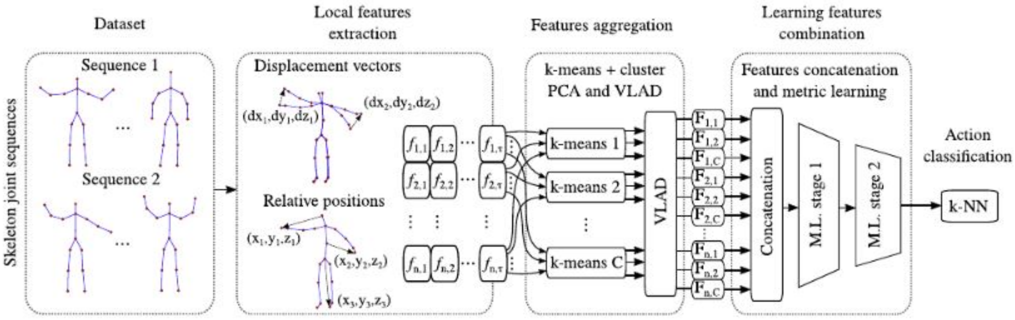


FIGURE 2.5: Learning features combination for HAR [8]

Another interesting work is the proposal of Sequence of the Most Informative Joints (SMIJ) [9], where each joint is compared in terms of the information it provides, and the joints are sorted with respect to the information content they provide (Figure 2.6).

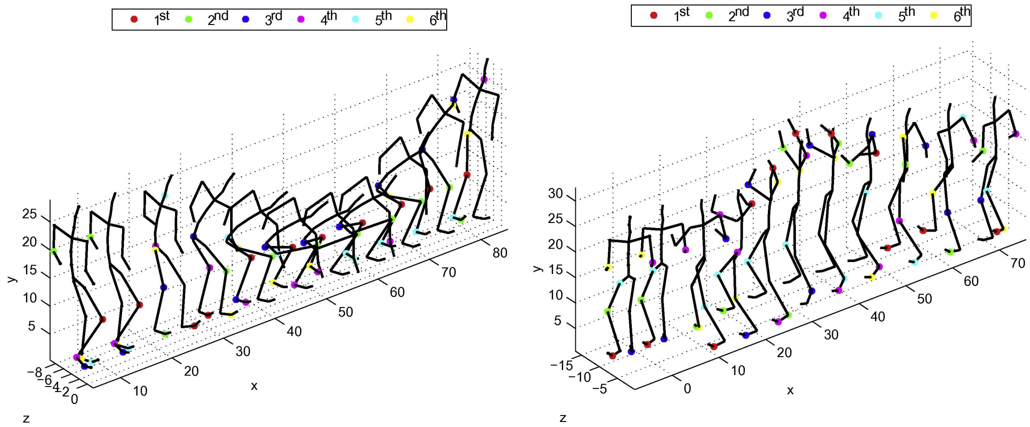


FIGURE 2.6: Demonstration of the most informative joints along the key frames of two different actions [9]

2.2 HAR Based on Inertial Sensor

The invent of low-cost, small and light-weight, wearable inertial sensors have further aided the research of HAR, as they provide very little hindrance to the person performing these actions and can be used in real life scenarios [30].

Qaiser et al. [10] studied the classification of arm action in cricket using inertial sensors. Figure 2.7 shows the utilized sensor positions. In this work they utilized mean, mode, standard deviation, peak to peak value, minimum, maximum, first and second derivative as features that were extracted from acceleration and angular velocity signals.

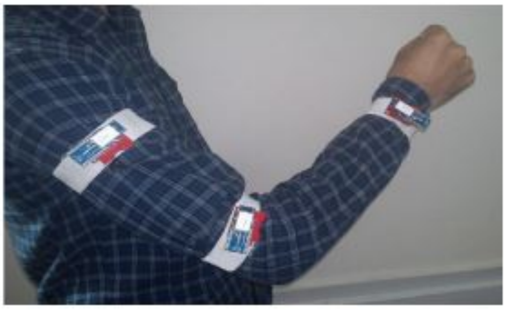


FIGURE 2.7: Sensor placements for classifying cricket actions [10]

Additionally, Guo et al. [11] evaluated the effect of task complexity on the accuracy of using Xsens MVN BIOMECH, which is an inertial sensor-based motion capture system (Figure 2.8). They performed experiments based on 11 tasks, and found that wrongly estimated foot separations and the initial system estimation error on Base of Support (BOS), are two major sources of instabilities and errors of BOS estimation.

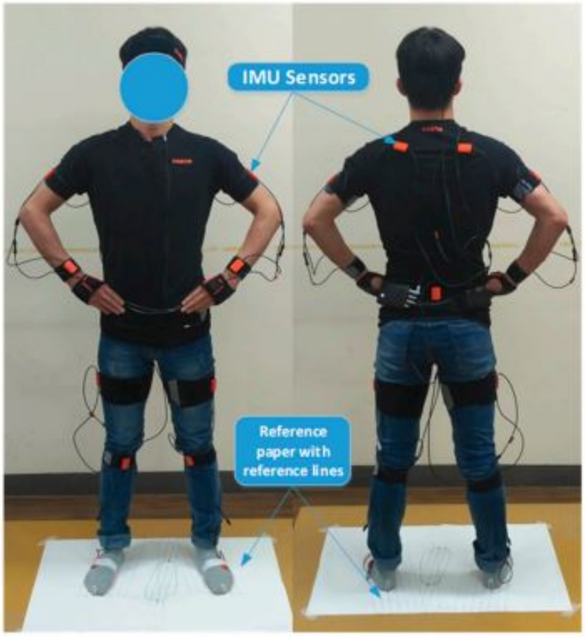


Figure 2.8: Xsens MVN BIOMECH body suit and footprint papers [11]

Ermes et al. [12] analyzed the use of inertial sensors in the detection of sports activities in controlled and natural environments. Figure 2.9 represents the sensors used for this task. In this work a hybrid classifier was used, which was composed of a tree structure possessing a priori knowledge and artificial neural networks, and 3 reference classifiers.

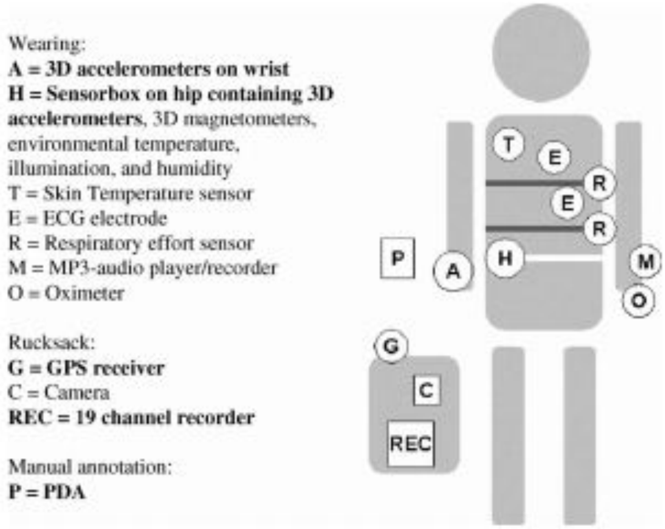

Wearing:
A = 3D accelerometers on wrist
H = Sensorbox on hip containing 3D accelerometers, 3D magnetometers, environmental temperature, illumination, and humidity
T = Skin Temperature sensor
E = ECG electrode
R = Respiratory effort sensor
M = MP3-audio player/recorder
O = Oximeter
Rucksack:
G = GPS receiver
C = Camera
REC = 19 channel recorder
Manual annotation:
P = PDA
T
E
R
E
R
P
A
H
M
O
G
C
REC

Figure 2.9: Data collection and annotation system [12]

Due to the wearable nature of these inertial sensors, one or more of them can be placed at different parts of the human body to fully grasp the movements. Attal et al. [31] reviewed the placement of these sensors on specific parts of the human body and provided a comparison of the obtained accuracy and the number of activities performed. For fall detection, sensors placed on the chest, waist, ankle and thigh were compared [32], whereas Prittikangas et al. [33] tested thighs, necklace and wrists for the recognition of activities such as drinking, ascending or descending stairs, watching TV and typing.

In addition to placing these sensors on the different parts of the human body, they can also be placed on accessories. Dang et al. [34] placed an inertial sensor to various positions on a cane that is used as a mobility aid for walking. Based on the movements of the cane, the walking distance was estimated. Similarly, Gellaerts et al. [13] instrumented a ski-mounted inertial sensor on the equipment of skiers to analyze cycle parameters and classified the movements in real time. Figure 2.10 shows the sky mounted inertial sensor used in [13].

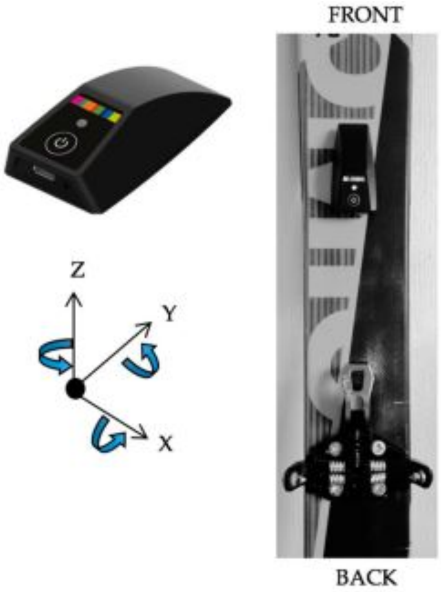


FIGURE 2.10: Ski mounted inertial sensor [13]

2.3 HAR Based on Sensor Fusion

Regarding the sensors used to acquire the actions, each of the sensors has their own advantages and short-comings, and a fusion of these sensors results in a higher action recognition performance [19]. 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. [14] 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. The data acquisition system they made use of is shown in Figure 2.11. On the other hand, Stein and McKenna [20] 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.

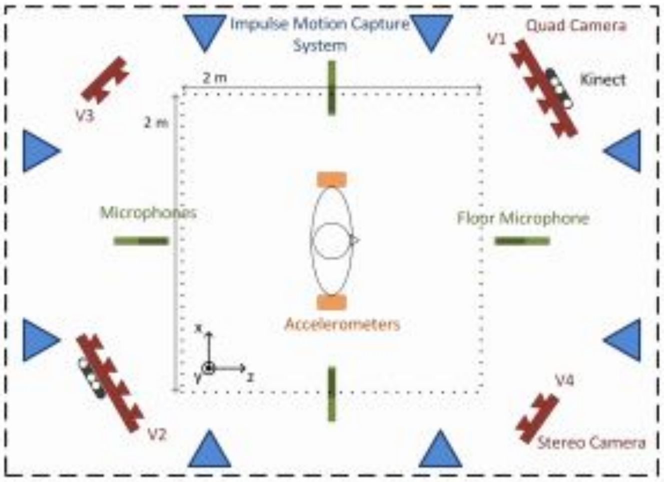


FIGURE 2.11: Data acquisition system with different modality sensors [14]

Chen et al. [15] 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. The classification performance they obtained for each action is shown in Figure 2.12, which show that sensor fusion results in a higher classification accuracy than by using each sensor individually.

