GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

HUMAN ACTION RECOGNITION USING SKELETON JOINT IN REAL TIME ENVIRONMENTS

Rajesh Kumar*[1], Ajay Kumar [2] & Kamlesh Singh [3]
Automation & Robotics Laboratory Ambalika Institute of Management & Technology, Lucknow, India

ABSTRACT

In this paper, we have to propose an effective approach for the Human activity recognition in the real time environment. We recognized several human activities using Kinect. Through the Kinect a 3D skeleton, joints data collected from the real time video in the analogous form of frames and skeleton, joints, orientation, rotation of all the joint angles from the any random selected frames. After extracting the frames we have implemented classification technique PCA (principal component analysis) with some data we have to classify all the activity models, however, we have to conclude the very less number of data (8-12%) to train our system from all the activity frames. After applying the PCA classification techniques we got excellent accuracy 93.6%. Finally, we observe that our proposed techniques are more accurate than other methods; therefore this technique is more suitable in real-time application such as robotics, human computer interface, in game player's activity recognition.

Keywords: *Human Activity, Kinect, Skeleton Joints, Principle Component Analysis, Gesture Recognition.*

I. INTRODUCTION

Since last decades, Robotics has been known as a science, which creates the Automatic interpretation between human activities and perception, and integrating them by using controls, machines, and electronics with the help of digital system that is computers. Recently Robotics has been used in more areas such as field robotics, service robotics, or human enlarge. The idea that has been proposed in my work targets for give robot skill to execute a task, identifying the human activities and learning with their movements. Many researchers has been demonstrated related work in large scale since the 1050 in these areas such as robotics research and computer vision. The goal of computer vision is to extract the information from a particular scene and design it. The recognition task can be classified in to three stages: feature representation, feature extraction and action classification. This paper aims to present a humanoid robot having the skill of observing, training, and representing actions operated by humans for generating new skills. This system has been implemented in such a way that it will distinguish different actions along with grants the permission to robot for regenerating their actions. Being able to identify and recognize human actions is most essential for many applications such as smart homes and assistive robots. Human robot interaction (HRI) has been implemented in the view of real world applications Human activity recognition is an important functionality in any intelligent system designed to support human daily activities. The measurement of image or camera motion and more on the labeling of the action taking place in the scene. We have to select most informative frame and design one another action feature, which are able to remove all the noisy frames and minimize the computational cost. Activity recognition activity detection does not provide the label however attempts to distinguish one activity from another by using classifier technique PCA/ANN for the action recognition and these are nonparametric classifiers having property to avoid the over fitting problems and advantages to take the large numbers of the classes. Our work mostly concentrates on the activity recognition of the videos, which are captured by the RGB camera. The video, which are taken by camera in 2dimension frames having RGB images in the sequential order. Many research in literature survey on the topic activity recognition for 2D videos. The spatial temporal approach has mostly used to measure the similarity between two activities. To find the accurate similarity calculation, the spatial temporal detection method and representation have proposed [5, 13]. NBNN and HMM methods widely used for the human activity recognition [5, 2]. In these approaches, human activity can explain by the combinations of the key joints and other points. On the other way, take advance technique RGBD cameras of the Microsoft kinect have to practically capture the RGB videos in real time as well as in depth map.

A. **Action recognition with a multilevel HDP-HMM:** In this paper [2], they classify human action, which occurs in the depth image sequences by using feature based on the Skelton joints position. The representation of the action classes are by the multi-level hierarchical Hidden Markov Model process (HHMMP). Nonparametric Markov model process allows inference of the hidden state from training data. The models parameter to each classes are formulated as the transformation shared based distribution, to promoting use of the unlabeled example during the borrowing and training information across the action classes. Further parameters are learn in discriminative way, we use normalize gamma process to representation of the HDP and margin base function for this purpose. The parameters from complex poster distribution induce by discriminative function by using sampling. The experiments have two different dataset, which show that action classes learning model by technique produces have good classification result. The experiment demonstrates utility of these approaches, intended to discriminative criterion, and applies the technique to classify action involving objects and humans.

B. **Graph-embedded system for action identification:** The Human action recognition applications [3], in the field of pattern recognition are an important issue remote surveillance with commercial video content indexing. However, the human action characterizes by the non-linear dynamics of operations and therefore it is not simply learned and recognized. The properly study proposed a silhouette base recognition technique, in which a process for k-NN to build the most powerful differential technique spatio temporal having used to classification reason. In first step ALPP (adaptive locality preserve projection) technique suggest to finding the low dimension subspace, which are linear in local data. The result is efficient to solve the overlapping problems in the human body structure different class of activity; temporal data are used to extract the non-base different activity method. At the end, the margin between the nearest neighbor metric grouping methods to give the most efficient spatio temporal is applied. The result of this work shows the proposed method which are able to human action recognition in the real time and the performances are better. In future work the approach can improve by this perspective order to apply, the methods for daily base application have more activity are needed. The silhouette information's as neglected the input , 3D information sources of ambiguity and limitation. The system capable for real time for applying data sources, in depth information or the many viewed images (take one activity sequences with many cameras from the many view) can improve the approaches. Besides that, the depth information it provides the delicate human activity for the recognition.

