HUMAN ACTION RECOGNITION USING RGB-DEPTH VIDEOS

A DISSERTATION

Submitted in partial fulfilment of the requirements for the award of degree of

Master of Technology

in

Computer Science and Engineering

by

AJAY YADAV
(14535002)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE 247 667 (INDIA)
May 27, 2016

DECLARATION

I declare that the work presented in this dissertation with title, **Human Action Recognition Using RGB-Depth Videos**, towards the fulfillment of the requirements for award of the degree of **Master of Technology** in **Computer Science & Engineering**, submitted to the **Department of Computer Science and Engineering**, **Indian Institute of Technology-Roorkee, India**, is an authentic record of my own work carried out during the period from **June 2015 to May 2016** under the guidance of **Dr. R. Balasubramanian**, Associate Professor, Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee.

The matter presented in this dissertation has not been submitted by me for the award of any other degree of this or any other institute.

Date:

Place: Roorkee **(Ajay Yadav)**

CERTIFICATE

This is to certify that the statement made by the candidate in the declaration is correct to the best of my knowledge and belief.

Date:
Place: Roorkee

Dr. R. Balasubramanian
Associate Professor
Department of Computer Science and Engineering
Indian Institute of Technology, Roorkee

ABSTRACT

Being able to detect and recognize human activities is essential for several applications, including personal assistive robotics.Many approaches have been discussed in the past. Normally 2D data has been used in past. But , nowadays due to availabilty of low cost 3D cameras like Kinect, it is easier to perform research on depth data. Mainly skeleton and depth data provides more reliable and accurate system.

In this Dissertation, a novel approach to detect the activities performed by a human has been implemented. This involves the extracting the frames from a given depth video and getting the skeleton of human in each frame using kinect camera. Simple skeleton feature are used, which are efficient and fast to classify the activities using multiclass svm.This approach gives a better accuracy in comparison to many approaches developed in the past.

ACKNOWLEDGEMENTS

I would never have been able to complete my dissertation without the guidance of my supervisor, help from friends, and support from my family.

I would like to express my deepest gratitude to my supervisor, **Dr. R. Balasubramanian**, for his excellent guidance, meaningful insights and moral support. He has been supportive since the day I began working on this dissertation and gave me the freedom I needed to explore this area of research on my own, while pointing me in the right direction in the times of need. His comprehensive knowledge in the area of Computer Vision and Machine Learning and hard working nature has been a constant source of inspiration.

I am also grateful to the **Dept. of Computer Science, IIT-Roorkee** for providing valuable resources to aid my research.

Finally, hearty thanks to my parents and siblings, who encouraged me in good times, and motivated me in the bad times, without which this dissertation would not have been possible.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Human Action Recognition is the process of labeling the sequence of activities performed in a video. A depth video is a sequence of images ,in which every pixel consists of normal 2-D information along with the depth information.RGBD sensors like Microsoft's Kinect device gives the Depth information in a video.More accuracy can be attained by using this depth information. A video is a sequence of images at a very small interval.Thus using depth information of all the joints can help in getting better accuracy.

Normally Human actions are unstructured and complex , therefore it is very difficult to recognize them easily.also every person has different traits and they perform same activity in a different way. There can be many actions like talking on phone , cooking , drinking water , brushing the teeth etc.

Previous ways of Human Action Recognition have been using 2-D videos or using RFID sensors .These methods were not so accurate because of occlusion, noises and image quality.Highest achieved accuracy using 2-D videos is up to 80%. while use of RFID tag is very complex and expensive because we have to fit RFID sensors on every part of the body.Any Human action can be divided into sequence of several activities.like talking on phone can be divided into three activities like picking phone , lifting it near your ear and then start talking .in these type of models , In which next activity depends on the current activity , we can use Markov model to recognize the actions easily,But Hidden Markov models assume that current activity depends only on previous activity, it doesn't consider the all previous activities.Hence improved model like hidden conditional random fields can be used to classify actions more accurately.But there is a problem with these models that they are generative and hence are not much efficient int terms of performance . Therefore, some fast and efficient approaches are required for real time detection of activities. Also, markov models are very much sensitive to the noise hence sometimes they can't fit the temporal pattern, therefore,they have less accuracy in comparison to new models like Fourier temporal pyramid [2] and dynamic time warping [10].

1.1 Types of features from a 3D video

Refer to the paper [12] , There are following features in a 3D video, which can be used to detect an actions.

1.1.1 3D silhouettes [1]

3D silhouettes of a RGBD image cane be extracted easily. In 3D silhouettes , bag of sample 3D points are collected which represents whole body structure of human. These features are normally used in single person action recognition.Histogram of oriented gradients can be collected from these silhouettes.A slight occlusion or noise can cause major change in these features , hence these features are used with some other features to get more accuracy.But to calculate these features , RGB part of video is also required.

1.1.2 Skeletal joints or body parts [2]

A human body can be represented as some segments connected by a set of joints. Hence, getting the 3D coordinates of all the joints is useful to recognize actions. RGBD sensors like Kinect device gives the skeleton information of a human body. main parts of body are torso,head,hands and legs. Specifically as stated in figure 3.1 there are 20 joints in a human body.Kinect device works well when human faces towards camera and also there is no occlusion.However, results may not be reliable when some part of human body is hidden.Many software packages like PrimeSense [13] can be used for tracking joints location using kinect 3d camera.

1.1.3 Spatio temporal features [3]

A video can be assumed as a volume of points in a time t. suppose that every joints' location is represented as x_i, y_i, z_i then we can calculate the number of pixels in time t area covered by x_i, y_i, z_i , if there is any movement in that area then there will be more number of points.Here joints' locations are spatio-temporal interest points(STIP). STIP's spatio-temporal features includes background information, Thus there can be noise or occlusion in the features.

1.1.4 Local 3-D Occupancy pattern [4]

We can transform coordinates x, y, x, t into 4D point sets.Thus , instead of using spatio temporal volume we can project each point as 4-D coordinates.The x, y, z space around joints can be divided into N_x, N_y, N_z sized cubic grid, The number of pixels inside each grid are counted and are normalized to get Local Occupancy Pattern(LOP) of the Human object interaction.

1.1.5 3-D optical flow [5]

Optical flow is measurement of velocity of flow of pixels in an image. It is used in motion detection , video segmentation , action detection.2-D optical flow easy to calculate. Let, we have two images at time t-1 and t, then optical flow can be calculated as $D = (x_t - x_{t-1}, y_t - y_{t-1})$. In 3-D video , it is very time taking to calculate optical flow because we have to process every image and depth also.Scene flow can be calculated by transforming the 2-D optical flow into 3-D optical flow using depth z. Let, focal length of the sensor be f ,then $X = (x - x_0)Z/f, Y = (y - y_0)Z/f$

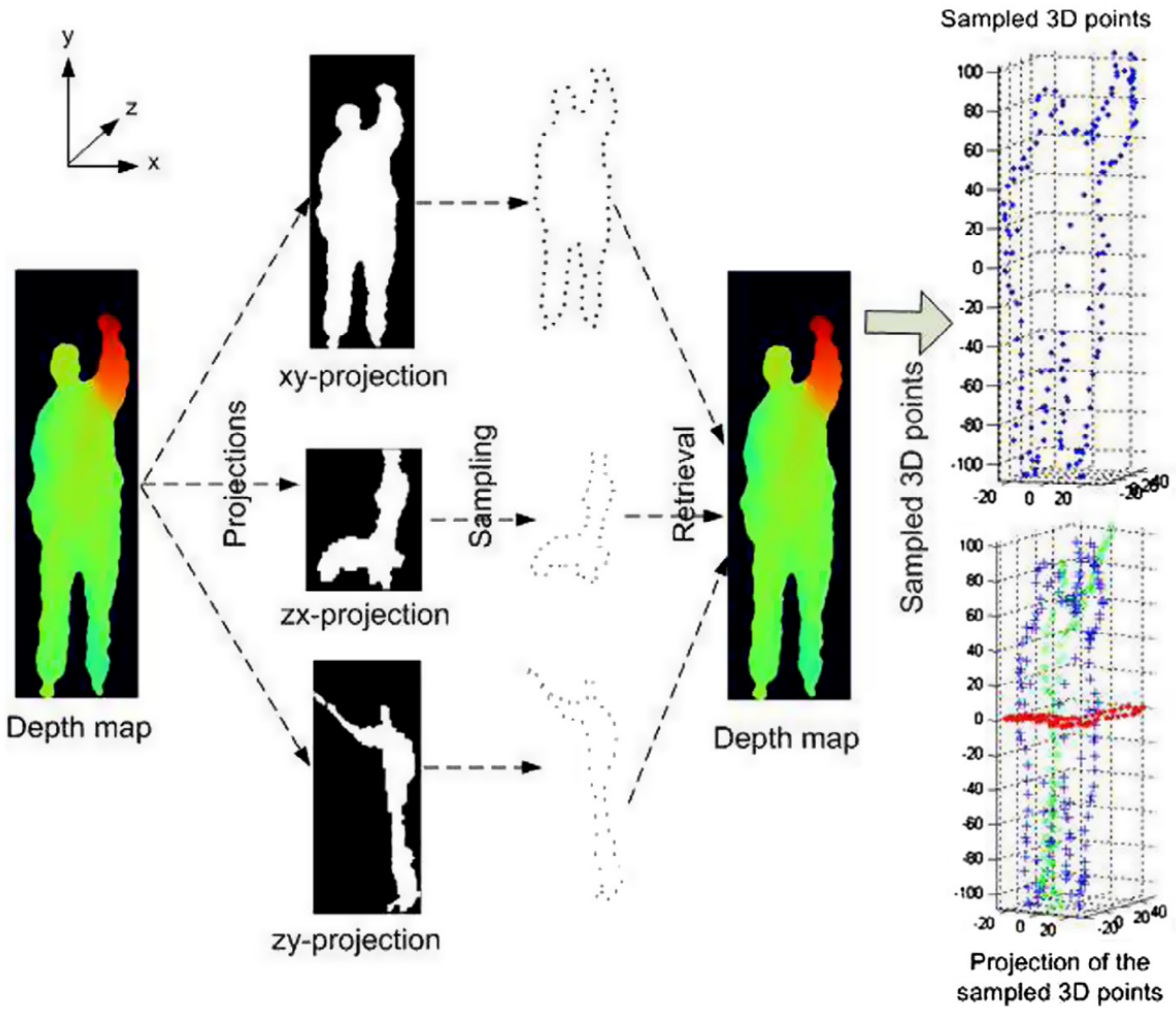

Figure 1.1: Example of 3D silhouettes feature extraction framework [1]

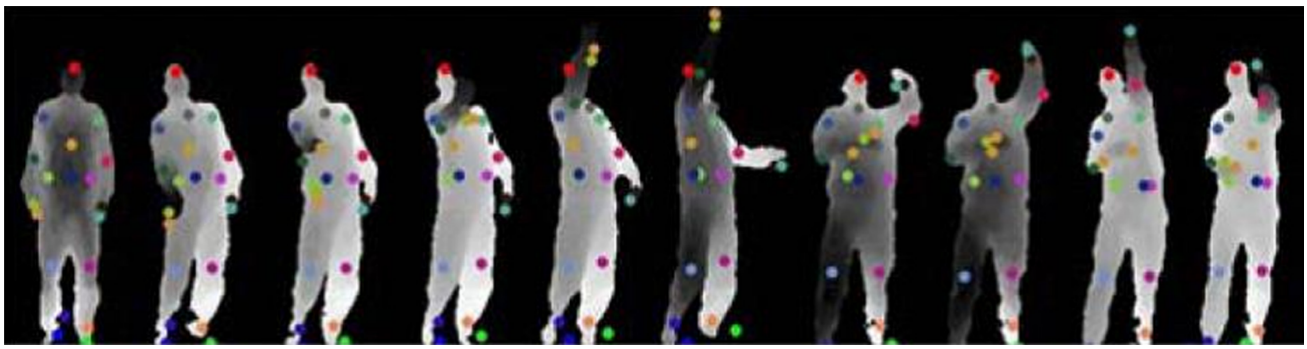

Figure 1.2: Displaying joints position from depth image sequence for the tennis serve action in MSRAction3D dataset [2]

where x_0, y_0 is the principal point of the sensor.Thus 3D scene flow can be calculated subtracting respective 3D vectors in subsequent frames.