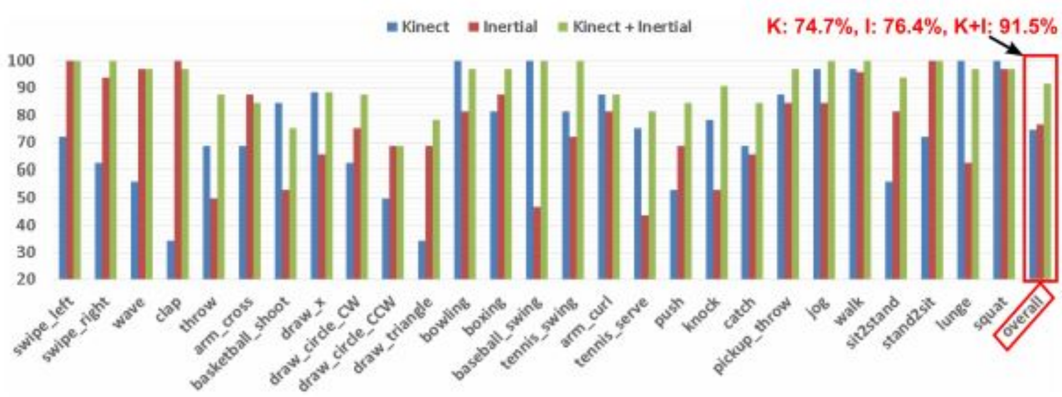


FIGURE 2.12: Classification performance for subject generic experiments [15]

Chapter 3

Sensors and Datasets for HAR

3.1 Sensors

HAR systems can be divided into three main categories (Figure 3.1) (i) sensor (inertial) based to detect movements of body parts (ii) camera or vision based that record video sequences and use computer vision algorithms to understand these videos, and (iii) radio based that understand human activities based on the information about utilized objects or change in environmental variables [16]. Figure 3.2 shows an RGB camera, and a MoCap system that can be used for HAR.

This thesis makes use of (i) Microsoft Kinect (a depth sensor) and (ii) MEMS inertial sensor, and so the rest of this section is dedicated to a description regarding these two sensors.

3.1.1 Microsoft Kinect

Microsoft Kinect (Fig. 3.3) is a commercially available, low cost RGB-D camera. It is manufactured with a built-in RGB camera, an infrared emitter and depth sensor,

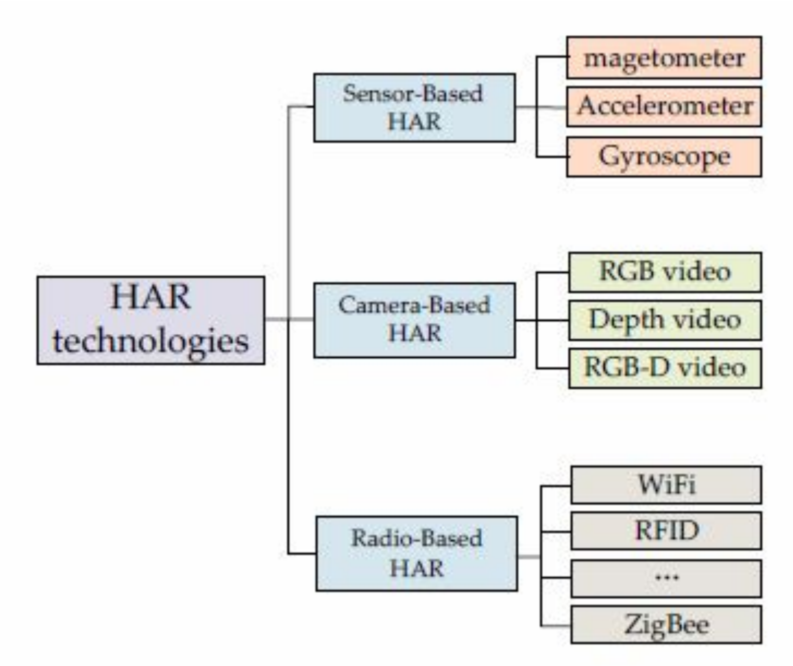
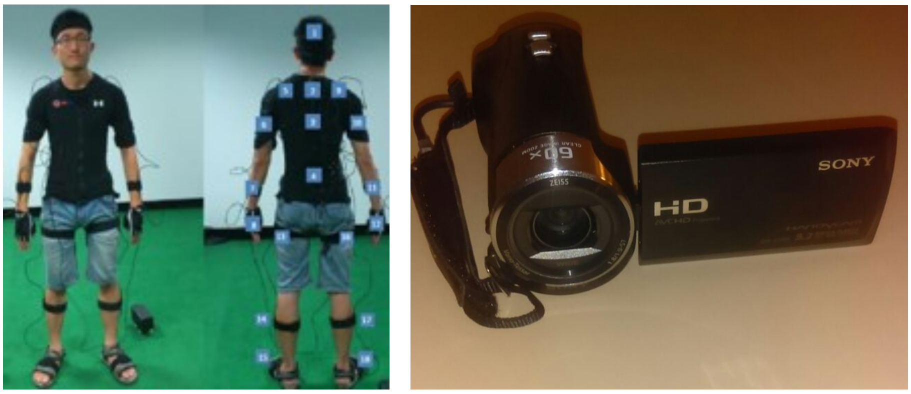


FIGURE 3.1: Categorization of human action recognition systems [16]



(a) Inertial MoCap system [35]

(b) RGB camera

FIGURE 3.2: Sensors used in human action recognition framework

a microphone, a tilt motor to set the camera angle and an LED light. Kinect captures color images with a resolution of 64 × 480 pixels and 16-bit depth images having a resolution of 320 × 240 pixels with a frame rate of 30 frames per second [17].

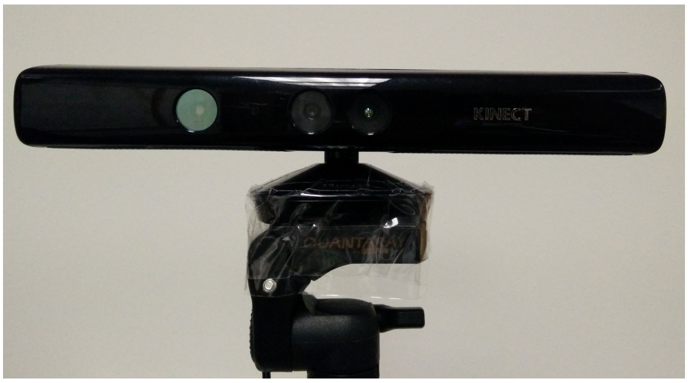


FIGURE 3.3: Microsoft Kinect [17]

Moreover, Kinect SDK, a publicly available support package can be used to track 20 body skeletal joints with their 3-D spatial coordinates (Figure 3.4).

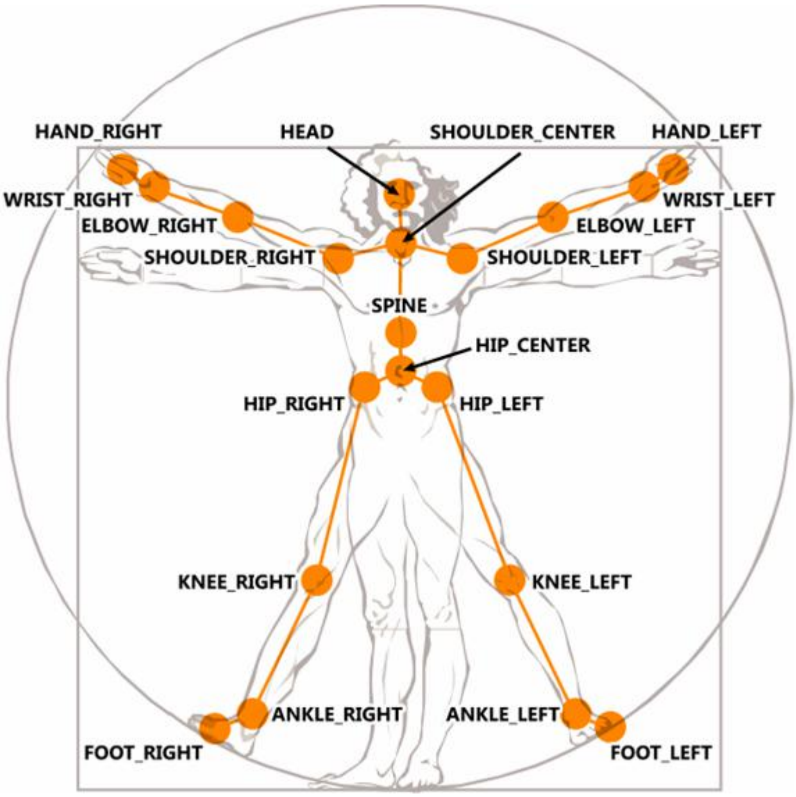


FIGURE 3.4: Joints tracked by Kinect SDK [18]

3.1.2 Inertial Sensor

The low-cost, wearable inertial sensor (Figure 3.5) considered in this work consists of 9-axis MEMS sensor that captures 3-axis acceleration, 3 axis angular velocity and 3-axis magnetic strength. The sampling rate of this sensor is 50 Hz and the measuring range is ±8g for acceleration and ±1000 degrees/second for rotation [17]. Figure 3.6 depicts an instance of a signal obtained from the inertial sensor.

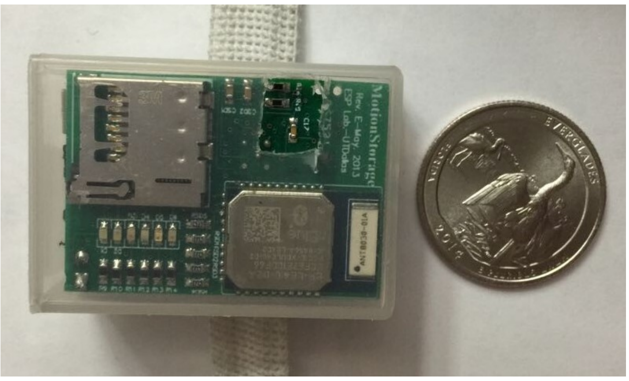
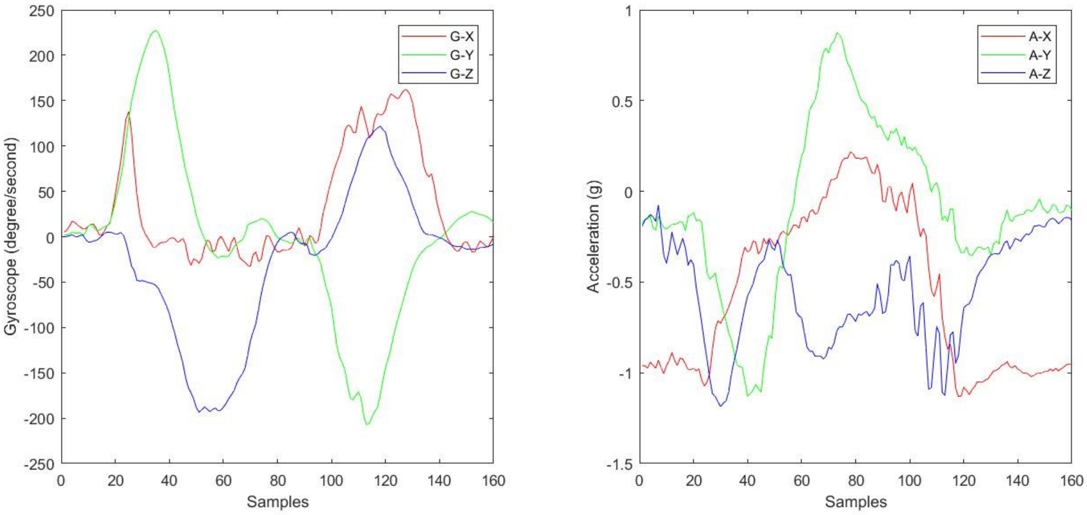


FIGURE 3.5: Inertial sensor [17]



(a)

(b)

FIGURE 3.6: Gyro measurements (a) and acceleration (b) obtained from the inertial sensor

3.2 Datasets

Table 3.1 presents a list of publicly available human action datasets that use sensors of more than one modality.