C. **Eigen Joints-based activity Recognition Using NBNN**: In this paper [5], we have the 3D positions of the joints of the body to offer an effective way to identify human actions.RGBD sensor and associated SDK with the release, in the real time the joints of the human body extracted with the acceptable accuracy. In this method, we differentiate on the basis of the position of the joints features a new type of action, are to combine information including speed operation, and offset the Eigen Joints. In multi class classification of the actions, the NBNN techniques are used. The recognize result on MSR 3D datasets to explained that the method performs state of art method. In the addition, they search that the how many image frames required in this approach to action recognition on Microsoft research 3D action data sets and we have to observed that the 13-28 frames are more than sufficient to achieve the result using entire video sequences. Compression between the SVM and NBNN shows non quantized descriptors and videos to part computation is much better for the activity recognition.

II. METHOD & MATERIAL

We have to build a new approach on the basis of the differences of the skeletal joint in spatial domain and temporal domains. In normalized data, we are applying principal component analysis (PCA) to finding the Eigen joints by reducing the noises and redundancy on joint differences. Thus, after extracting the set of frames we implements classification techniques Principal Component Analysis (PCA) with some variants for classify our all different gesture models. Accordance with the image classification we avoided quantization of the frames to class distances and descriptor, alternatively of the videos to videos distances. In addition the efficient method performed activity recognition by operated whole video sequences. The scope of these work is widely applied many number of the real world application like as human computer interaction, health issues, video surveillances and video search based on contents. The entire work mainly concentrate on video sequences of activities which captured by the RGB cameras.