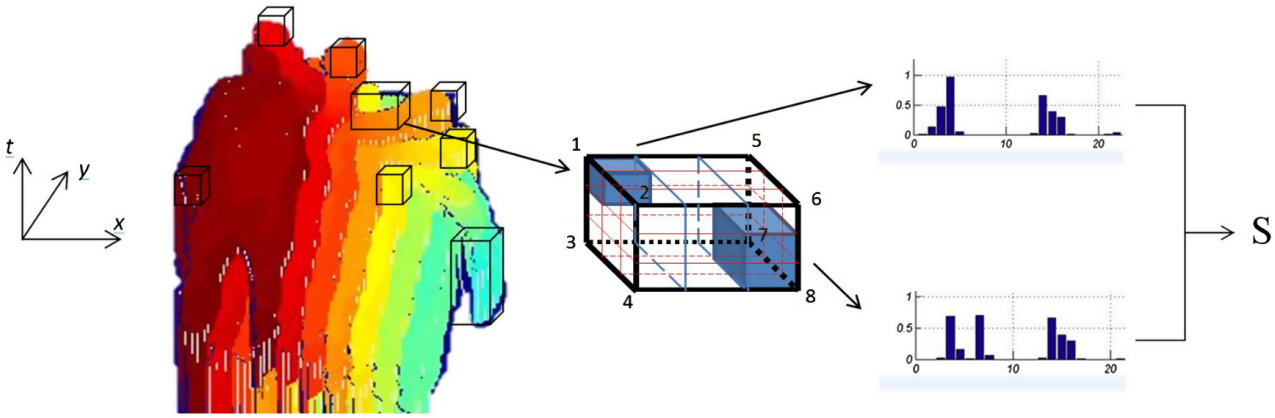

Figure 1.3: Framework for getting spatio-temporal features [3]

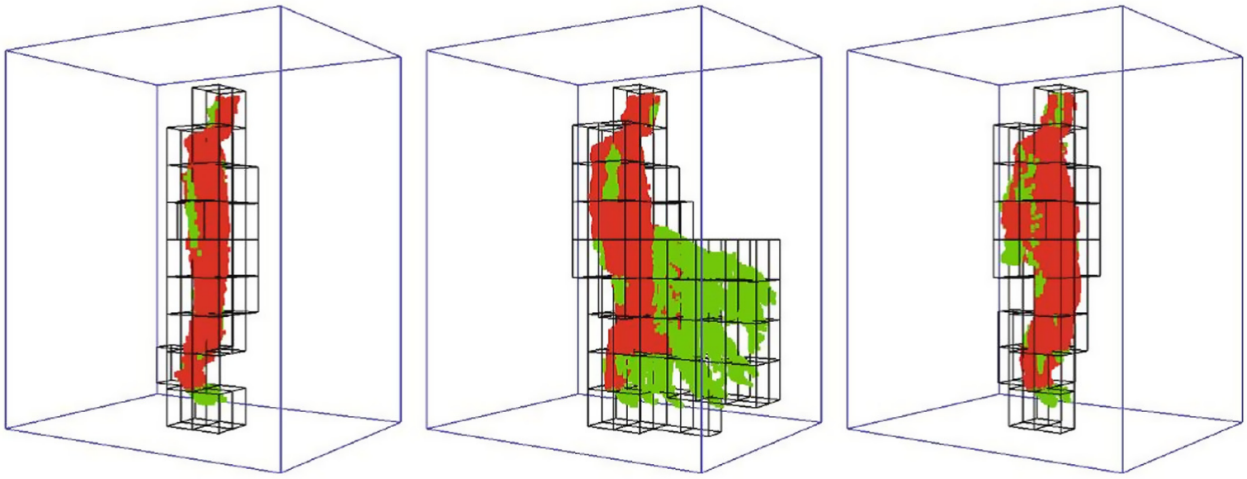

Figure 1.4: Example of spacetime occupancy pattern of action Forward Kick [4]

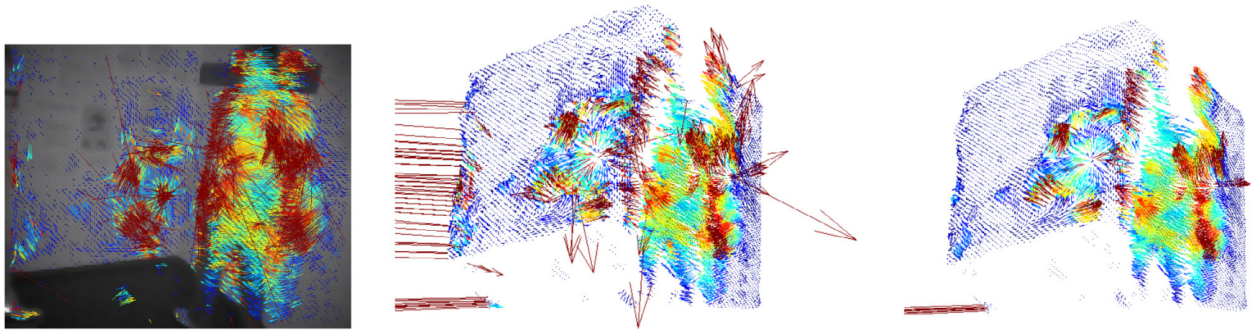

Figure 1.5: (a) Optical flow to calculate 2D velocity vectors , (b) Using point correspondences to calculate 3D velocity vectors , (c) median filter to smooth the component. big motion vector indicated by red and small one from blue. [5]

1.2 Kinect 3D Sensor Camera

Kinect 3D camera is a small black box , which captures the video along with depth of the object. Here depth means distance of object from the camera. It works due to integration of

three hardwares :

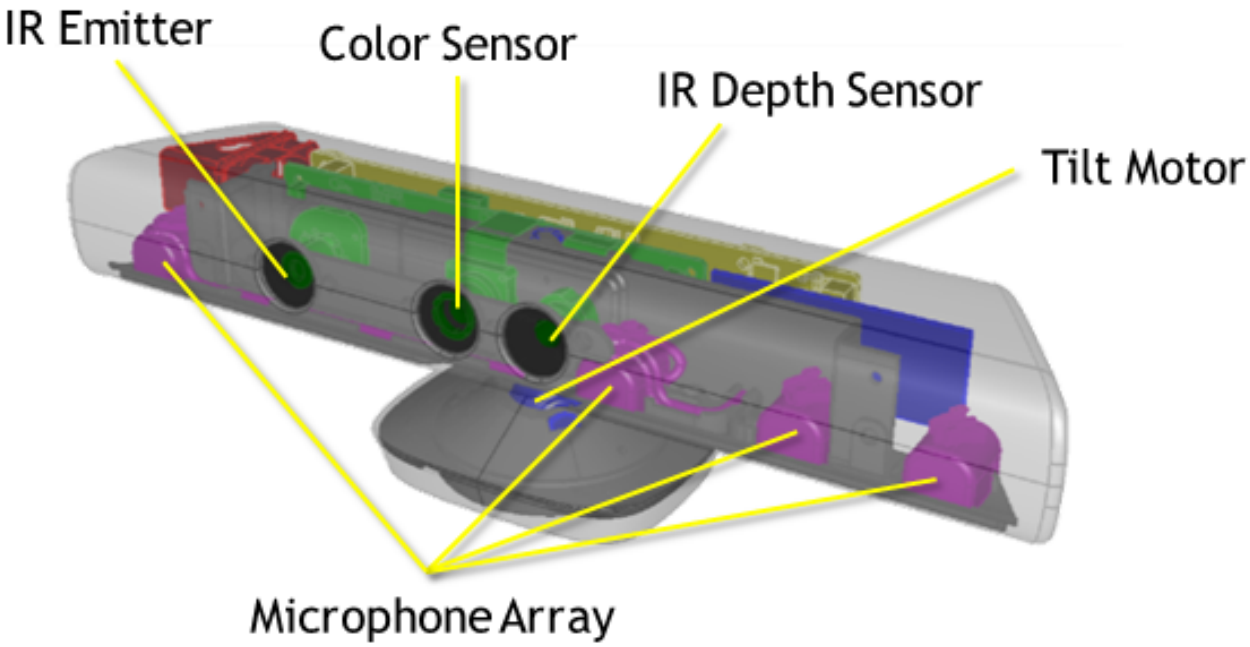

Figure 1.6: Kinect Camera and its component [6]

1.2.1 RGB Camera

This is a VGA camera ,which is used to compute the RGB pixel of any image , where R is for Red , G is for Green ,B is for Blue. This camera aids in facial recognition, gesture recognition and other tracking activities.

1.2.2 Depth Sensor

This is the infrared sensor along with CMOS (complimentary metal-oxide semiconductor) technology to capture the depth information of data regardless of light condition in the room.

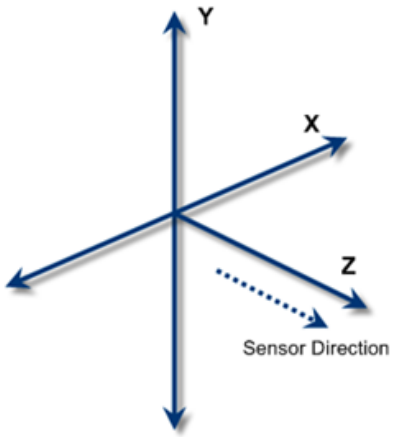

Figure 1.7: 3D coordinate system for Kinect [6]

1.2.3 Multi-array Microphone

This is an array of 4 microphones combined with each other to capture voices of players inside the room. They allow the players to control using voice from few feet.

1.3 Available Datasets

Dataset	Scenario	Class	Data source	Subjects	Samples
MSRAction3D[2]	gaming	20	skeleton + depth	10	567
HuDaAct[14]	human daily activities	12	color + depth	30	189
LIRIS[15]	human daily activities	10	grayscale + depth + color	-	828
CAD-60 Cornell Activity Datasets[16]	human daily activities	12	skeleton + depth + color	4	60
CAD-120 Cornell Activity Datasets[17]	human daily activities	10	skeleton + depth + color	4	120
Act4$\hat{2}$[18]	human daily activity	14	skeleton + depth + color	-	6844
MSRDailyActivity3D[2]	human daily activity	16	skeleton + depth + color(RGB)	10	320
UTKinectAction[19]	atomic actions	10	skeleton + depth + color (RGB)	10	200

Table 1.1: Publicly available kinect datasets.

After coming of Kinect RGBD sensor many datasets came rapidly.Table 1.1 refers to publicly available RGBD datasets for Human activities recognition.

1.4 Challenges

Human action recognition is trending topic in Computer vision.There are many application of Action recognition like Surveillance , Medical , Daily life automation , Robotics . But there are many challenges in building an efficient and reliable Action Recognition System.Few of them are:

1.4.1 Robustness

Robustness is one of the most important challenge in the Action recognition. Because of Environment variation it is not easy to detect the Objects.Due to this system is error prone.

Also different person perform same activity with different rates hence there is a problem in temporal matching of two actions. So to make a robust system temporal matching with less noise is needed.

Figure 1.8: Example of Environment change [4]

1.4.2 Various Activities

There are various activities performed by a person, hence it is very vital to detect and recognize all of them.To make a system which can recognize and detect all the activities with very less false rate is very necessary.

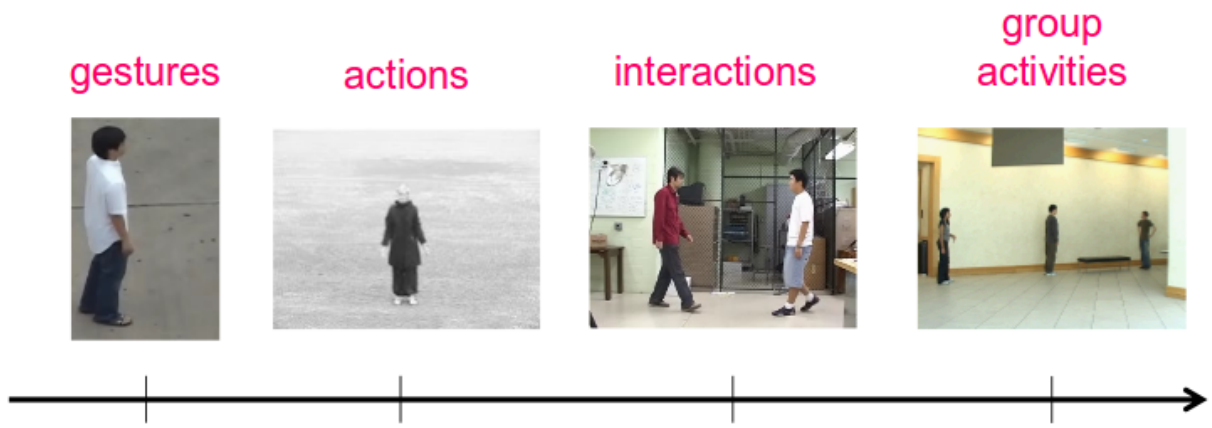

Figure 1.9: Level of Human Activities [4]

1.4.3 Various Objects

There can be multiple objects in a video and their interactions can be in different ways for same type of activities. Hence, Recognizing the action in such a scenario is not that easy.Hence Object to object interaction is more important thing to consider in Human action recognition.

