Human Activity Recognition in Real-Times Environments using Skeleton Joints

Ajay Kumar, Anil Kumar, Satish Kumar Singh,Rahul Kala

Robotics and Artificial Intelligence Laboratory, Indian Institute of Information Technology, Allahabd, India

***Abstract* — In this research work, we proposed a most effective noble approach for Human activity recognition in real-time environments. We recognize several distinct dynamic human activity actions using kinect. A 3D skeleton data is processed from real-time video gesture to sequence of frames and getter skeleton joints (Energy Joints, orientation, rotations of joint angles) from selected setof frames. We are using joint angle and orientations, rotations information from Kinect therefore less computation required. However, after extracting the set of frames we implemented several classification techniques Principal Component Analysis (PCA) with several distance based classifiers and Artificial Neural Network (ANN) respectively with some variants for classify our all different gesture models. However, we conclude that use very less number of frame (10-15%) for train our system efficiently from the entire set of gesture frames. Moreover, after successfully completion of our classification methods we clinch an excellent overall accuracy 94%, 96% and 98% respectively. We finally observe that our proposed system is more useful than comparing to other existing system, therefore our model is best suitable for real-time application such as in video games for player action/gesture recognition.**

***Keywords* — Human Activity, Kinect, Skeleton Joints, Principle Component Analysis, Artificial Neural Network, Gesture Recognition**

I. Introduction

Automatic activity recognition in real-time environment for Intelligent Device or Robotics has been automatic interpretation between human activities and perception. However, integrating them by using controls, machines, and electronics enhance its artificial intelligence efficiently. Recently, Robotics and intelligent devices has been used in more areas such as field medical robotics, service robotics, or human enlarge. The idea that has been proposed in our work targets for generate human like skills to execute a task, identifying the human activities and learning with their movements. Many researchers have been demonstrated related work in large scale since the 1050 in these areas such as robotics research and computer vision. The goal of smart device 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 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. Moreover, many research in literature survey on the topic activity recognition for 2D videos. The spatio temporal approach has mostly used to measure the similarity between two activities. To find the accurate similarity calculation, the spatio temporal detection method and representation have proposed [8, 16]. NBNN and HMM methods widely used for the human activity recognition [8, 5]. 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.

Many research works has been developed related to recognizing the actions of human. Human actions can be recognized in the form of skeleton [9], silhouettes [6], and in the form of images [7]. Visual surveillance technique is used for identifying packages of human actions [2]. Researchers used different techniques to recognize human actions such as hierarchical probabilistic approach [3], multi-modality representation of joints [4], HDP-HMM which is multi-level [5], Eigen-joint based method [8] using NBNN classifier. All the mentioned work is not reliable for real world application. Researchers have been explored different varieties of compact representation for human actions. However, authors has developed a technique to recognize the human actions which is independent of human action duration and starting location of actions [10]. A lot of work is developed by researchers to identify human actions from the sequence of images of captured human actions video [11], [12], [14] and [15]. In such type of technique, the main disadvantage is the prediction of parts of body in the sequence of actions. Several authors have developed method based on image processing using histograms [13] of 3D joints to recognize human actions. In [16] it has been demonstrated that how many frames are collected and required to execute action recognition. Apart from this, many other methodologies have been implemented in [17], [18], [19] and [20].

Temporal templates is a new approach to represent the human action[23] The representation of temporal templates is a stable vector images having every points vector values and task of the motion belonging in mage sequence. Two components of templates, sample the first value is a function of regency of motion in a sequence. After that we have to construct a recognition method which

DOI: 10.9781/ijimai.2016.379

has two matching with temporal templates against the stored action. The develop method automatically perform temporal templates segmentation and its run in real time and independent linear changes in speed of standard platform. In this paper [24]authors have to take only required frames which are enough to 90% correct activity recognition and we conclude that the 10-20 frames or 7-10 snippets are more than sufficient to perform similar than the entire videos based on experiments. These methods are used both form and motion features sampled densely over the image plane. The method investigates the question, how long video snippets are required to serve the basic unit for action recognition. The advantage of extracting both form further this method perform well on different database without changes any parameter. The main advantage of this paper the action recognizes well with a very short snippet of frames (frame rate 25Htz) which are not applicable for invariance to rotation scale and viewpoint. The Activity recognition using several hardware devices is used in several previous motivational research works. Several more human activity/action recognitions are discussed using gait patterns [28-33].

