

# Human Skeleton Matching for E-learning of Dance Using A Probabilistic Neural Network

Sriparna Saha<sup>1</sup>, Rimita Lahiri<sup>2</sup>, Amit Konar<sup>3</sup>

<sup>1,2,3</sup>Electronics & Tele-Communication Engineering Department

<sup>1,2,3</sup>Jadavpur University, Kolkata, India

<sup>1</sup>sahasriparna@gmail.com, <sup>2</sup>rimita.lahiri@yahoo.com, <sup>3</sup>konaramit@yahoo.co.in

Bonny Banerjee<sup>4</sup>, Atulya K. Nagar<sup>5</sup>

<sup>4</sup>Electrical & Computer Engineering Department

<sup>4</sup>The University of Memphis, United States of America

<sup>5</sup>Mathematics and Computer Science Department

<sup>5</sup>Liverpool Hope University, United Kingdom

<sup>4</sup>bonnybanerjee@yahoo.com, <sup>5</sup>nagara@hope.ac.uk

**Abstract**—With the growing interest in the domain of human computer interaction (HCI) these days, budding research professionals are coming up with novel ideas of developing more versatile and flexible modes of communication between a man and a machine. Using the attributes of internet, the scientists have been able to create a web based social platform for learning any desired art by the subject himself/herself, and this particular procedure is termed as electronic learning or e-learning. In this paper, we propose a novel application of gesture dependent e-learning of dance. This e-learning procedure may provide help to many dance enthusiasts who cannot learn the art because of the scarcity of resources despite having great zeal. The paper mainly deals with recognition of different dance gestures of a trained user such that after detecting the discrepancies between the gestures shown and actually performed by a novice; the user can rectify his faults. The elementary knowledge of geometry has been employed to introduce the concept of planes in the feature extraction stage. Actually, five planes have been constructed to signify major body parts while keeping the synchronous parts in one unit. Then four distances and four angular features have been obtained to provide entire positional information of the different body joints. Finally, using a probabilistic neural network the dance gestures have been classified after training the said network with sufficient amount of data recorded from numerous subjects to maintain generality.

**Keywords**— human computer interaction; e-learning of dance; Kinect sensor; probabilistic neural network

## I. INTRODUCTION

With the technological advancement in the field of human computer interaction (HCI), it has become quite easy to develop interactive models that provide effective ways of communication in terms of cost efficiency, time complexity and portability as well. Using the newly developed sensors with RGB camera embedded within the device itself, it is possible to detect different human body gestures and analyze the same by generating skeletal data, containing three dimensional body joints information in terms of position as well as orientation. Literally, the term ‘gesture’ signifies the different postures that a human structure embodies while displaying various actions as well as emotions. Dance is considered as one of the best ways of conveying human emotions. Instead of considering static uncorrelated postures

of a user at different time instances to convey his/her mental states, dance is chosen as a preferred alternative as it incorporates both facial expressions and body movements, which collectively enhances the ability of a user to convey his emotions. Moreover, with the smooth transition between different dance poses, it adds an aesthetic value which enables a layman to understand the message that is intended to be conveyed through that concerned act.

As an obvious consequence of such rapid progress of different engineering methodologies, social networking is a part and parcel of our daily life. This motivated researchers to delve further in the concerned domain and come up with ideas for building an immersive and collaborative environment to support more realistic and versatile interactions between human beings and a computer. To engage a larger percentage of the research fraternity in solving these real time problems, many popular multinational research organizations are coming up with innovative challenges. For example, Huawei 3DLife/EMC2 challenge [1] provides a platform for demonstrations of online human computer interactive systems and thus encourages the research community to solve the problems faced by the industry or improve the already acquired solutions to move closer to the desired output. With the recent advent of internet, e-learning program has emerged as one of the most popular internet accessory, which helps a certain human being to acquire different skills online with the help of internet, which wouldn't have been possible for the user to learn in person otherwise. It is needless to tell, machine intelligence plays an important role in these kinds of systems. In this paper, we propose a novel cost effective and fast e-learning procedure of detecting different dance postures embodied by an already trained subject, so that a novice can be able to learn from those postures and correct himself in case of any wrong gesture.

Literature show much research has already been done in the said area, but still there remains a discrepancy between the real world scenario and the artificially framed simulated environment, and our aim is to minimize that gap as far as possible. Saha *et al.* explored a completely different aspect of the same problem by presenting a novel gesture recognition algorithm to segregate between different emotive gestures of Indian classical dance [2]. But the work lacks an efficient

feature extraction stage, which is overcome in the proposed framework. Waldherr *et al.* developed an application of human robot interaction by employing camera based human movement tracking approach in order to control the movement of a robotic arm [3]. Essid *et al.* presented a new idea of multi modal dance collection based approach that facilitates the interaction between human beings in virtually framed environment [4]. As a response to the Huawei 3DLife challenge, Alexiadis *et al.* proposed a Kinect based approach for creating an online dance evaluating software, which measures the worth of a performance with respect to a standard one and generates feedback for further improvement [1].