1.4.4 Learning

Due to less number of videos , It is very challenging to learn all the actions by classifier.Although there are very efficient supervised learning techniques but Human efforts can not be used to label all type of actions hence need of designing unsupervised and interactive system is required.

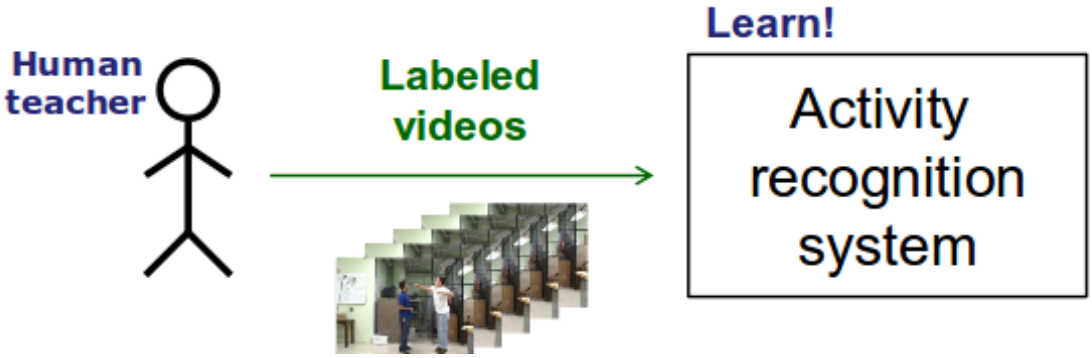

Figure 1.10: Learning through Human Teacher [4]

1.5 Problem Statement

1.5.1 Input

Video of Action performed by a subject.

1.5.2 Output

Tag of Action performed by the subject.

1.5.3 Objective

Main objective is to develop an efficient and robust Human Action Recognition System to recognize the action performed by different subjects in different environment.

1.6 Dissertation Overview

CHAPTER 1 gives an introduction of Human Action Recognition (HAR). It gives account of features extracted and the problem statement of this dissertation.
CHAPTER 2 gives the review of literature and related work done in this area. All the given approach have used the same dataset and configuration for evaluation.
CHAPTER 3 discuss the proposed method and techniques involved in the framework.
CHAPTER 4 gives the results and compares them to other already existing state of the art techniques.
CHAPTER 5 concludes and discuss the future aspect of the problem.

2

REVIEW OF LITERATURE

Human action Recognition is the most relevant research interest point of many researcher since 1990s. In 1990s research were mainly focused on recognizing human actions using 2D videos. But the accuracy is not that good because of occlusion ans noise.But with evolving of many low cost 3D camera sensor like Microsoft Kinect 3D camera research is mostly intended towards 3D videos which are normal 2D videos with Depth information for each pixel. With availability of many software packages like PrimeSense Tracking Package [13] ,Now it is very easy to track the joints location in a video using skeleton tracking. Many techniques used are based on only depth data along with videos , while some are based only on skeleton. In this dissertation also we use Skeleton features to evaluate the label of the activity performed. In this chapter first section presents all the techniques used in the past and are related to the work proposed in this dissertation, all the techniques are based on depth data only.

2.1 Related techniques for Human Action Recognition

Depth Map are space time based features , which gives local or global temporal pattern of data. They generally don't have much texture information as compared to Color Images.They are very sensitive to occlusion and hence a slight disturbance can make whole global feature noisy. Hence designing an efficient and reliable Human action recognition system is a challenging task. Hence, it motivates the researcher to find some semi-local , high gradient features which can work on these type of datasets.

2.1.1 Space-Time Volume based method using Depth map

Li et. al [1] present a method to recognize actions using the sequence of depth maps.The authors used the concept of bag of 3d points in graphical form to create an action graph of all the postures. In this action graph every node is a posture and graph links point towards the next probable posture. Experiments were conducted on MSRAction3D Datasets [2], results showed very good accuracy of around 90% with just 1% sampling of 3d points from depth maps. One disadvantage of this approach is that it lost the contextual information of the points in the images.And also it is severely affected by occlusion and noise.

Yang et. al [7] proposed a special Depth Motion Map feature to present aggregated temporal motion energy map.The authors calculate the Depth motion map for each frame then combine

all of them to form a final Depth Motion Map. Then Histogram of gradients is applied to find the DMM-HOG feature in all the three dimension projection. Then SVM is applied to extracted features. Accuracy achieved by this method is relatively good.However hand crafted projection might create problem in recognizing the activities from different views.This approach is efficient because HOG is applied only to last DMM found.

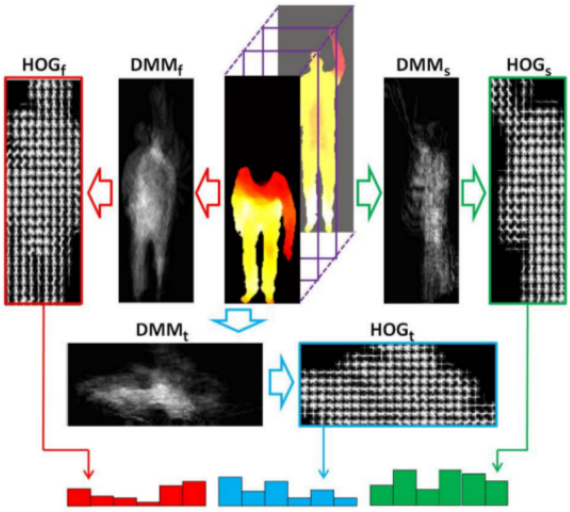

Figure 2.1: Framework proposed by Yang et. al [7]

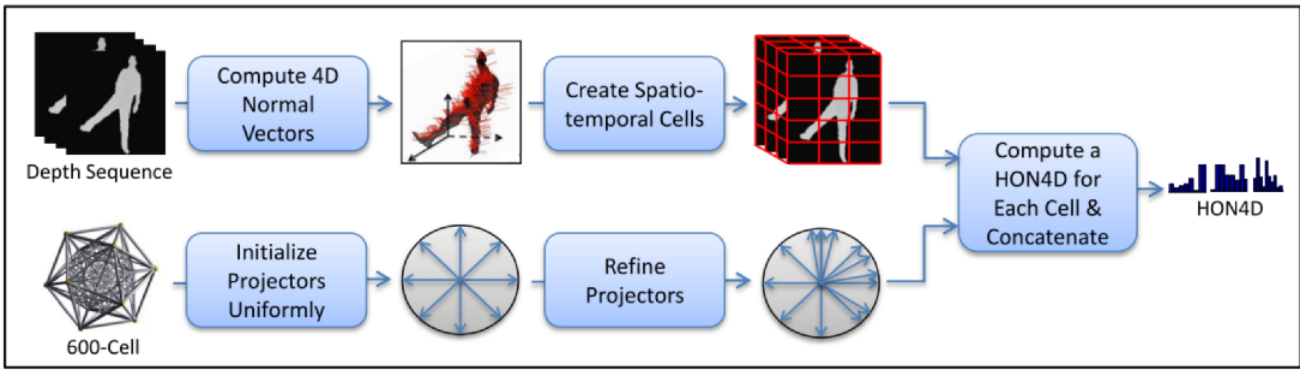

Figure 2.2: Framework proposed by Oreifej and Liu [8]

Most recently Oreifej and Liu [8] presented a different approach to find the histogram of normal gradient in 4D namely spatial coordinate, time and depth. Histogram of Normal gradient calculated along with all the dimension ans then normal vectors are projected on polychoron with 120 vertices and 600 faces.Then all the features are trained using SVM, it gives state of the art accuracy.

2.1.2 Skeleton based Approaches

Skeleton based approaches are solely based on the location of joints in the frame.It has been shown that temporal modeling with these location is very good. Hence many Researchers have

shown better accuracy by just using Skeleton feature.There are three type of Skeleton capturing techniques:

1. Active motion capture (MoCap) systems
2. Multiple view based color images
3. Single view based depth images

One metric to measure the goodness of these datasets is embedded noise. Overall MoCap datasets are cleanest among others. But , due to high cost of MoCap Systems it is not always economical to do the experiments hence due to invention of low cost cameras like Microsoft Kinect , Leapmotion , Single view depth images are used for experimentation purpose.

Skeleton-based Sequential Approaches

we are going to discuss recent work in this field but we can not ignore the seminar given by Campbell and Bobick [11]. They represents human actions as motion trajectory of joints location. They represents the low level feature of body motion obtained via projection of 3d joint trajectories.They calculate the phase space along each dimension using difference between consecutive location of joints. A static action is a point while a moving action is a curve. A certain action is identified by projecting subspaces a action feature in 2d subspaces. Due to phase based representation their approach is space variant and view variant.

Xia et. al [9] proposed new method to calculate feature called Histogram of 3d Joint Locations (HOJ3D). It include the spatial occupancy pattern relative to center point of human body , i.e. torso.They propose a modified spherical space system on the torso.Then partition the whole joint location into n bins.Linear Discriminant Analysis (LDA) helps to reduce the dimension by 1. Then K-means is used for vector quantization to create discrete features.Then these discrete features are trained via HMM to get the action label. They experiment the dataset on MSRAction3D [2] , and their method gives a much better accuracy.

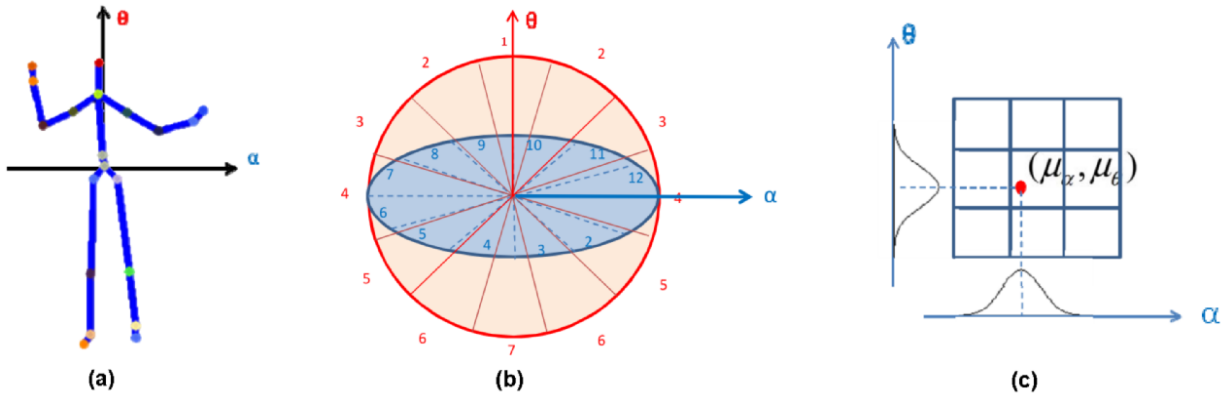

Figure 2.3: Basis axis for HOJ3D (a) and spherical coordinate system used in [9]. (c) The probabilistic method for binning as given in using Gaussian [9]

Wang et. al [2] proposes a action-let based technique in which they utilize both skeleton ans well as local occupancy patter to get more accuracy. The key idea of their approach , their are some actions for which the object interaction matters , hence local occupancy pattern (LOP) works to differentiate these actions.The LOP features are calculated based on the 3d cell around each joint to get the occupancy of pixels.For skeleton based feature they calculate the relative angle based on joint pairs and then uses the data mining to technique to mine relevant features only using threshold point.They proposes Fourier Temporal Pyramid features at each joint. In this feature they take only low frequency feature and discard the high frequency feature to remove noise.Then they use SVM model to detect the action. figure 2.4 gives the overview of whole approach.