A. Activity recognition in videos captured with single camera: In this paper [3] the author presents an approach to activity recognition, video matching and localization based on hierarchical code book model to local temporal video volume. This method based on "bag" of the video word representation and its does not require knowledge about the action, motion estimation, background subtraction or tracking. It's also robust to the temporal and spatial scale changes as some deformation. The algorithm code video as a compact set of the temporal and spatio volume, while consider spatio-temporal composition order to account of spatial and temporal context information. The hierarchy achieved by constructing a codebook of its video volumes. Then large context volume contains many spatiotemporal volume has to consider. These ensembles used to construct probabilistic models of video volume and its spatio-temporal composition. The algorithm applied in three available videos dataset for the action recognition over different complexity (Weizmann, MSR and KTH). The result is superior to other approach and better in case of single training data example and the cross-data activity recognition. The result is highly competitive over state of art method. However, major advantage of this approaches, does not required foreground and background tracking and segmentation, it is susceptible to the analysis on line real time. The proposed method could easily extend to the multi action localization and action retrieval by modifying inference mechanism. Since the proposed method code the video used to spatio-temporal volumes and its compositional information, it could not impose any type of constraints over the video content. Therefore, it may be extended unconstraint video matching and the content based search. The main advantage of the propose algorithm for activity recognition in the video it's not required a model of event. However, it has some drawback that does need to address in future work.

B. Activity recognition model with joint representation in 3D space : In this paper [4] recognized that the multi-level modality and perspective to representation of the joints and the 3D action recognition. In certain the RGB sequences and the depth images, construct the best differences motion history image and after recognize many perspective calculation for getting motion processes, then these histograms extract the gradient from the each calculation to explain the target motion, finally multiperspective and the multi projections joint represent, the recognition and discriminant model proposed challenges for human activity recognition. The MSR 3D actions mostly experiment to showing the difference between two motion images, two modalities history and the performances are better than the MHI at the same time, these descripted technique also efficient and very strong what's more to represent and recognize model to improve the further performances.

II. Proposed Methodology

We have to build a new approach on the basis of the differences of the skeletal joint in spatial domain and temporal domains. Moreover, after on 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) and Artificial Neural Network (ANN) respectively 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.

