KE 5208

SENSE MAKING & INSIGHT DISCOVERY

PROJECT REPORT

HUMAN ACTION RECOGNITION USING MULTIMODAL DEPTH CAMERA AND INERTIAL SENSOR

TEAM MEMBERS

Huang FuXing	A0163461J
Low Kang Jiang	A0074752B
Tey Peng Mok	A0163350N
Yeo Kun Song	A0035456E

1.0 PROJECT OBJECTIVE

The context of this project is to make sense of data from modality sensors for robust human action recognition. The UTD-MHAD human action datasets contains 27 actions performed by 8 subjects for 4 times each. Excluding 3 dataset which was corrupted. There are 861 data sequence recorded using wearable inertial sensor and depth camera. The objective of the project is to develop a robust algorithm to perform action classification using either single sensor or fusion of multiple sensors and evaluate the recognition performance for different cases.

2.0 TECHNOLOGY APPROACH

Development Platform	Jupyter Notebook(IPython)
Programming Language	Python 3.6
Cloud Computing Tool	Amazon Web Services(AWS)
Instance (Ubuntu)	m4.4xlarge (53.5 ECUs, 16 vCPUs, 2.4 GHz, Intel Xeon E5-2676v3, 64 GiB memory, EBS only)
Time Taken	Approximately 30 mins for the case of all 27 actions included
Classification Algorithm	Gaussian Naïve Bayes

Table 1: Overview of the technology tolls used and algorithm for UTD-MHAD actions recognition.

Table 1 shows an overview of the technology tool and classification algorithm used for this action recognition project. The algorithm has been developed in *Jupyter Notebook* using *Python 3.6*. As the classification modelling required intensive computation resources, we made used of AWS cloud computing services by created an *Ubuntu EC2* instance to provide necessary computation resource to train the model.

Generally, the project can be separated into a few different stages as follow:

1. Data processing
2. Classification model building using single sensor data
3. Classification model building using fusion of multimodal sensors data
4. Performance evaluation for all the models built

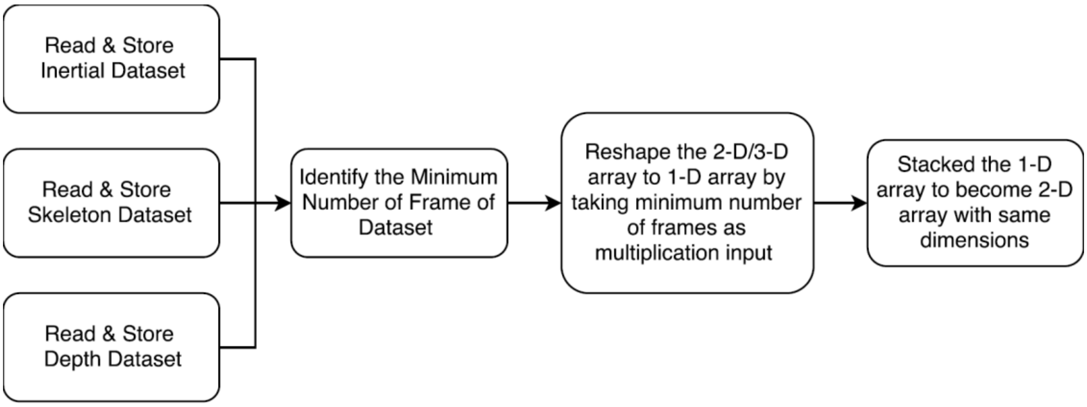

Read & Store Inertial Dataset
Read & Store Skeleton Dataset
Read & Store Depth Dataset
Identify the Minimum Number of Frame of Dataset
Reshape the 2-D/3-D array to 1-D array by taking minimum number of frames as multiplication input
Stacked the 1-D array to become 2-D array with same dimensions

Figure 1: Flow chart of the data processing step involved for three different type of sensors data

Figure 1 shows the flow chart of the data processing step involved for three different type of sensors data before feed into the classification machine learning algorithm. The process starts with the reading and storing and of all the 861 datasets. The minimum number of frame among the dataset for different sensors were identified to be used as an input parameter to reshape the 2-D/3D array into 1-D array. The reshaped 1-D arrays then have been stacked to become 2-D array to be fed into the machine learning algorithm. Table 2 shows the reshaped array for 3 different type of sensors data before feed into the Gaussian Naiye Bayes classification model.