Table 3.1: Publicly available multi-modal human action datasets: M:MoCap, R:RGB, D:Depth, A:Audio, I:Inertial (Adopted from [19])

Dataset	Modality					# Sub	# Act	# Seq	Year
	M	R	D	A	I				
UTD-MHAD [17]	0	1	1	0	1	8	27	861	2015
URFD [36]	0	2	2	0	1	5	>5	70	2014
TST Fall detection [37]	0	0	1	0	2	11	8	264	2014
Berkley MHAD [14]	1	12	2	4	6	12	11	660	2013
50 salads [20]	0	1	1	0	7	25	17	966	2013
ChAirGest [38]	0	1	1	0	4	10	10	1200	2013
Huawei/3DLife [39]	0	5	5	5	8	17	22	3740	2013

To govern the effectiveness of the proposed algorithm, it was tested using the University of Texas at Dallas Multi-modal Human Action Dataset [17]. This particular dataset has been chosen because it mimics the real world scenarios, as it comprises of actions that utilize the movement of different parts of the body rather than targeting only a certain group of joints.

The position of the inertial sensor is 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 the rest 6 actions), which makes the dataset fairly difficult and the robustness of the algorithm is a necessary requirement to achieve good results. For our purpose, we only use the skeletal and inertial signal information.

Section 3.2.1 provides a detailed explanation about the dataset.

3.2.1 University of Texas at Dallas Multi-Modal Human Action Dataset

The UTD-MHAD [17] is a publicly available dataset and comprises of data synchronized from RGB videos, skeleton joint positions and depth information obtained from Microsoft Kinect, and inertial signals, i.e. 3 axis linear accelerations and gyro measurements obtained from a wearable inertial sensor, i.e. IMU. There are a total

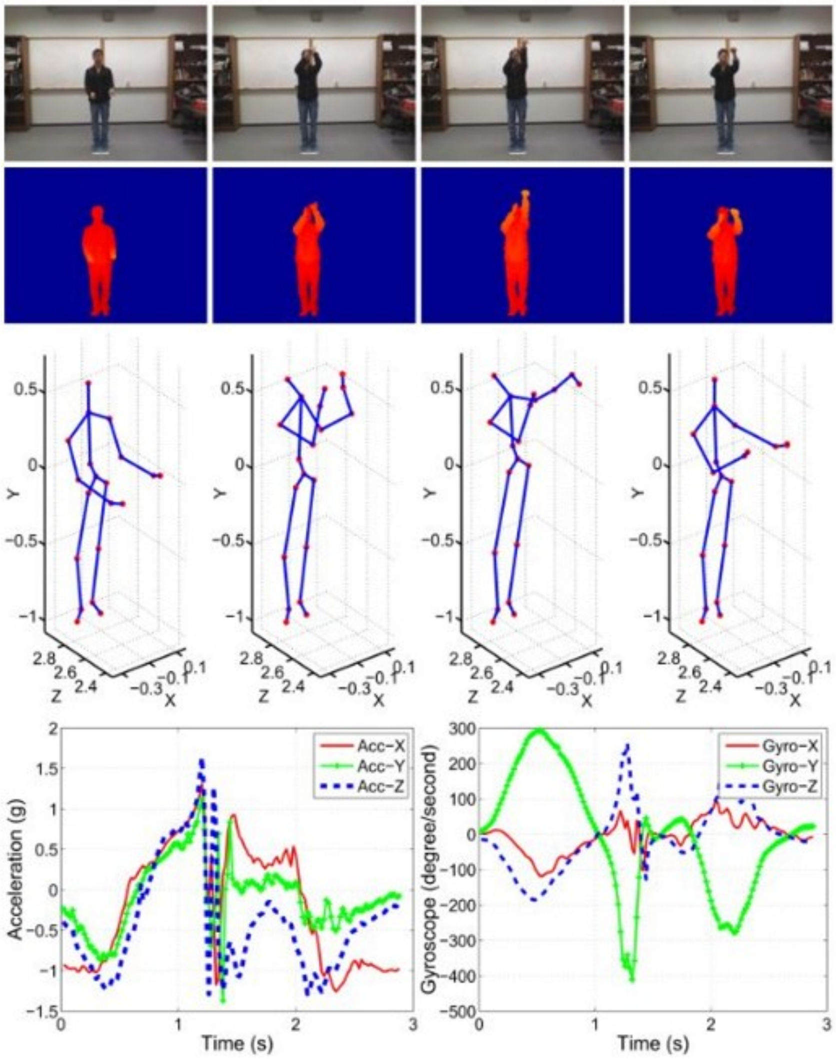


FIGURE 3.7: Data visualized as observed by different sensors in UTD-MHAD dataset [17]

of 27 registered actions, 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 total number of entries in the dataset is 861. The actions as visualized by different sensors are shown in figure 3.7, while the 27 registered actions in the dataset are given in figure 3.8

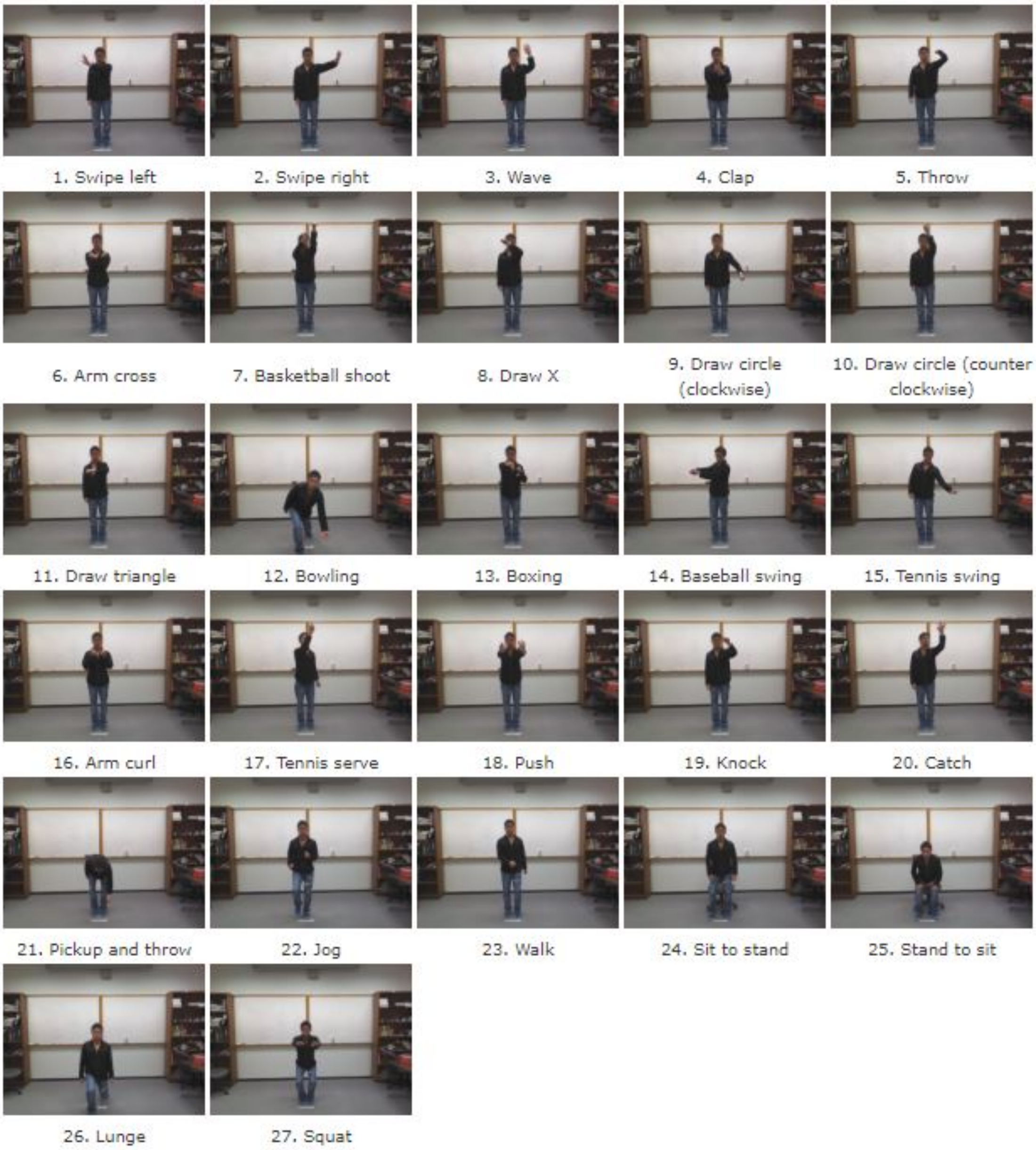


Figure 3.8: Actions present in the UTD-MHAD dataset [17]

Chapter 4

ANN Based Classifiers for HAR and Fusion of Them Using Logarithmic Opinion Pool

Variations in speed while performing an action, dissimilarities in the way two different people perform the same action, and noise due to jitters are the main complications that require a robust and precise classification algorithm for HAR.

The proposed algorithm (Figure 4.1) performs action classification by utilizing a depth and an inertial sensor. Feature extraction is performed on the frame-wise skeletal joint positions from the depth sensor, and linear accelerations and angular velocities are obtained from a wearable inertial sensor, i.e. IMU, located on different parts of the body. The feature extraction stage involves resizing all the signals from a particular sensor to the same size to reduce temporal variations. Moreover, after performing normalization on the skeletal joint positions and the extraction of mean and standard deviation from the inertial sensor measurements, individual classifiers ($Classifier_K$ and $Classifier_I$) are trained for each sensor.