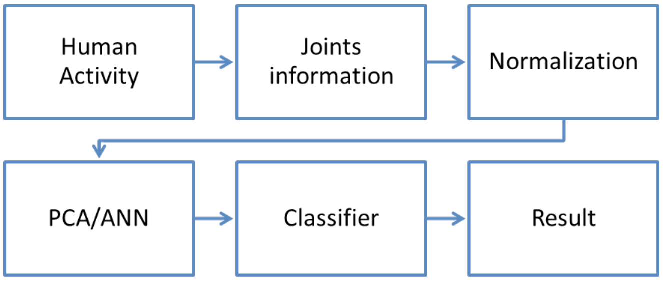


Fig.1 Graphical representation of proposed methodology

A. Human Activity:

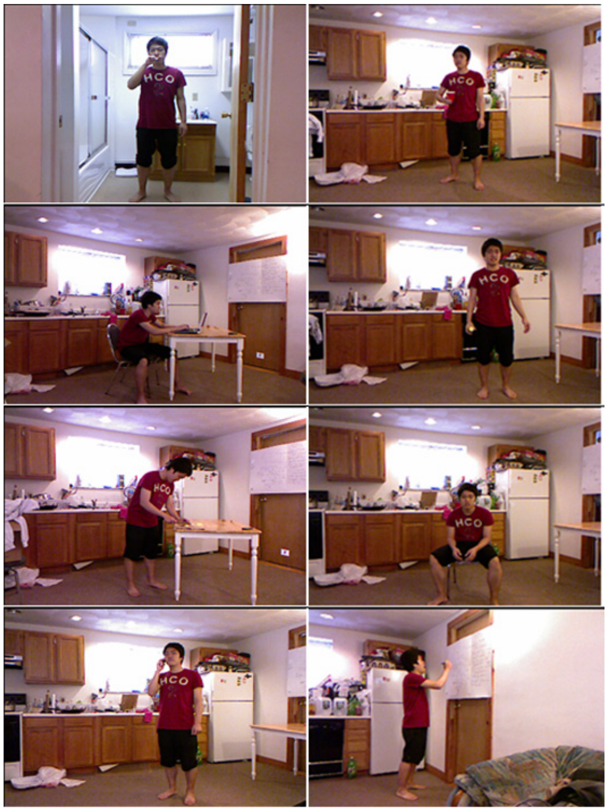


Fig.2- Corresponding activity and depth image of dataset top to down and left to right (1-Brushing teeth. 2-Working on computer. 3- Cooking (chopping). 4- Talking on the phone. 5- Drinking water. 6- Opening pill container. 7- Talking on couch. 8- Writing on whiteboard)[27].

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 .

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 its 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 analysed 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 analysed by the camera and creates the depth map of the object. Kinect could be able to analysed 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

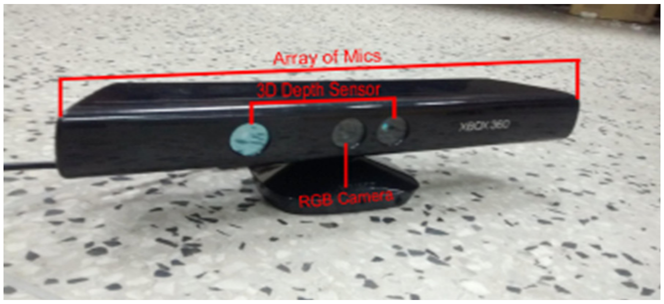


Fig 3 Microsoft kinect XBOX260

This method to finding the body position is not sufficient. Therefore to find them machine learning algorithm used which called randomize decision forest [21] 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: Out of various resolutions available we have to obtained only 640x480 resolution images.

(2) Depth image frame extraction: It is also obtained at the resolutions of 640x480 images.

(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 [22]. The below statements are in C language to show the work.

$$\text{Depth}= \text{depth point} << 3 \tag{1}$$

The depth values calculated in the millimetres 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

Default mode tracked the all 15 joints and in seated mode the person could be tracked only the upper body part having 10 joint positions is obtained. The most important uses of the depth map are background subtraction from the frames. That pixel which doesn’t belonging to region of the interests subtracted. So only that pixel remains which have redundant to zero intensity value.

(1) 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).

$$\text{Hand length} = 3^{*} \max\{\text{Xabs}(xh- xw),\text{Yabs}(yh- yW)\} \tag{2}$$

Where, the max is the maximum values and Xabs,Yabs is the absolute value. We have to set the minimum size of the hand is 95×95 pixel so same hand image will be determined even that both hands and wrist pixel values are coincide.

One more thing noted that in the kinect there have two camera RGB and other is infrared. There have some distance between them the region of that is depth image have not aligned. 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 pixels 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 are in the following range.

$$\text{Depth-cutoff} < \max(\text{hand\_depth}, \ \text{wrist\_depth}) \tag{3}$$

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 [26]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 grayscale 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.

$$\text{In}=(\text{I-Imin}) \ ((\text{new max}\llbracket\text{-newmin})\rrbracket)/\text{max}\llbracket\text{-min}\rrbracket+ \text{newmin} \tag{4}$$

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,

$$\text{In} = (\text{newmax-newmin})1/(1+e\hat{}(((\beta\text{-I}))/\alpha))+\text{newmin} \tag{5}$$