Sensors Type	Reshaped Array(1-D)
Inertial	minimum frame among all inertial dataset*6
Skeleton	20*3*minimum frame among all skeleton dataset
Depth	240*320* minimum frame among all depth dataset

Table 2: Reshaped array for 3 different type of sensors data before fit into the Gaussian Naiye Bayes classification model

Why use minimum frame?

The approach of using minimum frame of different type of sensors to be used as multiplication input is due to several reasons. First, it is more robust if the model to be able to recognize different action with less but concise input data as it can provide the shortest delay or even real-time recognition. Second, the computing resource needed to build the classification model by using minimum frame approach is the much lesser and the model can be trained in the shortest time.

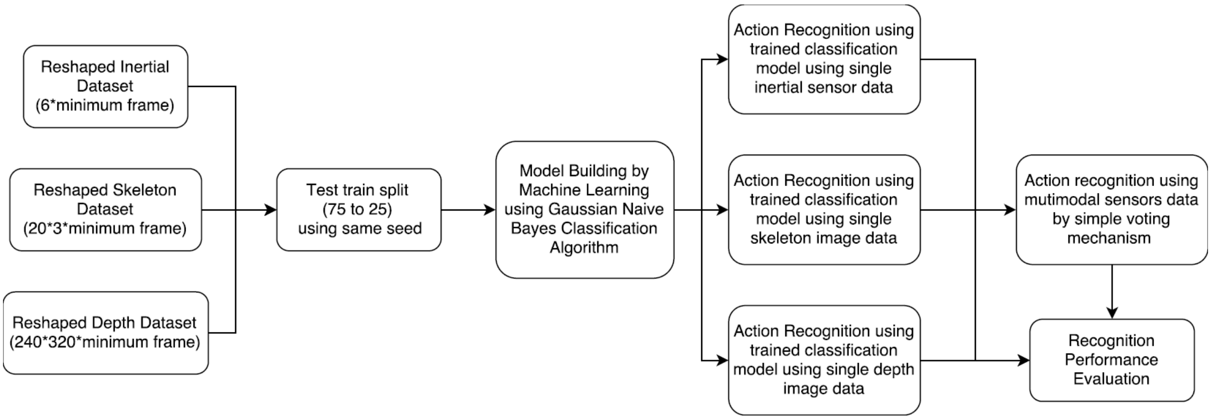

Reshaped Inertial Dataset (6*minimum frame)
Reshaped Skeleton Dataset (20*3*minimum frame)
Reshaped Depth Dataset (240*320*minimum frame)
Test train split (75 to 25) using same seed
Model Building by Machine Learning using Gaussian Naive Bayes Classification Algorithm
Action Recognition using trained classification model using single inertial sensor data
Action Recognition using trained classification model using single skeleton image data
Action Recognition using trained classification model using single depth image data
Action recognition using mutimodal sensors data by simple voting mechanism
Recognition Performance Evaluation

Figure 3: Flow chart for action recognition classification using single sensor model and fusion of multimodal sensors model

Figure 3 shows the classification flow chart for the action recognition project. Firstly, the reshaped dataset for different type of sensors have been split into training and testing set using the same seed in 75 to 25 proportion. Secondly, three models were built using the Gaussian Naïve Bayes classifier using AWS cloud computing *EC2* instance.