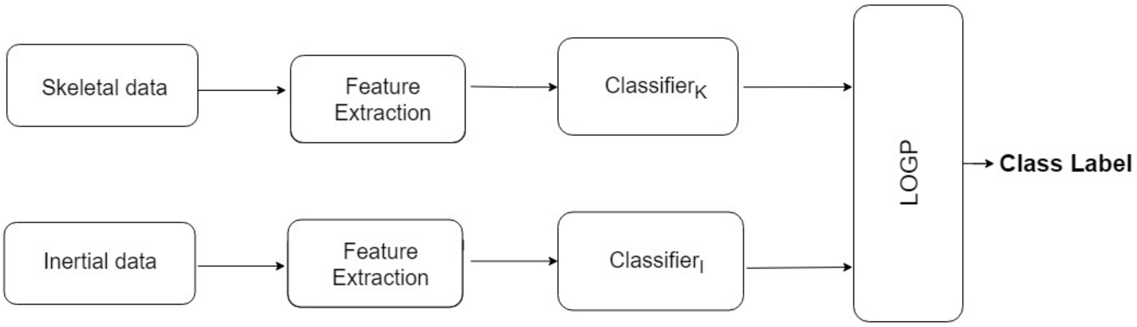


FIGURE 4.1: Overview of the proposed algorithm

Similar to the training phase, the testing phase involves feature extraction from depth and inertial data, and classifying them using the trained classifiers. Finally, a decision level fusion is performed on the outputs of the individual classifiers using Logarithmic Opinion Pool (LOGP [15] or LOP [40]), and a class label is assigned for the performed action. The implemented algorithm can be scaled up to include data from more than two sensors.

4.1 Feature Extraction

4.1.1 Feature Extraction from Depth Sensor data

The depth sensor provides $[x_{i,j}\ y_{i,j}\ z_{i,j}]$, the spatial coordinates of each tracked 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

$$I_K = \begin{bmatrix} x_{1,1} & y_{1,1} & z_{1,1} \dots x_{1,N} & y_{1,N} & z_{1,N} \\ x_{2,1} & y_{2,1} & z_{2,1} \dots x_{2,N} & y_{2,N} & z_{2,N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{M,1} & y_{M,1} & z_{M,1} \dots x_{M,N} & y_{M,N} & z_{M,N} \end{bmatrix} \tag{4.1}$$

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

Due to the variations in speed in performing actions, the total number of frames for each action may differ. To eliminate this temporal variation, the dimensions of the feature vectors should be comparable to each other. The literature suggests different approaches in this regard, including PCA [41], Locally Linear Embedding [42] and Dynamic Time Warping [43].

The proposed algorithm uses bi-cubic interpolation to reduce the temporal variations. Frequently used in image processing tasks, bi-cubic interpolation provides better results than nearest neighbor and linear interpolation, and a lesser processing time than B-Spline interpolation [44].

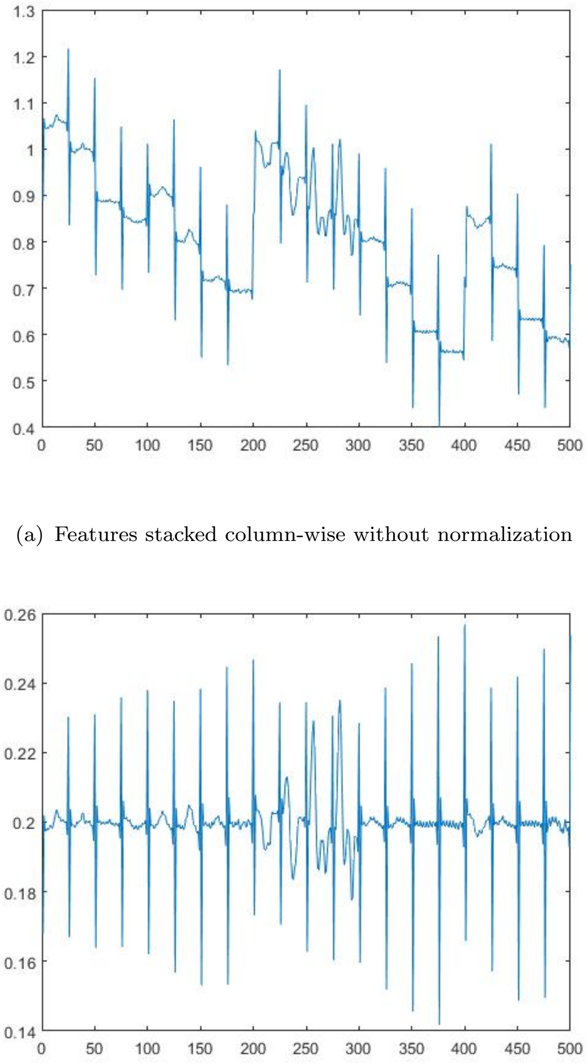
After the interpolation operation, the number of columns in I_K reduces to

$$\hat{N} = \lambda N_{min} \tag{4.2}$$

where N_{min}, a data dependent parameter, is the least number of frames amongst the entries from the training dataset, and λ is a scaling constant that helps in dimensionality reduction.

Each row of I_K is divided by its norm, 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 effect is shown in Figure 4.2, when the features are stacked with and without normalization.

The rows of the reduced matrix are stacked column-wise to produce a $20\hat{N} \times 1$ input vector to the classifier, labeled as $Classifier_K$. However, there is noise present in the form of spikes and for that Savitzky-Golay [45] filter is applied to reduce these spikes.

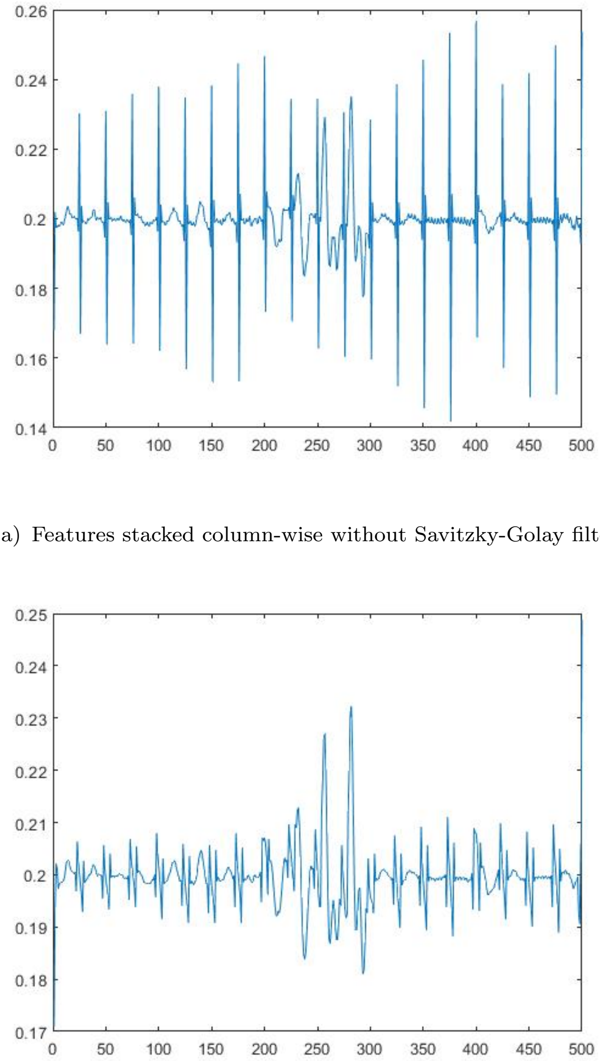


(a) Features stacked column-wise without normalization

(b) Features stacked column-wise with normalization

Figure 4.2: Illustration of the effect of normalization on the rows of I_K

Savitzky-Golay filter is a method of data smoothing and is based on local least-square polynomial approximation [46]. The parameters of the filter should be chosen in such a way that only the spikes are reduced, without compromising the information present in the signal. An illustration is shown in Figure 4.3.



(a) Features stacked column-wise without Savitzky-Golay filter

(b) Features stacked column-wise with Savitzky-Golay filter

Figure 4.3: Effect of using Savitzky-Golay filter

4.1.2 Feature Extraction from Inertial Sensor Data

A wearable inertial sensor, i.e. IMU, can be placed at any part of the body, and provides 3-axis linear acceleration and angular velocity 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, ω_i is the angular velocity and i depicts the respective axis. Then the data obtained from the inertial sensor can be represented as

$$I_I = \begin{bmatrix} a_{x,1} & a_{y,1} & a_{z,1} & \omega_{x,1} & \omega_{y,1} & \omega_{z,1} \\ a_{x,2} & a_{y,2} & a_{z,2} & \omega_{x,2} & \omega_{y,2} & \omega_{z,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{x,N} & a_{y,N} & a_{z,N} & \omega_{x,N} & \omega_{y,N} & \omega_{z,N} \end{bmatrix} \tag{4.3}$$

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 in the case of skeleton data, the inertial sensor data has different signal sizes. To reduce this variation, all the signals are resized using bi-cubic 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 data. Furthermore, the inertial sensor measurements are partitioned into temporal windows, of size $W \times 6$, and statistical features, i.e. mean and the standard deviation, are calculated for each window per direction, and are used as inputs to $Classifier_I$. The effect of changing the window length, W, on the classification accuracy is investigated in Chapter 5.

4.2 Feature Classification

After obtaining robust and subject-invariant features, individual classifiers are trained for each sensor. This section focuses on the classifiers used in this work, i.e. neural networks and provides a discussion regarding their implementation in the proposed algorithm.

4.2.1 Artificial Neural Network

A neural network (Figure 4.4) can model the relationship between an input vector and the target value. Neural networks are made up of many connected processors called neurons, the input neurons get activated through sensors perceiving the environment, while other neurons get activated from weighted connections with other neurons [47].

Neural networks have been used in tasks ranging from digit classification [48], plant classification [49] and face recognition [50] to music composition [51]. Their effectiveness is a direct reason for their popularity in the field of machine learning.

The network we make use of in this thesis has the following structure

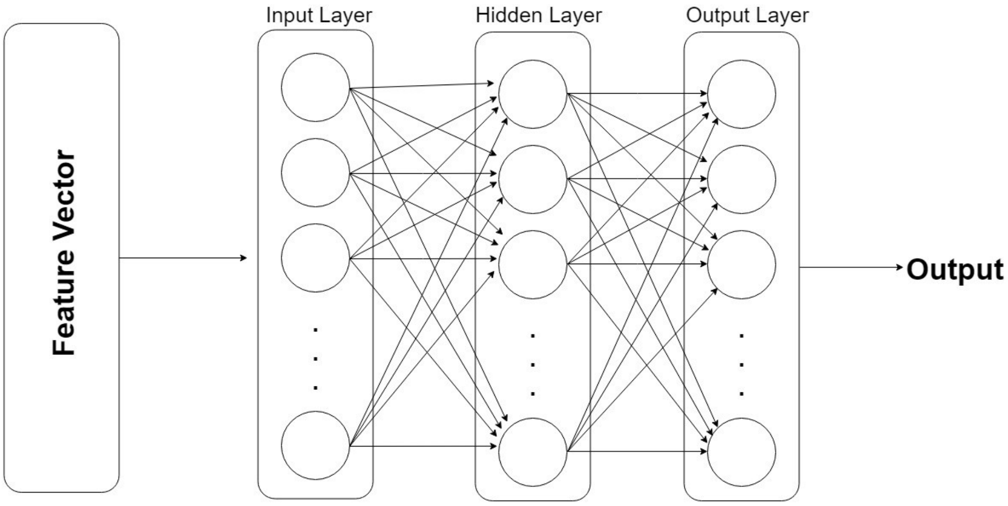


Figure 4.4: Structure of the proposed Neural Network classifier

In order to train the network this work makes use of conjugate gradient backpropogation algorithm due to its less memory requirements, as it makes use of the conjugate search directions and still guarantee quadratic termination [52].