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,

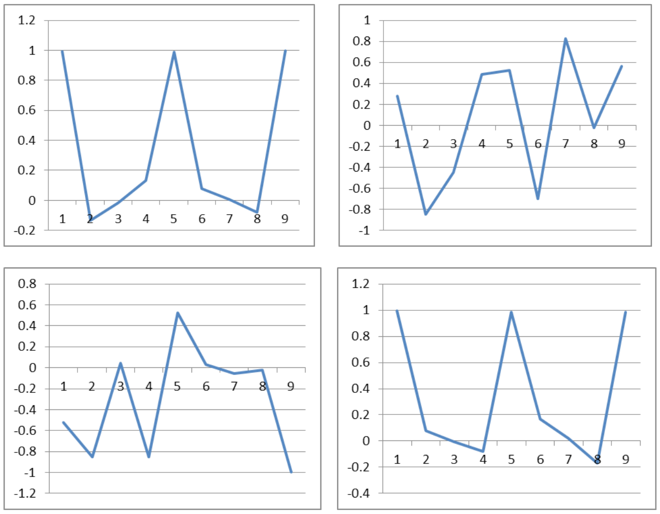
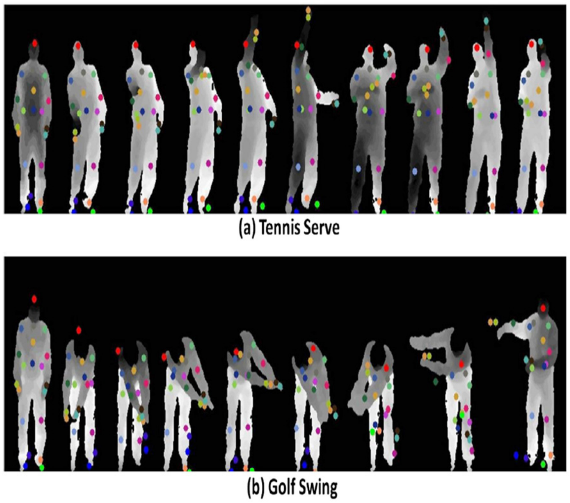


Fig 4. Normalized data joints from top to down and left to right(head, left shoulder, right shoulder, left hip).


(a) Tennis Serve

(b) Golf Swing

Fig 5. 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. [26]

N^2, N^2 pairwise comparison. Each comparison generates the 3 element (Du, Dv, Dd). in end Fnorm is with dimensions of 3*(N(N-1)/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.

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

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 V1 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 V1 having size of M×1. Then Eigen vectors have needs to mapped original higher dimension space, the Eigen vectors V1 multiplied by the original normalize data matrix A. there for the Eigenvectors U1 is given by U1=A.V1. 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 {λ1,λ2, λ3, λ4……………… λn} are eigen values across the eigen vectors U1 of the ϕ matrix column wise sorted eigenvalues according that matrices. If p is the retained principal components then the percentage of the variance retained/conserved are calculated as,

Variance=(∑_(j=1)^pλj)/(∑_(j=1)^pλj) (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.

However, We have to use following difference based classification techniques.

- 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 them ,

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,

$$= | A1 – A2 | + | B1 – B2 | \tag{11}$$

- 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 defind 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.

2) **Artificial Neural Network :**

This section elaborates decision making procedure of a well-known classification technique i.e ANN (artificial neural network).A classification procedure generally encountered when a predefined p class/group has to be assigned to an object over a number of calculating attributes associated with that object. ANN demonstration requires the processed data set mentioned earlier. We have chosen 70% from the activity dataset for training while 30% of the testing data set. We have developed eight classes having fourteen samples each afterwards ANN Mat lab toolbox has been utilized we have generated the final class i.e. target class.

The class having a maximum score to the remainder of the classes is determined of the output neuron. The back propagation performs the task of train/retrain artificial network. Feed forwarding action has been applied to an input provided to the artificial network. As a result we are getting better output for iterations S.

Errors in all the iterations S are analysed and again feed forwarded to the artificial neural network to adjust all the weights along with biases.

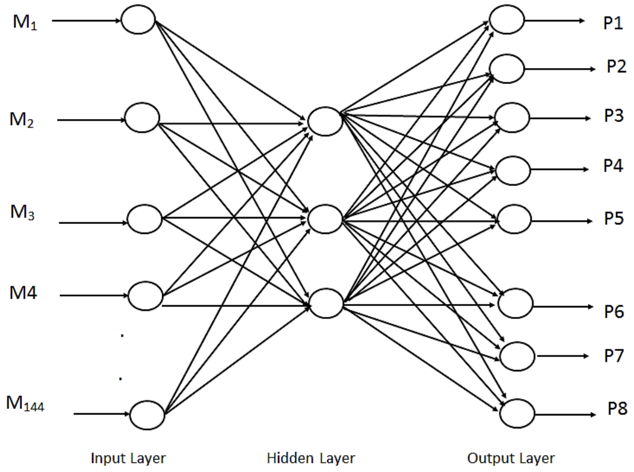


Fig.6 The functioning diagram of artificial neural network

Let Mij(s) be a weight for interface neuron i and j.