Xu *et al.* proposed a new feed forward neural network based approach to detect and recognize hand gestures in order to develop a virtually occurring driving training system called Self Propelled Gun (SPG) [5]. Murakami *et al.* extended the neural network based approach and used another variant of neural network named recurrent neural network to carry out finger alphabet recognition dynamically [6]. The above mentioned method has implemented the framework for a small vocabulary, but there is further scope of improvement in terms of extent of application of the designed framework. Mao *et al.* proposed a new algorithm that iteratively determines the structure of a probabilistic neural network (PNN) by recurrently calculating the value of a smoothing parameter using a genetic algorithm (GA) technique [7]. Wu *et al.* developed a novel leaf pattern recognition algorithm, used for plant classification purpose, by employing PNN [8].

As discussed above, several methodologies have been adopted in the said domain, but none of them has been able to provide optimal results for e-learning of dance gestures. This motivated us to enlighten the other drawbacks and improvise the methodologies as required. This paper has emphasized on mainly five different classical dance gestures corresponding to ‘anger’, ‘fear’, ‘happiness’, ‘sadness’ and ‘relaxation’. Moreover, due to space constraints, the diagrams presented in the experimental results section primarily show more emphasis on implementation of the proposed scheme on female subjects only, hence it is important to state that the proposed scheme has been tested by employing the same on male subjects as well, and in each case more or less likely results are obtained as in case of female participants. For further analysis of the above mentioned five key emotive gestures, feature extraction seems to be an area of utmost importance in skeleton based system analysis because inaccurate feature selection may have an adverse impact upon the system performance as a whole. Here, the Microsoft’s Kinect sensor [1], [2] plays a key role of registering the three dimensional RGB images, storing depth map at a predetermined frame rate and generating three dimensional positional information in terms of coordinates from a skeleton comprising of 20 different human body joints with the help of software development kit (SDK). An innovative feature space has been formulated using a plane based approach and the concerned feature set contains both positional and angular parameters to avoid unnecessary loss of information. Since hand and leg movements need to be focused on while analyzing any dance gesture, the feature set has been constructed such that it includes every information corresponding to hand and leg joints without loss of generality.

In this paper, the elementary geometric knowledge of creating planes and deriving perpendicular distances has been used. Five different planes have been created including a reference plane based on body core (formed by spine, hip center and shoulder center joints), as this part is relatively stationary while performing any dance movement and the other four functioning planes have been constructed corresponding to the arm and leg joints respectively. In the next step, the orientation of the planes corresponding to the concerned arm and leg joints have been derived with respect to the reference plane and the perpendicular distance from the spine to each of the plane is calculated while conducting the experiment for each subject, yielding a feature set comprising of four distance based and four orientation based features for each frame of each subject. Further, using a probabilistic neural network (PNN) algorithm [7]–[10], an appreciable matching accuracy of 89.7% has been attained. The proposed framework has not only generated better performance in terms of precision parameters but it has improved the hardware and time complexity by a noticeable extent. Although the proposed system has been implemented only for Indian classical dance postures, but it can be applicable for any dance form with clearly distinguishable gestures.

This rest of the paper is divided into four sections. The detailed overview of proposed framework along with the preliminary concepts has been discussed in Section II. Section III presents the experimental results obtained during different stages of the execution of the proposed model. Section IV concludes the discussion by combining our findings.

## II. OVERVIEW OF THE PROPOSED WORK

The steps required for the proposed work for e-learning of dance gestures are given in Fig. 1. The elaboration of the overall process is given in following sub-sections.

### A. Skeleton Formation Using Microsoft’s Kinect Sensor

Kinect sensor typically looks like a long bar like device with a set of IR (infrared) and RGB (red, green, blue) cameras embedded within it. The device was primarily developed as a game peripheral controller, but later it found extensive application in the domain of gesture recognition. Actually, the Kinect sensor generates three dimensional joint coordinates information by generating a skeletal image of the subject present in front of it within a threshold distance 1.2-3.5m using SDK toolkit. The IR cameras are primarily responsible for sensing and thus generating depth map after processing the data recorded from the user standing in front of it.

A sample RGB image with its corresponding skeleton obtained using Kinect sensor is provided in Fig. 2. For the proposed work, the required 15 joints are hip center (*HC*), spine (*S*), shoulder center (*SC*), elbow left (*EL*), wrist left (*WL*), hand left (*HL*), knee left (*KL*), ankle left (*AL*), foot left (*FL*), elbow right (*ER*), wrist right (*WR*), hand right (*HR*), knee right (*KR*), ankle right (*AR*), foot right (*FR*) are highlighted.

### B. Five Planes Construction

For the next step of feature extraction, we have created five planes.

- a) *Plane 1 ( $P_{ref}$ )*: For body core - hip center ( $HC$ ), spine ( $S$ ), shoulder center ( $SC$ ).
- b) *Plane 2 ( $P_{LA}$ )*: For left arm - elbow left ( $EL$ ), wrist left ( $WL$ ), hand left ( $HL$ ).
- c) *Plane 3 ( $P_{RA}$ )*: For right arm - elbow right ( $ER$ ), wrist right ( $WR$ ), hand right ( $HR$ ).
- d) *Plane 4 ( $P_{LL}$ )*: For left leg - knee left ( $KL$ ), ankle left ( $AL$ ), foot left ( $FL$ ).
- e) *Plane 5 ( $P_{RL}$ )*: For right leg - knee right ( $KR$ ), ankle right ( $AR$ ), foot right ( $FR$ ).