The search direction is determined according to Polak Ribiére [52] updates as

$$p_k = -g_k + \beta_k p_{k-1} \tag{4.4}$$

where β_k is defined as

$$\beta_k = \frac{\Delta g_{k-1}^T g_k}{g_{k-1}^T g_{k-1}} \tag{4.5}$$

where g_k represents the current gradient and g_{k-1} represents the previous gradient.

The output vector , O_α, (4.7) represents the probability distribution modeled by the expert or the classifier for each test case, and this is achieved by using a softmax activation function at the output layer, according to the following formula:

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} \tag{4.6}$$

where x_i is the input to the softmax function and C is the number of classes.

O_α is in the form of a $C \times 1$ vector and each entry represents the conditional probability of the label being assigned to the input sample o.

$$O_\alpha = [p_\alpha(1|o)\ p_\alpha(2|o)\ \ldots\ p_\alpha(C|o)]^T \tag{4.7}$$

where $\alpha \in \{1, 2\}$ is used to index each respective classifier.

4.3 Sensor Fusion

This work makes use of decision level fusion using Logarithmic Opinion Pool (LOP) for merging the model of the probability distributions produced by the individual classifiers. The reasons for employing decision level fusion include its flexibility to incorporate more sensors, classifiers can be trained independently from each other and a final decision can be made based on the trusts level of the classifiers. Moreover, if a sensor stops working due to any reason, the algorithm can be modified to rely on other available sensor(s) without breaking down.

4.3.1 Logarithmic Opinion Pool

LOP is employed to merge the individual posterior probabilities of the classifiers and estimate the global membership function

$$P(c|o) = \frac{1}{Z_{LOP}(o)} \prod_{\alpha} p_{\alpha}(c|o)^{w_{\alpha}} \tag{4.8}$$

where $\sum_{\alpha} w_{\alpha} = 1$, $w_{\alpha} \geq 0$ represents our confidence for each classifier α, $c \in [1, 2, ..., C]$ represents a class label, and a uniform distribution is assumed when fusing the sensors, i.e. $w_1 = w_2 = \frac{1}{2}$.

$Z_{LOP}(o)$ is a normalizing constant, defined as

$$Z_{LOP}(o) = \sum_{c} \prod_{\alpha} p_{\alpha}(c|o)^{w_{\alpha}} \tag{4.9}$$

however, it can be omitted to achieve computational efficiency.

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

$$Label = \underset{c=1...C}{\operatorname{argmax}}\ P(c|o) \tag{4.10}$$

According to Smith et al. [40], for LOP to model the true underlying conditional distribution effectively, the individual probabilities, p_{α}, should model the true underlying probabilities well, yet should be diverse.

Chapter 5

Experimental Results

5.1 Comparison with State-of-the-Art Methods

Table 5.1 provides a comparison of the results obtained with the proposed algorithm and compares them with the state-of-the-art results obtained on UTD-MHAD [17]. It shows the accuracies obtained from the classification of skeletal data alone, inertial data alone, and their fusion.

To test and compare the performance of the proposed algorithm, 8-fold cross-validation is performed, as in [53] and [15], by training the respective classifiers on 7 subjects and testing on the left out subject. This procedure has been repeated for every subject in turn, and the final accuracy is the average accuracy of all the 8 subjects.

TABLE 5.1: Recognition accuracies for subject-generic experiment. (W: Window length, λ: Dimensionality reduction constant)

Algorithm	Skeletal Accuracy	Inertial Accuracy	Fusion Accuracy
Chen et al. [15]	**74.7%**	76.4%	91.5%
Proposed Algorithm	72.0%	**88.5%**	**97.3%**

This implementation achieves slightly higher accuracy than previous implementation in [53] by 2.3%, while it beats the results obtained in [15] by 5.8%.

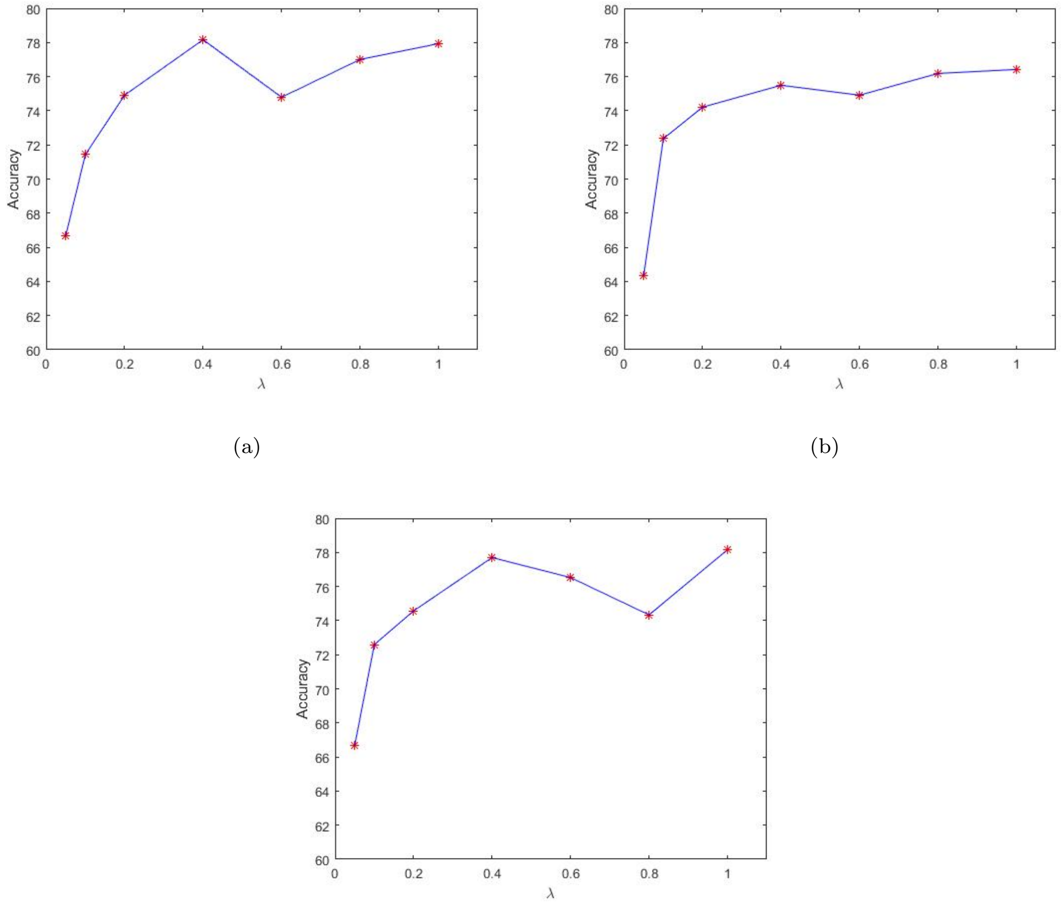
5.2 Performance Evaluation

5.2.1 Influence of λ and W

Figure 5.1 to Figure 5.3 depict the experimental results, and show the 8-fold cross-validation accuracy, with varying λ from (4.2) and window length, W, from the input to the inertial sensor classifier.

From Figure 5.1, it can be observed that increasing λ increases the 8-fold cross-validation accuracy. The reason for this can be allocated to the fact that decreasing λ results in a loss of information, and since the skeletal data comprises of the location of 20 joints (as opposed to one inertial sensor), significant information is lost when the dimension of the feature vector (the input to the neural network classifier for skeletal data) is reduced.

It is important to note the highest accuracy obtained from the skeletal data does not guarantee the highest fusion accuracy. The reason for this can be associated to the fact that the percentages obtained from the neural network classifier, i.e. p_α, from (4.7) for the skeletal case are low for a specific class, as compared to the ones obtained when using inertial data alone, and thus have a low contribution to the overall fusion accuracy. This can further be explained by the fact that it is very hard for a human to perform an action using only certain joints, while keeping the other joints stationary. Moreover, this can also be termed as one of the major sources of noise. In this work since we use one inertial sensor, the classification accuracy obtained from the inertial data does not face this problem, however, since Kinect tracks 20 joints, this problem is persistent.



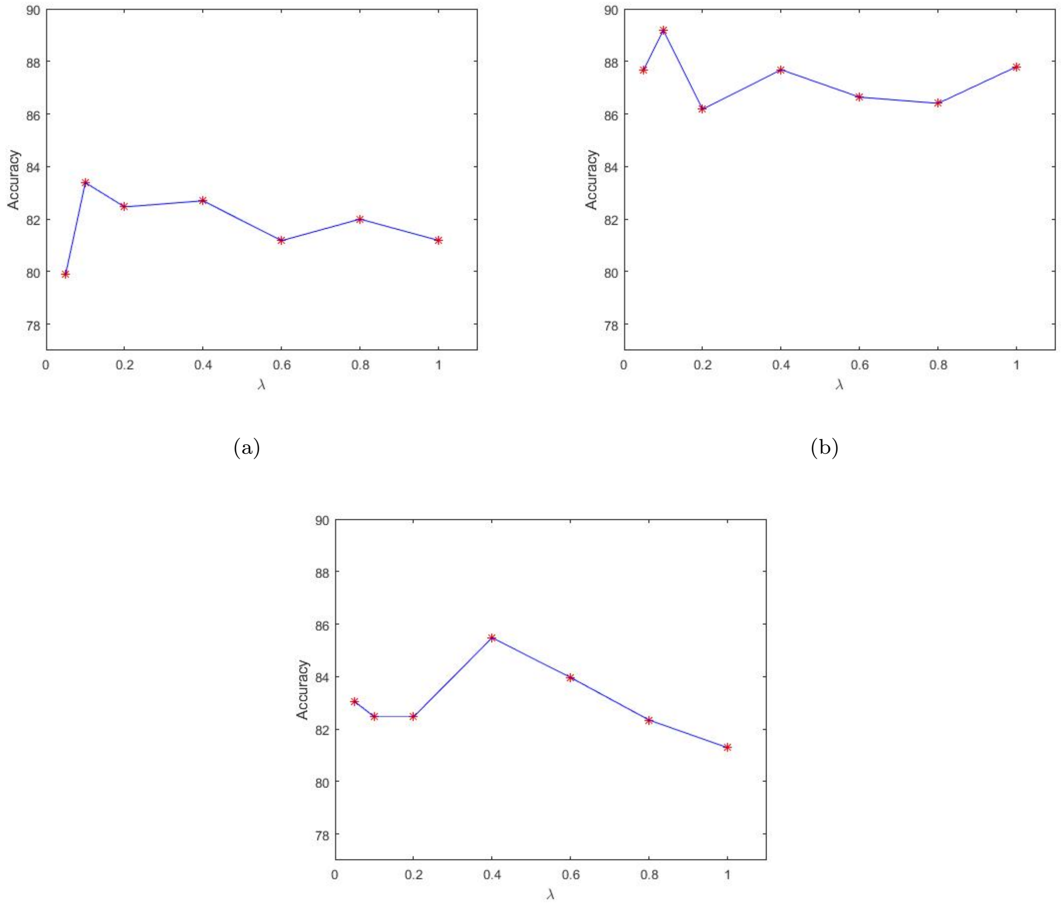
(a)

(b)

(c)

Figure 5.1: Plot of accuracy against λ for (a) $W = 3$, (b) W= 17 and (c) $W = 35$ when using skeletal data alone

Figure 5.2 represents the 8-fold cross-validation accuracy when using inertial data alone. This figure follows the general trend observed in the case of data fusion (Figure 5.3). The reason can be assigned to the fact that the percentages, p_α, obtained from the classifier when using inertial data alone are higher than when using skeletal data alone, and thus have a higher contribution the overall fusion accuracy, as mentioned earlier. However, due to the skeletal data providing valuable information, the fusion accuracy is higher than when using each of the sensors alone.