Then, the three classification models built are used to predict or recognize the action in test data and their performance has been evaluated using confusion matrix and in terms of classification accuracy. Different combinations of number of activities included has been used to compare the performance in terms of number of activities included. In the final stage of the project, the classification result of three sensors have been fused using simple voting mechanism to compare its performance to single sensor.

Why chose Gaussian Naïve Bayes Classifier?

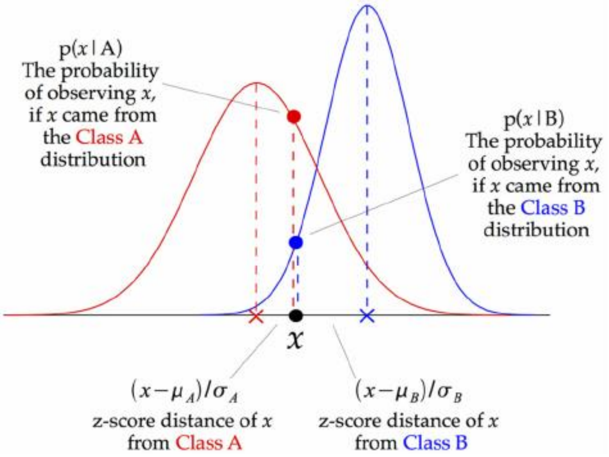


Figure 4: Gaussian Naïve Bayes Classifier

As mentioned, the classification model that is being used in the project is Gaussian Naïve Bayes(GNB) Classifier. As shown in Figure 4 for GNB classifier, the likelihood of the features is assumed to be Gaussian. The selection of GNB classifier is due to the following sense making assumptions:

1. Each action class is independent of the other known action class
2. Sensors reading for each action is considered as random continuous variable that follows a Normal(Gaussian) distribution given the class of action

In fact, according to the owner of the dataset, the dataset collected possesses large intra-class variations due to the following reasons:

1. Subjects performed the same action at different speed in different trials
2. Subjects had different heights
3. The same action was repeated in a natural way which made each trial slightly different

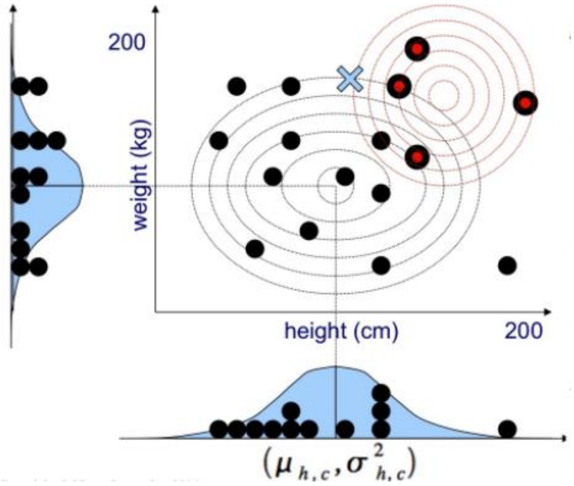


Figure 5: Human body proportion such as height and weight follow Gaussian distribution

As shown in Figure 5, human body proportion such as height, weight, arm length, etc. tends to follow the normal(Gaussian) distribution. This lead to the fact that the sensors reading collected from the subjects who performed the 27 actions also possesses a normal distribution due to the difference in the height and speed. Hence, heuristically we feel it made sense to use GNB Classifier for action recognition. Also, as compared to Neural Network Classifier and Random Forest Classifier, GNB classifier is much less intensive in terms of computational requirement.