Let the result generated with neuron j be Pj.

The error is evaluated with the help of equation 13

$$Ej(s) = K(s) – pj(s) \tag{13}$$

Where K, E and q represents desired output, error and output respectively weights can be modified with the help of equation 14

$$Mij(s+1) = Mij(s) + \alpha\beta i(s)pj(s) \tag{14}$$

Where α indicates a real constant number, also called learning rate and its range is less than 1. βjis evaluated with the help of equation 13 and 14 respectively.

$$Bj\ (s) = Ej(s)\phi j'(\mu j(s)) \tag{15}$$

$$\mu j\ (s) = \sum iMij\ (s)\ qj\ (s) \tag{16}$$

Where ϕ indicates transfer function.

In the architecture of neural network consist 3 type neuron layers input, hidden and output layer.In the feed forward network(FFN) the signals are flow from input units to output unit strictly in FFD. The processing of the data extend from multiple layers of the units but there are no any feedback connection available and in recurrent network consist of feedback connections. Activation value of units undergoes the relaxations process of network involves to stable state in that activation doesn't change more. In one other application changes of activation value of neuron significant, such as dynamical output of the network.

The neural network is configuring as the application combination of input produce the desire group of the output. There are various method set strength of connection, exit. One other way to set weight finally by using forward knowledge. One other method to train the neural network to feed it teach pattern and change it weight according some learning rule. The learning situation classified into three sorts in neural network.

III. Result analysis

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 analysed in matlab toolbox on version R2013a successfully. Further we have compared the result with existing related publications that has been done so far.

- *Class Variations among data*

Here we have done comparative analysis of inter class variations and intra class variations of the Cornell dataset of human activities. Intra class variation is defined as the variation exists within class and Inter class variation is defined as the variation exists between two different classes. A set of classes is said to be well if there is a low intra class variation and high inter class variation as our dataset classes are having the low intra class variation and high inter class variation mentioned in Fig. 7. Regions in red are a class different from the green region class. It has been clearly observed that both the classes is well separated to each other.

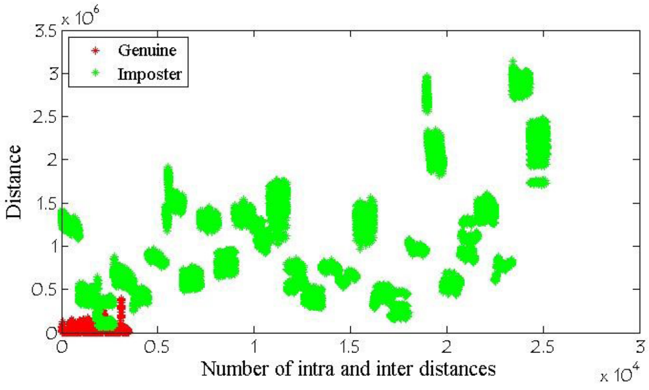


Fig. 7. intra and inter class distance distribution

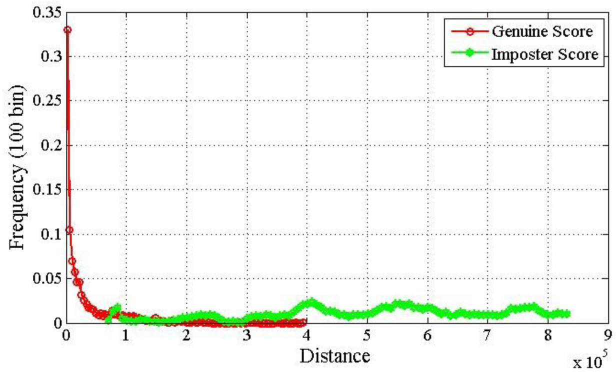


Fig. 8 intra and inter class distance histogram

- **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 %
ANN	92.5 %	94.5 %	98.2 %	95.1 %

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 %
ANN	92.5 %	94.5 %	98.2 %	95.1 %

- **Confusion matrix**

The given Confusion matrix is a tabular representation of classification between actual and predicted classes. Each row and column shows the different class of human activity. In confusion matrix correct entries are represented by diagonal cell and incorrect entries is represented by off-diagonal entries. Rows and column indicates the actual and predicted output of the classifier.