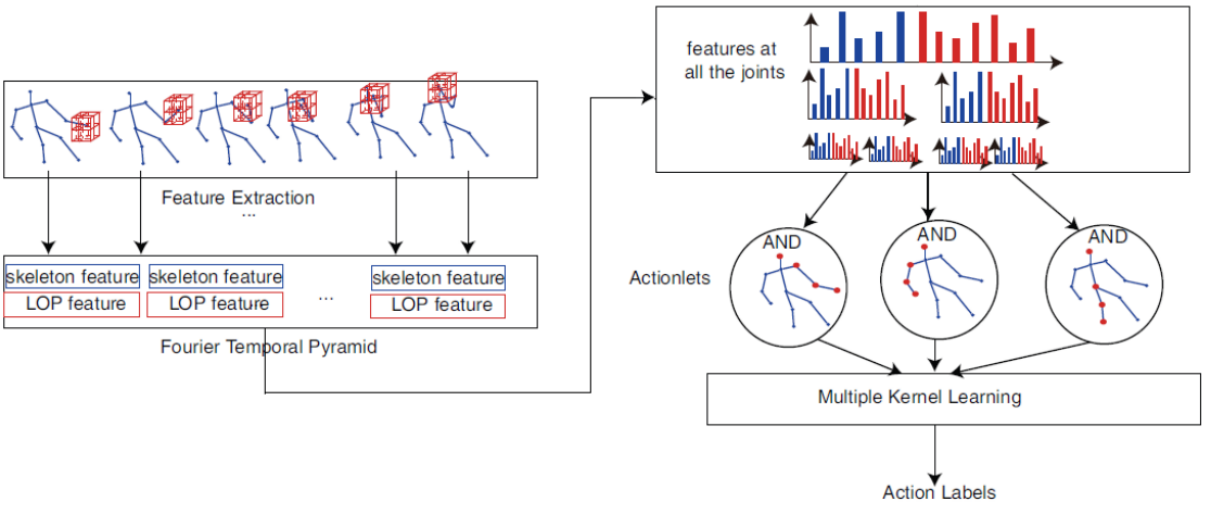

Figure 2.4: Framework for action-let proposed in [2]

Vemulapalli et. al [10] proposes a new skeleton representation using Lie Algebra and Lie Groups. They represent each angle as rotation matrix with respect to base axis, then this rotation matrix is converted to lie algebra and then to evaluate them lie group is used. Then they interpolate using time and velocity concept and get the equal number of features . Then they uses Dynamic Time Warping (DTW) to get nominal curve for each class then they warp the curves using DTW to warped curves for each class. Then they uses Fourier Temporal Pyramid proposed in [2] and then apply One vs All SVM to get the action class. They showed a better accuracy on MSRAction3D dataset.

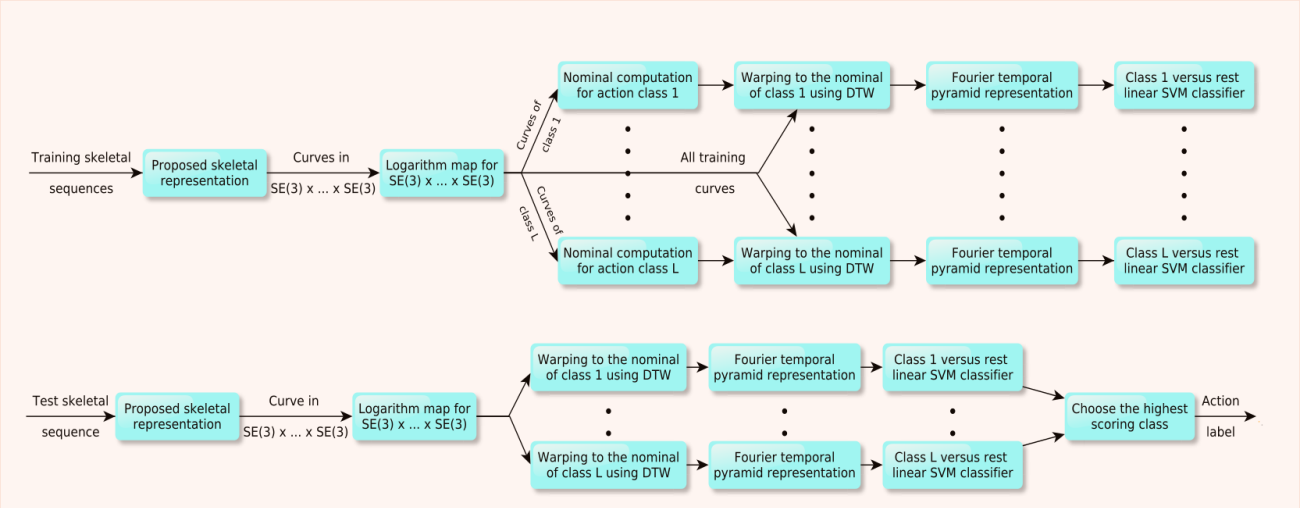

Figure 2.5: Framework for approach using Lie group proposed in [10]

2.2 Summary

In this chapter we present all the related techniques to our approach. Mostly techniques which are described are on Depth data , and Skeleton based approaches are more related to our approach. Firstly we describe Depth map based approaches, most of the approaches calculate spatio-temporal feature from given depth map. Then in second section we describe the approaches based on skeleton features , we give the account of skeleton capturing technique , then we describe the related approaches like Xia et. al [9] , who propose the technique using the spatial occupancy of the coordinates of joints. Then Wang et. al [2] is the original paper which gives dataset MSRAction3D [2]. Our approach is very close to this approach. and also Vemulapalli et. al [10] is closely related with our work.In next chapters we will discuss our approach.

3

AN APPROACH TO HUMAN ACTION RECOGNITION USING SKELETON DATA

We propose an approach for Human Action Recognition which is closely related to [2] . This approach is based on motion trajectory of joint location. As proposed in [11] , We can represent motion of body in form of motion trajectory of joints location. We use MSRAction3d Dataset [2] to show results of our technique.Figure 3.1 shows the framework of our approach.

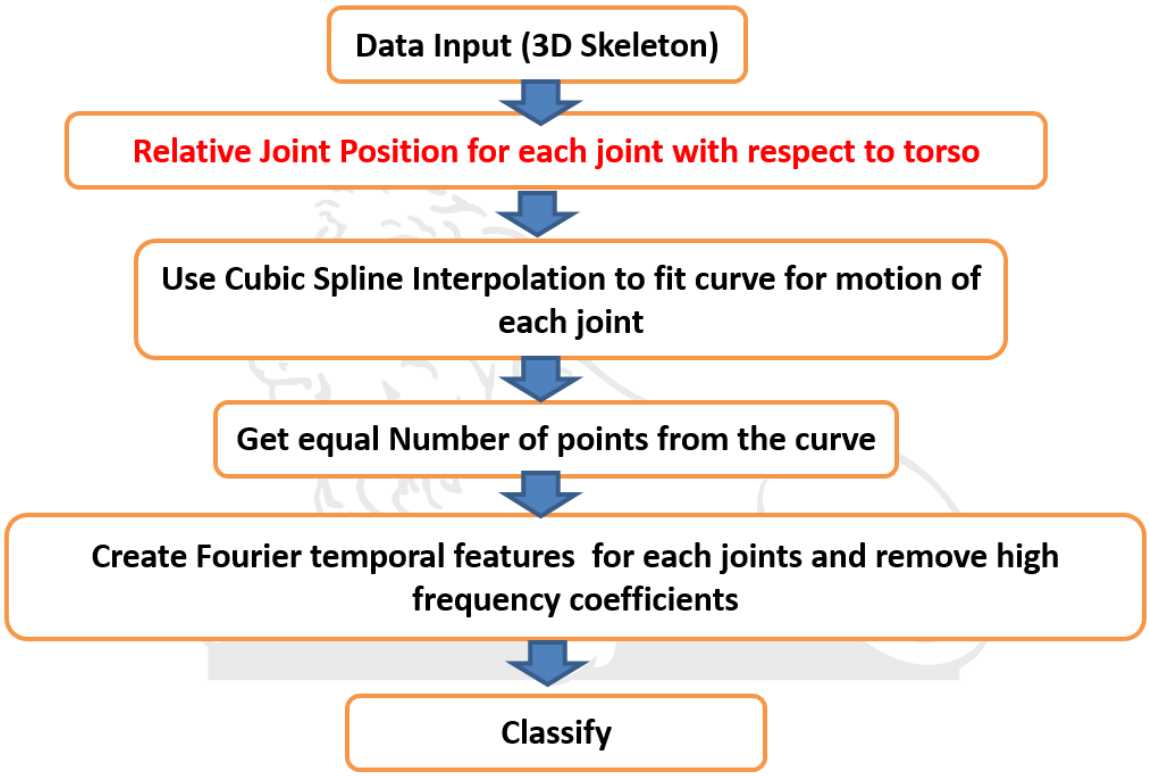

Figure 3.1: Framework for proposed approach

3.1 Skeleton Feature based on phase representation

Our feature extraction method is inspired from the technique proposed by Campbell and Bobick [11].They used joint angles to plot the curve of 2D projection of points , here we use raw skeleton data and make them relative to the center point of body i.e. torso. As shown in Figure

3.2 , Red represents the first Actor while Blue represents the second actor's movement of points of legs. But MSRAction3d dataset has same body part movement for same type of actions , like for Kicking only legs are moved hence all other joints coordinates are nearly stable. Hence, our technique works smoothly. Firstly we get the Raw data from the skeleton tracking system . Then , we calculate the relative coordinates location with respect tot torso. Now we interpolates the data as motion trajectory and fit a piecewise cubic polynomial to get the locus of points for each joints. Now we extract the same number of points from each video, hence same type of action would have the same type of trajectory hence we can easily recognize them.

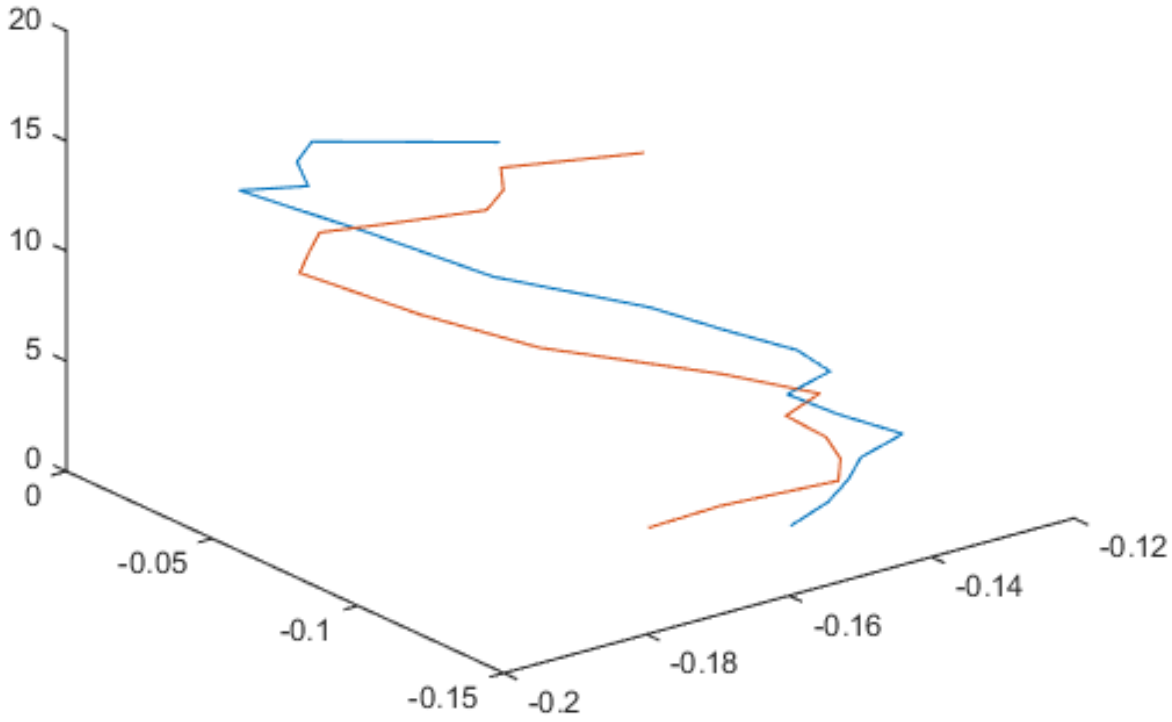

Figure 3.2: Red represents the first actor's motion while Blue represents the second actor's motion trajectory

3.2 Cubic Spline Interpolation

A Cubic Spline Interpolation is spline curve made by piecewise interpolation of m control points. Second order derivative of these points are set to 0 by end points. Hence a tridiagonal equation system of $m-2$ variables is considered.
A cubic spline is of the form