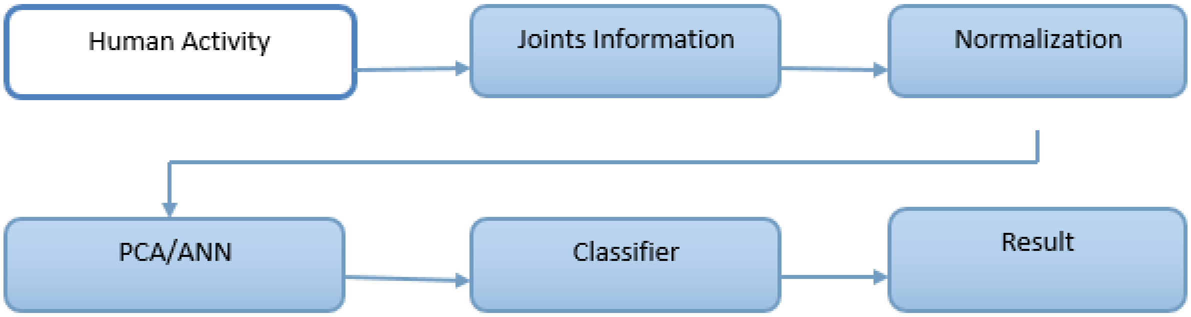


Fig.1 Graphical representation of proposed methodology

A. Human Activity

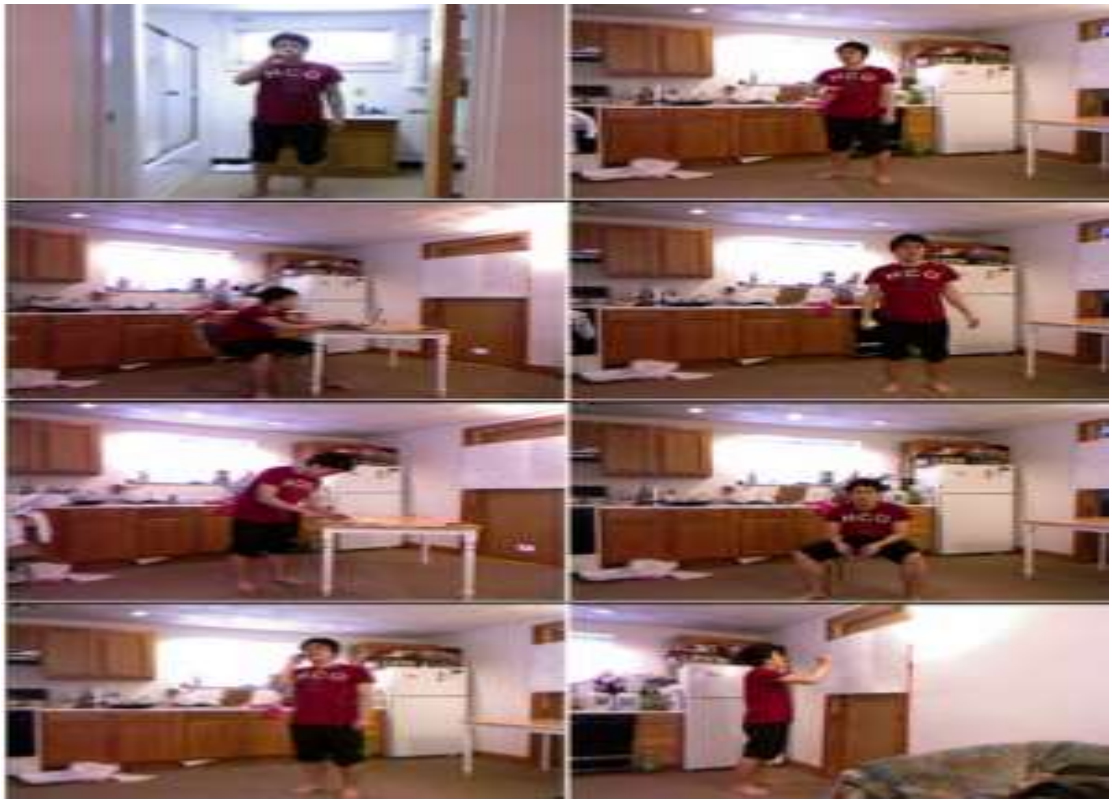


Fig.2- Corresponding activity and depth image of dataset top to down and left to right (1-Brushing teeth. 2-Working on computer. 3- Cooking (chopping). 4- Talking on the phone. 5- Drinking water. 6- Opening pill container. 7- Talking on couch. 8- Writing on whiteboard)[27].

Human activity recognition is an important functionality in any intelligent system designed to support human daily activities. the measurement of image or camera motion and more on the labeling of the action taking place in the scene. Moreover the several activity is performed in real-time environment shown in Fig 1 where subject person is perform daily routine work such as Brushing teeth, working on computer etc. while our kinet is capturing the activity with help of joint angles information. During our experiment we use all male dataset in kitchen and corridor environments. We are assuming that a kinet is mounted on wall in the front of subject user to capture the activity perform by user. However our kinet start recording the activity of humans from initial starting to until activity is not completed. Thus we have set of video image data for each individual activity. We have currently 25 joint angles for human body .

B. Joint information from Kinect

However, the total number of joints is 15. Where 11 joints have both the joint orientation value and the joint position value, and 4 have only positions value. The values of orientation and position are in following format
F = Po(1), Pp(1), Po(2), Pp(2)............Po(11), Pp(11)....Pp(15)

Where F is the frames and Po = orientation values of the joints which are 3×3 matrix stored and follows by
= 0, 1, 2, 3, 4, 5, 6, 7, 8, Co
Co is the Boolean confidence values which are o or 1

The joints which are used in to taking the data through the Microsoft kinect are given as follow: (a) Head (b) Neck (c) Torso (d) Left Shoulder (e) Left Elbow (f) Right Shoulder (g) Right Elbow (h) Left Hip (i) Left Knee (j) Right Hip (k) Right Knee (l) Left Hand (m) Right Hand (n) Left Foot (o) Right Foot .