3.0 PERFORMANCE EVALUATION

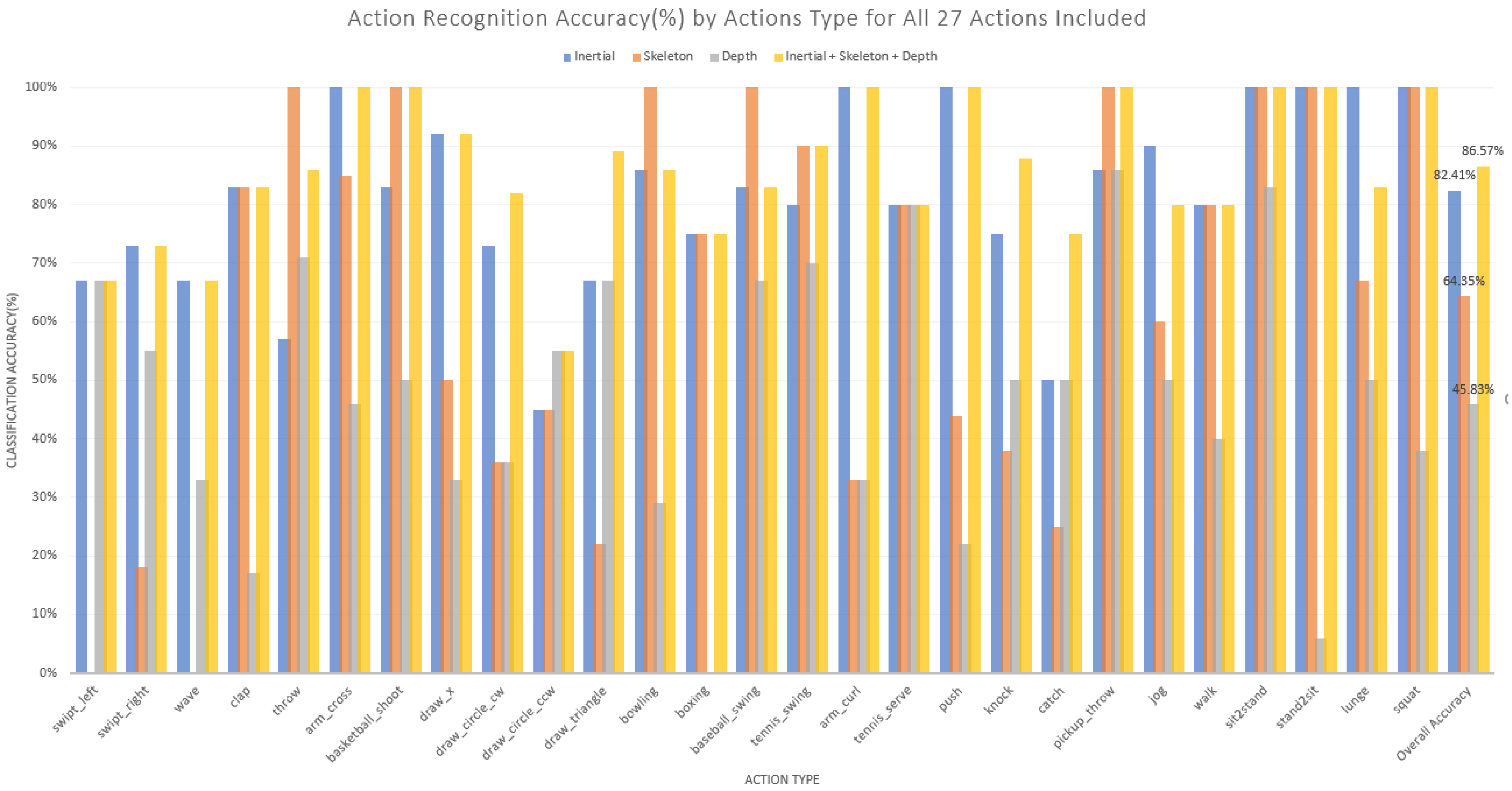


Figure 6: Action Recognition Accuracy (%) for 27 UTD-MHAD actions dataset

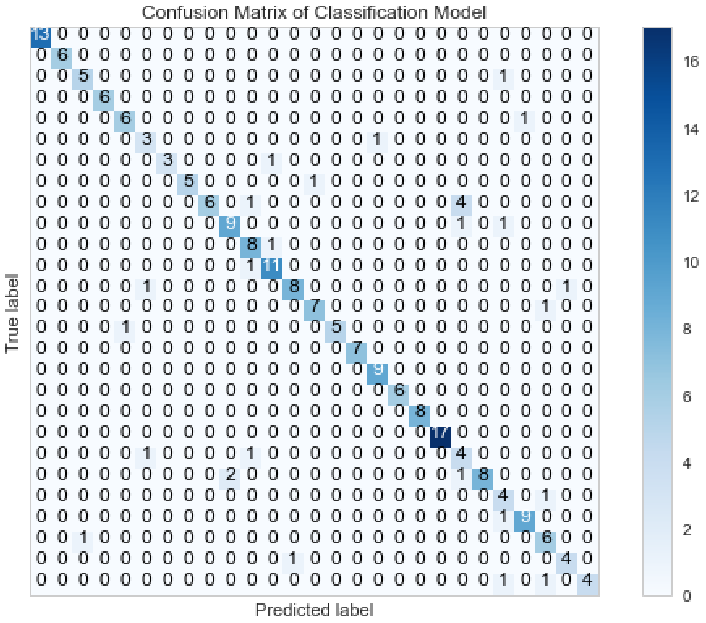


Figure 7: Confusion Matrix for 27 actions recognition using multimodal fusion sensors classifier

Figure 6 and 7 shows the action recognition accuracy and classification confusion matrix by action when all the 27 actions were included. The overall accuracies of classification using fusion of multimodal sensors improved the accuracy of the best single sensor(inertial) model by 4.16% which represent a 5.05% improvement. Given that the accuracy of single Inertial sensor is already high, it is important to notice that the overall recognition accuracy using multimodal sensors model has improved over 20% and 40% as compared to using Skeleton sensors and Depth sensors alone respectively.

Out of 27 actions, 4 actions have shown improvement in the recognition accuracy when using fusion of multimodal sensor data as compared to single sensor classification model. 18 actions have the same or no improvement in recognition accuracy as compared to the best single sensor classifier model among the three types of sensors. Interestingly, it is important to notice that the accuracies for 5 actions have dropped when using the fusion of multimodal sensors.

The accuracies of the action recognition for different number of actions included is shown in Appendices. As shows, the classification model achieved the highest accuracy when only 5 actions were included. The action recognition accuracy dropped when the number of action included increased as there are more action classes to be recognized. Nevertheless, the result shows that the fusion of three sensors increase the overall accuracy of action recognition for all the different settings of number of actions included.