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{3.1}$$

where (x_i, a, b, c, d) is a 5-tuple describing the parameters of $S_i(x)$.
Given our set of data Y and locations X, we wish to find n polynomials $S_i(x)$ for $i = 0, \ldots, n-1$ such that

$$S_i(x_i) = y_i = S_{i-1}(x_i), \quad i = 1, \ldots, n-1 \tag{3.2}$$

$$S_i'(x_i) = S_{i-1}'(x_i), \quad i = 1, \ldots, n-1 \tag{3.3}$$

$$S_i''(x_i) = S_{i-1}''(x_i), \quad i = 1, \ldots, n-1 \tag{3.4}$$

$$S_0''(x_0) = S_{n-1}''(x_n) = 0 \tag{3.5}$$

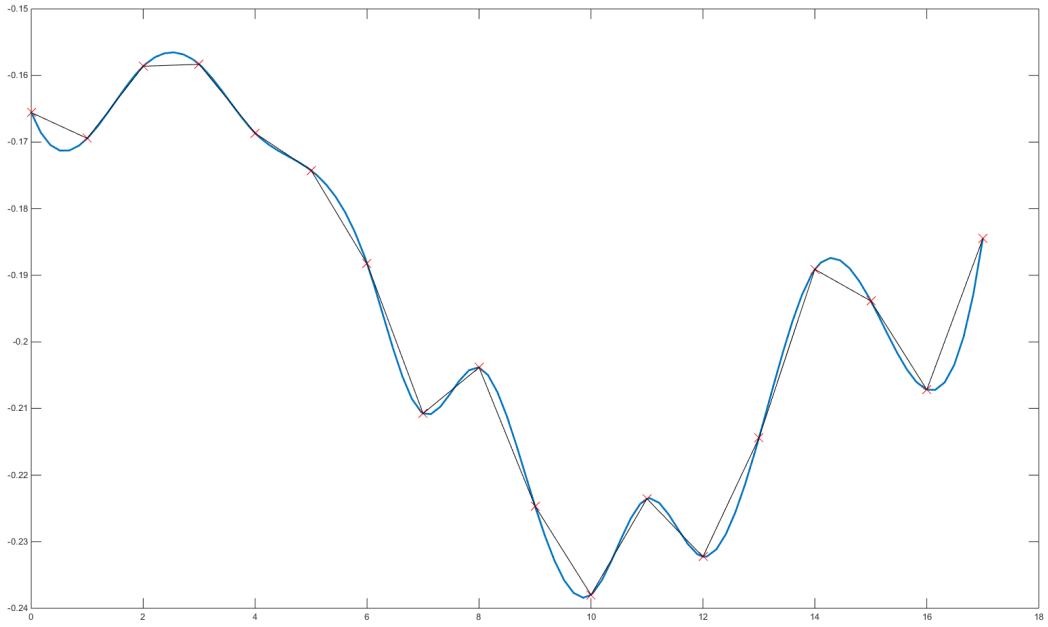

Figure 3.3: A cubic spline curve where blue curve is the spline interpolation and Red points are the original points

3.3 Fourier Temporal Pyramid [2]

This technique is taken from [2]. They propose pyramid of Fourier features by rejecting high frequency Fourier coefficients because they are noise. They take $1/4th$ of size of vector each time to concatenate them at each level.Most appropriate number of level is 3 for current proposed system.

Fourier Temporal Pyramid (FTP) is better than Dynamic Time warping(DTW) in terms of sensitivity to noises, hence gives better results than DTW.

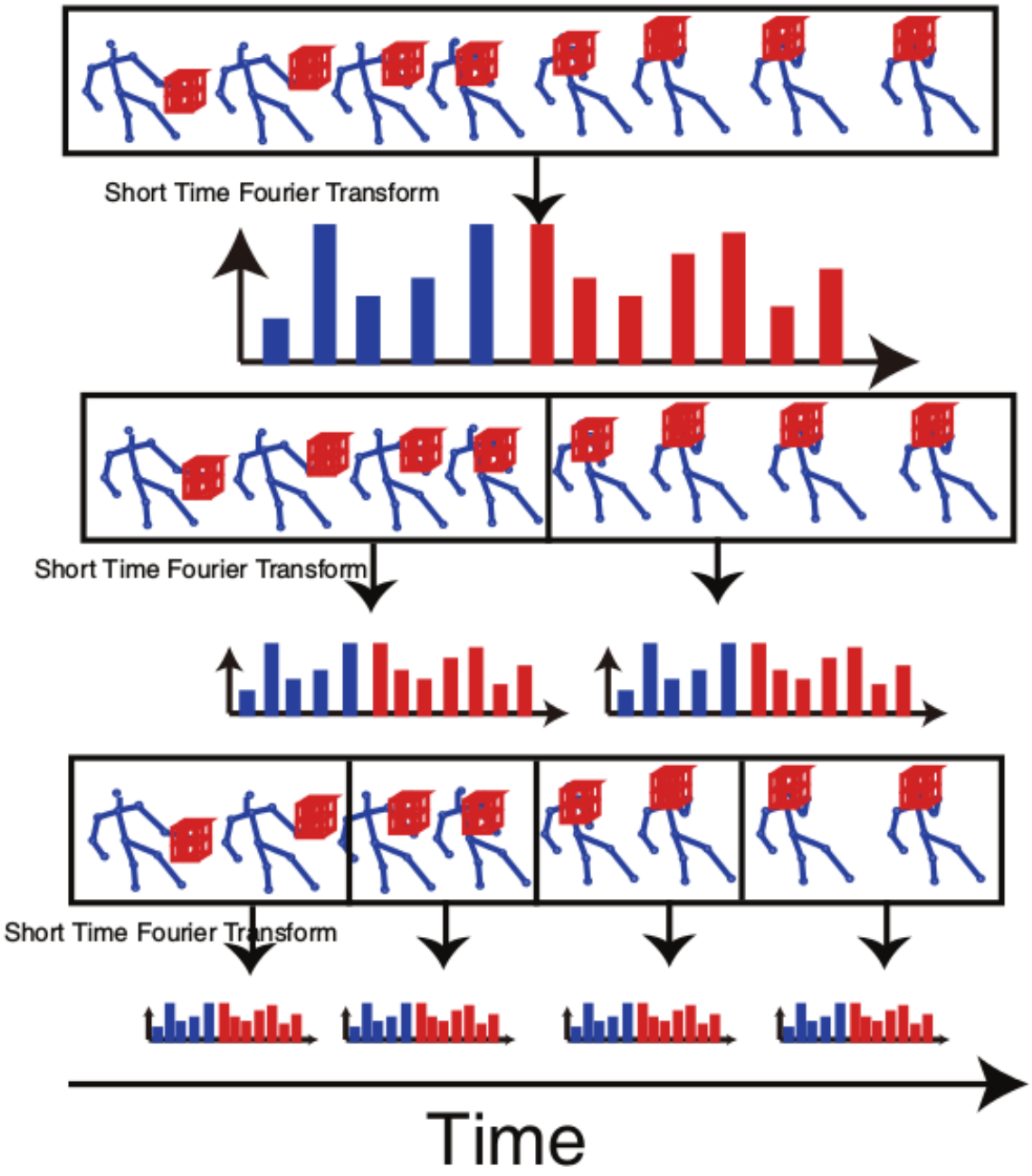

Figure 3.4: An example of Fourier Temporal Pyramid [2]

3.4 Learning and Classfication

We classfied our data using multiclass SVM using one vs one configuration.This gives us the best results. We also evaluate using Random forest classifier also but the accuracy is not that good. We use LIBSVM [20] package to classify the data.

$$L(w, b, \lambda) = \frac{1}{2}\sum_{i=1}^{d} w_i^2 - \sum_{j=1}^{n} \lambda_j \left\{ y_j(< w, x_j > +b) - 1 \right\},$$

and condition of KKT-Optimality are

$$\begin{aligned} \nabla_w L &= 0, \text{i.e., } w = \sum_{j=1}^{n} \lambda_j y_j x_j \\ \nabla_b L &= 0, \text{i.e., } \sum_{j=1}^{n} \lambda_j y_j = 0 \\ \lambda_j \{y_j(< w, x_j > +b) - 1\} &= 0, \text{ for all } j \leq n. \end{aligned}$$

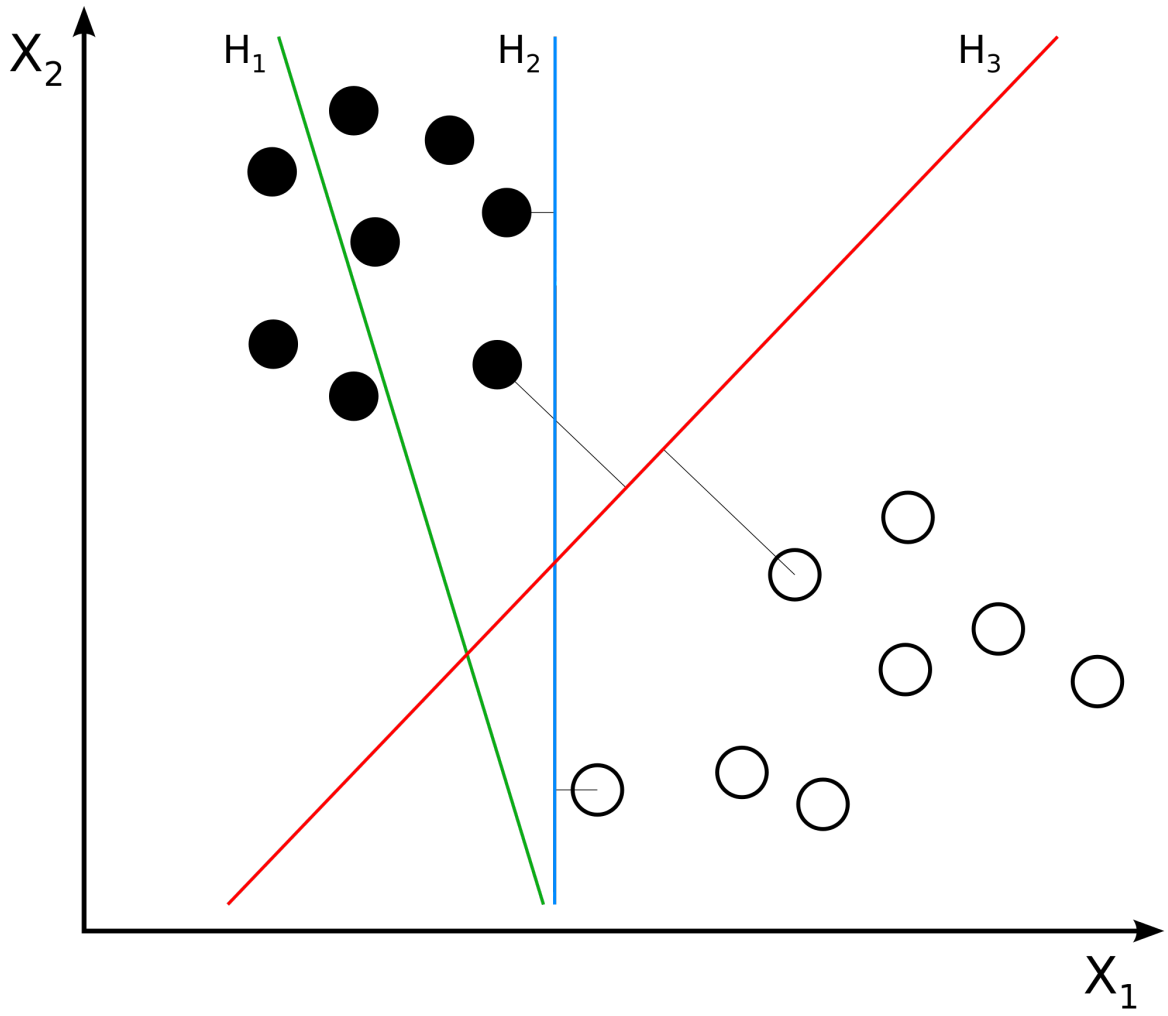

Figure 3.5: A multi class SVM [1]

4

RESULTS

In this we evaluate the proposed method on MSRAction3D dataset [2]. MSRAction3D dataset is captured by Kinect sensor. It has depth map and skeleton location of joints. It consist of 20 activities namely High right arm wave,Horizontal right arm wave,Right arm hammer,Right hand catch,Right fist forward punch,Right hand hight throw,Right hand draw cross,Right hand draw tick,Right hand draw clockwise circle,Front hand clapping,Two hand up wave,Right fist right side boxing,Forward bend,Right foot forward kick,Right foot side kick right,Jogging,Right hand tennis swing,Right hand tennis serve,Golf swing,Right hand pickup throw.We divide the whole activities into 3 sets .

Table 4.1: Action Split as proposed by [1].

AS1	AS2	AS3
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup & throw	Side boxing	Pickup & throw

4.1 Importance of Refining of features

In the chart 4.1 we can see that after using motion trajectory refinement the accuracy improves. Motion trajectory refinement helps to adjust the temporal pattern of different objects.

