Skeleton based Human Action Recognition using Kinect

Ayushi Gahlot
Dept of CSE
IMS Engineering college, Ghaziabad

Purvi Agarwal
Dept of CSE
IMS Engineering college, Ghaziabad

Akshya Agarwal
Dept of CSE
IMS Engineering college, Ghaziabad

Vijai Singh
Dept of CSE
IMS Engineering college ,Ghaziabad

Amit Kumar Gautam
Dept of CSE
IMS Engineering college, Ghaziabad

ABSTRACT
This paper covers the aspects of action recognition using Kinect technology by human skeletal tracking. Microsoft Kinect is one of the latest advancements in Computer Vision based HCI (Human Computer Interaction). The paper is focused on how the Kinect sensor captures the 3D information of a scene and recognizes the action being performed by the human body by retrieving the depth image information and real-time skeletal tracking. The Kinect technology has revolutionized the way humans interact with the machines. It has a wide range of applications areas. The paper also covers one of the proposed approach to skeletal based action recognition using Kinect.

Keywords
Microsoft Kinect sensor, action recognition, Skeletal tracking, HMM, Pose estimation

1. INTRODUCTION
Action recognition using the Kinect technology is an advanced way of interacting with machines. The Microsoft Kinect is used for retrieving 3D information of a scene and analyzing depth map and skeletal joint information of the human body. This helps the Kinect sensor to identify the type of action being performed by the person such as standing, walking, punching, sitting, waving etc. Action recognition using Kinect has wide range of application areas such as computer science, robotics, electronics, medical and many other commercial uses. People can play games by using their own body movements. Even in medical purposes, the doctors can operate a patient from a remote location by using Kinect. There are a large number of software applications and machines which are using Kinect to interact with humans. Action recognition using Kinect has been a great advancement in computer vision based HCI (Human Computer Interaction). The Kinect sensor senses the environment and generates a depth map for it. The human body is tracked using skeletal tracking by using the mean shift algorithm. In skeletal tracking, the Kinect sensor recognizes 24 joints in the human body which represent different body parts. Using the 3D joint information, the Kinect identifies the gestures and actions being performed by the human body [5] and then the machine responds according to the action input. An approach of real-time skeletal tracking using Kinect is discussed in other sections of the paper.

2. INTRODUCTION TO KINECT
The Microsoft Kinect is one of the most recent advancements in Computer Vision. The Kinect technology has emerged with great opportunities for multimedia computing by enabling 3D scene capturing. Kinect has revolutionized the way of playing games and doing various tasks such as handling of machines and applications. Kinect sensor recognizes the actions of the human body, i.e., the key technology behind Kinect is human-body language understanding, which means that the computer first recognizes and understands what the user is doing, before responding. The Kinect sensor directly senses the third dimension(depth) of the human body and also the environment.

The Kinect technology has wide availability and low cost which extends its applications areas to computer science, electronics engineering, robotics, medical field and many more. The Kinect effect has the potential to completely transform Human-Computer Interaction(HCI).

2.1 Kinect Sensor
The Kinect hardware contains a depth sensor, a color(RGB) camera and a four-microphone array as shown in Figure 1. The depth sensor consists of the IR(Infrared) projector along with the IR camera. The IR Camera is a monochrome complementary metal oxide semiconductor(CMOS) sensor. It is based on principle of structured light. The IR projector is an IR laser which passes through a diffraction grating, turning into set of IR dots. The IR projector, IR camera and the projected IR dot pattern have a relative geometry which is known. If a dot in the image matches a dot in projector pattern, it can be reconstructed in 3D.

The Kinect sensor produces a depth map for the IR image. The depth value is encoded with gray values, therefore, the darker the pixel closer the point is to the camera. If no depth values are available (indicated by black pixels), then the points may be too far or too close to be computed. The depth values produced by Kinect may be inaccurate due to invalid calibration between the IR projector and IR camera. This error in calibration may arise due to heat, vibration or drift in the IR laser. This problem can be addressed by using various recalibration techniques.

The four array microphones are used for speech and voice recognition. Figure 2 shows some of the specifications of Microsoft Kinect Sensor.

