Rotation and Scale Invariant Posture Recognition using Microsoft Kinect Skeletal Tracking Feature

Samiul Monir*, Sabirat Rubya†, Hasan Shahid Ferdous‡
*Department of Computer Science and Engineering
University of Liberal Arts Bangladesh, Dhaka, Bangladesh
Email: samiul.monir@ulab.edu.bd
†Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
Email: sabirat@csebuet.org
‡Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
Email: hasan.shahid@csebuet.org

***Abstract*—Human posture identification for motion controlling applications is becoming more of a challenge. We present a posture classification system using skeletal-tracking feature of Microsoft Kinect sensor. Posture recovery is carried out by detecting the human body joints, its position, and orientation at the same time. Angular representation of the skeleton data makes the system very robust and avoids problems related to human body occlusions and motion ambiguities. The implemented system is tested on a class of relatively common postures comprising hundreds of human pose instances by different people, where our classifier shows an average accuracy of 94.9% , 96.7% and 96.9% for linear, exponential and priority based matching systems respectively.**

***Index Terms*—Human Posture Recognition, Kinect, Skeleton Tracking.**

I. Introduction

Posture inference is an important problem in computer vision. It stands at the junction of various important applications ranging from surveillance to multimedia information retrieval, human computer interaction (HCI), health-care and biometrics [1]. Moreover, motion sensing and posture identification based computer gaming has gained much popularity. Postures are among the primary ways through which a human being interacts with reality. Human postures refer to any position that the human body can take intentionally or habitually. For example, it is possible for a human observer to disambiguate actions such as walking, running, standing, sitting, etc., from just a single pose. Therefore, the possibility of building an automated system capable of receiving and classifying this type of information has been one of the most fascinating topic for the research community in recent years.

There are two main approaches for recovering human pose: model-based or direct approach and learning-based or indirect approach [2]. Model-based approaches assume that a parametric body model is known. Then a cost function representing the pose is optimized through predicting and updating position variables incrementally. Learning-based approaches do not require a detailed description of a human body model. The model is learned by training examples representing typical human postures and the new poses are interpolated by searching and comparing. In our work we adopt the learning-based approach, where the body joints for training postures are stored as feature vectors and later searched for finding the best match with a test posture.

Real-time depth sensing systems have recently been very popular in the video game industry by implementing virtual systems where the human body works as a controller. The Microsoft XBOX Kinect system is an example of such a natural user interface (NUI) that is capable of parsing a depth-map stream at thirty frames per second. This depth data is utilized to estimate the positions of twenty predefined body joints that constitute a wireframe skeleton of a moving user in its field of view. Subsequent algorithmic processing can then attempt to understand the users motion in order to interactively control the game play. We employ an effective system of posture recognition that uses this skeleton tracking feature of Kinect. As we know that different postures are taken by keeping the bones and joints of the body in different positions and angles, therefore, tracking the skeleton's location and retrieving joints' information from it provides the most reliable and straightforward approach for posture recognition. In the proposed method, a set of vectors and angles manipulated from the skeletal data represents the body postures. A classifier is then trained by saving these angles for different postures. When a test posture has to be identified, three different distance metrics are applied to it for comparing with the training postures. For every test image, some rotations have been applied on the vectors obtained from the skeletal points so that it becomes independent of the orientation of the Kinect. Afterwards, the system outputs three maximum matched postures from the training database for each metric with a value containing how well the reference posture is harmonized with the test posture. Performance of the proposed system is measured considering the outcome of all of these three techniques.

Many existing approaches have tried to recognize postures following learning based methods [2]. Most of their

978-1-4673-5119-5/12/$31.00

recognition accuracy is affected by the high variability that exists within the same postures performed by different people. Besides, complex foreground and background, multiple objects in the scene, and moving camera or background make the posture recognition problem more complex. Using Kinect instead of other sophisticated tools, we have addressed most of the problems that existing systems face and provided a solution that is advantageous for the following reasons:

1) The proposed algorithm is linear in time.
2) Direct use of skeletal-tracking feature instead of depth or image data guarantees the effectiveness of the proposed method. Discovering body joints from image requires more effort as well as provides less accurate data.
3) In most of the typical approaches, subjects are continuously monitored by camera. Many people may feel uncomfortable with this. Since we dont use image or video streams, this monitoring is not required and hence privacy of the target users is ensured.
4) The deviation in output that is caused in many systems when different people take same postures is diminished by acquiring vector and angle information. An additional step of rotating the vectors obtained from the test images ensures that the system is independent of the orientation of the sensor, thus making our system both rotation and scale independent.