As configuration given in the [1], we test the results in 3 configuration:

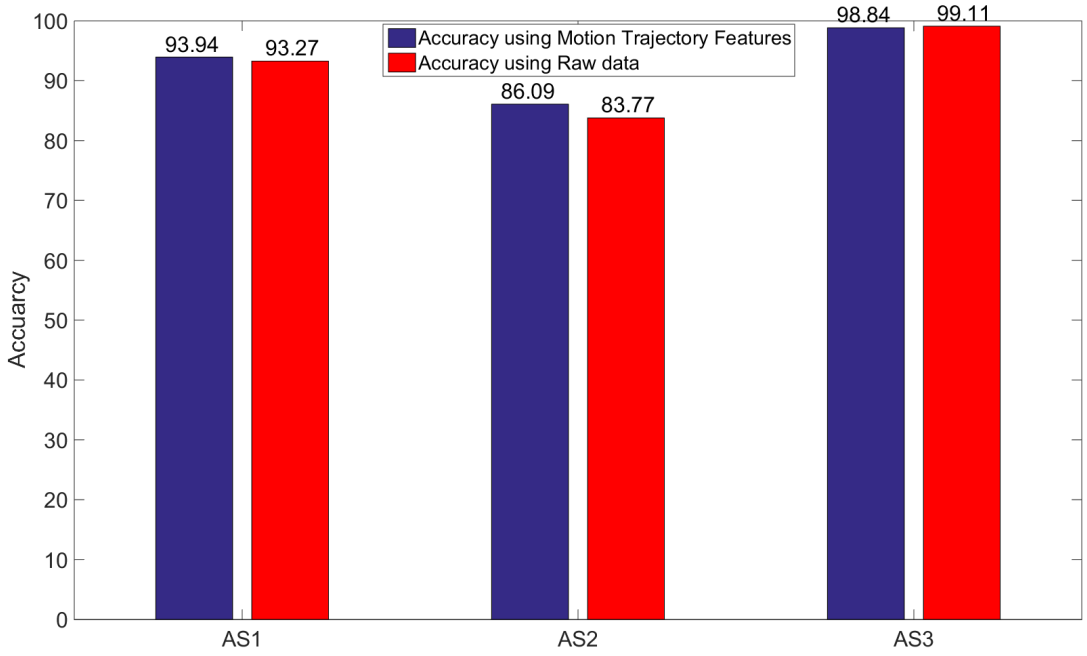

Figure 4.1: Blue bar is the accuracy after refinement, while Red bar is accuracy without using motion trajectory features

4.2 Test One

In this evaluation we take 1/3rd of dataset for training and 2/3rd of dataset for testing. We used sampling to create the training and testing set. Our algorithm gives a good result overall.

Approach	AS1	AS2	AS3	Overall
Bag of 3D Points [1]	89.5%	89%	96.3%	91.6%
Proposed	93.75%	96%	99%	96.25%

Table 4.2: Results of Test One , in which 1/3rd of data is for training and rest for testing [1]

Figure 4.2: Confusion Matrix corresponding to Test One (AS1)

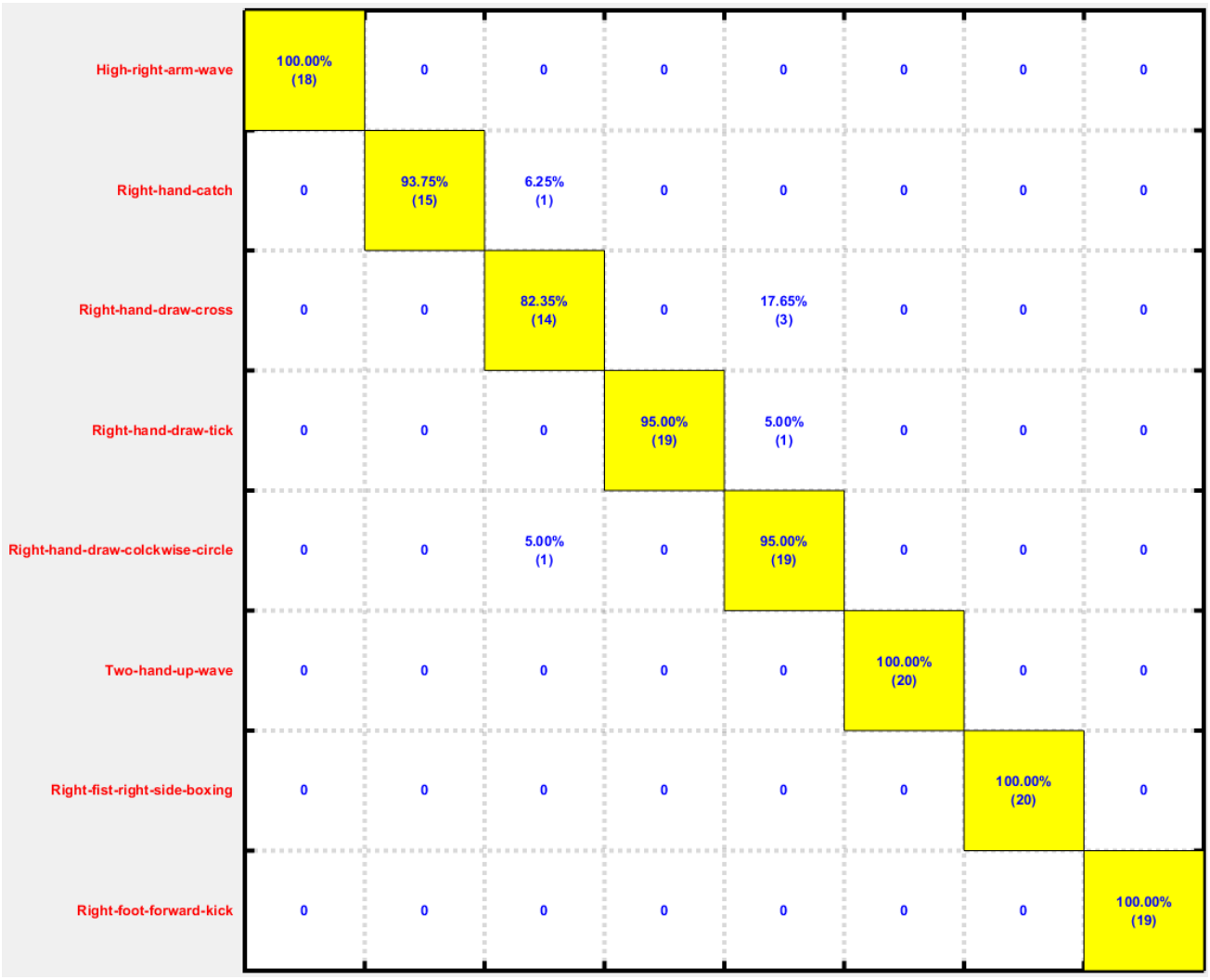

Figure 4.3: Confusion Matrix corresponding to Test One (AS2)

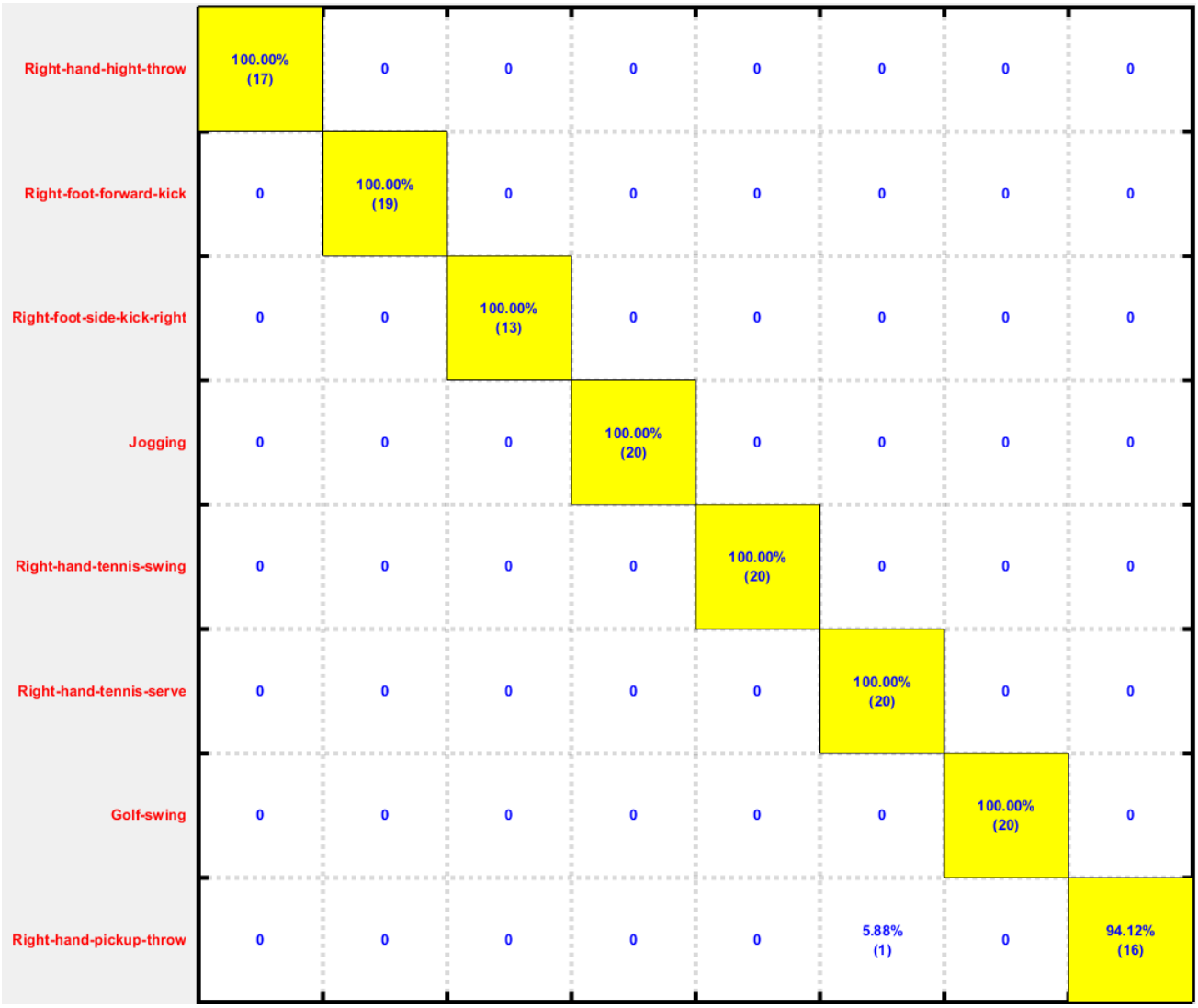

Figure 4.4: Confusion Matrix corresponding to Test One (AS3)

4.3 Test Two

In this evaluation we take 2/3rd of dataset for training and 1/3rd of dataset for testing. We used sampling to create the training and testing set. Our algorithm gives a good result overall.

Approach	AS1	AS2	AS3	Overall
Bag of 3D Points [1]	93.4%	96.9%	96.3%	94.2%
Proposed	95.94	96.05	97.33	96.44%

Table 4.3: Results of Test Two , in which 2/3rd of data is for training and rest for testing [1]