(a)

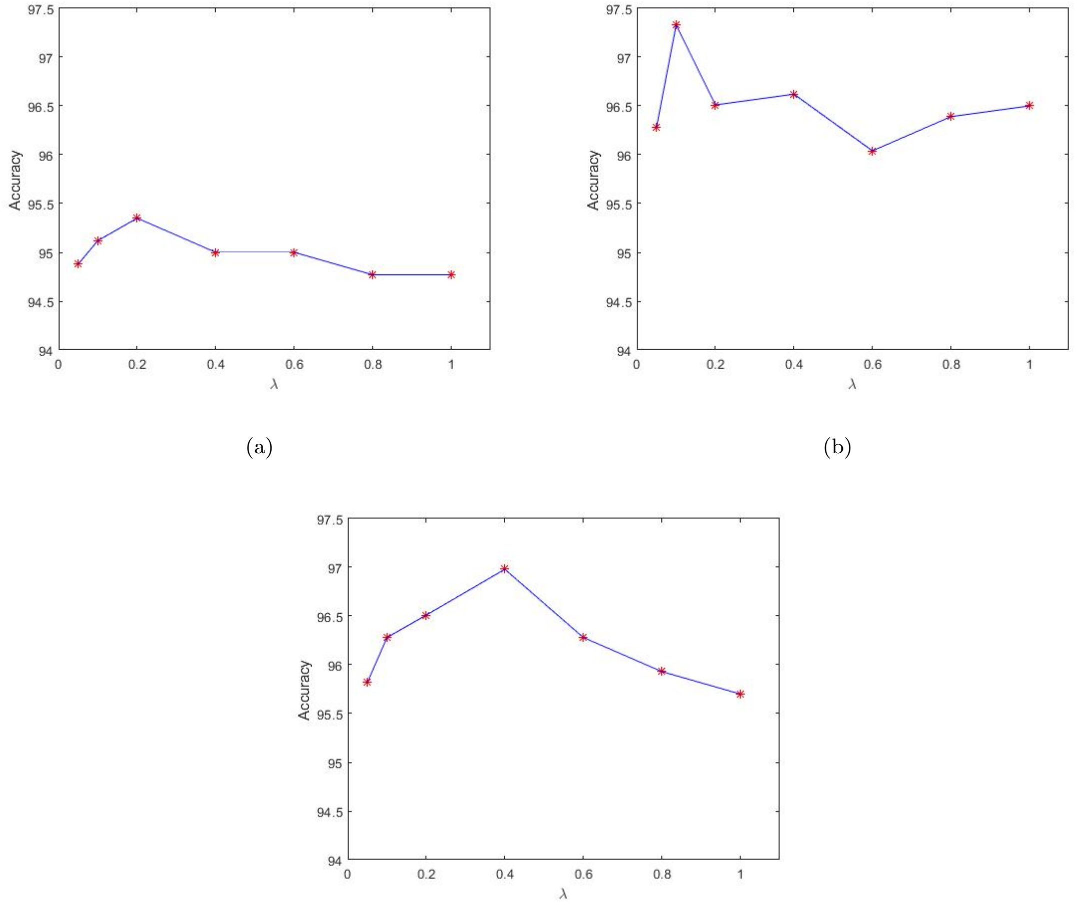
(b)

(c)

FIGURE 5.2: Plot of accuracy against λ for (a) $W = 3$, (b) W= 17 and (c) $W = 35$ when using inertial data alone

Increasing W from 3 to 17 increases the accuracy of the classification in Figure 5.2. The reason for this can be accounted to the fact of over-fitting. Increasing W results in less number of windows and hence a feature vector of a lower dimension, which does not over-fit to the training data. However, it should also be noted that decreasing the dimensionality a lot can result in a loss of information, and hence the classification accuracy can be decreased, as observed when changing W from 17 to 35.

Figure 5.3 represents the 8-fold cross-validation accuracy for the decision level case. From the figure, it can be observed that increasing λ increases the accuracy up to a point, and then the accuracy is decreased.



(a)

(b)

(c)

FIGURE 5.3: Plot of accuracy against λ for (a) W = 3, (b) W= 17 and (c) W = 35 when using decision level fusion

5.2.2 Comparison of Subject-Based Accuracies

Figure 5.4 to Figure 5.6 depict the accuracy of the proposed algorithm with respect to each sensor used along with the fusion accuracy, for each of the 8 subjects. These charts show the highest (Figure 5.4), the intermediate (Figure 5.5) and the lowest accuracy (Figure 5.6). These figures represent the accuracy obtained from having different λ and W values.

Figure 5.4 represents the case which achieved the highest 8-fold cross-validation accuracy of 97.3%. In terms of each subject, fusion accuracy of subject 8 was the lowest (93.5%), while that of subject 1, subject 2 and subject 3 were the highest and similar, around 99.1%. Moreover, inertial data achieved a higher recognition performance than skeletal data for all the subjects.

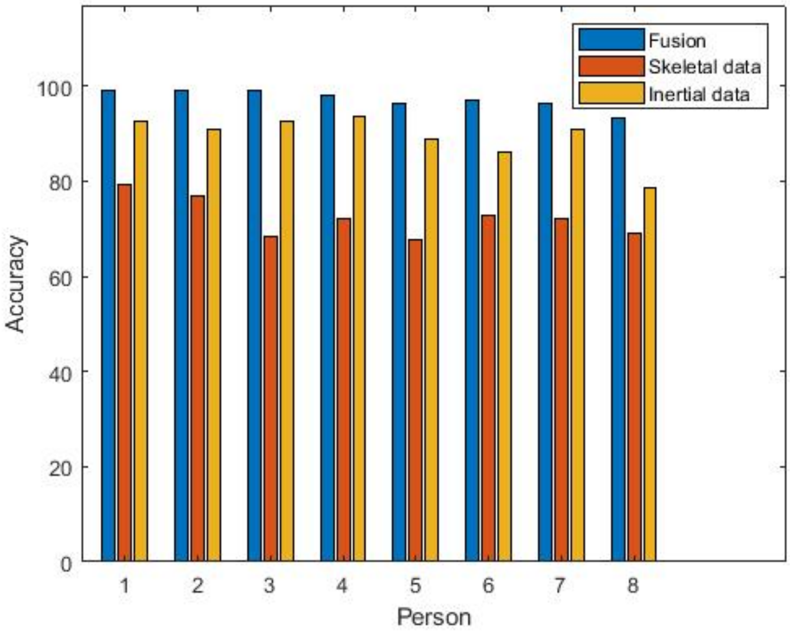


Figure 5.4: Subject-based accuracies for the case that achieved the highest fusion accuracy of 97.3% at $\lambda = 0.1$ and $W = 17$

Figure 5.5 represents the case that achieved an intermediate accuracy of 95.7%. In this case, subject 8 achieved the lowest accuracy of 90.7%, while subject 2 achieved the highest accuracy of 98.1%. The skeletal data had a higher classification accuracy than the inertial data for subjects 3, 4 and 8.

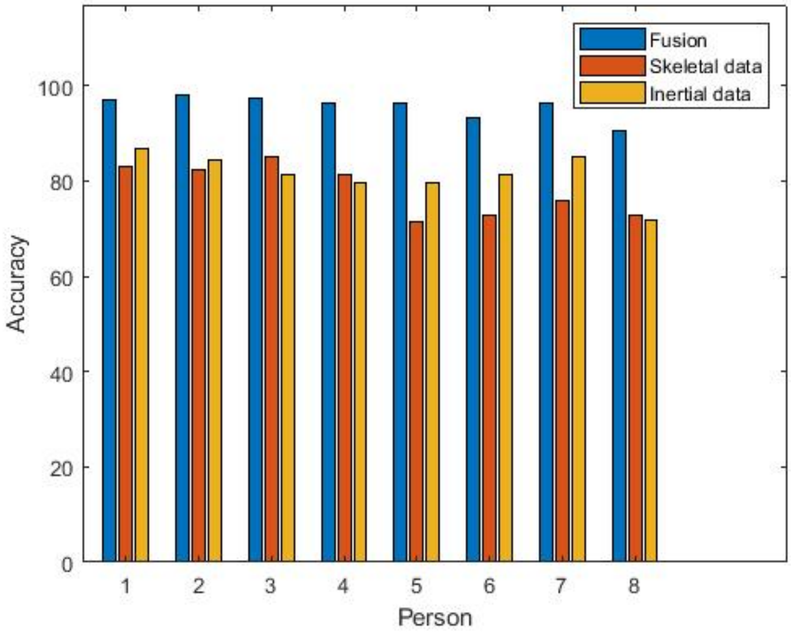


Figure 5.5: Subject-based accuracies for the case that achieved the intermediate fusion accuracy of 95.7% at $\lambda = 1$ and $W = 35$

Figure 5.6 represents the case that achieved the lowest 8-fold cross-validation accuracy of 94.8%. In this case, subject 2 and subject 3 achieved the highest fusion accuracy of 99.1%, while subject 8 achieved the lowest fusion accuracy of 85.0%. In terms of the individual classifiers, skeletal data obtained a higher accuracy for subject 1, while both skeletal and inertial data achieved an equal accuracy of 75.7% in the case of subject 8.

From the bar charts, it can be seen that the fusion accuracy is always higher than that of using each sensor alone. This is due to the fact that each sensor has its own limitations, and the redundancies encountered when using a sensor of a particular modality can be overcome by a sensor of a different modality, and vice versa.

An interesting observation is that the inertial sensor measurements, in a majority of the cases, achieved a higher accuracy than the skeletal measurements. This is due to the noise that is caused by the movement of joints in the case of the skeletal data.

An example could be that the subject moves their legs while performing the action 'swipe right' with their right arm. Since only one inertial sensor is used, it does not encounter this type of noise.