Different settings in terms of different number of actions included have also been conducted and evaluated as shown in Appendices. Notice that the fusion of multimodal sensors has improved the accuracies of case of 5 actions included to 100% and the overall accuracies has been improved for other cases as well.

4.0 COMPARISON TO DATASET OWNER'S METHODOLOGY

The problem of data having different dimensions presents a stumbling block to training a classifier. In our case, the inconsistent dimension is the number of time frames.

For inertial data, we followed the owner’s method of binning the time frames into 6 intervals and extract the median, mean, variance and standard deviation of each bin as features. However, the result was only insignificantly better than our method.

For Depth data, the owner worked around this restriction by applying Depth Motion Map (DMM) as a feature representation method. The Depth data was re-represented in three - top, side and front - views and the difference between each frame is summed up into a single frame. Regardless of the number of time frames it has, each file is represented as three frames or DMMs. PCA was applied and the top n-components which accounted for 95% of variance are extracted as features for classifier training.

The Skeleton data is derived from the Depth data and we used it for the comparison between our method and the owner's.

Our results show that using DMM as a workaround for feature representation does give a better accuracy (69% compared to 64%). Using DMM+PCA for feature extraction also showed improved results (70%). This could be because information from all the frames are represented by the DMM. However, the improvement is not significant, and it also means that our method of trimming frames only results in a drop of 8.5% accuracy in exchange for speed. Considering real-time application, the speed gained should be more desirable than the accuracy lost.

