Automatic Multiple Kinect Cameras Setting for Simple Walking Posture Analysis

Suttipong Kaenchan, Pornchai Mongkolnam, Bunthit Watanapa, and Sasipa Sathienpong
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok 10140, Thailand
E-mails: 54441344@st.sit.kmutt.ac.th, {pornchai|bunthit|sasipa.kal}@sit.kmutt.ac.th

Abstract—**We propose an automatic setting of multiple skeletal tracking Kinect cameras, in lieu of mere using a single camera, to capture a human skeleton because of possible viewing occlusions. Using multiple cameras from different angles gives a more complete whole body; however, more required steps are needed in combining multiple skeletons into one final skeleton. One camera is used as a reference for the other cameras to transform their coordinates into the reference camera's coordinate system. Once every view is in the same coordinate, one skeleton is able to be composed. Due to camera's sensory errors, nevertheless, the supposedly same joint of the skeleton, which is obtained from the transformations, may not be exactly located at the same position. Therefore, average joints are used for the composed skeleton. The skeleton is then used to analyze the walking posture of a human subject in order to check whether or not the walking is balanced.**

Keywords-balanced walking; composed human skeleton; Kinect cameras; skeletal tracking; walking posture analysis

I. INTRODUCTION

Kinect cameras [1] have been used in various applications since its first launch in 2010. They are used in entertainment, exercises, simple medical applications, and physical therapies, to name a few. Most applications rely on just one camera due to its easy implementation and simple required tasks of the applications. However, with more complicated applications like a human walking analysis, it is necessary to use two or more cameras to capture many views of the same subject in a scene because a view from one camera may not be able to capture the whole subject. For instance, a camera could not capture all fifteen joints of a human skeleton when the person sits with two legs under a desk while viewing from the top of the desk. Therefore, we propose the use of multiple cameras simultaneously capturing multiple views of the subject, a human in particular, so that a more complete human skeleton could be obtained for applications needing robust details as in walking posture analysis.

An automatic setting of multiple cameras is presented by transforming different local coordinate systems obtained from different cameras into one global coordinate system. However, due to inherent cameras' capture errors, the fifteen skeletal joints in the global coordinate system, which are obtained after the coordinate transformations from each camera, may not perfectly coincide. Therefore, averages of those fifteen joints are used to best represent the skeleton. The obtained skeleton is subsequently used for a simple walking posture analysis. In a perfect standing posture, a line connecting the center of mass and the middle of left and right hips of the human skeleton forms a vertical vector, and another line connecting left and right hips forms a horizontal vector. Any deviation from that posture results in a somewhat imperfect position, called tilt. To help better see the degree of the deviation, a visualization tool is presented with four levels of deviation toward the front, back, left, or right directions from the perfect posture using a color variation ranging from green to yellow, orange, and red.

II. BACKGROUND

A. Related Work

L. Xia et al. [2] proposed the technique for detecting humans using Kinect's depth information rather than using the usual RGB images from a video in which there was a problem of ability to differentiate the humans from a background partly due to clothes, lighting conditions, and visual background complexity. J. D. Huang [3] used a Kinect camera in the rehabilitation system for students in one special education school with muscle atrophy and cerebral palsy. Suggestions were provided to users whether therapeutic exercises were proper. One problem of such system was that one camera was incapable of capturing blocked views of human body parts. N. Pattanotai et al. [4] proposed the technique to manually set up multiple Kinect cameras in order to compose multiple views of a human skeleton into one whole skeleton. They claimed that using one camera was not practical enough in the situation where some skeletal joints were visibly blocked from the camera. Because the manual setting of measuring instruments was used, consequently reading errors and inconvenience were unavoidable. Nevertheless, this work had inspired both the work of S. Kaenchan et al. [5] whose work was to find ways to compose a skeleton without using any measuring instruments and our work which has extended it to handle an automatic multiple-camera setting and a simple walking posture analysis.