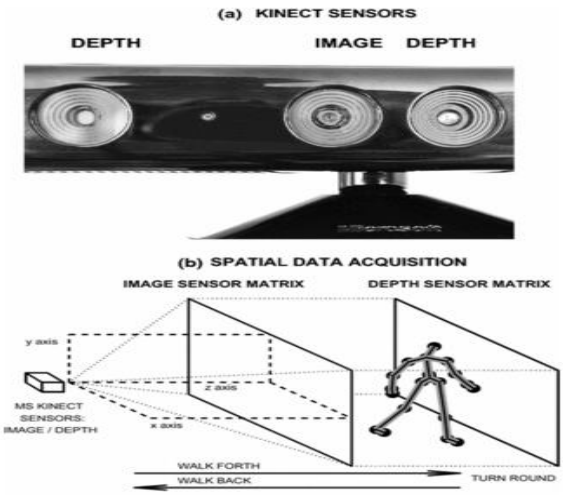

Figure 1:Kinect Sensor

SENSOR ITEM	SPECIFICATION RANGE
Viewing angle	43% vertical by 57% horizontal field of view
Mechanized tilt range(vertical)	±28%
Frame rate (depth and color stream)	30 frames per second(FPS)
Resolution, depth stream	QVGA (320*240)
Resolution, color stream	VGA (640*480)

Figure 2: Microsoft Kinect Specifications

2.2 Kinect Skeletal tracking

The Skeletal tracking is described by large number of dimensions. These dimension sets describe unique individuals, their sizes, shapes, postures, motions, hair. clothing. etc.

In skeletal tracking, the human body is represented by the combination of a number of joints which represent body parts such as head, shoulders, neck and arms. All joints are represented by their 3D coordinates [1]. Here, we are treating the segmentation process of depth image as a per-pixel classification task. If we evaluate each pixel separately, it will avoid combinatorial search.

For training, realistic depth images of humans of different sizes and shapes, with variety of poses, are generated, from large databases. Also, we train a randomized decision forest classifier to avoid over fitting. Now the spatial nodes of the per-pixel distribution obtained, are computed using mean sift algorithm, which results in 3D joint proposals. The optimized implementation of this algorithm runs in 5ms per frame on XBOX 360 GPU (Graphical Processing unit). Therefore, the whole pipeline of Kinect Skeletal tracking is as follows:

Step1- To perform per-pixel body part classification

Step2- To hypothesize the body joints by using mean shift algorithm to find a global centroid of local modes of density (probability mass).

Step 3- Mapping of hypothesized joints to the skeletal joints and fit a skeleton, using temporal continuity and prior knowledge.

3. HUMAN ACTION RECOGNITION

Human action recognition is gaining importance in the past few decades as it provides a wide variety of applications as in surveillance, robotics, security, patient monitoring and other systems that involves human-computer interaction.

Action recognition means the recognition of an action by using a system that t analyzes the video to acquire knowledge about the action and uses this knowledge to identify similar actions. Ryoo and Aggarwal has classified human activities into four categories- actions, gestures, interactions, and group activities.

Action recognition comprises of many actions like standing, walking, punching, sitting, waving etc. The method of action recognition is a tedious task as there can be many variations in human body movement [13]. Secondly, every individual has a different body shape, size and motion gestures and the interpretation of these actions can be different [7]. There can be other problems that can be introduced during the recognition process like variation in illumination, introduction of noise, shadow etc.

Traditional approaches were not able to eliminate these defects. With the introduction of depth camera these issues can be resolved as it can improve segmentation result by combining color, depth and motion. The depth camera is available at reasonable price and is easy to use. Also, the computer vision algorithms provide the more relevant segmentation results. With the combination of depth camera and computer vision algorithm we can direct the actions into 3D coordinate system which makes the recognition process efficient [4], [8].

A proposed approach for human action recognition is Skeletal tracking that involves Real time tracking based algorithm. The input for the method is the depth data which is collected from a Kinect sensor [3]. A skeleton tracking algorithm is used for the continuous detection of the joints (24) in human body as shown in Figure 3. For the identification of similar actions, spherical angles between joints [1] and angular velocities are measured. Then a motion energy based method is used to incorporate horizontal symmetry. At the end, HMM (Hidden Markov Model) performs the action recognition.

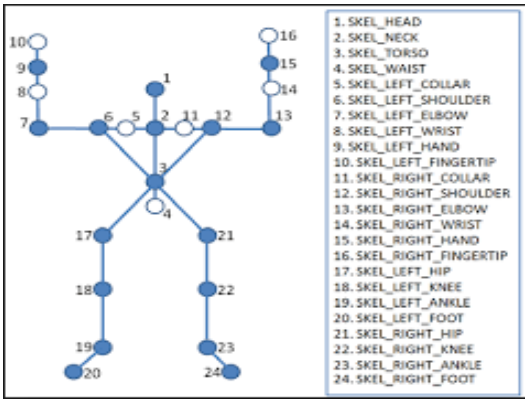

Figure 3: Joints of human body that are considered in action recognition

4. A PROPOSED APPROACH FOR HUMAN ACTION RECOGNITION

Skeleton tracking based action recognition method is proposed in this section. The following steps are used for the action recognition:

4.1 Pose estimation

The initial step in action analysis process is the pose estimation method [2]. This step is applied to make the process invariant to differences in appearance, body shapes and various interpretations of same actions.