The rest of the paper is organized as follows - Section II discusses prior research works in detecting human postures. Section III illustrates the overview where Section IV describes our proposed approach in detail. The experimental evaluation and performance measurements of our technique are represented in Section V. Finally, Section VII concludes with directions to future works discussed in Section VI.

II. Related Works

Controlling and directing a virtual character is an important problem in animation movies and computer games. It requires the detection of varied postures and gestures made by the player in front of the screen. For this reason, a great deal of research on posture detection has been done based on surveillance camera, depth image, or video data.

Boulay *et al.* has proposed an improved way to detect human posture from video sequence [3]. Their method was based on two steps. At first they recognize human postures from *blob* using horizontal and vertical projections on the reference axis of the images and then using a learning phase, made by videos from different actors and 3D human models. Their proposed method was somewhat dependent on camera position and also could emphasize only on a few key postures.

An improved solution for detecting human action using low-level video features was provided in [4]. They presented a novel approach consisting segmentation and recognition of human action sequence using a hierarchical topic model. But with their model, it is hard to make explicit difference between human and other objects. Jiang *et al.* [5] had shown that human postures could be recognized using convex programming. A multi-stage linear programming was used instead of image segmentation in their proposed technique. Although the system performed better than many existing graph-cut or belief propagation methods but it was quite sensitive to human clothing and other unnecessary features if those were available with the figures of the images. Again, silhouette histogram was utilized to detect human postures in [6]. Hsieh *et al.* [6] extracted human silhouette using background subtraction and then mapped it into three different polar coordinate systems. However, presence of shadow in background or other outdoor environments persuaded this method to perform very poor. In our work, this dependency of feature vectors has been overcome by using the skeletal data that can be effectively distinguished from other irrelevant features.

Posture detection by tracking skeletons from static image and video data has been another topic of research interest over the years. Junior *et al.*, using artificial neural network, presented that they could automatically detect human skeleton in a single image, based on a 2D model combined with anthropometric data [7]. Lee and Cohen proposed the use of an adaptive approach, where a human model was used to synthesize the image regions corresponding to human forms and thus separating the person from the background [8]. A technique for self-initializing by kinematic tracker was adopted in [9] that automatically discovered appearance of human models from a video sequence. But these approaches have their drawbacks, most of these use static image data or video sequence to determine human skeleton, which requires a huge computational overhead to extract features.

Researchers have also been interested in vision based posture recognition in recent years. They have been trying to investigate a topological approach in order to define descriptors suitable to estimate location of body parts and orientation of body articulations. For example, eCushion - an eTextile device for sitting posture monitoring system is developed by Xu *et al.* to analyze orientation independent postures [10]. Although their proposed system has achieved a great success rate of 92% but this system works only for sitting postures only.

After Microsoft introduced Kinect sensor in 2010, researchers and developers all over the world have concentrated to work with it in different area of technology [11], [12]. Some related Kinect based works include hand posture detection [13], American Sign Language recognition [14], and seated posture detection [15]. Although Microsoft has introduced Kinect as a gaming tool, it is also proving its capability as a strong, inexpensive, off-the-shelf and widely available sensor device. Raptis *et al.* [16] discussed about 'gesture recognition' using relative angles between body joints. They plotted these angles against time frame and used signal matching with noise reduction techniques to identify dance gestures. Using its multimodal sensors, Kinect can also be used for stroke rehabilitation, both in a clinical setting and as a tool to aid stroke survivors in their exercises at home [17]. Forward and backward movement and an optimization solution for hand articulation are also provided using Kinect [18]. Kinect has made a remarkable progress in medical science too. Many of the medical image exploration system use Kinect as an

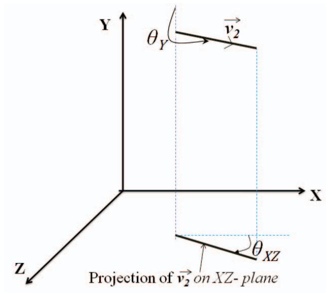


Fig. 1: Angle calculation for a vector.

invisible controller. Human pose detection using Kinect has gained extra focus for conducting these medical operations successfully.