Technically, fifteen skeletal joints are obtained by the API of the OpenNI framework [6] which is an open source SDK used for the development of 3D sensing middleware libraries and applications of body motion tracking and hand gestures for

978-1-4673-5324-3/13/$31.00 ©2013 IEEE

Kinect camera. H. Lee and L. Chou [7] detected gait instability in the elderly person who was susceptible to falling by measuring the inclination between an upright direction and the line connecting the center of mass (COM) and center of pressure (COP) located at a foot. The work had shown that COM could very well be used in a walking analysis. A. A. Manasrah [8] used the Kinect camera to capture human skeletal joints from which the body segments such as shank, thigh, etc., were defined. Thereafter each segment was given a different weight and a different distance from the to-be-computed COM. In our work we use the COM and the middle of the left hip and right hip of the human skeleton to find any inclination or tilt from the upright direction in the walking posture analysis.

B. *Geometric Transformation*

A coordinate system transformation [9] is used to transform one coordinate system into another desired coordinate system. For instance, each object's local coordinate system in a scene is transformed into one global coordinate system for a simpler manipulation on those objects. The transformation can be performed, as in

$$Q' = M \cdot Q, \tag{1}$$

where Q is the coordinate matrix before applying the transformation matrix M to it, getting the new coordinate matrix Q'. The transformation matrix M is obtained by multiplying the translation matrix T to the rotation matrix R_{xyz} based on the so-called X-Y-Z fixed angle rotation [10] with θ_x, θ_y, and θ_z being the rotating angles about the X, Y, and Z axes, respectively, as in

$$M = T \cdot R_{xyz}(\theta_x, \theta_y, \theta_z) \cdot \tag{2}$$

After replacing the matrix M in (1) and using s for sin(), c for cos(), t_x, t_y, and t_z for the translation distances in the X, Y, and Z axis directions, respectively, the newly transformed coordinates are obtained, as in

$$\begin{aligned} x' &= x(c\theta_z c\theta_y) + y(c\theta_z s\theta_y s\theta_x - s\theta_z c\theta_x) \\ &+ z(c\theta_z s\theta_y s\theta_x - s\theta_z c\theta_x) + t_x \end{aligned}, \tag{3}$$

$$\begin{aligned} y' &= x(s\theta_z c\theta_y) + y(s\theta_z s\theta_y s\theta_x + c\theta_z c\theta_x) \\ &+ z(s\theta_z s\theta_y c\theta_x - c\theta_z s\theta_x) + t_y \end{aligned}, \tag{4}$$

$$z' = x(-s\theta_y) + y(c\theta_y s\theta_x) + z(c\theta_y c\theta_x) + t_z \cdot \tag{5}$$

III. Methodology

In order to compose a whole skeleton, we need to know each camera's orientation and distance relative to a reference coordinate system. The superimposed skeleton is obtained as an average of all skeletons from those cameras.

A. *Cameras' Orientations*

One reference coordinate system is used as a global coordinate to which other local coordinate systems are transformed. In Fig. 1, Camera A's coordinate system is used as the reference coordinate system for the other coordinate systems from Cameras B and C. For instance, each skeletal joint's coordinate (x, y, z) viewed from Camera B is transformed into the Camera A's coordinate system. Similarly, each skeletal joint's coordinate (x, y, z) viewed from Camera C is transformed into the Camera A's coordinate system. All the transformations are done according to the aforementioned geometric transformation. Consequently, there are three separate sets of the fifteen skeletal joints in the same coordinate system of Camera A. Then these three sets of values are averagely composed into one whole skeleton to be used for a further analysis.

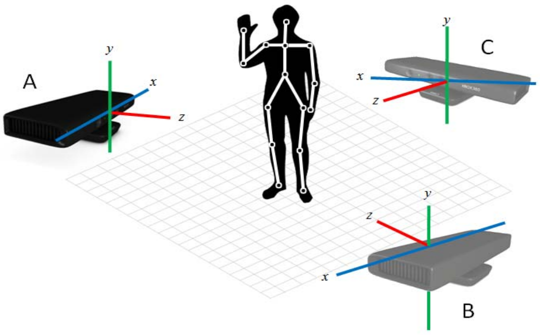


Figure 1. Multiple local coordinate systems of Cameras A, B, and C.