TABLE III.
CONFUSION MATRIX OF ACTIVITY USING ANN

Gesture type	Brushing teeth	Working on the computer	Cooking	Talkig on the phone	Drinking Water	Opening pill Container	Talking on the couch	Writing on the whiteboard
Brushing teeth	**100%**	0%	0%	0%	0%	0%	0%	0%
Working on the computer	0%	98%	2%	0%	0%	0%	0%	0%
Cooking	0%	0%	99%	1%	0%	0%	0%	0%
Talking on the phone	0%	0%	2%	97%	1%	0%	0%	0%
Drinking Water	0%	0%	2%	0%	98%	0%	0%	0%
Opening pill Container	0%	0%	0%	0%	0%	100%	0%	0%
Talking on the couch	0%	0%	0%	0%	3%	0%	97%	0%
Writing on the whiteboard	0%	0%	0%	0%	0%	0%	0%	100%

- **Accuracy assessment**

Accuracy assessment is one of the important task 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.

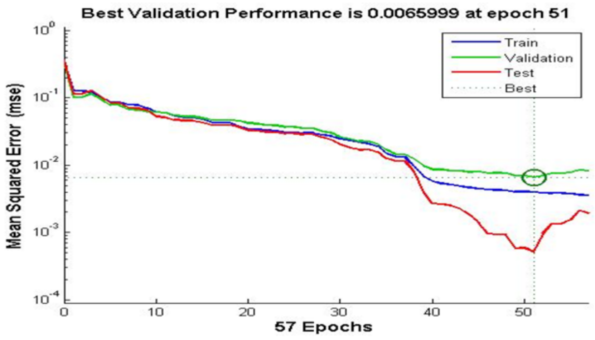


Fig. 9 Performance Analysis with MSE.

IV. Conclusion And Future Work

In this research work we proposed a activity recognition technique which are based on the Eigen joint captured by the Microsoft kinect camera for different distance based classifier technique. Three distances are used in my work and compared the result by the other classification techniques. We have to see that the 10-12 % frames are more than enough to recognize the activity with the best accuracy from the activity video. In other computer vision systems, mostly the human activity recognitions are highly depends on context. The activities are different in every situation. Besides, what composes an activity is a part of the perception. It would be interesting to debate we have a wide range of scenarios can define a common set of action or not. Any recognition problem is finding the key to strong representative facility Pattern sets. Action a 4D event, coordinates (3D) points in the human body shape is travelling with direction of time. On abstraction facility level, variance for style changes, anthropometry changes, see the dressing changes Speed changes, noise tolerance, feature extraction and ease of calculation speed etc. For dynamic time system are problem to be seen2D or 3D representations by various trade -offs of the details that are associated with human motion another interesting problem is therefore, above all possibility to obtaining a group of common properties, which will act better in view of the challenging problem. Other interesting aspects are the questions that must be taken in part to system to better performance receive. Example ,variance On the scene by finding invariant features, feature extraction level can be controlled working with different approaches in the training level of the training include samples dataset or view level classifier using invariant matching technique. Like this, the complexity of the system which blocks up that way must be introduced the best performance of integrated system is better topic matter for investigate. Speed, accuracy, and robustness are typical contradictory of vision systems. That preference should be given to the order depending on the application area of these three performances measures. In the end, whether it will work satisfactorily, which is possible to build a common identification system battle for various applications to answer thought-annoying question is which are still exploring abroad amount of analysis in that area is needed.

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.

V. Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

References

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

[2] Ye, Juan, Simon Dobson, and Susan McKeever. "Situation identification techniques in pervasive computing: A review." Pervasive and mobile computing8.1 (2012): 36-66.

[3] Augustyniak, Piotr, et al. "Seamless tracing of human behavior using complementary wearable and house-embedded sensors." Sensors 14.5 (2014): 7831-7856.

[4] Liu, An-An, et al. "Coupled hidden conditional random fields for RGB-D human action recognition." Signal Processing 112 (2015): 74-82.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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

[11] 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.

[12] Kao, Wei-Chia, Shih-Chung Hsu, and Chung-Lin Huang."Human upper-body motion capturing using Kinect." Audio, Language and Image Processing (ICALIP), 2014 International Conference on.IEEE, 2014.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2847-2854

[18] 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.