Our proposed system uses this 3D skeleton data, which is one the most essential features in Kinect, to track human posture. Hence, both the feature extraction and matching steps become linear in time unlike the other state-of-the-art methods that use complex pattern matching algorithms.

III. Background Knowledge

Microsoft has provided a software development kit for Kinect through which the low-level data streams from the Kinects video, microphone, and depth sensors can be accessed. This SDK provided by Microsoft is capable of tracking skeletal data, too. It can track the skeleton image of one or two people moving within the Kinects field of view. We use this feature for recognizing different human postures.

A. Feature Selection

Kinect uses anatomical data and depth image to find joints in a body. It can track two human bodies at the same time and identify twenty joints on each body. We build a set of fourteen vectors from these twenty joint points and use them as features for our matching procedure.

B. Impact of using Skeletal Data

The position of different joints of the body is readily available from the skeletal data which removes an extra step of using complex algorithms for human body parts recognition from images [19]. Again, many traditional methods require high-resolution images from expensive sensors for accurate identification of postures. Kinect, on the other side, is low in cost and skeletal data provided by it can be used without hampering one's privacy, as no video/image data is used.

IV. Posture Recognition Process

In this paper, we discuss three different methods for identification of human postures. We also focus on the steps to make the system scaling and rotation independent.

A. Unique Identification of Vector Positions Representing Skeletal Joints

We know that position of any vector can be uniquely identified by calculating two angles. We apply this to find out the position and orientation of the fourteen vectors built from the twenty joint points. Some parts of the body are not considered in calculation (head, wrists, and feet) as they contribute little to distinguish postures. To uniquely identify a vector, two angles are measured for each of these vectors-

$$\theta_Y = \text{Angle between the vector and positive } Y\text{-axis}$$

$$\theta_{XZ} = \text{Angle between the projection of the vector on } XZ \text{ - plane and } X\text{-axis.}$$

Fig. 1 demonstrates these angles for the vector joining shoulder center and right shoulder joints which is labeled as $\vec{v}_2$. Thus, we keep track of fourteen pairs of angles for each training posture in the database.

B. Matching Procedure

Three different techniques are adopted - linear, exponential, and priority based or weighted matching. Linear method is the simplest one but exponential method shows much better performance than the other two. Priority based matching performs more accurately for some specific class of postures.

To start with, fourteen pair of angles is calculated from the skeletal data provided by the test posture in the same way, where each pair is in the form $<\theta_Y, \theta_{XZ}>$. These values are then compared to all the training postures for maximum match. This step is followed in all of the three techniques.

Linear Matching: According to linear matching system, the sum of difference of the twenty eight angles between a training posture and the test posture is calculated. Let, the sum of all the angles for a particular posture P is θ_P. Therefore,

$$\theta_P = \sum_{i=1}^{14} (|\theta_Y^{v_i}| + |\theta_{XZ}^{v_i}|) \tag{1}$$

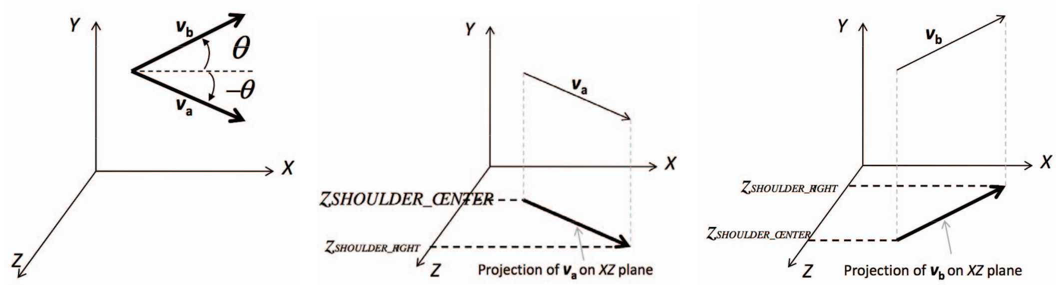
Where v_i represents the i^{th} vector from the fourteen vectors. The sum of difference between a reference and a test posture then can be represented by Eq. 2,

$$sum_{ref} = |\theta_{ref} - \theta_{ts}| \tag{2}$$

Where, ref and ts stand for a reference posture from the database and the test posture respectively. The best matched reference pattern is the one for which the value of sum_{ref} is minimum. If S is the set of all the reference postures available in the training database, then the reference posture having the minimum sum of difference with the test posture can be calculated by Eq. 3.

$$P_{min} = \min_{ref \in S} (sum_{ref}) \tag{3}$$

Thus P_{min} will be the output of the linear method as the best matching recognized posture. This process has some drawbacks in the cases where test pattern is quite similar to



(a) Determining direction of rotation (b) Calculation for clockwise rotation (c) Calculation for anti-clockwise rotation

Fig. 2: Minimizing effect of rotation on the system.

more than one of the reference patterns. Exponential matching, discussed next, solves this problem to some extent.