In order to find the needed translation distance values t_x, t_y, and t_z and rotation angles θ_x, θ_y, and θ_z of Camera B to the reference Camera A as can be seen in Fig. 2, the following steps are used.

Step 1: Use initial values of the translation distance and the rotation angles, which are set to be in a suitable range in order to get a fast convergence. Also set a current sum of distance differences to some known large value.

Step 2: Perform a coordinate transformation and calculate the distance difference of each joint, using the Euclidean distance between the reference joint and the corresponding transformed joint.

Step 3: Sum up the distance differences of all fifteen joints and compare the new sum with the current sum. If the new sum is smaller than the current sum, then update the current sum to the new sum.

Step 4: Gradually change the translation distance values and the rotation angle values.

Step 5: Repeat Step 2 to Step 4 until the values exceed the range limit. Then use the transformation with the smallest sum.

Step 6: Repeat Step 1 to Step 5 for the remaining cameras with respect to the reference Camera A.

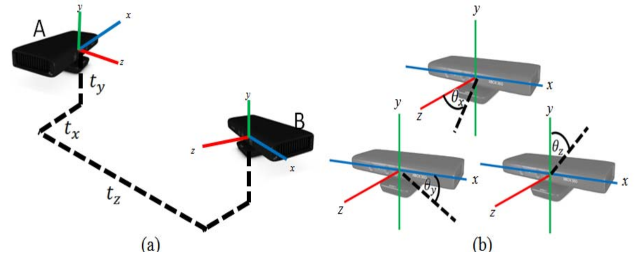


Figure 2. Example of the (a) translation distance values and (b) rotation angles between Camera A and Camera B.

B. Skeletal Composition

All cameras' orientations are identified after the coordinate transformation of each camera to the reference camera is performed, one after another; for instance, the transformation is done first for the reference Camera A and Camera B, thereafter the transformation is done for the reference Camera A and Camera C. In this example there would be three obtained skeletons, specifically, one directly from the reference Camera A, another transformed one from Camera B, and the other transformed one from Camera C. In general, the Kinect camera gives a confidence level of each joint's quality, ranging from zero to one. In our work, only the joints with the highest confidence equal to one are used in composing one whole skeleton which could be computed, as in

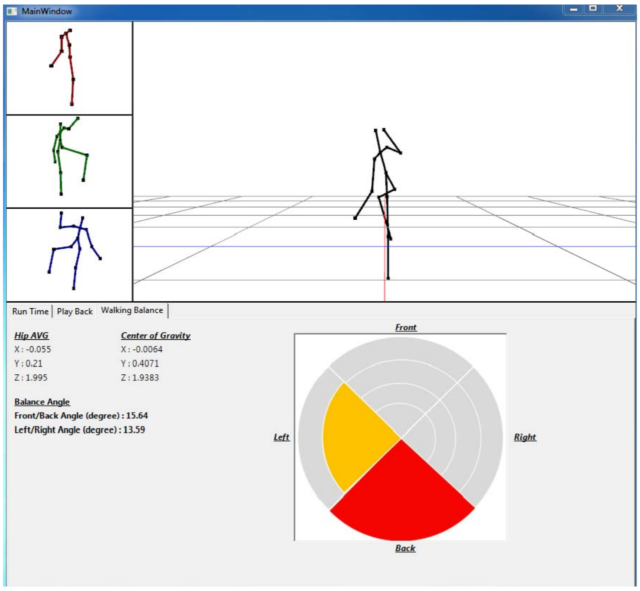
$$J_i = \frac{\sum_{k=1}^{n_i} J_{i,k}}{n_i}, \tag{6}$$

where J_i is the average position of the ith joint; $J_{i,k}$ is the position of the ith joint with the confidence equal to one from Camera k, and n_i is the number of cameras having the confidence equal to one at the ith joint. The missing joints are not used to compute the average.