It can be argued that instead of trimming the trailing frames, we should trim the front as there are probably more static frames where the subjects are standing still. Surprisingly, our results show that it gave a marked 13% decrease in accuracy (from 64% to 51%). This suggests that the front frames are more relevant for gesture prediction.

We have also tried trimming frames at equal interval. Intuitively, it is a better approach since the neighbour of each trimmed frame are intact and the general sequence of frames is not abruptly disrupted. However, we found that there is no marked improvement in overall accuracy compared to the presented method.

5.0 CONCLUSIONS

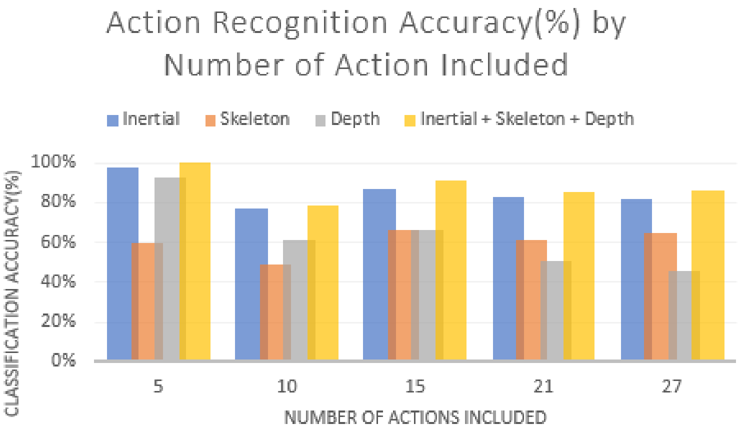
In conclusion, this report has provided a notable approach to improve the overall recognition accuracy of actions using fusion of multimodal sensors data. Gaussian Naïve Bayes Classifier has been proved to be able to classify the human actions with the assumption that the data captured by the sensors are normally distributed. Minimum frame approach has been used to standardize the 1-D array to be fit into the classification model to reduce the computational requirement which can lead to ensure less delay on the recognition process or even real-time action recognition. It is important that not all actions accuracy can be improved by the fusion of multimodal of sensors data. Indeed, the accuracy of some actions were dropped hence the accuracies are very much action dependent. The dataset owner's methods on feature extraction has a higher overall accuracy for single sensors, but the improvements are largely insignificant. However, these incremental improvements could be amplified through sensor fusion to give a significant overall improvement. Nevertheless, fusion mechanism proved to be able to improve the overall accuracy of human action recognition in line with the experimental result shown by the dataset owner.

6.0 REFERENCES

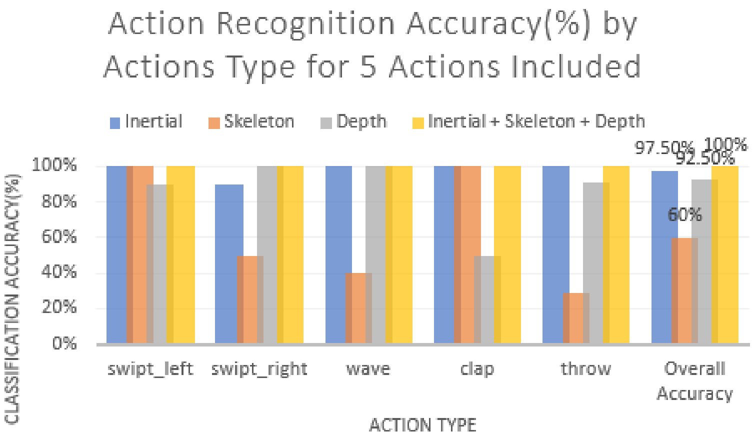
C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor", *Proceedings of IEEE International Conference on Image Processing*, Canada, September 2015.