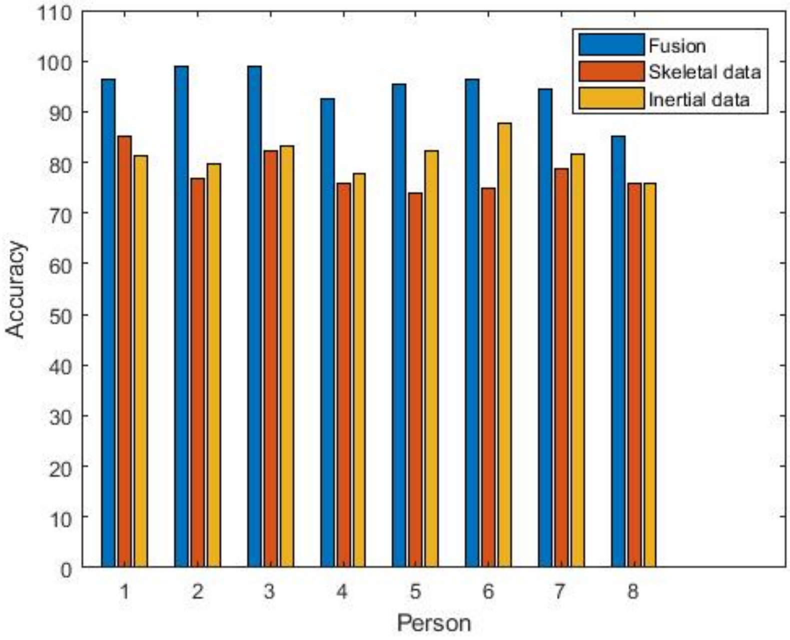


Figure 5.6: Subject-based accuracies for the case that achieved the lowest fusion accuracy of 94.8% at $\lambda = 1$ and $W = 3$

Lastly, each subject has different body dimensions and the way they perform a particular action is almost unique. This reason can be held accountable for a difference in accuracies for different subjects. It can be seen that subject 8, in almost all of the cases, achieved a lower accuracy than the rest.

5.2.3 Action-Based Recognition Performance

Figure 5.7 to Figure 5.9 represent the confusion matrices obtained with using different values of λ and W. These confusion matrices represent the action based performance of the proposed algorithm, i.e. the recognition accuracies for each action.

Figure 5.7 represents the confusion matrices obtained when using skeletal data, inertial data, and fusion of both, for the case that achieved the highest fusion accuracy of 97.3%. In this case λ was 0.1, while W was set to be 17.

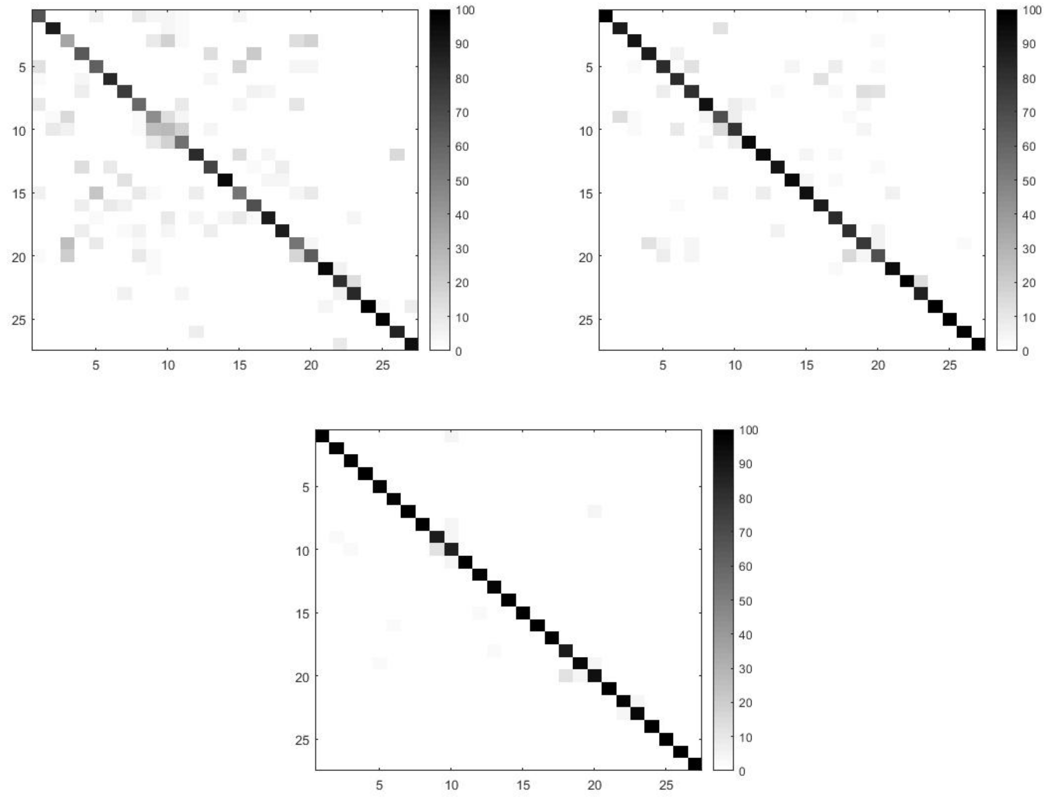


FIGURE 5.7: Confusion Matrix of Skeletal (top left), Inertial (top right) and Fusion (bottom), for the case that achieved the highest fusion accuracy of 97.3% at $\lambda = 0.1$ and $W = 17$

Figure 5.8 represents the confusion matrices obtained when using skeletal data, inertial data, and fusion of both, for the case that achieved an intermediate fusion accuracy of 95.7%. In this case, λ was set to be 1 and W was set to be 35.

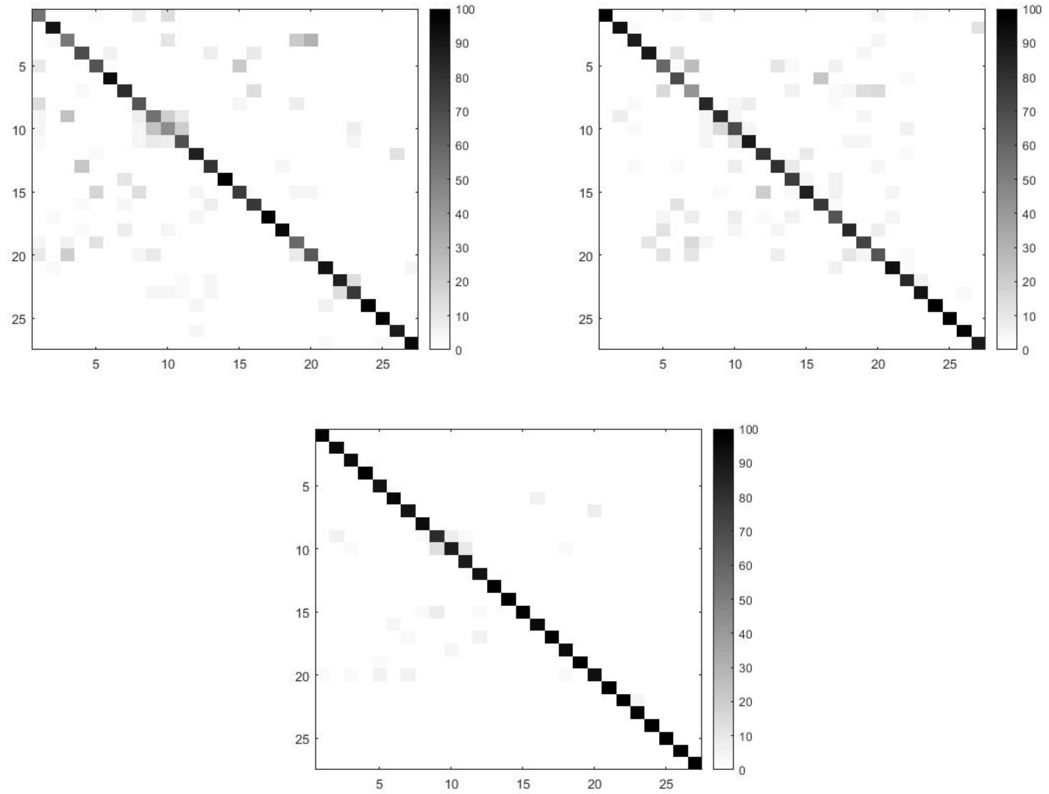


Figure 5.8: Confusion Matrix of Skeletal (top left), Inertial (top right) and Fusion (bottom), for the case that achieved the intermediate fusion accuracy of 95.7% at $\lambda = 1$ and $W = 35$

Figure 5.9 represents the case that achieved the lowest fusion accuracy of 94.7%. In this case λ was set to be 1 and W was set to be 3.

The mis-classifications when using skeletal data alone were higher than when using inertial data alone. Moreover, the cases where both skeletal data and inertial data had mis-classifications resulted in the fusion case also having mis-classifications, such as drawing circle clockwise (activity 9) and drawing circle counter-clockwise (activity 10).

Looking at the confusion matrices, it can be seen that for every case, the number of mis-classifications after the fusion of the two sensors are much less than the number of mis-classifications for each individual sensor.

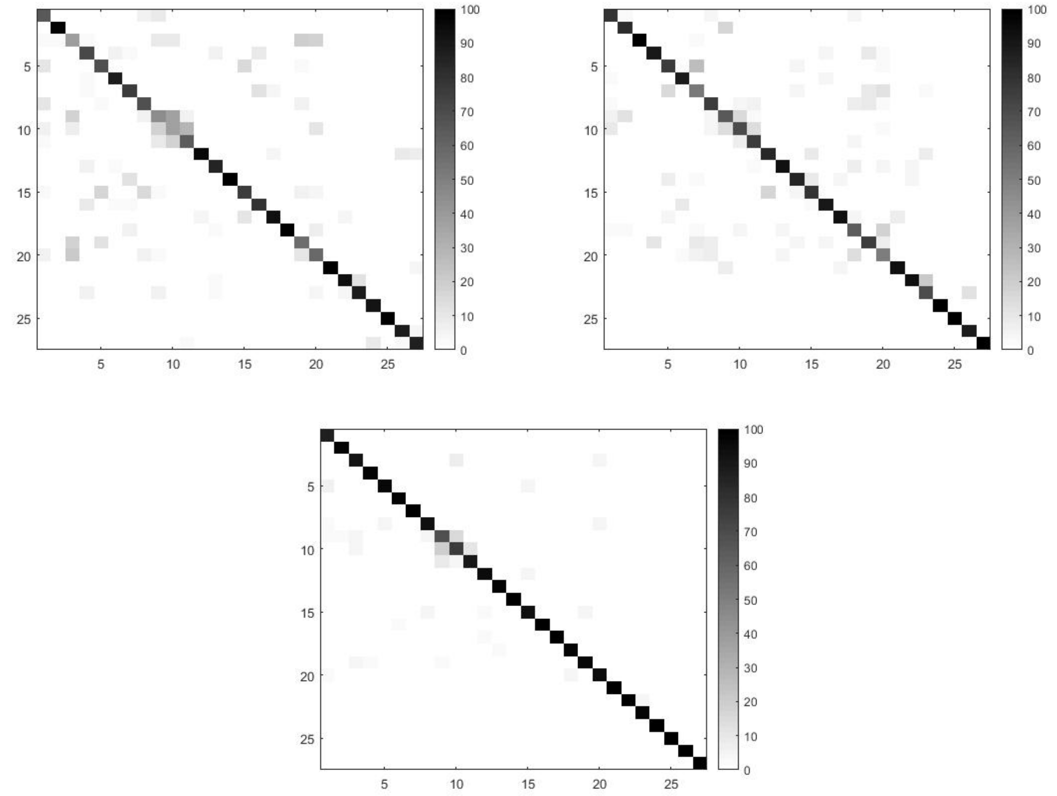


Figure 5.9: Confusion Matrix of Skeletal (top left), Inertial (top right) and Fusion (bottom), for the case that achieved the lowest fusion accuracy of 94.8% at $\lambda = 1$ and $W = 3$

Actions such as drawing circle in a clockwise direction (action 9) and drawing the same circle in a counter-clockwise direction (action 10), jogging (action 22) and walking in place (action 23), and throw (action 5) and catch (action 20) had a lot of mis-classifications amongst them due to them being of similar pattern. This observation was made across different values of λ and W.