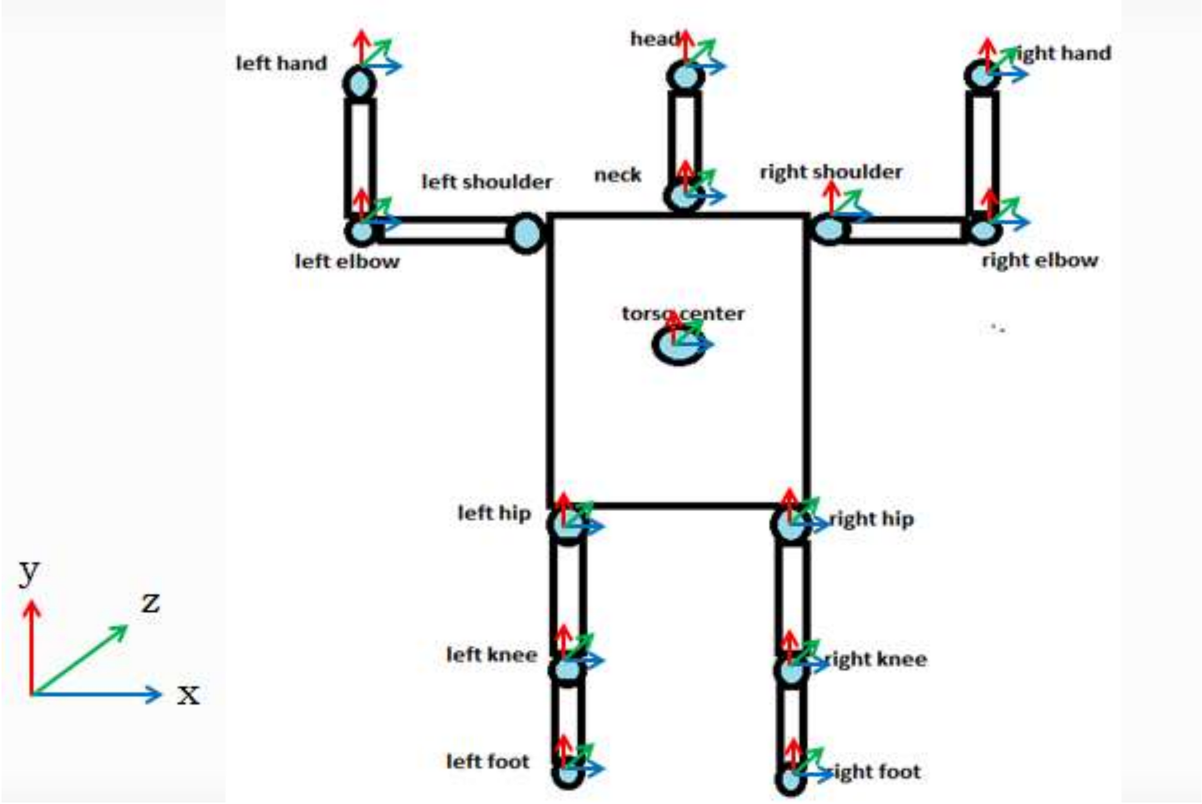


Fig 3. Joint Information

1. *Differences between orientations and joint positions :*

The joints positions value is more accurate than the joints angle. The explanation of this point we have to compare two methods of computing a hand position in game. First method is taking the position of hand joint in API. The other method is to take torso position and orientation, both shoulder and elbow joints angle. Typically the result of hand position will different from from one returned by API because avatar will have will different lengths than model used in skeleton API (specially consider previous mentioned point that skeleton allow body segments length to vary the time where avatar model in game have fix length).the position will be match if segment length match exactly same at all time. The hand positions compute using angle driven method typically noisier other than the hand positions return direct from API. The result we have recommend by using joint position when possible.Unfortunally this method is not practical for game which based on avatar that's have need to drive using the joints angle. In that case the available option is to use either joint angle and deal the noisy hand and feet, one more way to use the joint position are constraint in post processing which have compute modify the joint angle better tuned avatar.

2. **Acquisition Module :**

Microsoft Kinect - An active stereo device which is having basically three important modules. Which Shown in Fig. 3

- RGB Camera
- Depth Sensor
- Array of four mics(provides four streams of the data).
- Color Stream: In this stream it's provide captured live video stream.
- Depth Stream: It's captured the each pixel with the depth information having image acquired by the kinect sensor.

Skeletal: Stream: In our work we have to take data of different– different person for perform different activity through this we get the X, Y and Z coordinates for 15 skeleton joints which are head, neck, right shoulder, right elbow, right hand, left shoulder, left elbow, left hand, torso Centre, right hip, right knee, right foot, left hip, left knee and left foot.

Audio Stream: Kinect has 4 mics array which are used to capture and gives 4 channel audio. The most important stream provided is the skeleton stream. Kinect has the capability to infer the body positions. It is capable to do so with the help of structured light with which it created the depth map which was discussed earlier and a machine learning algorithm. The light used by the kinect is infrared laser which. This pattern of light is projected on the object which is then analyzed by infrared camera on the kinect to create the depth map. Kinect can be initialized In these streams the most important stream is skeleton stream. Kinect is capable to find the body position. It's also capable to do more with help of structure light which are created the depth map discussed in machine learning algorithm. The lights which are used in kinect are the infrared laser light. The infrared lights are projected on object which analyzed by the kinect camera to create depth map. The light which is used in kinect is infrared light. This infrared light projected on object which is analyzed by the camera and creates the depth map of the object. Kinect could be able to analyzed capture colour of the frames under different resolutions and different speeds. • 12 FPS: 1280x960 RGB • 15 FPS: Raw YUV 640x480 • 640x480 • 30 FPS: 80x60, 320x240, 640x480

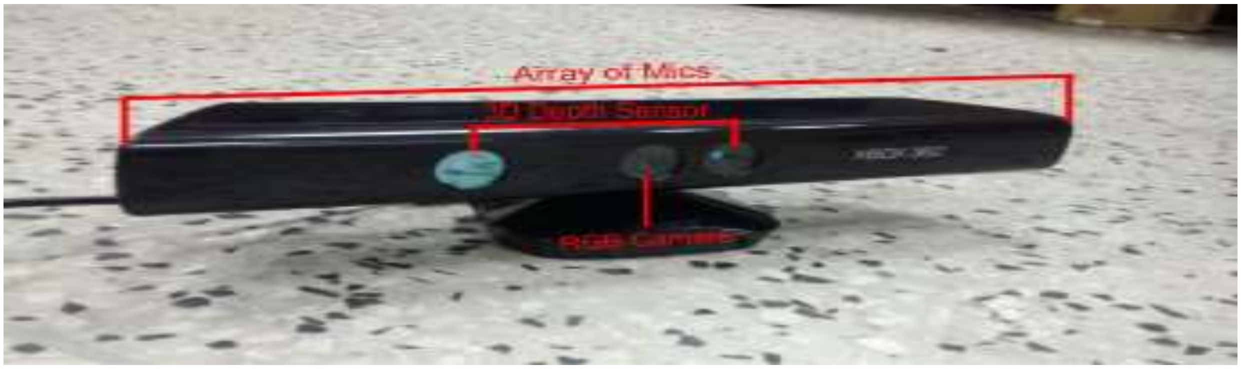


Fig 4. Microsoft kinect XBOX260

This method to finding the body position is not sufficient. Therefore to find them machine learning algorithm used which called randomize decision forest [15] which are trained by using more than thousand samples which has been skeletons associated. These algorithms learned to find the 3D joint position gives depth image. Microsoft Kinect used for the image acquition.

C. Image Acquisition

In this section we describe that the how static activity image is obtained in the real time. For the acquisition of the activity gesture, Microsoft kinet used. The following steps are in image acquisition:

(1) Colour image frames extraction.
(2) Depth image frame extraction.
(3) Skelton data used to track and extract the activity of user
(4) Background subtraction

For calculation of distance to the pixel from kinect sensor actual value of the pixel are shifted 3 places in depth frame (depth point) to. Right [13]. The below statements are in C language to show the work.

