Gesture Recognition from Two-Person Interactions Using Ensemble Decision Tree

Sriparna Saha, Biswarup Ganguly and Amit Konar

Abstract The evolution of depth sensors has furnished a new horizon for human–computer interaction. An efficient two-person interaction detection system is proposed for an improved human−computer interaction using Kinect sensor. This device is able to identify twenty body joint coordinates in 3D space among which sixteen joints are selected and those have been adapted with certain weights to form four average points. The direction cosines of these four average points are evaluated followed by the angles made by x, y and z axes, respectively, i.e., twelve angles have been constructed for each frame. For recognition purpose, ensemble of tree classifiers with bagging mechanism is used. This novel work is widely acceptable for various gesture-based computer appliances and yields a recognition rate of 87.15%.

Keywords Human–computer interaction · Kinect sensor · Direction cosines · Ensemble decision tree

1 Introduction

Human body tracking is a well-studied topic in today's era of human–computer interaction (HCI) [1], and it can be formed by the virtue of human skeleton structures. These skeleton structures have been detected successfully due to the smart progress of some devices, used to measure depth (e.g., Sony PlayStation, Kinect sensor). Human body movements have been viewed using these depth

S. Saha (✉) · B. Ganguly · A. Konar
Electronics & Tele-Communication Engineering Department,
Jadavpur University, Kolkata, West Bengal, India
e-mail: sahasriparna@gmail.com

B. Ganguly
e-mail: biswarupgangulyee24@gmail.com

A. Konar
e-mail: akonar@etce.jdvu.ac

P.K. Sa et al. (eds.), *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Advances in Intelligent Systems and Computing 518,
DOI 10.1007/978-981-10-3373-5_29

sensors which can provide sufficient accuracy while tracking full body in real-time mode with low cost.

In reality action and reaction activities are hardly periodic in a multi-person perspective situation. Also, recognizing their complex aperiodic gestures are highly challenging for detection in surveillance system. Interaction detections like pushing, kicking, punching, exchanging objects are the essence of this work. Here, two-person interactions have been recognized by an RGB-D sensor, named as Kinect [2, 3].

Park and Aggarwal [4] have presented two-person interactions via natural language descriptors at a semantic level. They have adapted linguistics for representing human actions by forming triplets. Static poses and dynamic gestures are recognized through hierarchical Bayesian network. Yao et al. [5] have indicated velocity features to posses the best accuracy while recognizing single-person activity. A comparative study between SVM and multiple-instance learning boost classifier is drawn. Yun et al. [6] have recognized several interactions acted by two persons using an RGB-D sensor. The color and depth image information is extracted from the skeleton model captured from the sensor. Six different body pose features are considered as the feature set where the joint features work better than others. Saha et al. [7] have proposed a superior approach of two-person interaction model where eight interactions have been modeled and recognition is achieved through multi-class support vector machine (SVM) with rotation invariance case. In the last two literature survey works, the right person is in action while the left one is initially static and gradually reacting to the situation. But in our approach, we built such a model that any person can act or react according to the interaction delivered by any person. Thus, the possible number of gesture interaction reduces in such a case. If two persons are asked to perform eight interactions, the total interactions would be twenty ($=^{8}C_{2} - 8$) and here lies the novelty of our work.

Here, we have implemented Kinect sensor to identify eight interactions using skeleton structures with some predefined features. As Kinect sensor captures twenty body joint coordinates in 3D, out of which sixteen have been adapted with some specified weights to form four mean points. The direction cosines of these four average points, i.e., twelve angles per frame, are the feature set of our proposed work. For a specific interaction, 3s video stream is captured to detect the skeleton. Ensemble decision tree (EDT) [8, 9] with bagging technology is employed for recognition purpose with a accuracy of 87.15%.

2 Fundamental Ideas

Kinect [2, 3] is a combination of a camera, an infrared (IR) emitter–receiver, a microphone block and a tilt motor. The RGB camera captures three-dimensional data at 30 frames per second in a 640 × 480 resolution. The IR camera estimates the reflected beam and measures the depth of the subject from the Kinect sensor in 1.2–3.5 m range.