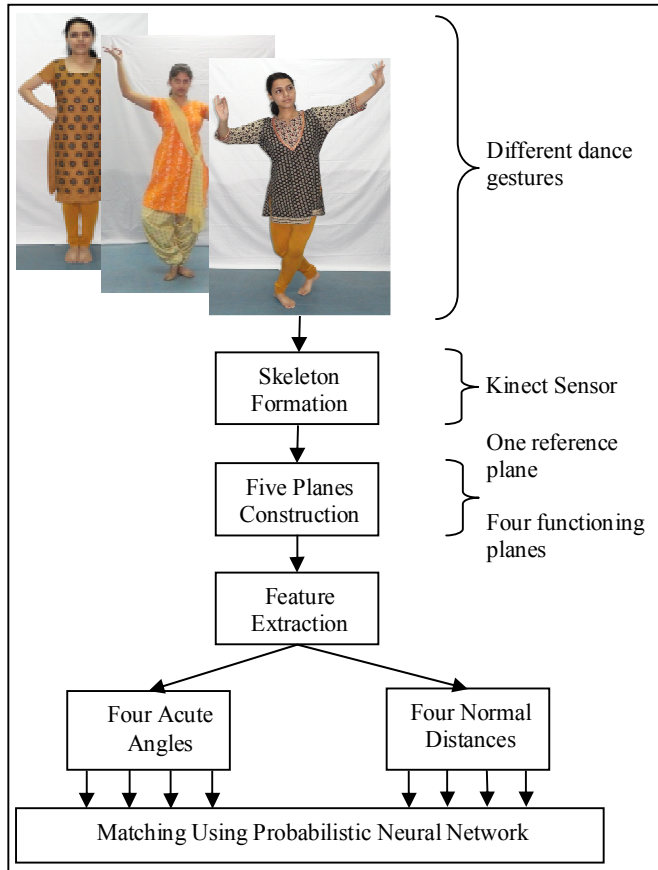


Fig. 1. Flowchart of the proposed work for e-learning of dance.

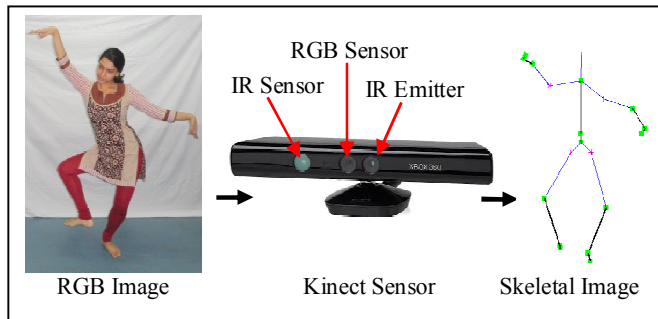


Fig. 2. The sample images to depict the skeleton construction procedure using Kinect sensor.

$P_{ref}$  is nearly constant for a specific gesture, thus it is taken as the reference plane. The rest of the planes ( $P_{LA}, P_{RA}, P_{LL}, P_{RL}$ ) are used as functioning planes. Dance can be considered as an amalgamation of several rhythmic body movements that collectively express an art, conveying a story behind it. Since body movements are involved in this case, hence body parts obviously play an important role in this art. For transitional movements, the body parts majorly involved are the arms and the legs. The Kinect sensor provides three-dimensional joint coordinates of three joints corresponding to each arm and leg as well. To accommodate all three joint information in a single entity, a plane concept has been introduced, such that a single functioning plane can singlehandedly include all the joints information. The pictorial representation of these planes is given in Fig. 3.

### C. Feature extraction

Let there be  $M$  number of subjects and each subject is depicting  $E$  different emotions using body gestures through dance. To demonstrate each  $e$  ( $1 \leq e \leq E$ ) gesture, we need  $F$  number of frames and from each frame  $f$  ( $1 \leq f \leq F$ ),  $R=8$  features are extracted. Now for a specific subject  $m$  ( $1 \leq m \leq M$ ) for a specific frame  $f$  for a specific emotion  $e$ , the extracted features are:

- a) *Feature 1 ( $e_{f,1}^m$ )*: Acute angle between  $P_{ref,f}^{e,m}$  and  $P_{LA,f}^{e,m}$ .
- b) *Feature 2 ( $e_{f,2}^m$ )*: Acute angle between  $P_{ref,f}^{e,m}$  and  $P_{RA,f}^{e,m}$ .
- c) *Feature 3 ( $e_{f,3}^m$ )*: Acute angle between  $P_{ref,f}^{e,m}$  and  $P_{LL,f}^{e,m}$ .
- d) *Feature 4 ( $e_{f,4}^m$ )*: Acute angle between  $P_{ref,f}^{e,m}$  and  $P_{RL,f}^{e,m}$ .
- e) *Feature 5 ( $e_{f,5}^m$ )*: Normal distance between  $S_f^{e,m}$  and  $P_{LA,f}^{e,m}$ .
- f) *Feature 6 ( $e_{f,6}^m$ )*: Normal distance between  $S_f^{e,m}$  and  $P_{RA,f}^{e,m}$ .
- g) *Feature 7 ( $e_{f,7}^m$ )*: Normal distance between  $S_f^{e,m}$  and  $P_{LL,f}^{e,m}$ .
- h) *Feature 8 ( $e_{f,8}^m$ )*: Normal distance between  $S_f^{e,m}$  and  $P_{RL,f}^{e,m}$ .