C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition", N. Multimed Tools Appl (2017) 76: 4405, February 2017.

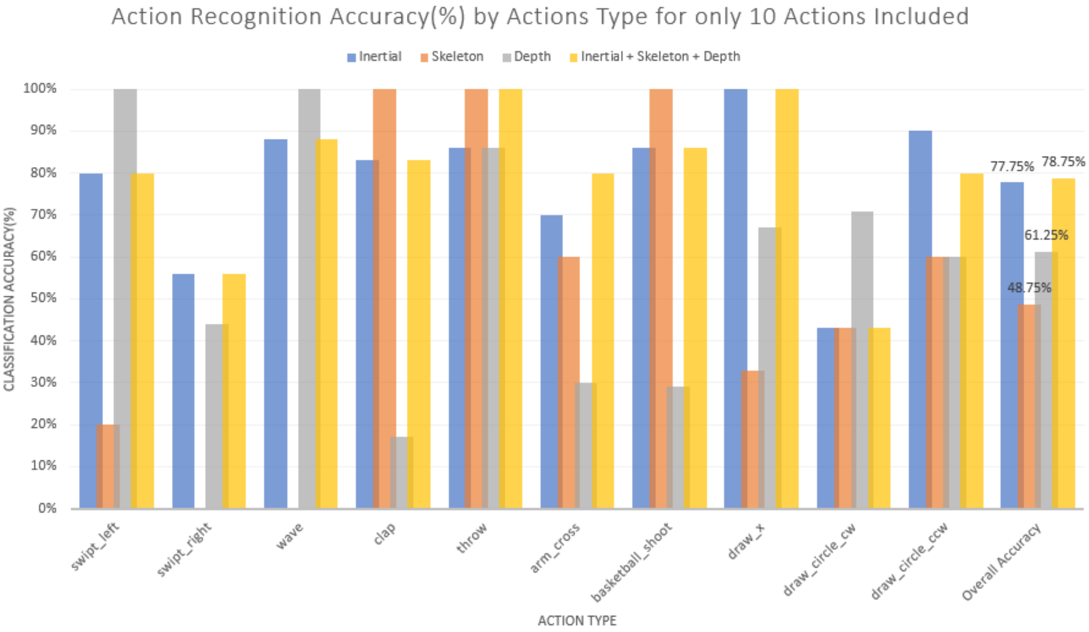
7.0 APPENDICES



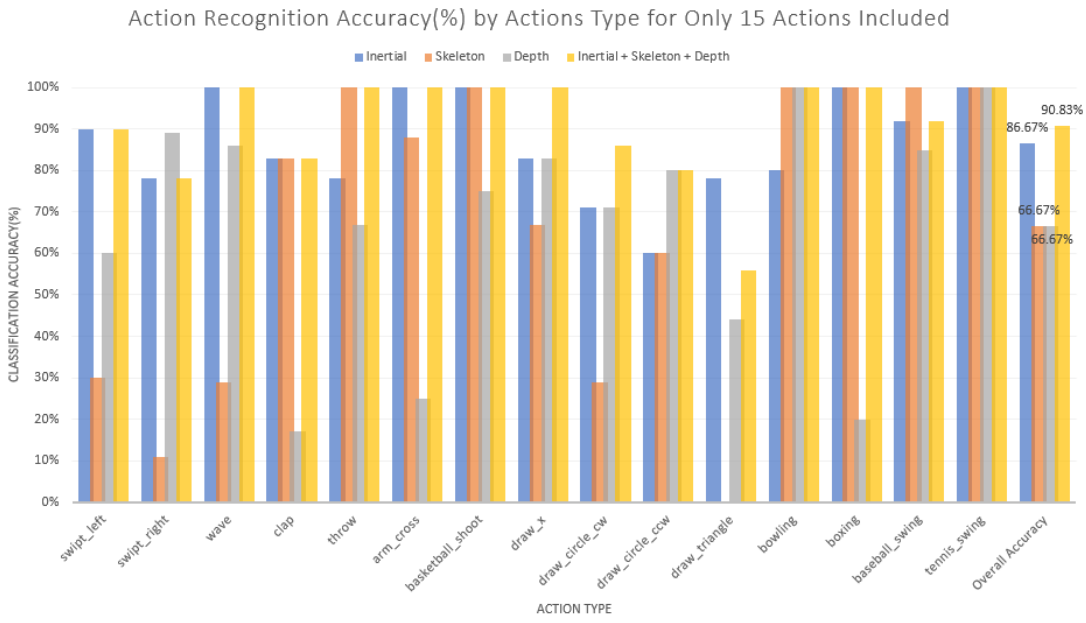
Action Recognition Accuracy(%) by Number of Action Included