C. Walking Posture Analysis

Three cameras are placed facing a human subject in different directions. The human must be away from each camera 0.8 to 4.0 meters. Any skeletal position closer than 0.8 meter or farther than 4.0 meters could not be captured [11]. Once the composed skeleton is obtained, the middle point between its left hip joint and right hip joint is computed, resulting in the so-called center hip point. Then the COM point is computed using the previously mentioned method. As a result, two vectors are formed, one connecting the left hip joint and the right hip joint, another connecting the center hip point to the COM point. Let's call these two vectors a Hip_{level} vector and a $Hip_{mid \to COM}$ vector, respectively.

In a normal standing posture, the Hip_{level} vector is parallel to a level surface, and the $Hip_{mid \to COM}$ vector is perpendicular to the surface. Under an unbalanced condition, the $Hip_{mid \to COM}$ vector would tilt toward front, back, left, or right in some degrees. Fig. 3 illustrates the composed view of a human's walking, which is obtained from three different views from three Kinect cameras. In addition, the upper body's unbalance is shown in four concentric circles with four different levels indicating different degrees of unbalance. The walking human in the figure tilts greatly toward back (four levels in red) and significantly toward left (three levels in orange).



Hip AVG
X : -0.055
Y : 0.21
Z : 1.995

Center of Gravity
X : -0.0064
Y : 0.4071
Z : 1.9383

Balance Angle
Front/Back Angle (degree) : 15.64
Left/Right Angle (degree) : 13.59

Figure 3. Example of a walking posture analysis.

There are four levels of unbalance degree. The higher the degree, the larger the radius, and represented in green (0-9 degrees), yellow (10-19 degrees), orange (20-29 degrees), and red (30-90 degrees), in ascending order of degrees. Eight possible directions of tilt include front, front left, front right, back, back left, back right, left, and right. Fig. 4 shows two unbalance examples of back right tilt and front left tilt in different degrees.

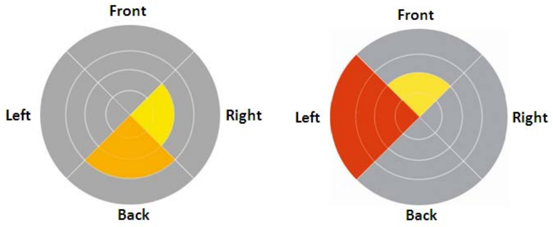


Figure 4. Examples of the upper body's imbalances.

IV. Results

We set up three Kinect cameras that wirelessly send the skeletal joints to one computer which acts as a processing unit used for composing the whole skeleton and analyzing a walking posture. Three thousand frames of video capture are used in each experiment. Note that the cameras operate at thirty frames per second (30 fps). Once the composed skeleton is obtained, two further analyses are performed as follows.

A. Accuracy of Skeletal Composition

We calculate at each joint the difference distance (in centimeter) between the composed skeleton and the skeleton from each camera in one of our experiments as shown in Table I. For instance, the composed Head joint is about 2.00 cm, 6.29 cm, and 10.25 cm away from the corresponding Head joints of Camera A, Camera B, and Camera C, respectively. Overall, this table demonstrates discrepancies of each camera's captures, possibly due to its sensory accuracy and surrounding noises.

TABLE I. Distance from Composed Skeleton of Each Camera

Joint	Distance (cm.) from the Composed Skeleton		
	Camera A	***Camera B***	***Camera C***
Head	2.00	6.29	10.25
Neck	2.57	3.41	10.78
Torso	5.62	3.24	13.25
Shoulder (left)	3.61	2.96	11.00
Elbow (left)	7.50	5.25	7.16
Hand (left)	12.62	13.06	8.51
Shoulder (right)	2.21	4.54	10.87
Elbow (right)	7.91	7.88	10.63
Hand (right)	16.72	9.80	19.19
Hip (left)	9.52	3.44	16.26
Knee (left)	5.72	4.09	14.11
Foot (left)	5.96	2.27	13.79
Hip (right)	8.56	5.12	16.78
Knee (right)	5.70	4.91	14.06
Foot (right)	4.78	2.48	12.97