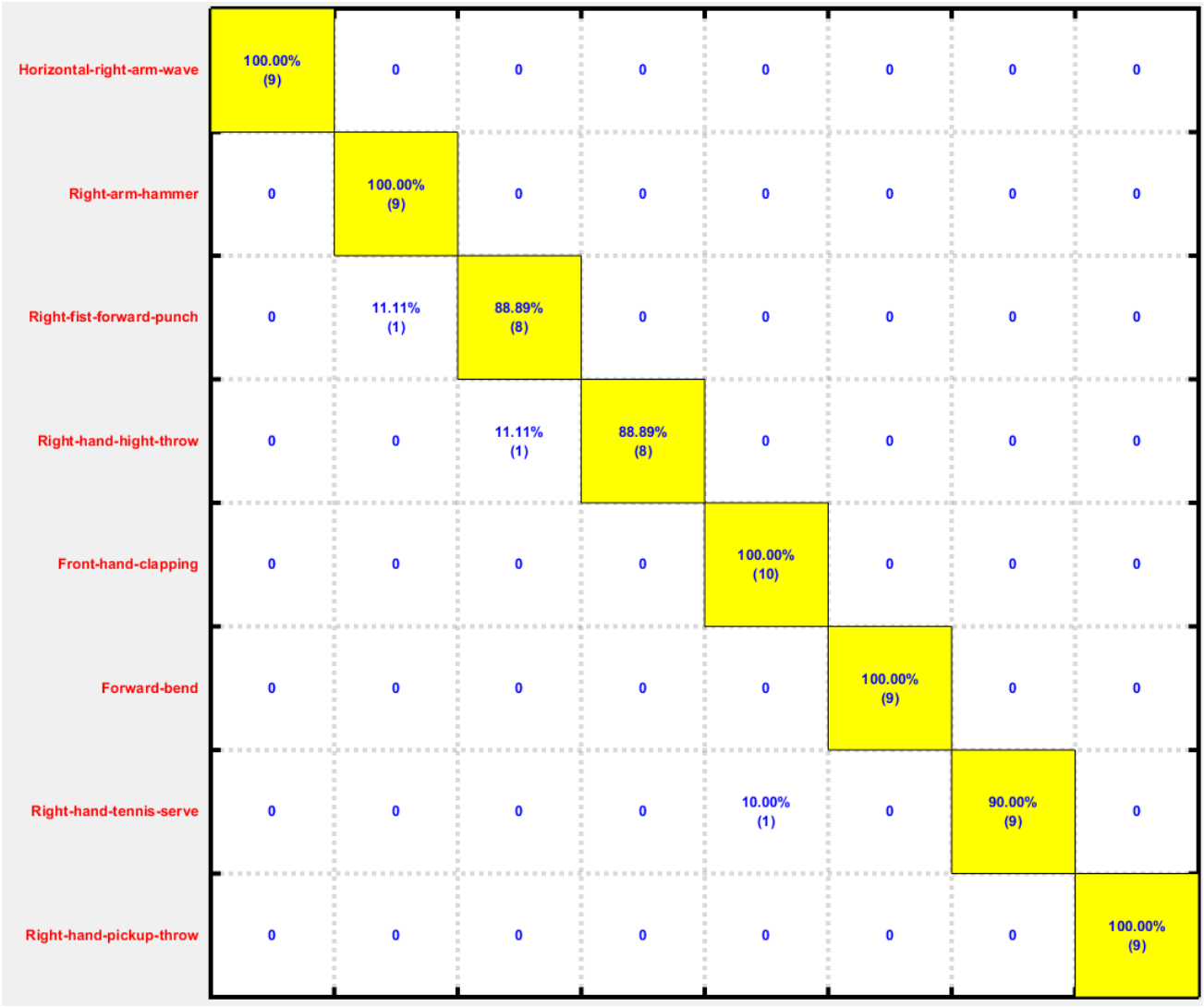

Figure 4.5: Confusion Matrix corresponding to Test Two (AS1)

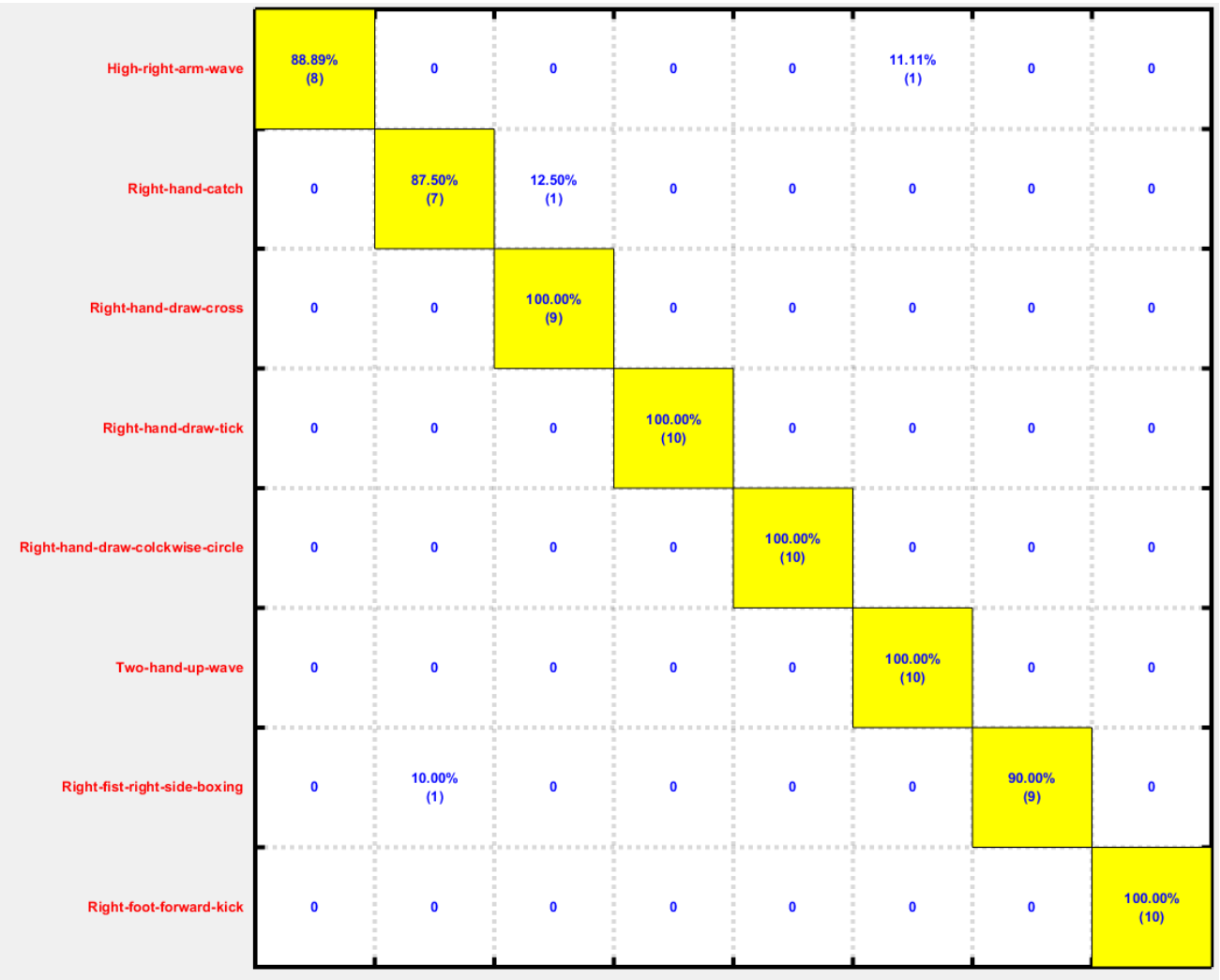

Figure 4.6: Confusion Matrix corresponding to Test Two (AS2)

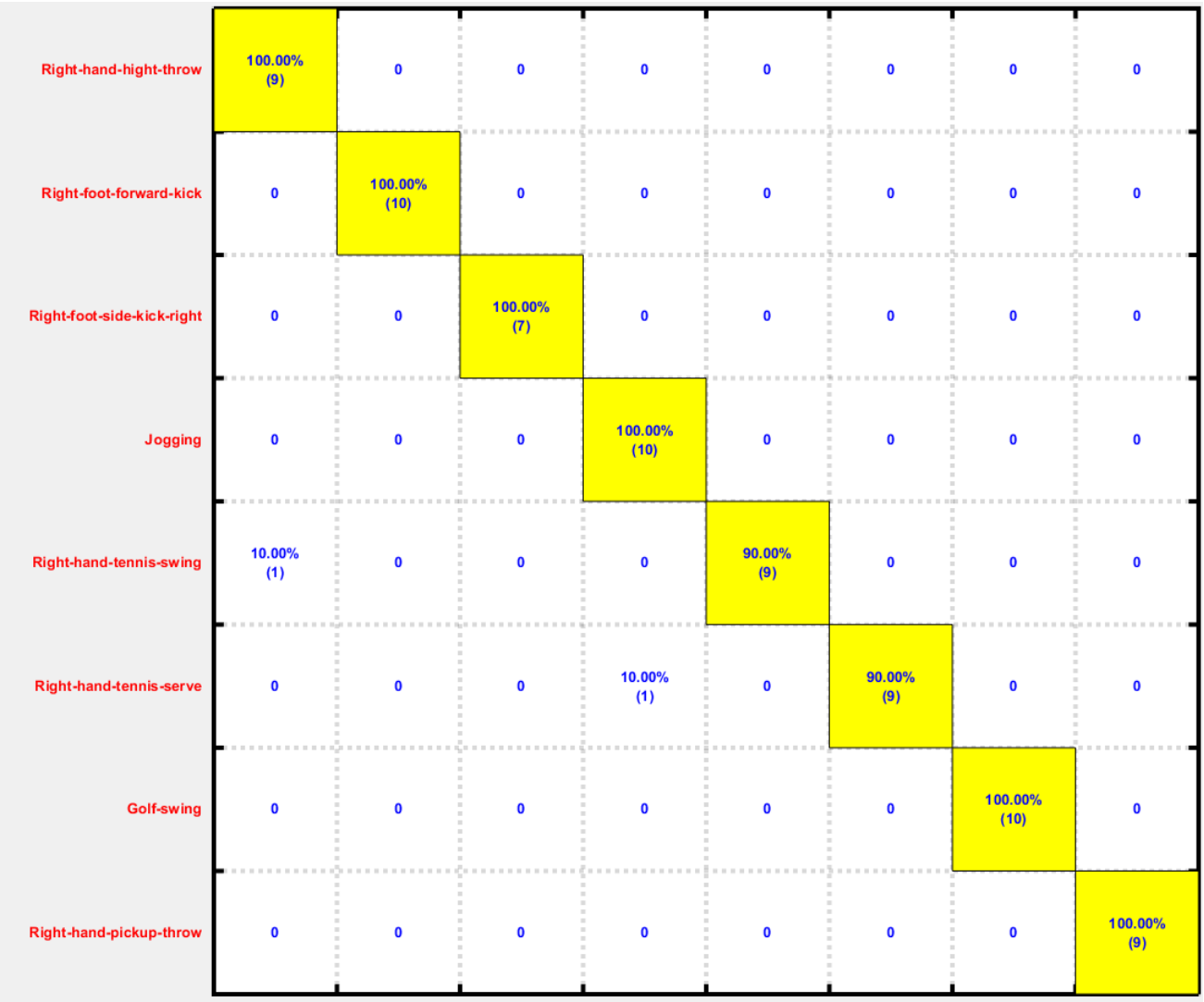

Figure 4.7: Confusion Matrix corresponding to Test Two (AS3)

4.4 Cross Subject Test

In this evaluation we take 1/2nd subjects as training subjects and 1/2nd subject as testing subjects.Then we evaluate the algorithm using SVM classifier.The algorithm gets a much improved accuracy of 92.95% compared to original paper [1] accuracy . 74.7%.

Approach	AS1	AS2	AS3	Overall
Bag of 3D Points [1]	72.9%	71.9%	79.2%	74.7%
Proposed	93.93	86.08	98.84	92.95%

Table 4.4: Results of Cross Test , in which 1/2 of subjects are train subjects and rest are test subjects [1]

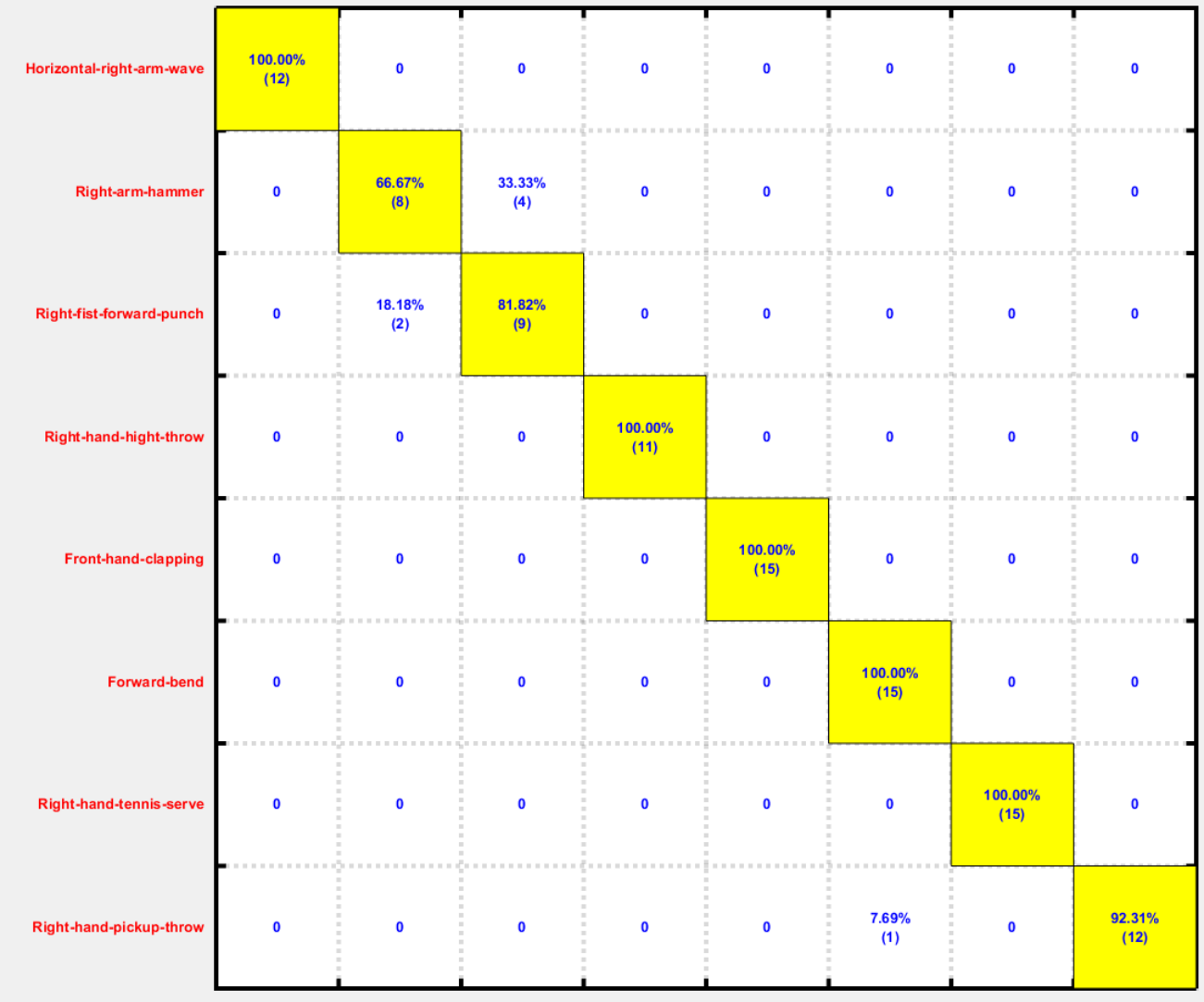

Figure 4.8: Confusion Matrix corresponding to Cross Test(AS1)