A natural question may arise regarding the reason behind the requirement of both angular and distance-based features. Actually, using only one sort of features, it is not feasible to provide complete positional information about different body joints. As position not only includes translational motion but orientation is also very important for dance gestures. While dancing, the body parts of a dancer do not move following a linear displacement; those parts are also realigned during execution of different gestures. Hence it is very important to include both distance-based as well as orientation-based features in order to register the extent of realignment since that has quite an impact on the resulting output. For better understanding, the formation of five planes corresponding to the skeletal images from Fig. 2 is given in Fig. 3.

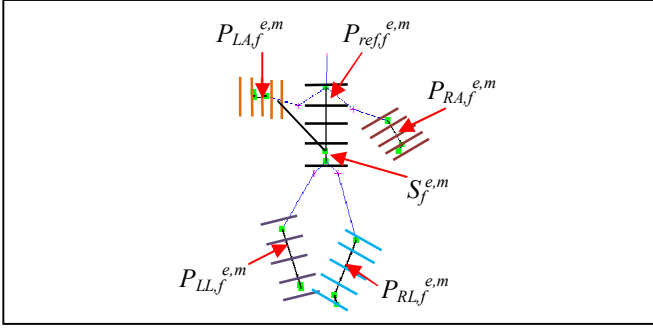


Fig. 3. Five planes formed for skeletal images from Fig. 2.

Then the angle  $\alpha$  between those two planes (the reference plane and any one of the functioning planes) is calculated using (3).

$$\alpha(\vec{n}_1, \vec{n}_2) = \arccos \frac{|x_1 \times x_2 + y_1 \times y_2 + z_1 \times z_2|}{\sqrt{x_1^2 + y_1^2 + z_1^2} \times \sqrt{x_2^2 + y_2^2 + z_2^2}} \quad (3)$$

Let the equation of a plane  $p$  is

$$ax + by + cz = d \quad (4)$$

also the co-ordinate of a point  $S$  is  $[A_1, B_1, C_1]$ . Then the normal distance  $nd$  from  $S$  to plane  $p$  is  $nd$ .

$$nd = \frac{aA_1 + bB_1 + cC_1 - d}{\sqrt{a^2 + b^2 + c^2}} \quad (5)$$

For better visualization Fig. 4 is provided which shows the procedure for acute angle and normal distance measurement. The pseudo code to calculate these features is given in Table I.

TABLE I. PSEUDO-CODE FOR FEATURE EXTRACTION.

<b>Input:</b>	3D joint co-ordinates to form reference plane: $J_{ref,1}$ , $J_{ref,2}$ and $J_{ref,3}$ . Any three 3D joint co-ordinates: $J_1$ , $J_2$ and $J_3$ . 3D joint co-ordinate of spine $S$ .
<b>Output:</b>	One acute angle feature say $\alpha$ and one normal distance say $nd$ .
<b>Procedure:</b>	
<b>Begin</b>	
	$n_{ref} = \text{cross}(J_{ref,1} - J_{ref,2}, J_{ref,1} - J_{ref,3})$
	$d_{ref} = J_{ref,1}(1) \times n_{ref}(1) + J_{ref,1}(2) \times n_{ref}(2) + J_{ref,1}(3) \times n_{ref}(3)$
	$n = \text{cross}(J_1 - J_2, J_1 - J_3)$
	$d = J_1(1) \times n(1) + J_1(2) \times n(2) + J_1(3) \times n(3)$

$$flag_1 = ((n_{ref}(1)) \times (n(1))) + ((n_{ref}(2)) \times n(2)) + ((n_{ref}(3)) \times n(3))$$

$$flag_2 = (((n_{ref}(1))^2) + ((n_{ref}(2))^2) + ((n_{ref}(3))^2))^{1/2}$$

$$flag_3 = (((n(1))^2) + ((n(2))^2) + ((n(3))^2))^{1/2}$$

$$\alpha = (\arccos(|flag_1| / (flag_2 \times flag_3))) \times 180 / \pi$$

$$flag_4 = ((n(1)) \times (S(1))) + ((n(2)) \times (S(2))) + ((n(3)) \times (S(3))) - d$$

$$flag_5 = (((n(1))^2) + ((n(2))^2) + ((n(3))^2))$$

$$nd = |flag_4 / flag_5|$$

**End.**

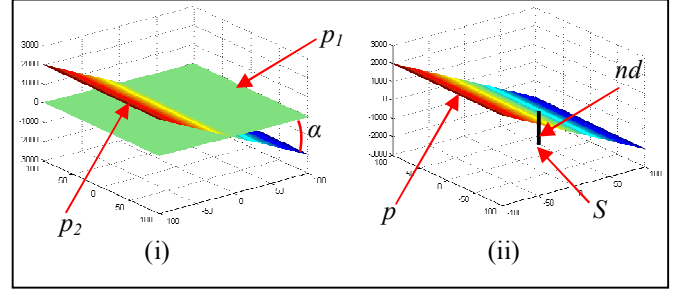


Fig. 4. Pictorial view of procedure for calculation of acute angle and normal distance.