Depth= depth point << 3 (1)

he depth values calculated in the millimeters and the range of the Microsoft kinect sensor is 0.4 meters to 8 meters. If the object is closer or outer than this range can't be resolved.

Skelton streams are the most important features of Microsoft kinect. Its provide the position and location of the persons whether they tracked or not. The skeletons which are not tracked are given to zero value returned. Kinect tracked the skeleton in two modes:

- Seated mode
- Default mode

Background subtraction: Kinect is used to extract region of the interest and background subtraction. The data available in the form of skeleton to get position of the left hand i.e (xh ,yh) and left wrist (xw, yw).

Hand length = 3* max{Xabs(xh- xw),Yabs(yh- yW)}
(2)

Where, the max is the maximum values and Xabs, Yabs is the absolute value.

Now what meant that some pixels are in both images but that's would be at different location. Let's assume that pixel A is in color image located at (x, y), and the pixel in depth images located at (x + δx, y + δy) i.e this should be slightly shifted. Therefore, after taking the depth region of interest (ROI) frames, the each separate pixels mapped in the color frame. That's every pixel in depth image, we have to find location in the color image and then color intensity. Then the background subtraction process is done. Now we have to set pixels and depth information about that, replace depth information corresponding to color information and that's in the following range.

Depth-cutoff <max(hand depth, wrist depth)
(3)

All the pixel value, which has the depth value, is above than the depth cutoff shown in the black color (zero intensity). Now the resultant image that's shown without the unwanted background pixel. Kinect programmed to only track the skeleton, which are nearest to kinect sensor. Therefore, its recognize the activity of one users at one time.

D. Normalization Technique:

Normalization [12] is a process to change the range of the pixels intensity value. Normalization also called the histogram stretching or contrast stretching. In general field of data processing such as the image processing, it refers to as the dynamic range expansion. The purpose of the dynamic range expansion in various applications usually to bring image or the other type of the signal into the range, which are normal or more familiar to sense hence the terms normalization. Often motivation is to achieve consistency in the dynamic range for set of data, images or signal to avoid the mental distractions or fatigue. Normalization transform the n-dimensional gray scale images with the intensity values in range (min, max) into the new image with the intensity value in range (new min, new max). All the elements are scaled in the range of -1 to +1 in the normalization technique. The main advantage to normalization is to remove the infraclass variation between the data if the same activity is performed by different persons.

In=(I-Imin) ((new max[{-newmin)}])/max[{-min}]+ newmin (4)

Furthermore, we have the intensity range between 50-180 of the images and we have to normalize that into the range of 0-255. In the processing of normalization we have to subtract 50 from given intensity values, then the final range are between 0-130. After that, all pixel intensity multiplied 255/130 to making in a given range. Normalization technique is a nonlinear process, these are done when the values are not in a linear relationship, in that case following formula are used,

$$In = (newmax-newmin)1/(1+e^{(((\beta-I))/\alpha)})+newmin \qquad (5)$$

Where α defines the width of input intensity range and the β defines the intensity around which range is centered. In our experiments, we are trying to normalize the Fc based on videos, which are taken by the Microsoft kinect camera, which based on entire activity videos. As given in Fig. 4 every frames we have N joints that's may result in large feature dimensions Fcc, Fcp, Fci containing N(N - 1)/2,