Ensemble decision tree is a well-known pattern recognition classifier [8, 9]. Here, 'tree' classifier is employed as base classifier. The predictions of each base classifier are associated to determine the class of the test samples. Bagging mechanism is carried out for classification. In bagging, classifiers are trained by various datasets, obtained from the original dataset via bootstrapping. The divergence among the weak learners is examined by this re-sampling process, repeated T times. Then majority voting is taken to predict the class of the unknown. Here, T has been chosen to be 100, and bootstrap size (n) is taken 30% of the total dataset.

3 Proposed Algorithm

The block diagram of the proposed algorithm is given in Fig. 1. Suppose for our proposed algorithm, number of subjects to be chosen are N and number of actions to be executed for a single person are G. Thus, the total interactions possible between the two persons are GG. Now when one specific subject n ($1 \leq n \leq N$) is asked to perform a particular action g ($1 \leq g \leq G$), we have captured a total number of T frames. Now for each tth ($1 \leq t \leq T$) frame, we have twenty 3D body joints information out of which sixteen joints are selected for this suggested work. These selected joints are shoulder left (SL), shoulder right (SR), elbow left (EL), elbow right (ER), wrist left (WL), wrist right (WR), hand left (HaL), hand right

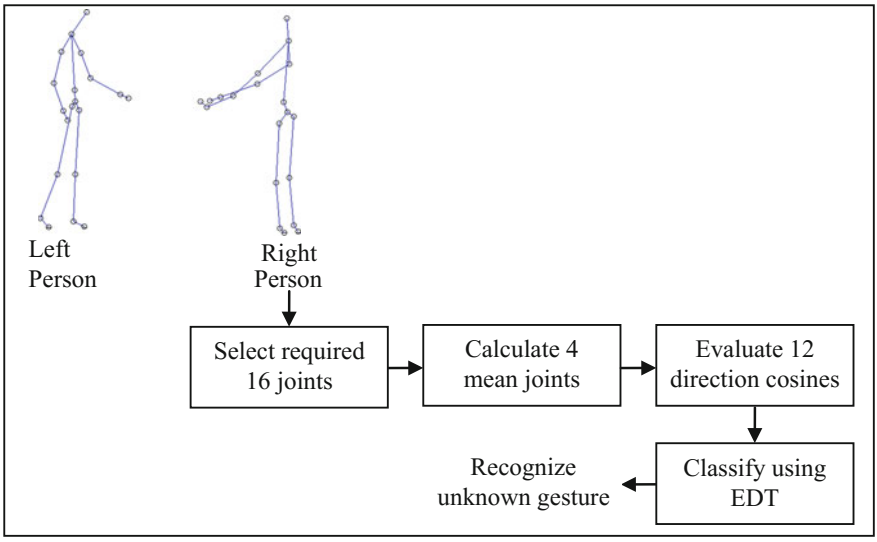

Fig. 1 Block diagram for gesture recognition for two-person interactions (as same steps need to be followed for both the persons; thus, steps required to be followed for *right* person are given only)

(*HaR*), hip left (*HL*), hip right (*HR*), knee left (*KL*), knee right (*KR*), ankle left (*AL*), ankle right (*AR*), foot left (*FL*) and foot right (*FR*).

Now four mean joints are formulated taking four body joints from arm or leg at a time. But while calculating these mean joints, we have given weights to the corresponding body joints based on the distances between them [7]. Considering left arm, the distance between *SL* and *EL* is the highest, while the distance between *WL* and *HaL* is the lowest. The weightage given to the sixteen body joints should be according to these ratios. Finally, the four mean joints become,

$$J_1^t = \frac{w^{SL} \times SL^t + w^{EL} \times EL^t + w^{WL} \times WL^t + w^{HaL} \times HaL^t}{4} \tag{1}$$

$$J_2^t = \frac{w^{SR} \times SR^t + w^{ER} \times ER^t + w^{WR} \times WR^t + w^{HaR} \times HaR^t}{4} \tag{2}$$

$$J_3^t = \frac{w^{HL} \times HL^t + w^{KL} \times KL^t + w^{AL} \times AL^t + w^{FL} \times FL^t}{4} \tag{3}$$

$$J_4^t = \frac{w^{HR} \times HR^t + w^{KR} \times KR^t + w^{AR} \times AR^t + w^{FR} \times FR^t}{4} \tag{4}$$