The feature vector ( $G_e^m$ ) obtained for  $e^{\text{th}}$  gesture for  $m^{\text{th}}$  subject for all frames is given in (6).

$$G_e^m = [e_{1,1}^m \quad e_{2,1}^m \quad \dots \quad e_{f,4}^m \quad \dots \quad e_{F-1,8}^m \quad e_{F,8}^m]^T \quad (6)$$

#### D. Matching of Unknown Gesture Using Probabilistic Neural Network

Neural networks are often used for classification purpose in pattern recognition problems, different neural networks are based upon different learning rules. For example, the back propagation neural network (BPNN) employs the heuristic rule by modifying the parameters by small amounts at each iteration and thus improving the system performance. Hence BPNN takes a lot of time for training purpose, besides that it has a tendency to get trapped at local minima. So instead of using back propagation network, in this case probabilistic neural network (PNN) [7]–[10] has been preferred because of the latter's ease of use and low time complexity.

A PNN is basically a mathematical implementation of a statistical algorithm termed as Kernel Discriminant Analysis, in which the different actions are arranged in a multilayered feed forward network (FFNN) with three layers as shown in Fig. 5. The structure of a PNN is very similar to a BPNN; only the sigmoid activation function has been replaced with a statistically derived exponential function in such a way that the resulting network assures optimal convergence to Bayes decision strategy as the dimension of the training set increases. The most prominent advantage of PNN over BPNN is that new training samples can be added without disturbing the already adapted weights of the existing neurons.

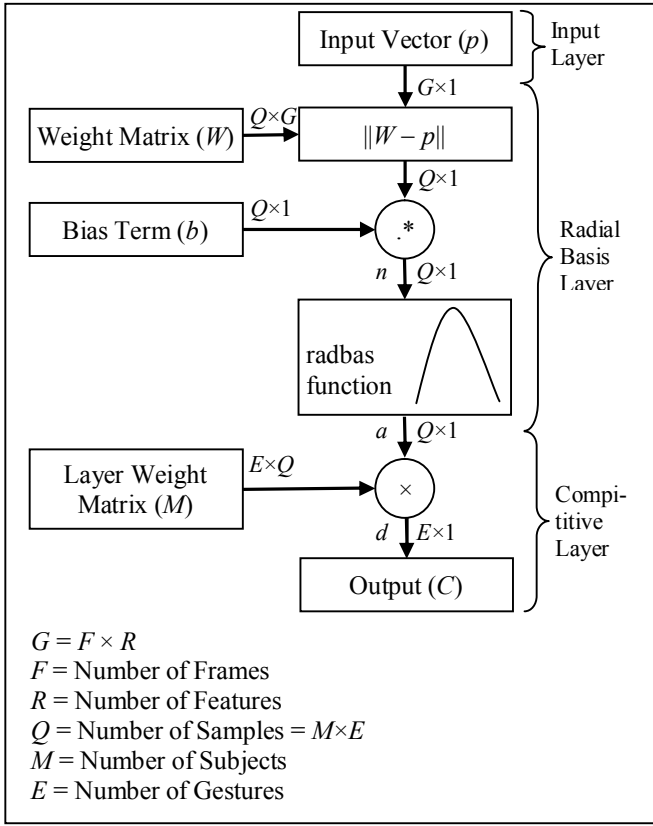


Fig. 5. Architecture of different layers for PNN.

In this paper, the PNN is comprised of three layers, an input layer, radial basis layer and a competitive layer. Once an input vector is fed to the already trained network, the first layer computes the closeness of the fed input with respect to weight vectors in terms of distance measures, and those distances are scaled using the nonlinear radial basis function. Basically, the radial basis layer sums up the contributions of each class and generates a vector formed by probabilities and the competitive layer assigns the fed input pattern to the class having the highest probability. The network structure of PNN is as follows. Fig 5. provides a detailed overview of the network structure used for the current paper.

*a) Input Layer:* The input vector denoted by  $p$  is of the dimension  $G \times 1$  in this case  $G = F \times R$ . In this way, all the features of all the frame for a specific gesture corresponding to a particular subject has been accommodated in a single feature vector.

*b) Radial Basis Layer:* The distances between the fed input vector and weight vectors of the trained network is determined. Here, a dot product measure has been adopted to signify distance with respect to the rows of weight matrix  $W$  having dimension of  $Q \times G$ . So, a matrix of dimension  $Q \times 1$  is formed such that  $i^{\text{th}}$  element of the vector  $\|W-p\|$  is actually the dot product of  $i^{\text{th}}$  row of  $W$  and the input vector. Here  $Q$  is the number of samples taken in the training phase and  $G$  is having the same dimension as the feature vector.

$$\|W-p\|_i = W(i,:) \cdot p \quad (7)$$

where  $i$  denotes the  $i^{\text{th}}$  term.