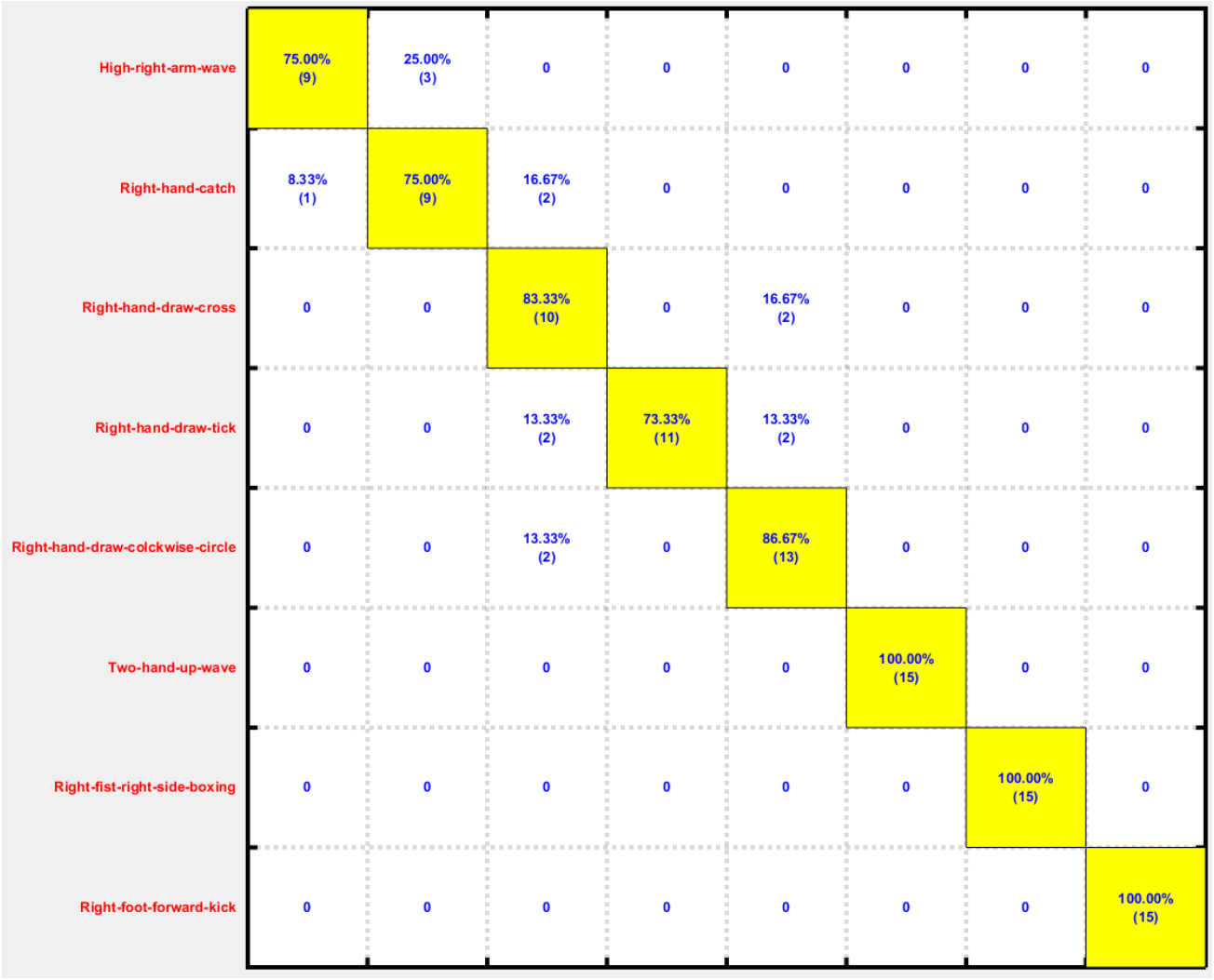

Figure 4.9: Confusion Matrix corresponding to Cross Test(AS2)

Approach	Overall Accuracy
Bag of 3D Points [1]	74.7%
Histograms of 3D joints [9]	78.97%
Random forests [21]	90.90%
Proposed	92.95%

Table 4.5: Overall results of Cross Test , in which 1/2 of subjects are train subjects and rest are test subjects as given in [1]

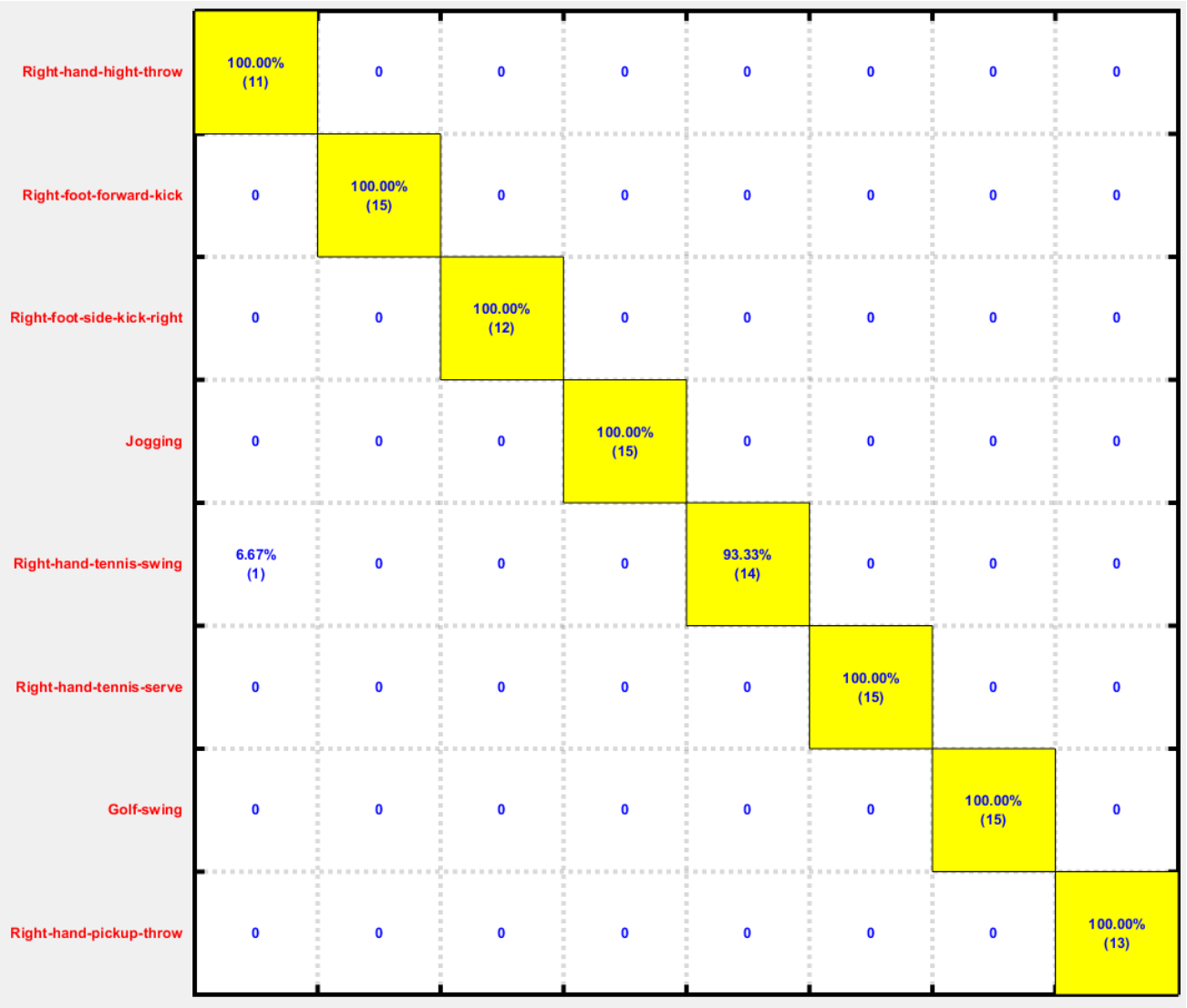

Figure 4.10: Confusion Matrix corresponding to Cross Test(AS3)

5

CONCLUSION AND FUTURE WORK

In this Dissertation, we presented a novel technique to determine the human action performed with the help of Kinect sensor. We represented whole motion as a trajectory of joints then interpolated them along the time to get the nominal features. We used skeleton features only to get a better accuracy than the algorithms in same category with different approach using the same dataset. Our approach beats many state of the art techniques with specific configuration given by Li et.al [1]. We compared our approach with other approaches.

In the future work, our algorithm is not for multiple person and also action are disjoints . There is a scope of building approach to determine the online actions in which all the actions are chained and multiple people involved. Also, we can use object interaction to get the more accurate results. Our algorithm works on MSRAction3d [2] dataset , which is very complex dataset but it is very clean dataset. Hence we can use some preprocessing to get more clean dataset to improve accuracy of the System.

BIBLIOGRAPHY

[1] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 9–14, June 2010.

[2] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297, June 2012.

[3] Lu Xia and JK Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2834–2841, 2013.

[4] Antonio W. Vieira, Erickson R. Nascimento, Gabriel L. Oliveira, Zicheng Liu, and Mario F. M. Campos. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings*, chapter STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences, pages 252–259. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[5] A. Swadzba, N. Beuter, J. Schmidt, and G. Sagerer. Tracking objects in 6d for reconstructing static scenes. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–7, June 2008.

[6] $https://msdn.microsoft.com/en-us/library/jj131033.aspx$.

[7] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1057–1060, New York, NY, USA, 2012. ACM.

[8] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 716–723, June 2013.

[9] Lu Xia, Chia-Chih Chen, and JK Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 20–27. IEEE, 2012.

[10] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, June 2014.

[11] L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 624–630, Jun 1995.

[12] J.K. Aggarwal and Lu Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70 – 80, 2014. Celebrating the life and work of Maria Petrou.

[13] PrimeSense Inc. *Prime Sensor NITE 1.3 Algorithms notes*, 2010. Last viewed 19-01-2011 15:34.

[14] Bingbing Ni, Gang Wang, and Pierre Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *In Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

[15] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, Christophe Garcia, and Bülent Sankur. Evaluation of video activity localizations integrating quality and quantity measurements. *Comput. Vis. Image Underst.*, 127:14–30, October 2014.

[16] Jaeyong Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from rgbd images. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 842–849, May 2012.

[17] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *Int. J. Rob. Res.*, 32(8):951–970, July 2013.

[18] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. Human daily action analysis with multi-view and color-depth data. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *Computer Vision ECCV 2012. Workshops and Demonstrations*, volume 7584 of *Lecture Notes in Computer Science*, pages 52–61. Springer Berlin Heidelberg, 2012.

[19] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *CVPR Workshops*, pages 20–27. IEEE, 2012.

[20] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[21] Y. Zhu, W. Chen, and G. Guo. Fusing spatiotemporal features and joints for 3d action recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 486–491, June 2013.