The position of three joints is taken into consideration for the post estimation: left shoulder, right shoulder and right hip. It consists of joints near torso areas whose position does not change during action execution. This gives the correct estimation of the subject's pose [15].

4.2 Action representation

Action representation is done to handle the differences in body type, appearances and execution of actions in humans [6]. For this purpose, joint angles are measured as shown in Figure 4.

Posture	Features	Pos1	Pos2	Pos3	Pos4	Pos5
Standing	Coordinate	100	100	100	100	100
	7 angles	100	6.45	100	100	100
	9 angles	100	28.95	51.95	100	100
	17 angles	100	93.06	100	100	100
Sitting	Coordinate	100	22.09	0	100	100
	7 angles	100	100	100	100	100
	9 angles	100	100	86.96	100	100
	17 angles	100	100	100	100	100
Bending	Coordinate	100	0	0	100	100
	7 angles	100	15.79	0	7.22	100

Figure6: Calculation of 3D coordinates and angles for different postures

All angles are computed using the Torso joint as a reference. Figure 6 shows the calculated angles and coordinates of different actions [10]. The proposed action representation is computed by the use of only a subset of the supported joints. Only the joints corresponding to the upper and lower body limbs were considered after experimental evaluation. The joints are Right shoulder, Right elbow, Right wrist, left shoulder, left elbow, left wrist, left knee, Left foot, Right foot and Right knee.

4.3 Horizontal symmetry

There can be an issue in the action recognition process that concerns the implementation of the same gesture with either a right or left body part. A motion energy based approach is proposed to address this problem [14]. This method identifies and apply symmetries to both the common upper limb movement and whole body actions.

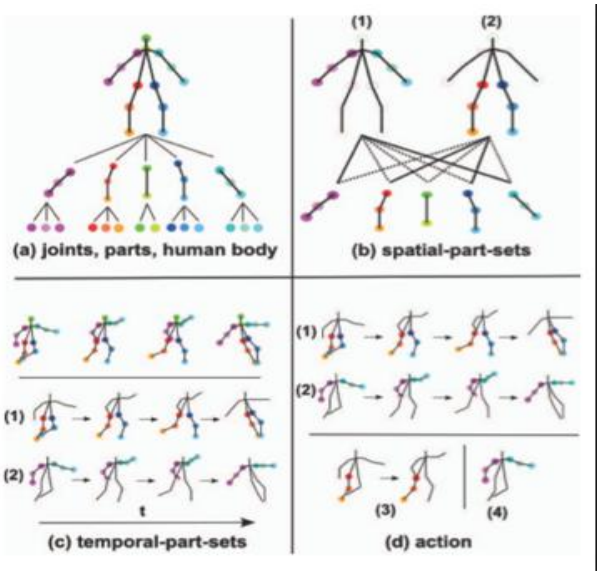

Figure 4: Action presentation

4.4 HMM Based Recognition

HMM based recognition offers stability in modelling recognition issues that exhibit an inherent temporality [9]. In this we employ a set of j HMMs, where we have an individual HMM for every supported action a_j .The action observation sequence (say s) acts as input for each HMM, and it returns a posterior probability P(a_j /s) representing fitness of the observation sequence. For implementation, the first order HMMs which are fully connected, allow all possible hidden state transitions, are used to map the low-level features to the high-level actions. For all hidden states, GMMs (Gaussian Mixture Models) are observed [11]. The Baum−Welch (or

Forward−Backward) algorithm is used for training and the Viterbi algorithm is used for evaluation. Also, the total number of hidden states of the HMMs is considered to be a free variable. In this way, the developed HMMs are implemented.

5. PROS AND CONS

There are various advantages and shortcomings in different types of action recognition techniques. Some of them are listed in Table 1

6. RESULTS

Detection or tracking of one and more people moving in the field of view of sensor, using the tracking of parts of the sensor. The results calculated have shown 90% accuracy in recognition process. Figure 5 shows how a human body action is recognized.