In the next step, an element wise multiplication is carried out between  $\|W-p\|$  and the bias vector ( $b$ ) of the same dimension as  $\|W-p\|$  to generate a vector  $n$  of dimension  $Q \times 1$  using the equation,

$$n = \|W-p\| \cdot b \quad (8)$$

Further, each element of  $a$  is replaced with its radial basis output. The radial basis function is defined as,

$$\text{radbas}(n) = e^{-n^2} \quad (9)$$

Finally, the output of radial basis layer is obtained as,

$$a = \text{radbas}(\|W-p\| \cdot b) \quad (10)$$

*c) Competitive Layer:* Firstly, there is no bias term in the competitive layer. The output matrix of radial basis layer ( $a$ ) is multiplied with a layer weight matrix ( $M$ ) to generate a  $E \times 1$  dimensional vector  $d$ , depending upon which a competitive function (func) assigns 1 to the largest value of  $d$ .  $M$  is a matrix of  $E \times Q$  dimension containing the class information of the training samples. Basically, it is a sparse matrix comprising of target vectors with one 1 in each column. Depending upon the class label, 1 is assigned to the corresponding row of the particular sample column. In this case the index 1 suggests the number of samples classified by the network. To obtain the output vector in a properly organized structure there are several functions that can transform it into desired form in different languages. Finally the output of the competitive layer  $C$  is of  $E \times 1$  dimension with 1 assigned to corresponding class. Suppose the input vector belongs to  $e^{\text{th}}$  class then the  $e^{\text{th}}$  element of  $C$  is assigned 1 and others are assigned as 0.

TABLE II. PSEUDO-CODE FOR PROBABILISTIC NEURAL NETWORK.

<b>Input:</b>	Number of training samples $Q$ , Weight matrix $W$ , Input vector $p$ , Bias $b$ , Number of classes $E$ .
<b>Output:</b>	Each test gesture is assigned a class label in the vector $C$ of dimension $E \times 1$ .

**Procedure:**

**Begin**

**For**  $i=0$  to  $i \leq Q$

$$\|W-p\| [i] = W [i] \cdot p$$

**End.**

$$b = 1$$

$$n = \|W-p\| \cdot b$$

$$a = \exp(-n^2)$$

Calculate  $M$  if  $i^{\text{th}}$  sample belongs to  $j^{\text{th}}$  class then 1 is assigned in  $j^{\text{th}}$  row and  $i^{\text{th}}$  column.

$$d = M \times a$$

Calculate  $\max(d)$  and 1 is assigned to the corresponding index indicating the class label, generating class label matrix  $C$  of dimension  $E \times 1$




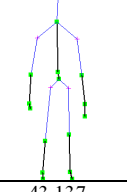
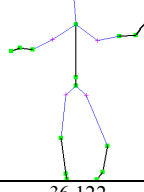
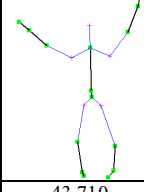



**End.**

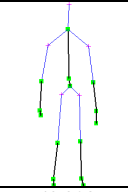
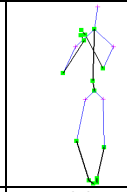
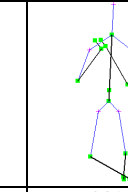



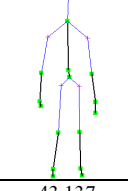
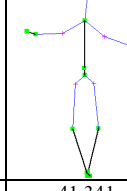
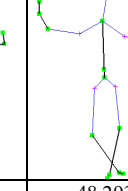



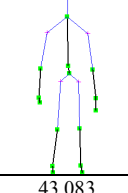
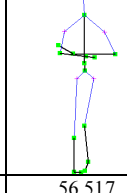
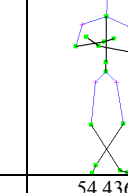
### III. EXPERIMENTAL RESULTS




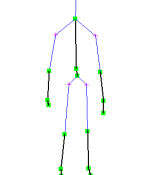
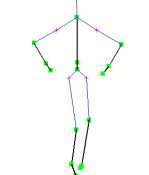
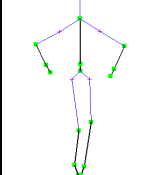
Here for these proposed work, we have taken data from 40 subjects (i.e.,  $M=40$ ). Each subject is asked to perform 5 gestures (i.e.,  $E=5$ ) each showing one emotion (e.g. ‘anger’, ‘fear’, ‘happiness’, ‘sadness’ and ‘relaxation’) through Indian classical dance. Thus total number of samples,  $Q=M \times E=40 \times 5=200$ . Each dance video is consisting of 4s duration. As Kinect sensor has sampling frequency rate is 30fps, thus  $F=4 \times 30=120$ . The number of rows in feature vector  $G_e^m$  is  $F \times R=120 \times 8=960$  as number of features  $R=8$ . Thus the dimension of  $G$  is  $960 \times 1$ .

The acquired RGB and skeletal images from Kinect sensor with corresponding feature space for 5 gestures for 25<sup>th</sup> subject is given in Table II. As it is clearly seen from Table III, the starting gestures of all the videos are almost same. Table IV shows the confusion matrix for the proposed system for all the 5 gestures considered in this paper.

TABLE III. RESULTS OBTAINED FROM 25<sup>TH</sup> SUBJECT FOR 5 GESTURES.