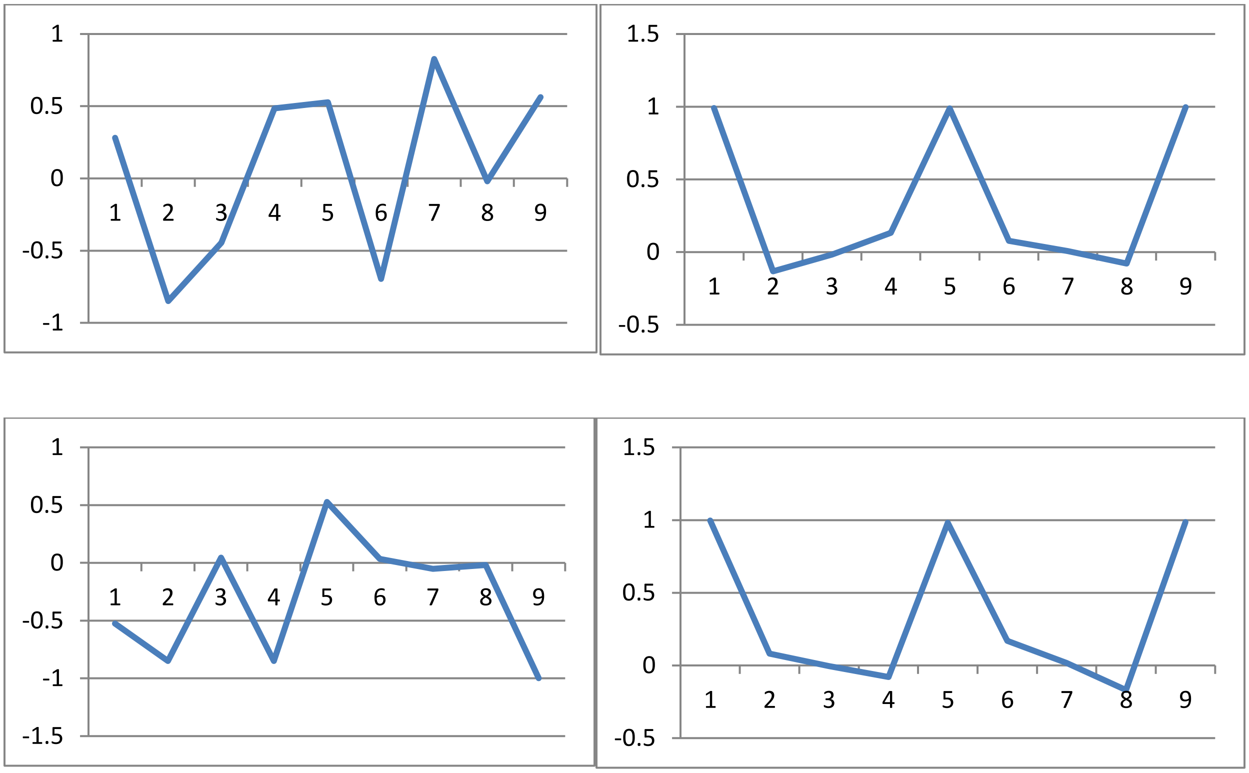
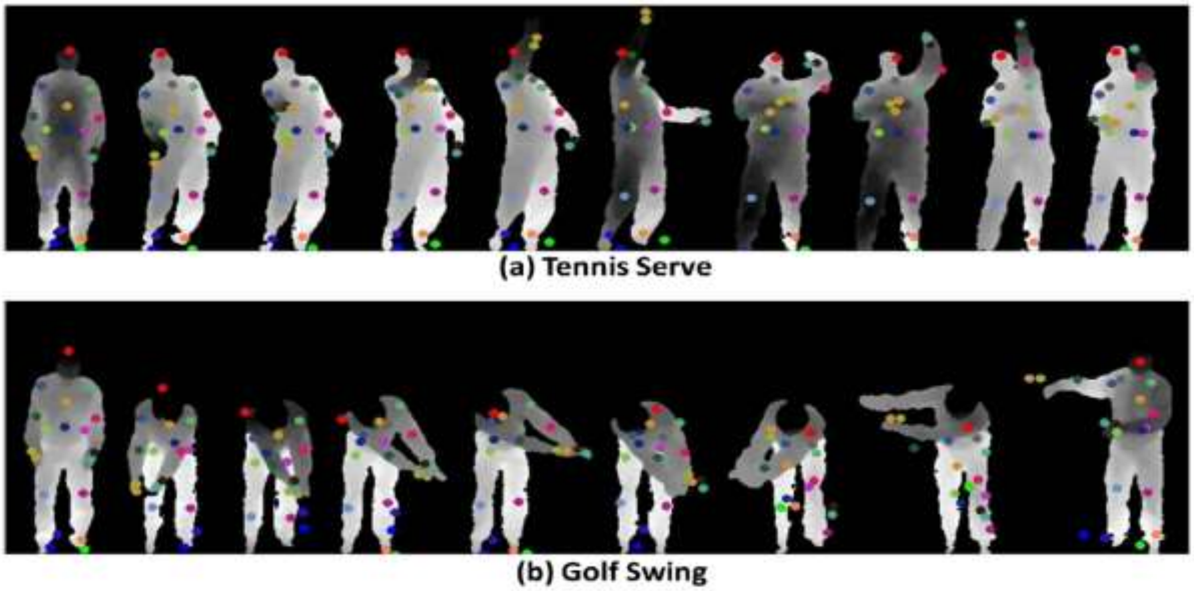


Fig 5. Normalized data joints from top to down and left to right (head, left shoulder, right shoulder, left hip)



(a) Tennis Serve

(b) Golf Swing

Fig 6. sequences of sampled depth images and skeleton joints of activity (a) Tennis Serve along with (b) Golf Swing. Each depth image consists of 20 joints. [11]

N^2, N^2 pair wise comparison. Each comparison generates the 3 element (Du, Dv, Dd). in end Fnorm is with dimensions of 3*(N(N1)/2 + N^2 + N^2).

However, if we have to find 20 skeletal joints in the every frame then the Fnorm have the 2970 length. As the skeletal joints have earlier high level information recover by depth maps, these larger dimensions having noises and redundant. After that we have to apply principal component analysis to reduce the noises in generalized Fnorm. The final representations are Eigen joints which are action descriptor of each frame. We are observing that the most of Eigen values are covered by first few leading eigenvector.

E. Classification Methods:

We have to use Principal component analysis for the feature extraction from the kinect dataset and different classification technique like Euclidean, negative and Manhattan as a classifier and neural network technique

1) Principal Component Analysis

Principal component analysis is a feature extraction technique from the data sets. New set of variable component are generated that's principal component. In the terminology of information, we want to find out significant information in images with several steps as givn followings.

Step1: *Zero Mean*: Suppose we have the data X has N variable and M observation to find the mean of data across M observation to find a mean vector $\bar{X}$. After that subtract the mean from data i.e.

$$A=X-\bar{X} \tag{6}$$

Now the data have zero mean.

Step2: *Find covariance matrix:* The next step is find the covariance matrix from the above data having size of the data matrix A is N×M. The covariance matrix given by $C=A.A^T$ have the size N×N. If the variable larger i.e. of order of ten thousand, then the covariance matrix will be very large. After that find Eigen vector will be computationally tough task have Eigen vector V_1 will be size of N×1.So slightly different method to utilized to finding Eigenvectors which is next.