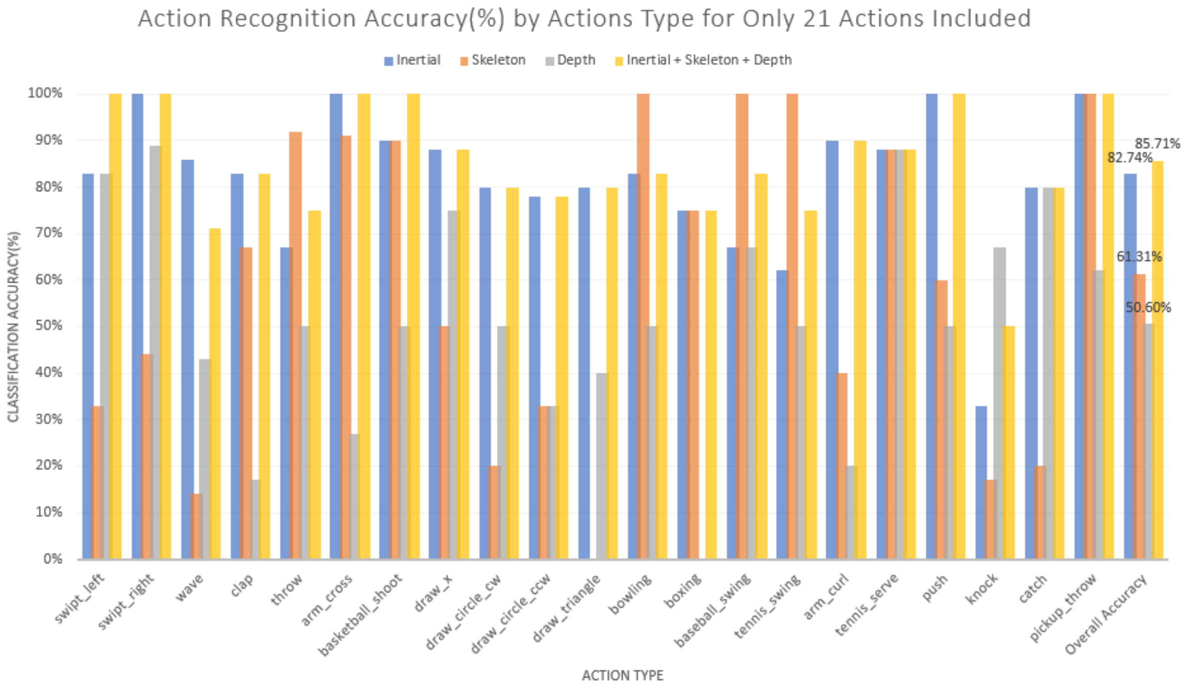
Action Recognition Accuracy(%) by Actions Type for 5 Actions Included



Action Recognition Accuracy(%) by Actions Type for only 10 Actions Included



Action Recognition Accuracy(%) by Actions Type for Only 15 Actions Included



Action Recognition Accuracy(%) by Actions Type for Only 21 Actions Included