Gesture Type	Parameters	Frame No. 20	Frame No. 60	Frame No. 100
Anger	RGB Image			
	Skeletal Image			
	Features	43.137 48.299 7.519 6.309 0.097 0.103 0.014 0.032	36.122 84.639 8.566 58.871 0.131 1.817 0.097 0.176	43.710 21.098 9.697 70.102 0.060 0.088 0.116 0.067
Fear	RGB Image			

	Skeletal Image			
	Features	43.208 45.853 5.451 6.319 0.096 0.103 0.029 0.032	53.377 34.056 12.129 14.143 0.556 0.096 0.135 0.095	56.810 31.081 64.756 16.067 0.288 0.083 0.454 0.088
Happ- iness	RGB Image			
	Skeletal Image			
	Features	43.137 45.106 6.123 7.317 0.096 0.103 0.024 0.032	41.341 69.122 19.586 14.253 0.296 0.368 0.149 0.125	48.293 57.917 31.005 15.996 0.306 0.281 0.232 0.119
	RGB Image			
Sad- ness	Skeletal Image			
	Features	43.083 41.159 6.414 6.305 0.097 0.102 0.021 0.032	56.517 77.769 21.857 60.849 0.128 0.301 0.130 0.023	54.436 75.685 24.690 38.096 2.516 1.809 0.224 0.200

Relaxation	RGB Image			
	Skeletal Image			
	Features	43.024 4.011 5.992 6.287 0.097 0.101 0.033 0.032	59.213 59.390 69.337 66.884 0.107 0.210 0.004 0.040	60.313 63.818 68.453 70.384 0.128 0.136 1.002 0.104

Along with classification accuracy, misclassification rate should also be checked with equal importance while evaluation of the performance of a classifier. Table IV presents a confusion matrix of the five classes taken under consideration. The large values at the diagonal positions signify high accuracy with significantly lower misclassification rates.

TABLE IV. CONFUSION MATRIX FOR THE PROPOSED SYSTEM USING PNN.

Actual Class	Predicted Class				
	Anger	Fear	Happiness	Relaxation	Sadness
Anger	<b>89.792</b>	3.984	2.561	1.874	1.819
Fear	2.311	<b>90.517</b>	3.549	1.195	2.428
Happiness	3.058	2.329	<b>89.911</b>	2.079	2.623
Relaxation	1.899	2.718	1.235	<b>91.437</b>	2.711
Sadness	1.567	1.934	2.539	3.877	<b>90.083</b>

We have validated our proposed work's performance with back propagation neural network (BPNN) [11], recurrent neural network (RNN) [6], feed forward neural network (FFNN) [12], ensemble classifier using binary tree (ECBT) [13], linear support vector machine (LSVM) [14], support vector machine with radial basis function (RBFSVM) [2], k-nearest neighbor (kNN). Neural networks are biologically inspired classification algorithms. The main building blocks of such algorithms are simply neuron like processing units termed as nodes, arranged in layers. Connections between nodes of adjacent layers are denoted in terms of weights which are updated iteratively during the learning phase. ECBT aims to classify multiple classes by dividing the original dataset into binary class subsets and developing a binary model for each such newly formed subset. This algorithm is based on the principal of adaptive boosting (AdaBoost) technology. LSVM attempts to classify simply constructing hyper-plane, while RBFSVM, another variant of SVM, that classifies the data by designing a non-linear Gaussian radial basis function based kernel such that the resulting algorithms easily fits into a maximum margin hyper-plane in a transformed feature space.

kNN belongs to the category of instance based learning, where the outputs are generated in terms of membership, such that a data sample is assigned a class label depending upon the votes of its 'k' nearest neighbors after employing majority voting strategy. The performance metrics include accuracy, precision, sensitivity and specificity. The comparison results for these metrics are given in Table V. Fig. 6 shows computational complexity values obtained for the above mentioned algorithms in second unit for a Windows 7 PC with 2GB RAM.

We have used paired *t*-test for statistical comparison where PNN acts as the reference classifier and it is compared with accuracy values of other six standard classifier results given in Table V.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{((\sigma_1^2 + \sigma_2^2)/2)}} \quad (11)$$

Here,  $\bar{x}_1$  and  $\sigma_1^2$  denote the mean and standard deviation of the proposed algorithm respectively and these variables are the same for any one out of the six competitive algorithms.

TABLE V. COMPARISON OF PROPOSED WORK WITH EXISTING LITERATURES.

Algorithms	Accuracy	Precision	Sensitivity	Specificity	<i>t</i>
PNN	0.917 (0.077)	0.904 (0.035)	0.926 (0.011)	0.925 (0.030)	
BPNN	0.883 (0.095)	0.891 (0.080)	0.908 (0.031)	0.895 (0.030)	0.273 (+)
RNN	0.866 (0.036)	0.878 (0.042)	0.877 (0.031)	0.891 (0.098)	0.584 (+)
FFNN	0.827 (0.062)	0.812 (0.037)	0.829 (0.050)	0.834 (0.084)	0.899 (+)
ECBT	0.850 (0.093)	0.845 (0.066)	0.858 (0.021)	0.833 (0.029)	0.549 (+)
LSVM	0.726 (0.056)	0.756 (0.093)	0.715 (0.088)	0.757 (0.085)	1.977 (+)
RBFSVM	0.789 (0.041)	0.793 (0.086)	0.803 (0.012)	0.798 (0.079)	1.443 (+)
kNN	0.747 (0.068)	0.763 (0.0684)	0.754 (0.030)	0.748 (0.041)	1.636 (+)

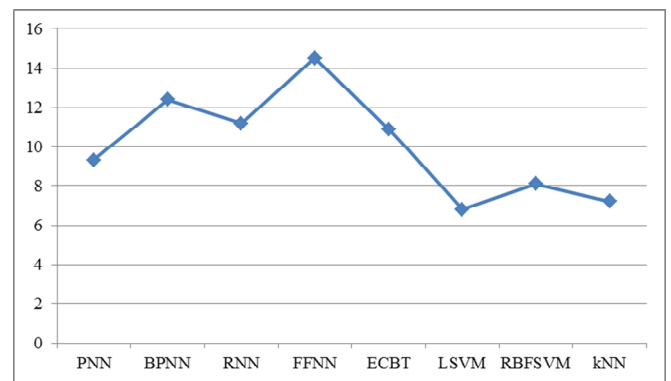


Fig. 6. Computational complexity values for the competitive algorithms.