In order to measure how accurate each camera is when compared to the composed skeleton, the root-mean-square error (RMSE) of each camera's skeleton is computed, as in

$$RMSE_k = \sqrt{\frac{\sum_{i=1}^{15} (J_i - J_{i,k})^2}{15}}, \qquad (7)$$

where J_i is the ith composed skeletal joint; $J_{i,k}$ is the kth camera's ith joint. The RMSE is 0.115, 0.103, and 0.187 for Camera A, B, and C, respectively. The lower the value, the closer of the camera's skeleton to the composed skeleton. In an ideal experiment where there are no sensory discrepancies, the RMSE of each camera would be zero because each transformed skeleton would place perfectly right on the composed skeleton.

Fig. 5 shows the different skeletal views of a standing human. The superimposed skeleton is averaged and shown as the whole skeleton. There are two missing joints from Camera B, and the side view from Camera C is rather difficult for us to apprehend. In light of more comprehensive views from various cameras, the whole skeleton is obtained and used to help do some postural analyses of our interest.

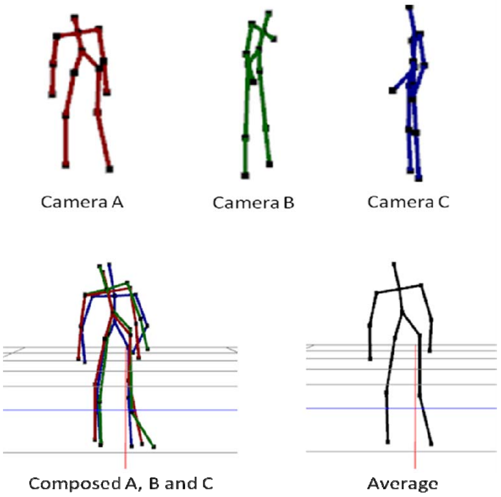


Figure 5. Example of the composed skeleton.

B. Multiple Cameras versus One Camera

To show the effectiveness of our work, we performed the experiment in which one of the three cameras could not capture a complete human skeleton at all times. Consequently, we could do the comparison between the complete, composed skeleton obtained from all three cameras and the incomplete skeleton obtained from one of the three cameras. Fig. 6 (a) shows the complete skeleton of a walking human. Its walking posture analysis results in the front left tilt with more emphasis to the left as shown in Fig. 6 (b). On the contrary, the incomplete skeleton results in no indication of any tilt as can be seen in Fig. 6 (c) and Fig. 6 (d), respectively.

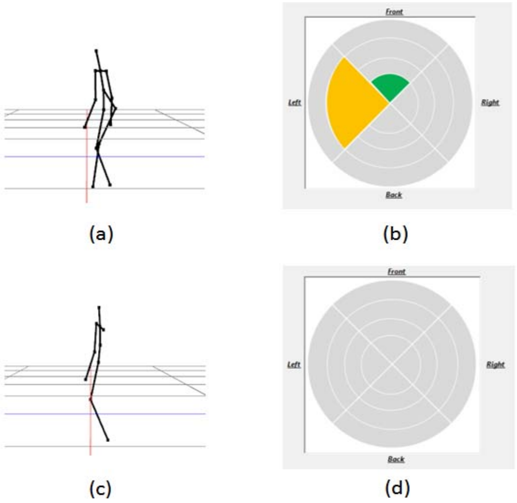


Figure 6. Views and analyses of multiple cameras (a, b) versus one camera (c, d).

V. Conclusions

We propose the automatic setting of multiple Kinect cameras. The cameras are simply placed in different angles toward a human to capture its skeleton. A complete skeleton comes with fifteen joints in total when a Kinect camera could entirely capture a human figure. In practice, however, a camera could not correctly capture those fifteen joints at one time, possibly due to various reasons; for instance, the human turns one side to the camera, or a furniture blocks some parts of the human. Therefore, the use of multiple cameras is needed in order for the more complete skeleton could be captured.