Exponential Matching: As we know, the exponential function ($f(x) = e^x$) is used to model a relationship in which a constant change in the independent variable gives a greater change (percentage increase or decrease) in the dependent variable, and y increases faster as x increases. In this matching process, exponential function of the difference in individual angles is calculated and summed up to find out sum_{ref}.

$$\theta_P^{v_i} = |\theta_Y^{v_i}| + |\theta_{XZ}^{v_i}| \tag{4}$$

$$sum_{ref} = \sum_{i=1}^{14} \exp(|\theta_{ref}^{v_i} - \theta_{ts}^{v_i}|) \tag{5}$$

As a result, postures having considerable difference in angles for one or two vectors will have a greater value of sum_{ref} which can help to distinguish the test posture more accurately. The best matched posture is then determined using Eq. 3.

Weighted Matching: As already discussed, the proposed system is scaling independent. Weighted matching is applied so that the system becomes partially rotation independent. This technique prioritizes some vectors over others in such a way that effect of rotation is normalized for them. From the fourteen vectors, we take six vectors that make the torso and assign a higher priority to the values of θ_Y and θ_{XZ} for these vectors. This is done to make sure that sum_{ref} is greater while comparing the skeleton data of a rotated body to a dissimilar posture and comparatively smaller while comparing to a similar reference posture. Thus, for priority based recognition, Eq. 1 is replaced by Eq. 6. Other steps of matching can be done using either linear or exponential method. We can refer them as weighted linear and weighted exponential method respectively.

$$\theta_P = \sum_{i=1}^{6} P_1(|\theta_Y^{v_i}| + |\theta_{XZ}^{v_i}|) + \sum_{i=7}^{14} P_2(|\theta_Y^{v_i}| + |\theta_{XZ}^{v_i}|) \tag{6}$$

where, $P_1 > P_2$ for weighted matching and $P_1 = P_2$ for linear matching techniques.

Following three different methods for comparing postures widens the possibility of identifying apparently similar postures distinctly.

C. Incremental Learning System

When no posture is matched to a particular test pattern in any of the above discussed three techniques, then this posture is inserted in training database as a new one for further use. The decision of whether a test posture is new or not is taken based on a threshold value of sum_{ref}, denoted as sum_{th}. This value is determined by experimenting on different new postures and taking average value from a number of runs. The condition for a test posture to be recognized as new pattern is therefore, $sum_{ref} > sum_{th}$ for all the three methods.

D. Making the System Rotationally Invariant

When the subject is not standing/sitting parallel to the Kinect, his torso will make an angle with the X-axis. We assume that for the training postures the angle between the torso and the positive X-axis is apparently zero. When a test posture has to be classified we find the angle between the subjects torso and positive X-axis according to Kinect coordinate system. The vector joining the shoulder center and the right shoulder is taken as representative of the torso because when the torso rotates along Y-axis, this vector is guaranteed to rotate. Next, we find out the intended angle using the following equation:

$$\theta = \cos^{-1}\left(\frac{v_{1x}}{|\vec{v_1}|}\right) \tag{7}$$

Where, θ is the angle between $\vec{v_1}$ and positive X-axis and $\vec{v_{1x}}$ denotes the x component of $\vec{v_1}$. As the formula of dot product has been applied to find out the angle, $0° \leq \theta \leq 90°$. For both $\vec{v_a}$ and $\vec{v_b}$ in Fig 2a, $\theta = 30°$ according to Eq. 7 where it should be -30° for $\vec{v_a}$. Therefore, we must also calculate whether the rotation is clockwise or anti- clockwise. The difference of z-component of the joint point for shoulder center and shoulder right is taken into account for this purpose, and θ is negated when this difference value is found to be negative (Fig 2b). Eq. 8 illustrates this situation:

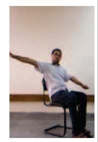
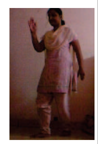
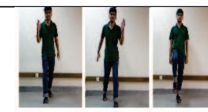
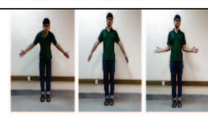
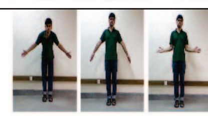
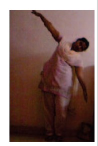
$$Z_{SHOULDER\_CENTER} - Z_{SHOULDER\_RIGHT} < 0 \Rightarrow \theta = -\theta \tag{8}$$

Thus we can calculate the value and direction of the angle between $\vec{v_1}$ and the positive X-axis. The fourteen vectors from the test image should now be rotated along Y-axis to cancel out the effect of rotation of the test posture with respect to the reference postures.

TABLE I: Average accuracy (%) for each category of posture

	Standing	*Sitting*	*Moving*	*Bending*	*Musicians*	*Sportsmen*	*Rotated*
Linear Method	92.51	86.29	89.92	87.22	91.98	93.67	88.33
Exponential Method	98.89	89.25	96.08	96.33	97.71	96.23	91.90
Weighted Method	98.89	89.56	97.91	96.02	96.48	96.98	95.67

TABLE II: Matched reference patterns for different test postures in four categories

Posture Category	Test Posture	Matching Method	Matches before applying Rotation	Matches after applying Rotation	Test Posture	Matches before applying Rotation	Matches after applying Rotation
Standing		Exponential					
		weighted					
Sitting		Exponential					
		weighted					
Moving		Exponential					
		weighted					
Bending		Exponential					
		weighted					

V. Experimental Results

The primary objective of the experiments with the Kinect posture recognition system is to determine whether the accuracy of the joint information recorded by the programs is adequate for use to distinctly identify human postures in different categories, how well a posture is matched to a test pattern with variations in testing conditions, the independency of the system on position and orientation of subject with respect to the Kinect sensor.

$$P = \sum_{i=1}^{3} (f_i \times match(ref_i, test)) \tag{9}$$

The sample database contains about 100 postures categorized in four different types: standing, sitting, moving and bending. These categories are considered as example to compare the performance of our matching algorithms in different categories of postures. A set of experiments has been conducted in which subjects of different height and gender queries for a match. Three maximum matched postures (if exist) are displayed in each of the three proposed techniques with the sum of difference value. Then the accuracy of the output is found based on human visual perception. We locate the position (1, 2, 3 ..) of the posture that human eye considers to be the best match. The performance of each method for a particular test posture is evaluated by Eq. 9.

Where, ref_i and $test$ represents the i^{th} matched posture and the posture under query respectively, f_i denotes the fitness value for the matched patterns, which decreases gradually with the position of the best matched pattern. For now, we assume that, $f_1 = 1.0, f_2 = 0.9, f_3 = 0.8$. The $match$ function is defined as following.

$$match(ref, test) = \begin{cases} 1, & \text{if } ref \text{ and } test \text{ are matched} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Six subjects of different gender and height with different clothing under different lighting and background scenarios have performed each type of experiment 5 times. The average accuracy of the three schemes is shown in Table I. The proposed method shows average accuracy rate of 89.1%, 96.7% and 96.9% respectively for linear, exponential and weight based recognition systems. The effect of rotational changes has been shown in Table II and the systems output before and after applying rotation matrix on test data demonstrates the difference in result. It is notable that the calculation excluding the orientation shows erroneous outputs when test subjects posture is same as a reference with different orientation.

When comparing data collected at distances that varied from 1.3 meters to 3.5 meters, postures are most clearly distinguishable at the region of 2 to 2.5 meters. When the subject is too close or too far away from the sensor, the inconsistency of the joint positions make the phases of calculation somewhat more difficult.

The Kinect SDK documentation specifies that it is intended for use only with standing, not seated figures. However, the result does not show much deviation in accuracy for the sitting postures. Thus we can conclude that the system is efficient for tracking seated postures too.

VI. Future Works

The performance shown by the experiments have highlighted the abilities of the system not only to recognize postures, but also the possibility to be tuned up in a dynamic context. In future, we plan to make an automated arrangement that will recognize human postures in a specific time interval to make a sequence of postures. This sequence can then be deployed for many real-time applications such as tutoring exercise or dance moves, stroke rehabilitation, detection of injuries, etc. This characteristic may be used to apply the proposed method in crowded environment by which many security systems will be benefitted.