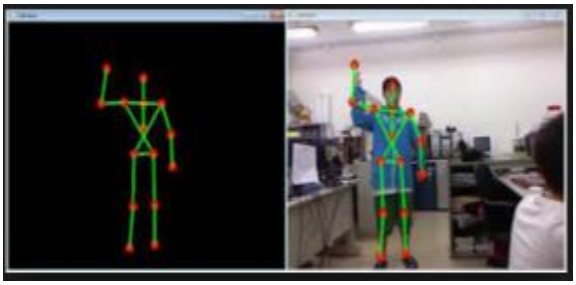

Figure 5: Detection of human action by skeleton view

S.no	Techniques	Pros	Cons
1	OpenNI/NITE	1.Various application, very popular. 2.Have skeleton tracking. 3. Available for most languages.	1. Difficult to install. 2. Calibration posture is in need.
2	Libfreenect	1. Several applications. 2. Available for most languages. 3.Any OS compatible applications.	1.No skeleton tracking. 2. Very difficult to install.
3	CL NUI	1.Ability to capture the broad range of body movement. 2. Small jitter.	1.Cannt perform motion prediction. 2.No learned soft constraints to handle the case of severe occlusion.
4	Microsoft Kinect SDK	1.Widely known in the community of robotics. 2.Easy to install, fairly widespread.	1.Poor high level API. 2.Support for windows only.
5	Evoluce SDK	1.Easy to install. 2. Ready to use methods for action recognition. 3. Have skeleton tracking.	1. Support for Window 7 only. 2. Calibration posture is in need. 3.Available for C, C# and C++

Table 1: Comparison between different action recognition techniques

7. CONCLUSION

Human action recognition is one of the main area of research these days. In this paper, we have shown skeleton based technique for action recognition. This approach is done with the help of depth maps. The action data that is stored with the help of kinect is then mapped into 3D coordinate system. This technique aims to provide an application that uses gestures to interact with virtual objects in the augmented reality application. It provides a way to use the gesture based interaction to manage operations in a virtual environment.

8. FUTURE SCOPE

Researches in action recognition are at the starting level, a more efficient system may require mating of several disciplines. The incorporation of large no of individual would also give a new direction to the research of action recognition. To discover the scopes of action recognition, working on true surveillance videos, movies is required. A lot of challenging work still remains in the field of action recognition.

9. ACKNOWLEDGMENTS

Our thanks to the experts who have guided us and Department of Computer Science, IMS Engineering College.

10. REFERENCES

[1] Gu, J., Ding, X., Wang, S., Wu, Y.: Action and gait recognition from recovered 3-d human joints. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans. on 40(4), 1021–1033 (2010)

[2] J. Shotton et al., ‘‘Real-Time Human Pose Recognition in Parts from a Single Depth Image,‘‘ Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), IEEE CS Press, 2011, pp. 12971304.

[3] S. Izadi et al., ‘‘Kinect Fusion: Real-Time Dynamic 3D Surface Reconstruction and Interaction,‘‘ Proc. ACM SIGGRAPH, 2011.

[4] W. Li, Z. Zhang, and Z. Liu, ‘‘Action Recognition Based on A Bag of 3D Points,‘‘Proc. IEEE Int’l Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB), IEEE CS Press, 2010, pp. 914.

[5] Z. Ren, J. Yuan, and Z. Zhang, ‘‘Robust Hand Gesture Recognition Based on Finger-Earth Movers Distance with a Commodity Depth Camera,‘‘ Proc. 19th ACM Int’l Conf. Multimedia (ACM MM), ACM Press, 2011, pp. 10931096.

[6] Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Eff ective codebooks for human action representation and classification in unconstrained videos. Multimedia, IEEE Transactions on 14(4), 1234–1245 (2012)

[7] Isaac Cohen, Hongxia Li, “Inference of Human Postures by Classification of 3D Human Body Shape”, International Workshop on Analysis and Modeling o f Faces and Gestures, pp. 74 – 81, 2003.

[8] D. M. Gavrila and L. S. Davis, “Towards 3-D Model-Based Tracking and Recognition of Human Movement: a Multi-View Approach”, International Workshop on Automatic Face- and Gesture-Recognition”, pp.272-277, 1995.

[9] Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proc. of the IEEE 77(2), 257–286 (1989)

[10] T. F. Syeda-Mahmood, M. Vasilescu, and S. Sethi, “Recognizing action events from multiple viewpoints,” IEEE Workshop on Detection and Recognition of Events in Video, pp. 64–72, 2001.

[11] M. Z. Uddin, N. D. Thang, J.T. Kim and T.S. Kim, Human Activity Recognition Using Body Joint-Angle

Features and Hidden Markov Model. ETRI Journal, vol.33, no.4, Aug. 2011, pp.569-579.

[12] H. Fujiyoshi and A. Lipton. Real-time human motion analysis by image skeletoniation. In IEEE Workshop on Applications of Computer Vision, pages 15-21, Princeton, 1998.

[13] Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In Human Motion Workshop, (with ICCV), 2007.

[14] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," Computer Vision and Image Understanding, vol. 104, no. 2, pp. 249–257, 2006

[15] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 928–934, 1997.