Visualization in Engineering
a SpringerOpen Journal

RESEARCH ARTICLE Open Access

Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring

SangUk Han[1], Madhav Achar[2], SangHyun Lee[3*] and Feniosky Peña-Mora[4]

Abstract

Background: For construction management, data collection is a critical process for gathering and measuring information for the evaluation and control of ongoing project performances. Taking into account that construction involves a significant amount of manual work, worker monitoring can play a key role in analyzing operations and improving productivity and safety. However, time-consuming tasks involved in field observation have brought up the issue of implementing worker observation in daily management practice.

Methods: In an effort to address the issue, this paper investigates the performances of a cost-effective and portable RGB-D sensor, based on recent research efforts extended from our previous study. The performance of an RGB-D sensor is evaluated in terms of (1) the 3D positions of the body parts tracked by the sensor, (2) the 3D rotation angles at joints, and (3) the impact of the RGB-D sensor's accuracy on motion analysis. For the assessment, experimental studies were undertaken to collect motion capture datasets using an RGB-D sensor and a marker-based motion capture system, VICON, and to analyze errors as compared with the VICON used as the ground truth. As a test case, 25 trials of ascending and descending during ladder climbing were recorded simultaneously with both systems, and the resulting motion capture datasets (i.e., 3D skeleton models) were temporally and spatially synchronized for their comparison.

Results: Through the comparative assessment, we found a discrepancy of 10.7 cm in the tracked locations of body parts, and a difference of 16.2 degrees in rotation angles. However, motion detection results show that the inaccuracy of an RGB-D sensor does not have a considerable effect on action recognition in the experiment.

Conclusions: This paper thus provides insight into the accuracy of an RGB-D sensor on motion capture in various measures and directions of further research for the improvement of accuracy.

Keywords: Motion capture; Action recognition; Motion classification; RGB-D sensor; Machine learning

Introduction

During a construction project, data collection is critical to the evaluation and control of ongoing project performances. The complexity of construction environments and the dynamics of moving equipment and human resources, however, often pose a challenge in undertaking such tasks on a jobsite. Particularly, the time-consuming tasks required for worker monitoring can give rise to the issue of implementing field observation in a daily management practice (Johnson and Sackett 1998). For efficient field data acquisition, research efforts have thus been made to investigate and propose available sensing devices—such as cameras, laser scanners, and the combination of sensors (e.g., ultra wideband and physiological status monitoring devices)—for the tracking of human movements and the analysis of construction activities (Cheng et al. 2013; Gong and Caldas 2011; Peddi et al. 2009; Gonsalves and Teizer 2009). The previous studies provide valuable insight into the analysis of human postures and actions, but further research is still needed for the capture of an articulated motion and the modeling of its kinematics.

* Correspondence: shdpm@umich.edu
[3]Department of Civil & Environmental Engineering, University of Michigan, Ann Arbor, MI 48109, USA
Full list of author information is available at the end of the article

Along this line, an RGB-D sensor—such as the Microsoft Kinect sensor—has gained great attention as a cost-effective and readily available device for motion capture.

Since it was released in 2010, the Kinect has been actively studied as a motion capture device to record the movement of human subjects. In this regard, action recognition techniques—in particular—have been explored for the detection of specific actions using the motion capture data for use with operation and safety analysis in construction. For example, Weerasinghe et al. (2012) propose a Kinect-based tracking framework for the localization of workers and the analysis of their movement patterns, which could potentially be used for productivity measurement. For operation analysis, Escorcia et al. (2012) also present an action recognition technique to classify construction workers' actions based on the color and depth information from a Kinect. On the other hand, Ray and Teizer (2012) utilize a Kinect for the pose analysis of construction workers to classify awkward postures based on ergonomic rules during safety and health training, and Han et al. (2013) study the unsafe action detection of workers for safety behavior monitoring with motion capture data from a Kinect. These studies have thus demonstrated the great potential of the Kinect to gather motion information from a jobsite, as well as the great potential of its applications to construction management. To validate the proposed approach, however, the prior work has mainly focused on the performances of motion classification and detection rather than the accuracy of estimated postures and actions (e.g., 3D human skeleton models). The results in the studies suggest that pose estimation is computationally verified to a certain extent, but the accuracy of the Kinect solely when used for motion capture still remains unexplored. Taking into account that one of the main uses of the Kinect is to estimate 3D body skeletons of humans and track their movements over time, the thorough assessment of a Kinect-based motion capture system will thus help elucidate: (1) up to what degree of accuracy a Kinect sensor can detect and track the 3D positions of body parts; (2) to what research areas the Kinect can potentially be applied, depending on the accuracy; and (3) which processes of motion analysis cause computational errors for the debugging of action recognition systems.