For all the cases ‘+’ significance is obtained which indicates that  $t$  value of 49 degrees of freedom is significant at a 0.05 level by two-tailed  $t$ -test. For this test we have taken into account the accuracy values.

#### IV. CONCLUSION

This paper provides a cost effective and easy to use way of e-learning of dance. Although the experiment has primarily emphasized on Indian classical dance gestures only, but it is applicable to learning of any kind of dance form. While conducting the experiments, maximum accuracy obtained is 91.7% which is fairly good in this domain. Moreover, the feature extraction strategy includes introduction of planes passing through body joints. The feature extraction strategy adopted in the present work is very efficient and has little scope of error, which reduces the error complexity of the entire system as a whole. Employing this strategy has not only improved the performance of the proposed framework, but it has also become easier to modify the network as per requirement of the specific application. Most importantly, the presented framework is very easy to install, so it can be recommended for real world applications due to its ease of use.

Since in dance forms facial expression plays an important role, there is ample scope of applying image processing based gesture recognition strategy that would take care of the facial expression as well.

#### ACKNOWLEDGMENT

The research work is supported by the University Grants Commission, India, University with Potential for Excellence Program (Phase II) in Cognitive Science, Jadavpur University and University Grants Commission (UGC) for providing fellowship to the first author.

#### REFERENCES

- [1] D. S. Alexiadis, P. Kelly, P. Daras, N. E. O’Connor, T. Boubekeur, and M. Ben Moussa, “Evaluating a dancer’s performance using kinect-based skeleton tracking,” in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 659–662.
- [2] S. Saha, S. Ghosh, A. Konar, and A. K. Nagar, “Gesture Recognition from Indian Classical Dance Using Kinect Sensor,” in *Computational Intelligence, Communication Systems and Networks (CICSyN), 2013 Fifth International Conference on*, 2013, pp. 3–8.
- [3] S. Waldherr, R. Romero, and S. Thrun, “A gesture based interface for human-robot interaction,” *Auton. Robots*, vol. 9, no. 2, pp. 151–173, 2000.
- [4] S. Essid, X. Lin, M. Gowing, G. Kordelas, A. Aksay, P. Kelly, T. Fillon, Q. Zhang, A. Dielmann, and V. Kitanovski, “A multi-modal dance corpus for research into interaction between humans in virtual environments,” *J. Multimodal User Interfaces*, vol. 7, no. 1–2, pp. 157–170, 2013.
- [5] D. Xu, “A neural network approach for hand gesture recognition in virtual reality driving training system of SPG,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, vol. 3, pp. 519–522.
- [6] K. Murakami and H. Taguchi, “Gesture recognition using recurrent neural networks,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1991, pp. 237–242.
- [7] S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L. Xiang, “A leaf recognition algorithm for plant classification using probabilistic neural network,” in *Signal Processing and Information Technology, 2007 IEEE International Symposium on*, 2007, pp. 11–16.
- [8] L. Shang, D.-S. Huang, J.-X. Du, and C.-H. Zheng, “Palmprint recognition using FastICA algorithm and radial basis probabilistic neural network,” *Neurocomputing*, vol. 69, no. 13, pp. 1782–1786, 2006.
- [9] D. F. Specht, “Probabilistic neural networks for classification, mapping, or associative memory,” in *Neural Networks, 1988., IEEE International Conference on*, 1988, pp. 525–532.
- [10] D. F. Specht, “Probabilistic neural networks,” *Neural networks*, vol. 3, no. 1, pp. 109–118, 1990.
- [11] S. Saha, M. Pal, A. Konar, and R. Janarthanan, “Neural Network Based Gesture Recognition for Elderly Health Care Using Kinect Sensor,” in *Swarm, Evolutionary, and Memetic Computing*, Springer, 2013, pp. 376–386.
- [12] M. Leena, K. Srinivasa Rao, and B. Yegnanarayana, “Neural network classifiers for language identification using phonotactic and prosodic features,” in *Intelligent Sensing and Information Processing, 2005. Proceedings of 2005 International Conference on*, 2005, pp. 404–408.
- [13] S. Saha, M. Pal, A. Konar, and J. Roy, “Ensemble Classifier-Based Physical Disorder Recognition System Using Kinect Sensor,” in *Computational Advancement in Communication Circuits and Systems*, Springer, 2015, pp. 169–175.
- [14] S. Saha, S. Datta, A. Konar, and R. Janarthanan, “A study on emotion recognition from body gestures using Kinect sensor,” in *Communications and Signal Processing (ICCSP), 2014 International Conference on*, 2014, pp. 56–60.