We also plan to use our devised system for experiments related to early detection of autism. As we know, many autistic children repeatedly show some abnormal behaviors that can be identified by their postures. They can be distinguished from others if we can notice their postures continuously under our system, gather proper information from their poses and map this information to relate with autistic activities.

VII. Conclusion

The emergence of affordable real-time depth sensors has opened new opportunity for interactivity. We have set out a novel posture detection method using such a sensor. Compared to less accurate image or video stream based posture recognition systems, skeletal-tracking offers balance in usability and cost. A substantial increase in noise may occur, in particular when we are faced with the fact that the world population is the user. The noise stems from the sensor, users' physique, clothing, and complicated motion patterns. However, the proposed method is more efficient and effective than previous methods for posture matching as a large target point set is involved. We use angular representation of vectors to match the template and target images, and this representation facilitates matching objects with large appearance variations. Overall, while more investigations can be done into the capabilities and limitations of the device, this research has determined that Kinect has much potential for being used in many interactive applications that require posture identification.

Acknowledgement

This research is partially supported by the CodeCrafters-Investortools Research Grant for the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology.

References

[1] S. Aoki, Y. Iwai, M. Onishi, A. Kojima, and K. Fukunaga, "Learning and recognizing behavioral patterns using position and posture of human body and its application to detection of irregular states," *Systems and Computers*, vol. 36, pp. 45–56, November 2005.

[2] A. Agarwal and B. Triggs, "Recovering 3d human pose from monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 44–58, January 2006.

[3] B. Boulay, F. Bremond, and M. Thonnat, "Posture recognition with a 3d human model," *ICDP*, vol. 3, pp. 135–138, 2005.

[4] D. Buchsbaum, K. R. Canini, and T. L. Griffiths, "Segmenting and recognizing human action using low-level video features," in *Annual Conference of the Cognitive Science Society*, July 2011.

[5] H. Jiang, Z.-N. Li, and M. S. Drew, "Human posture recognition with convex programming," in *International Conference on Multimedia and Expo*, July 2005.

[6] C.-H. Hsieh, P. S. Huang, and M.-D. Tang, "The recognition of human action using silhouette histogram," in *Australasian Computer Science Conference (ACSC 2011)* (M. Reynolds, ed.), vol. 113 of *CRPIT*, (Perth, Australia), pp. 11–16, ACS, 2011.

[7] H. S. Junior and S. R. Musse, "Automatic detection of 2d human postures based on single images," in *International Conference on Graphics, Patterns and Images*, May 2011.

[8] M. W. Lee and I. Cohen, "A model-based approach for estimating human 3d poses in static images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, pp. 905–916, June 2006.

[9] C. Mcintosh and G. Hamarneh, "Human limb delineation and joint position recovery using localized boundary models," in *In IEEE Workshop on Motion and Video Computing*, p. 1, 2007.

[10] W. Xu, Z. Li, M.-C. Huang, N. Amini, and M. Sarrafzadeh, "ecushion: An etextile device for sitting posture monitoring," in *2011, International conference on Body Sensor Networks (BSN)*, May 2011.

[11] B. Alkire, "Character animation and gesture-based interaction in an immersive virtual environment using skeletal tracking with the microsoft kinect," September 2011.

[12] M. Tang, "Recognizing hand gestures with microsofts kinect," September 2011.

[13] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3d tracking of hand articulations using kinect," in *International Conference on Computer Vision*, November 2011.

[14] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the kinect," in *13th International Conference on Multimodal Interfaces*, November 2011.

[15] "Seated posture and gesture recognition with kinect," October 2011. http://channel9.msdn.com/coding4fun/kinect/Seated-posture-and-gesture-recognition-with-Kinect.

[16] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," pp. 147–156, August 2011.

[17] K. LaBelle, "Evaluation of kinect joint tracking for clinical and in-home stroke rehabilitation tools," December 2011.

[18] J. wook Kang, D. jun Seo, and D. seok Jung, "A study on the control method of 3-dimensional space application using kinect system," *International Journal of Computer Science and Network Security*, vol. 11, pp. 55–59, November 2011.

[19] A. Kar, "Skeletal tracking using microsoft kinect," December 2010.