where w is the respective weights given to the body joints according to the superscript values. These weight values are same irrespective of time or frame number. The weights have been adopted in such a fashion that $w^{SL} + w^{EL} + w^{WL} + w^{HaL} \approx 1$, $w^{SR} + w^{ER} + w^{WR} + w^{HaR} \approx 1$, $w^{HL} + w^{KL} + w^{AL} + w^{FL} \approx 1$, $w^{HR} + w^{KR} + w^{AR} + w^{FR} \approx 1$ [10]. From each J_i ($1 \leq i \leq 4$) bearing 3D coordinate information, direction cosines ($\cos\alpha_{Ji}^t$, $\cos\beta_{Ji}^t$, $\cos\gamma_{Ji}^t$) are evaluated followed by the angles α, β, γ; that the mean joints make with positive x, y and z axes, respectively, using Eqs. (5)–(7).

$$\alpha_{Ji}^t = \cos^{-1} \frac{x_{Ji}^t}{\sqrt{(x_{Ji}^t)^2 + (y_{Ji}^t)^2 + (z_{Ji}^t)^2}} \tag{5}$$

$$\beta_{Ji}^t = \cos^{-1} \frac{y_{Ji}^t}{\sqrt{(x_{Ji}^t)^2 + (y_{Ji}^t)^2 + (z_{Ji}^t)^2}} \tag{6}$$

$$\gamma_{Ji}^t = \cos^{-1} \frac{z_{Ji}^t}{\sqrt{(x_{Ji}^t)^2 + (y_{Ji}^t)^2 + (z_{Ji}^t)^2}} \tag{7}$$

where x, y, z represents the 3D axes. Thus, for a specific subject n to interact with a particular action g, we have a total of twelve angles (three angles from each J_i) for a particular frame, which forms the feature space of our modeled work. For each

action performed by each person, we have T number of frames and twelve features have been extracted per frame. Therefore, the dimension of the feature space becomes $T \times 12$. Since the total training dataset is composed of N subjects and eight interactions, the total dimension is $N \times 8 \times T \times 12$. Whenever an unknown interaction is delivered by two persons, we segregate the two actions performed by two subjects and each subject's body gestures are recognized using EDT already specified in Sect. 2.

4 Experimental Results

For the proposed work, we have granted $GG = {}^{8}C_2 - 8 = 20$ two-person interactions while each person is showing $G = 8$ actions, namely approaching, departing, exchanging, hugging, shaking hands, punching, pushing and kicking. The number of subjects in the dataset is taken as $N = 70$ in the age group 25–35 yrs, and each interaction is taken for 3s, i.e., $T = 30$ fps $\times$ 3 s = 90 frames. The calculation procedure of cosine angles is explained in Table 1. The weights, by which the joints are adjusted, are $w^{SL} = w^{SR} = 0.271$, $w^{EL} = w^{ER} = 0.449$, $w^{WL} = w^{WR} = 0.149$, $w^{HaL} = w^{HaR} = 0.131$, $w^{HL} = w^{HR} = 0.348$, $w^{KL} = w^{KR} = 0.437$, $w^{AL} = w^{AR} = 0.119$, $w^{FL} = w^{FR} = 0.096$ [10]. The recognized interaction GG is approaching–hugging, i.e., the left person gesture is approaching, while the same for right person is hugging. The comparison of proposed method is done with support vector machine (SVM), k-nearest neighbor (k-NN) and back-propagation neural network (BPNN) as given in Fig. 2.

Table 1 Procedure for calculation of feature space for the subject in the right side from Fig. 2