Step3: *Finding Eigen vector*: To easily finding the Eigen vector/principal component of the covariance matrix mapped from a lower dimension subspace. Instead of $C=A.A^T$ are used. Consider M is less than the N then that's give a very small covariance's matrices having the size of M×M and every Eigen matrix V_1 having size of M×1. Then Eigen vectors have needs to mapped original higher dimension space, the Eigen vectors V_1 multiplied by the original normalize data matrix A. there for the Eigenvectors U1 is given by $U_1=A.V_1$. Now these Eigen vector have size of N×N.

Step4: *Dimension Reduction*: Now, the ϕ Eigen matrix in columns sorted in the descending order from the Eigen values. First p columns taken as principal components. Then the size of ϕ is N×p. For reducing the dimension of the data these operation having used,

$$A´ = \phi^{T}T; \tag{7}$$

Where the size of matrix A´ is p×M that are reduced dimensionality.

Step5: *Variance Conservation:* To find the how many principal components having sufficient to represented the data to less losses method for variance conservation [21]. We are more interested to retaining the many components conserve the 99% variances' of the data. If p= 0 means no principal components retained, then 0% of data retained. To generalized, consider that $\{\lambda_1,\lambda_2, \lambda_3, \lambda_4 \ldots\ldots\ldots\ldots\ldots\ldots\lambda_n\}$ are Eigen values across the Eigen vectors U_1 of the ϕ matrix column wise sorted eigenvalues according that matrices. If p is the retained principal, components, and then the percentage of the variance retained/conserved are calculated as,

$$\text{Variance} = \frac{\sum_{j=1}^{p} \lambda \text{j}}{\sum_{j=1}^{p} \lambda \text{j}} \tag{8}$$

Our target is to choose smallest value of p such as variance>= 0.99.

Step 6: *Reconstruct the data*: the original data could be constructed by A´ as,

$$\bar{A} = \phi A´ \tag{9}$$

It's noted that the $\bar{A}$ have dimension N×M have an approximation of original data A.

- ***Euclidean distance***

If the two points A, B is having Cartesian coordinates A (A1 , A2 , A3,…………An) & B (B1 , B2 , B3………Bn), then the Euclidean distance between then,
D (A, B) = D (B, A) =√ (〖〖(A〗 _1-B_1)〗 ^2+ 〖〖(A〗 _2-B_2)〗 ^2+ 〖〖(A〗 _n-B_n)〗 ^2) If Euclidean vector is a position of points a Euclidean n-space. So A & B are Eigen vector.
‖ A ‖ = √(〖A1〗 ^2+ 〖A2〗 ^2… 〖An〗 ^2) = √(A.A)
A vector also described a line segment from origin of Euclidean space to the point at that space

- ***Manhattan distance***

Manhattan distance basically the distance between two points which is measured along with axes at right angles.
D (A, B) = ‖ A – B ‖ = ∑_ (i=0)^n 〖| Ai-Bi|〗
(10)

Where (A, B) are vectors, A = (A1, A2, A3… An) & B = (B1, B2, B3… Bn), for example distance between points (A1, A2) & (B1, B2) is the
, = | A1 – A2 | + | B1 – B2 |
(11)

- ***Negative distance***

Negative distances are the weighted function which applies to input to get the weighted inputs. For example if we have to random weight matrix A and B then the negative distance X define as
X = -sqrt(sum(A - B)ˆ2) (12)

We use several distances based classifiers to compare the distance between dataset. Therefore, the variance of data actions more efficiently evaluated.

III. RESULT & DISCUSSION

In this section, results of proposed system have been demonstrated considering the different parameters such as performance and percentage accuracy of classification. We have also calculated the errors in terms of mean square error (MSE) of training samples of dataset

Dataset used Cornell based dataset having video sequences of human activities in the form of RGB images has been captured by Microsoft Kinect camera associated with depth map. Where each frame consists of fifteen skeleton joints available in world coordinates. All the action videos are of thirty Hz, each of 640×480 resolution. Dataset is having eight different activities in different environment on a single subject. All the eight activities have been selected from the human common activities as depicted in Fig. 2. • Platform used All the proposed system for automatic human action recognition using Eigen vectors have been implemented and analyzed in matlab toolbox on version R2013a successfully.

Platform used All the proposed system for automatic human action recognition using Eigen vectors have been implemented and analyzed in matlab toolbox on version R2013a successfully. Further we have compared the result with existing related publications that has been done so far.

Result comparisons here are the details of all the comparative results by using different techniques such as Euclidean distance, Negative distance, Manhattan distance and ANN shown in Fig. 9. We have trained the dataset with 15 frames, 25 frames and 30 frames and tested by 3 samples from each class. While we are getting more accuracy when number of samples are increased from 3 to 6 as Fig. 10 has achieved more accurate result when number of sample is increased to 6.