Each camera's coordinate system is transformed to the reference coordinate system which is chosen from one of the cameras. After all the transformations are complete, the composed skeleton is formed by averaging all the transformed joints, one by one. The composed skeleton is subsequently used in the walking posture analysis to find any upper-body tilts that may occur due to the human's poor physical condition. We have shown that our work could correctly identify tilt or unbalance of the human while walking and could be used in a simple walking posture analysis.

A. *Future Work*

This work would provide a solid extension for other Kinect-related researches, such as those found in human gesture recognitions [12], fall detections for the elderly people, physical therapy for home or hospital use, office syndrome monitoring, and smart bedrooms for the elderly people [13]. All the aforementioned work must rely on multiple Kinect cameras in order to capture a more complete human skeleton so that a detailed and reliable analysis could be performed. This work could very well be extended to accommodate those researches with the multiple-view capability.

We would like to improve further the accuracy of the composed skeleton, which is caused by the camera's sensory captures, by regularly adjusting the transformation matrix so that a better matrix is used instead of the one-time computed matrix. Currently the transformation matrix of each camera is found and subsequently used until our software application stops. The drawback of such method is that any discrepancy embedded in the matrix, even slightly, is used for the rest of the application. To improve it, at a regular interval, the new transformation matrix should be computed and used dynamically to adjust the current transformation matrix.

Acknowledgment

We wish to acknowledge the support from the Data and Knowledge Engineering Laboratory (D-Lab) of School of Information Technology at King Mongkut's University of Technology Thonburi (KMUTT).

References

[1] Microsoft, Xbox 360 + Kinect, [Online]: http://www.xbox.com/en-GB/Kinect/Home, [2013, June 23].

[2] L. Xia, C. Chen, and J. K. Aggarwal, "Human detection using depth information by Kinect," Computer Vision and Pattern Recognition Workshops (CVPRW), June, 2011, pp. 15-22.

[3] J. D. Huang, "Kinerehab: a kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities," The proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility, October, 2011, Scotland, UK, pp. 319-320.

[4] N. Pattanotai, P. Mongkolnam, and B. Watanapa, "Compositing human skeleton from motion captures using multiple Kinect sensors," The 4th National Conference on Information Technology (NCIT2012), April, 2012, Cha-am, Thailand, [in Thai].

[5] S. Kaenchan, P. Mongkolnam, B. Watanapa, and S. Sathienpong, "Multiple Kinect camera setting for compositing a human skeleton without using measuring instruments," The 9th National Conference on Computing and Information Technology (NCCIT2013), May, 2013, Bangkok, Thailand, pp. 354-360, [in Thai].

[6] OpenNI, Introducing OpenNI, [Online]: http://openni.org, [2013, June 23].

[7] H. Lee and L. Chou, "Detection of gait instability using the center of mass and center of pressure inclination angles," Archives of Physical Medicine and Rehabilitation, Vol. 87, April, 2006, pp. 569-575.

[8] A. A. Manasrah, "Human motion tracking for assisting balance training and control of a humanoid robot," Graduate School Theses and Dissertations, University of South Florida, 2012, pp. 13-22.

[9] F. S. Hill and M. K. Stephen, Computer Graphics using OpenGL, 3rd Edition, Pearson Education, 2007, p. 245.

[10] K. Shoemake, "Euler angle conversion," in P. Heckbert, Graphics Gems IV, Academic Press Professional, 1994, pp. 222–229.

[11] Microsoft, Kinect Sensor, [Online]: http://msdn.microsoft.com/en-us/library/hh438998.aspx, [2013, June 23].

[12] O. Patsadu, C. Nukoolkit, and B. Watanapa, "Human gesture recognition using Kinect camera," The proceeding of International Joint Conference on Computer Science and Software Engineering (JCSSE2012), May-June, 2012, Bangkok, Thailand, pp. 28-32.

[13] B. Ni, N. C. Dat, and P. Moulin, "RGBD-camera based get-up event detection for hospital fall prevention," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March, 2012, Kyoto, Japan, pp. 1405-1408.