Chapter 6

Conclusion and Future Work

In this study, a HAR system based on the idea of sensor fusion has been presented. The main problem with using one modality of the sensor arise from the limitations of that particular sensor, and hence this results in a lower performance of action recognition. For this purpose, this work incorporates data from two different types of commercially available sensors, mainly an RGB-D camera and a wearable inertial sensor. The proposed algorithm classifies the data acquired from the different sensors into one of the labeled action classes. To fuse these individual classifications and obtain a final classification of the performed action, this work makes use of decision level fusion by estimating the individual underlying probability distributions. For this part, LOP is utilized as the fusion algorithm. The algorithm has been tested on UTD-Multimodal Human Action dataset, as it contains actions involving the movement of different joints in the case of the depth sensor and just one inertial placed on the different parts of the body to classify different actions. The results show that the resulting classification accuracy after the fusion operation is performed is much higher than using each of the individual sensors alone.

One of the main problems in the field to HAR is the way an action is performed, and the speed with which it is performed. In other words, actions performed at different times by the same or different people will be performed differently and hence classifying these actions can be a difficult task. This work employs bi-cubic interpolation to reduce the temporal variations between the performed actions as a pre-processing step. Experiments have been conducted with different values of the scaling constant, i.e. λ, for dimensionality reduction for the case of depth sensor data, and the data is divided into window length, W, for the case inertial sensor data. The results show the effect of these parameters on the obtained accuracy.

This thesis uses a single hidden-layer neural network as the classification algorithm. The probability distributions of the performed action is obtained using a softmax function in the output layer of the network. The reason for choosing such a simple structure is to provide low training times and real time working capabilities of the algorithm. The reason for choosing a neural network as the classification algorithm is based on its success in a number of machine learning problems present in the literature, and the results in this thesis further show its capabilities. It is observed that using a neural network to classify each sensor's data gave a 8-fold fusion classification accuracy of 97.3%.

A critical reader of this thesis may question the use of two sensors and argue that utilizing more sensors may result in a much higher fusion accuracy. This is taken as a future work, to evaluate the performance of the proposed algorithm when additional sensors of the same or different modalities are added. Moreover, it will be interesting to find the saturation point, if any, after which adding more sensors does not further increase the classification accuracy and might even end up decreasing the accuracy due to the idea of the curse of dimensionality.

Bibliography

[1] A. Kläser, *Learning human actions in video.* PhD thesis, PhD thesis, Université de Grenoble, 2010.

[2] A. Ahmadi, F. Destelle, D. Monaghan, K. Moran, N. E. O'Connor, L. Unzueta, and M. T. Linaza, "Human gait monitoring using body-worn inertial sensors and kinematic modelling," in *SENSORS, 2015 IEEE*, pp. 1–4, IEEE, 2015.

[3] S. Nazir, M. H. Yousaf, J.-C. Nebel, and S. A. Velastin, "A bag of expression framework for improved human action recognition," *Pattern Recognition Letters*, vol. 103, pp. 39–45, 2018.

[4] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1293–1301, 2015.

[5] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE transactions on cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[6] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[7] Y. Du, W. Wang, and L. Wang, "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] D. C. Luvizon, H. Tabia, and D. Picard, "Learning features combination for human action recognition from skeleton sequences," *Pattern Recognition Letters*, vol. 99, pp. 13–20, 2017.

[9] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24–38, 2014.

[10] S. B. Qaisar, S. Imtiaz, F. Faruq, A. Jamal, W. Iqbal, P. Glazier, and S. Lee, "A hidden markov model for detection & classification of arm action in cricket using wearable sensors.," *J. Mobile Multimedia*, vol. 9, no. 1&2, pp. 128–144, 2013.

[11] L. Guo and S. Xiong, "Accuracy of base of support using an inertial sensor based motion capture system," *Sensors*, vol. 17, no. 9, p. 2091, 2017.

[12] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE transactions on information technology in biomedicine*, vol. 12, no. 1, pp. 20–26, 2008.

[13] J. Gellaerts, E. Bogdanov, F. Dadashi, and B. Mariani, "In-field validation of an inertial sensor-based system for movement analysis and classification in ski mountaineering," *Sensors*, vol. 18, no. 3, p. 885, 2018.

[14] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pp. 53–60, IEEE, 2013.

[15] C. Chen, R. Jafari, and N. Kehtarnavaz, “A real-time human action recognition system using depth and inertial sensor fusion,” *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773–781, 2016.

[16] Z. Wang, Z. Yang, and T. Dong, “A review of wearable technologies for elderly care that can accurately track indoor position, recognize physical activities and monitor vital signs in real time,” *Sensors*, vol. 17, no. 2, p. 341, 2017.

[17] 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 *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 168–172, IEEE, 2015.

[18] H.-H. Hsu, Y. Chiou, Y.-R. Chen, and T. K. Shih, “Using kinect to develop a smart meeting room,” in *Network-Based Information Systems (NBiS), 2013 16th International Conference on*, pp. 410–415, IEEE, 2013.

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

[20] S. Stein and S. J. McKenna, “Combining embedded accelerometers with computer vision for recognizing food preparation activities,” in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 729–738, ACM, 2013.

[21] F. Dellaert, S. M. Seitz, C. E. Thorpe, and S. Thrun, “Structure from motion without correspondence,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2, pp. 557–564, IEEE, 2000.

[22] B. D. Lucas, T. Kanade, *et al.*, “An iterative image registration technique with an application to stereo vision,” 1981.

[23] B. C. Ustundag and M. Unel, "Human action recognition using histograms of oriented optical flows from depth," in *Advances in Visual Computing*, pp. 629–638, Springer International Publishing, 2014.

[24] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1995–2006, 2013.

[25] Y. Wang, X. Jiang, R. Cao, and X. Wang, "Robust indoor human activity recognition using wireless signals," *Sensors*, vol. 15, no. 7, pp. 17195–17208, 2015.

[26] T. Hachaj, M. Piekarczyk, and M. R. Ogiela, "Human actions analysis: Templates generation, matching and visualization applied to motion capture of highly-skilled karate athletes," *Sensors*, vol. 17, no. 11, p. 2590, 2017.

[27] A. A. Chaaraoui, J. R. Padilla-López, F. J. Ferrández-Pastor, M. Nieto-Hidalgo, and F. Flórez-Revuelta, "A vision-based system for intelligent monitoring: human behaviour analysis and privacy by context," *Sensors*, vol. 14, no. 5, pp. 8895–8925, 2014.

[28] A. Manzi, P. Dario, and F. Cavallo, "A human activity recognition system based on dynamic clustering of skeleton data," *Sensors*, vol. 17, no. 5, p. 1100, 2017.

[29] Y. Ming, G. Wang, and C. Fan, "Uniform local binary pattern based texture-edge feature for 3d human behavior recognition," *PloS one*, vol. 10, no. 5, p. e0124640, 2015.

[30] K. Altun, B. Barshan, and O. Tunçel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," *Pattern Recognition*, vol. 43, no. 10, pp. 3605–3620, 2010.

[31] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31314–31338, 2015.

[32] H. Gjoreski, M. Luštrek, and M. Gams, "Accelerometer placement for posture recognition and fall detection," in *2011 Seventh International Conference on Intelligent Environments*, pp. 47–54, IEEE, 2011.

[33] S. Pirttikangas, K. Fujinami, and T. Nakajima, "Feature selection and activity recognition from wearable sensors," in *International symposium on ubiquitious computing systems*, pp. 516–527, Springer, 2006.

[34] D. C. Dang and Y. S. Suh, "Walking distance estimation using walking canes with inertial sensors," *Sensors*, vol. 18, no. 1, p. 230, 2018.

[35] J. Lv and S. Xiao, "Real-time 3d motion recognition of skeleton animation data stream," *International Journal of Machine Learning and Computing*, vol. 3, no. 5, p. 430, 2013.

[36] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer methods and programs in biomedicine*, vol. 117, no. 3, pp. 489–501, 2014.

[37] S. Gasparrini, E. Cippitelli, S. Spinsante, and E. Gambi, "A depth-based fall detection system using a kinect® sensor," *Sensors*, vol. 14, no. 2, pp. 2756–2775, 2014.

[38] S. Ruffieux, D. Lalanne, and E. Mugellini, "Chairgest: a challenge for multimodal mid-air gesture recognition for close hci," in *Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 483–488, ACM, 2013.

[39] L. Sun and K. Aizawa, “Action recognition using invariant features under unexampled viewing conditions,” in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 389–392, ACM, 2013.

[40] A. Smith, T. Cohn, and M. Osborne, “Logarithmic opinion pools for conditional random fields,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 18–25, Association for Computational Linguistics, 2005.

[41] A. Mackiewicz and W. Ratajczak, “Principal components analysis (pca),” *Computers and Geosciences*, vol. 19, pp. 303–342, 1993.

[42] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[43] M. Müller, “Dynamic time warping,” *Information retrieval for music and motion*, pp. 69–84, 2007.

[44] D. Han, “Comparison of commonly used image interpolation methods,” *ICCSEE, Hangzhou, China*, pp. 1556–1559, 2013.

[45] S. Orfanidis, “Introduction to signal processing, prentice hall,” *Englewood Cliffs, NJ*, 1996.

[46] R. W. Schafer, “What is a savitzky-golay filter?[lecture notes],” *IEEE Signal processing magazine*, vol. 28, no. 4, pp. 111–117, 2011.

[47] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural networks*, vol. 61, pp. 85–117, 2015.

[48] P. Sermanet, S. Chintala, and Y. LeCun, “Convolutional neural networks applied to house numbers digit classification,” in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 3288–3291, IEEE, 2012.

[49] S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L. Xiang, "A leaf recognition algorithm for plant classification using probabilistic neural network," in *Signal Processing and Information Technology, 2007 IEEE International Symposium on*, pp. 11–16, IEEE, 2007.

[50] X. Peng, N. Ratha, and S. Pankanti, "Learning face recognition from limited training data using deep neural networks," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 1442–1447, IEEE, 2016.

[51] D. Eck and J. Schmidhuber, "A first look at music composition using lstm recurrent neural networks," *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*, vol. 103, 2002.

[52] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. De Jesús, *Neural network design*, vol. 20. Pws Pub. Boston, 1996.

[53] Z. Fuad and M. Unel, "Human action recognition using fusion of depth and inertial sensors," in *Lecture Notes in Computer Science*, vol. 10882, pp. 373–380, Springer International Publishing, 2018.