Table I Results Comparisons With 3 Samples Using Pca

Methods Vs Results	Tested frames(15 for each class)	Tested frames(25 for each class)	Tested frames(30 for each class)	Aggregate results
Euclidean distance	79.2 %	83.3 %	91.66 %	84.72 %
Negative distance	75.0 %	79.2 %	83.3 %	79.5 %
Manhattan distance	79.2 %	83.2 %	87.5 %	83.3 %

Table II. Results comparisons with 6 samples using pca

Methods Vs Results	Tested frames (15 for each class)	Tested frames(25 for each class)	Tested frames(30 for each class)	Aggregate results
Euclidean distance	87.5 %	91.6 %	95.8 %	91.66 %
Negative distance	87.5 %	91.6 %	93.7 %	90.9 %
Manhattan distance	87.5 %	91.6 %	93.7 %	90.9 %

• **Accuracy assessment** Accuracy assessment is one of the important tasks for analyzing the accuracy of proposed system. Fig. 4.5 is showing the mean square error at different epoch levels as the graph has achieved the best performance at epoch level 51 in terms of validation that is 0.0065999 MSE.

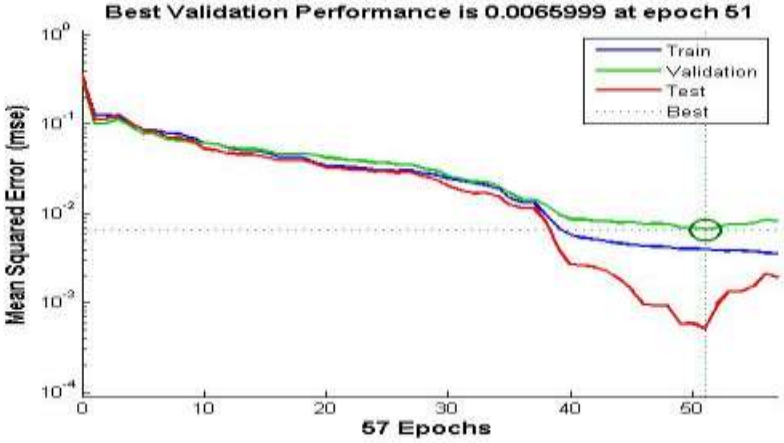


Fig. 7 Performance Analysis with MSE.

IV. CONCLUSION

In our recognition approach we have to recognize the activity of single object. My approach assumes that single person have present and only one act perform at a time. In future we have to extend the work for the multi object to perform different activities at a time, and each human activity recognize separately. The technique can be advanced to angle changes of different types of activity by different camera in the dataset

REFERENCES

1. *Cedras, Claudette, and Mubarak Shah. "Motion-based recognition a survey."Image and Vision Computing 13.2 (1995): 129-155*

2. *Raman, Natraj, and Stephen J. Maybank."Action classification using a discriminative multilevel HDP-HMM." Neurocomputing 154 (2015): 149-161.*

3. *Foggia, Pasquale, GennaroPercannella, and Mario Vento."Graph matching and learning in pattern recognition in the last 10 years." International Journal of Pattern Recognition and Artificial Intelligence 28.01 (2014): 1450001.*

4. *Li, Y. F., Jianwei Zhang, and Wanliang Wang. Active sensor planning for multiview vision tasks.Vol. 1. Heidelberg: Springer, 2008.*

5. *Xiaodong Yang; YingLiTian, "EigenJoints-based action recognition using Naïve-Bayes-Nearest-Neighbor," Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on , vol., no., pp.14,19, 16-21 June 2012.*

6. *Xia, Lu, Chia-Chih Chen, and J. K. Aggarwal. "View invariant human action recognition using histograms of 3d joints." Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on.IEEE, 2012.*

7. *Vantigodi, S.; Babu, R.V., "Real-time human action recognition from motion capture data," Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on , vol., no., pp.1,4, 18-21 Dec. 2013.*

8. *Vantigodi, Suraj, and VenkateshBabuRadhakrishnan. "Action recognition from motion capture data using meta-cognitive rbf network classifier." Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2014 IEEE Ninth International Conference on. IEEE, 2014.*

9. *Kao, Wei-Chia, Shih-Chung Hsu, and Chung-Lin Huang."Human upperbody motion capturing using Kinect." Audio, Language and Image Processing (ICALIP), 2014 International Conference on.IEEE, 2014.*

10. *Liang, Yan, et al. "Action Recognition Using Local Joints Structure and Histograms of 3D Joints." Computational Intelligence and Security (CIS), 2014 Tenth International Conference on.IEEE, 2014.*

11. *Ijjina, Earnest Paul, and C. Krishna Mohan. "Human action recognition based on motion capture information using fuzzy convolution neural networks."Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on.IEEE, 2015.*

12. *H. Gunes, M. Piccardi, Automatic temporal segment detection and affect recognition from face and body display, IEEE Trans. Syst. Man Cybern. B 39 (1)(2009) 64–84.*
13. *H. Liu, M. Sun, R. Wu, S. Yu, Automatic video activity detection using compressed domain motion trajectories for H.264 videos, J. Visual Commun.Image Represent. 22 (5) (2011) 432–439.*
14. *H. Pirsiavash, D. Ramanan, Detecting activities of daily living in firstperson camera views, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2847-2854.*
15. *K. Schindler, L. Gool, Action snippets: how many frames does human action recognition require? in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.*
16. *L. Xia, C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: IEEE CVPR Workshop on Human Activity nderstanding from 3D Data, 2012.*