[19] 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.

[20] Z. Zhang, D. Tao, Slow feature analysis for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2012) 436–450.

[21] Andrew Ng. Cs229 machine learning autumn 2013. http://cs229. stanford. edu. Accessed: 2014-06-20.

[22] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, J. Li, Hierarchical spatio-temporal context modeling for action recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2004–2011.

[23] A. Bobick, J. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Anal. Mach. Intell. 23 (3) (2001) 257–267.

[24] 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.

[25] O. Boiman, E. Shechtman, M. Irani, In defense of Nearest-Neighbor based image classification, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[26] Yang, Xiaodong, and YingLiTian. "Effective 3d action recognition using

[27] eigenjoints." Journal of Visual Communication and Image Representation 25.1 (2014): 2-11.

[28] Online Access http://pr.cs.cornell.edu/humanactivities/data.php

[29] Semwal, Vijay Bhaskar, et al. "Biped model based on human Gait pattern parameters for sagittal plane movement." Control, Automation, Robotics and Embedded Systems (CARE), 2013 International Conference on. IEEE, 2013.

[30] Semwal, Vijay Bhaskar, Manish Raj, and G. C. Nandi. "Biometric gait identification based on a multilayer perceptron." Robotics and Autonomous Systems 65 (2015): 65-75.

[31] Semwal, Vijay Bhaskar, et al. "Biologically-inspired push recovery capable bipedal locomotion modeling through hybrid automata." Robotics and Autonomous Systems 70 (2015): 181-190.

[32] Gupta, Jay Prakash, et al. "Human activity recognition using gait pattern."International Journal of Computer Vision and Image Processing (IJCVIP) 3.3 (2013): 31-53.

[33] Semwal, Vijay Bhaskar, Manish Raj, and G. C. Nandi. "Multilayer perceptron based biometric GAIT identification." Robotics and Autonomous Systems. Available online 21 (2014).

[34] Gupta, Jay Prakash, et al. "Analysis of Gait Pattern to Recognize the Human Activities." arXiv preprint arXiv:1407.4867 (2014).

Ajay Kumar received the M.Tech degrees in information technology from Indian Institute of Information Technology Allahabad, India in 2015. He is the recipient of the GATE scholarship from the Ministry of Human Resource Development, Government of India. His area of interests are Control System, Control Robotics, Image Processing, Digital Electronics.

Anil Kumar received the M.Tech degrees in information technology from Indian Institute of Information Technology Allahabad, India in 2015. He is author of more than four research papers in the field of Robotics and Artificial Intelligence. He is the recipient of the GATE scholarship from the Ministry of Human Resource Development, Government of India. His area of interests are robot motion planning, Temporal Logic, Path Planning.

Satish Kumar Singh (M'11-SM'14) is currentlyworking as an assistant professor in Indian Institute of Information Technology, Allahabad, India. He has completed his Ph.D., M. Tech. & B. Tech in 2010, 2005 and 2003 respectively. He is having more than 10 years of experience in academic and research institutions. He has several publications in international journal and conference proceedings of repute. He is member of various professional societies like, IEEE and IETE etc. He is an Executive Committee Member of IEEE Uttar Pradesh Section. He is serving as editorial board member and reviewer for many international journals. His current research interests are in the areas of digital image processing, pattern recognition, multimedia data indexing and retrieval, watermarking and biometrics.

Dr. Rahul Kala received the B.Tech. and M.Tech. degrees in information technology from the Indian Institute of Information Technology and Management, Gwalior, India in 2010. He received his Ph.D. degree in cybernetics from the University of Reading, UK in 2013. He is currently working as an Assistant Professor in the Indian Institute of Information Technology, Allahabad, India in the Robotics and Artificial Intelligence Laboratory. He is the author of three books and over 70 papers. His recent book is on robotic planning is entitled Intelligent Planning for Mobile Robotics: Algorithmic Approaches (IGIGlobal Publishers, 2013). He is a recipient of the Commonwealth Scholarship and Fellowship Program from the UK Government; the Lord of the Code Scholarship from RedHat and the Indian Institute of Technology Bombay; and the GATE scholarship from the Ministry of Human Resource Development,Government of India.