3D coordinates obtained						Direction cosines			
SL^{48}	−0.423	0.440	3.048	J_1^{48}	−0.109	0.049	0.736		
EL^{48}	−0.468	0.199	2.933					$\cos\alpha_{J1}^{48}$	−0.146
WL^{48}	−0.411	−0.009	2.878					$\cos\beta_{J1}^{48}$	0.066
HaL^{48}	−0.388	−0.086	2.858					$\cos\gamma_{J1}^{48}$	0.986
SR^{48}	−0.258	0.407	3.289	J_2^{48}	−0.059	0.029	0.812		
ER^{48}	−0.233	0.110	3.161					$\cos\alpha_{J2}^{48}$	−0.072
WR^{48}	−0.212	−0.136	3.321					$\cos\beta_{J2}^{48}$	0.036
HaR^{48}	−0.218	−0.179	3.386					$\cos\gamma_{J2}^{48}$	0.996
HL^{48}	−0.336	0.050	3.039	J_3^{48}	−0.091	−0.090	0.743		
KL^{48}	−0.379	−0.451	2.980					$\cos\alpha_{J3}^{48}$	−0.120
AL^{48}	−0.407	−0.808	2.865					$\cos\beta_{J3}^{48}$	-0.120
FL^{48}	−0.355	−0.876	2.840					$\cos\gamma_{J3}^{48}$	0.985
HR^{48}	−0.260	0.037	3.179	J_4^{48}	−0.072	−0.094	0.783		
KR^{48}	−0.272	−0.486	3.202					$\cos\alpha_{J4}^{48}$	−0.091
AR^{48}	−0.374	−0.807	2.865					$\cos\beta_{J4}^{48}$	−0.119
FR^{48}	−0.340	−0.844	2.975					$\cos\gamma_{J4}^{48}$	0.989

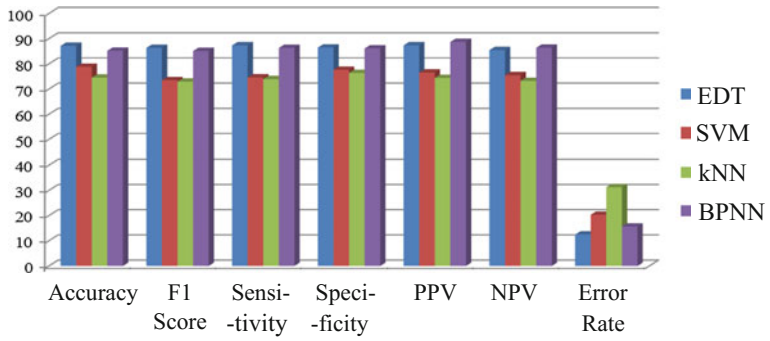

Fig. 2 Performance analysis with standard classifiers

5 Conclusion

Our proposed scheme recognizes twenty two-person interactions to explore an improved human–computer interaction in an efficient technique using Kinect sensor. Previous research [7] has been gone through calculating only average points, but in our proposed work, the body joints have been adapted with some weights based upon the distance between different body joints of human anatomy. We have obtained a better recognition rate of 87.15%. Therefore, this proposed method discovers its application from vision-based gesture recognition to public place surveillance. In the upcoming days, we will figure out to enhance our dataset comprising of some complicated interactions between two persons and recognized them with some statistical models like hidden Markov model, hidden conditional random field.

Acknowledgements 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. S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, 2015.
2. M. R. Andersen, T. Jensen, P. Lisouski, A. K. Mortensen, M. K. Hansen, T. Gregersen, and P. Ahrendt, "Kinect depth sensor evaluation for computer vision applications," *Tech. Rep. Electron. Comput. Eng.*, vol. 1, no. 6, 2015.
3. T. T. Dao, H. Tannous, P. Pouletaut, D. Gamet, D. Istrate, and M. C. H. B. Tho, "Interactive and Connected Rehabilitation Systems for E-Health," *IRBM*, 2016.
4. S. Park and J. K. Aggarwal, "Event semantics in two-person interactions," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, vol. 4, pp. 227–230.

5. A. Yao, J. Gall, G. Fanelli, and L. J. Van Gool, "Does Human Action Recognition Benefit from Pose Estimation?.," in *BMVC*, 2011, vol. 3, p. 6.
6. K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 2012, pp. 28–35.
7. S. Saha, A. Konar, and R. Janarthanan, "Two Person Interaction Detection Using Kinect Sensor," in *Facets of Uncertainties and Applications*, Springer, 2015, pp. 167–176.
8. T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.
9. T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple classifier systems*, Springer, 2000, pp. 1–15.
10. R. Drillis, R. Contini, and M. Maurice Bluestein, "Body segment parameters 1," *Artif. Limbs*, p. 44, 1966.