This paper evaluates the performance of the Kinect sensor on motion capture and action recognition for construction worker monitoring. An experimental study is undertaken to compare the accuracy of a Kinect with a commercial marker-based motion capture system, VICON, which has been used as the ground truth in prior work (Dutta 2012; Stone and Skubic 2011; Fernández-Baena et al. 2012). A VICON tracks the 3D locations of reflective markers attached to body parts with multiple cameras (e.g., 6 or 8 cameras), thereby minimizing occlusions and producing accurate tracking results. Extended from our previous work (Han et al. 2012), this paper performs the error analysis based on: (1) the estimated 3D positions of body joints, (2) the recomputed 3D rotation angles at particular joints, and (3) the effect of the motion capture accuracy on motion detection. The rest of this paper is organized as follows. Background section provides a background on the Kinect sensor and its performance evaluation. Methods section demonstrates a research methodology used to compute and analyze the three types of errors for the comparative study. Experiment section describes the experimental process for the collection of motion capture datasets with both a Kinect and a VICON. Results, including the error analysis, are presented and discussed in Results and discussion section. Finally, Conclusion section summarizes the findings of this study and suggests the direction of future research.

Background

This section summarizes the pros and cons of an RGB-D sensor (i.e., Kinect) for motion capture, and reviews previous work on the performance evaluation of a Kinect motion capture system. Based on the literature review, further research efforts required in this domain are identified.

An RGB-D sensor for motion tracking and analysis

The Kinect sensor was initially developed as a motion-sensing device for video gaming. A Kinect consists of two main components—one is a RGB camera that produces images at a 640 × 480 resolution, while the other is a depth sensor that measures the depth information of the image (Rafibakhsh et al. 2012). In addition, the depth sensor is comprised of both a projector and an infrared (IR) camera, all of which projects a structured IR light onto the scene and measures the depth by analyzing the distortion of the IR light (Weerasinghe et al. 2012; Khoshelhan 2011). Accordingly, the Kinect allows not only for the 3D reconstruction of a scene with point clouds but also for the 3D skeleton extraction of a human subject as combined with the motion capture solutions (e.g., OpenNI, Microsoft Kinect for Windows SDK, iPi Soft Motion Capture). In terms of the image processing for motion capture, the measured depth can be used for the building of 3D human models through 2D pose estimation (i.e., 2D skeletons with depth), as well as for the direct inference of 3D poses by integrating the depth into the pose estimation process. On the other hand, the use of IR light brings about constraints in the practical application of a Kinect to a field setting. For example, the Kinect's sensitivity of IR light to sunlight may cause unreliable motion capture outcomes in an outdoor environment, and its operating ranges for motion capture are also known to be limited (e.g., 0.8–4 m) (Weerasinghe et al. 2012; Han et al. 2013). Nevertheless, previous studies report that the operating distance for object tracking can be extended up to 10 m (Rafibakhsh et al. 2012) and 7.5 m (Ray and Teizer 2012) from a camera; hence, further investigation is required to clarify the range issue. Though limited to indoor applications, the Kinect still has the following notable advantages for motion sensing: (1) it requires no additional body attachment (e.g., markers, a special suit), which allows for worker observation without the interference of ongoing work; (2) the cost of a sensor (e.g., approximately 150–250 USD) is quite competitive, compared with other motion capture systems (e.g., approximately 96–120K USD for a marker-based VICON system) (Han et al. 2013); (3) the minimum number of sensors for motion tracking is only one Kinect; and (4) it provides an easy-to-use and easy-to-carry means for data collection in a field setting.

Previous work on the performance evaluation of an RGB-D sensor

For motion capture, performances of the Kinect can broadly be evaluated in terms of the functionalities such as the depth measured by a sensor and body part positions estimated by motion capture solutions. This section summarizes the previous work on depth measurement and discusses issues in the pose estimation assessment.

A principal function of the Kinect sensor is to compute the depth (i.e., the distance from a sensor) as a laser scanner does. Due to its low cost compared with that of a laser scanner (e.g., 10–130K USD) (Golparvar-Fard et al. 2011), previous studies have investigated the accuracy and resolution of Kinect depth data for the 3D

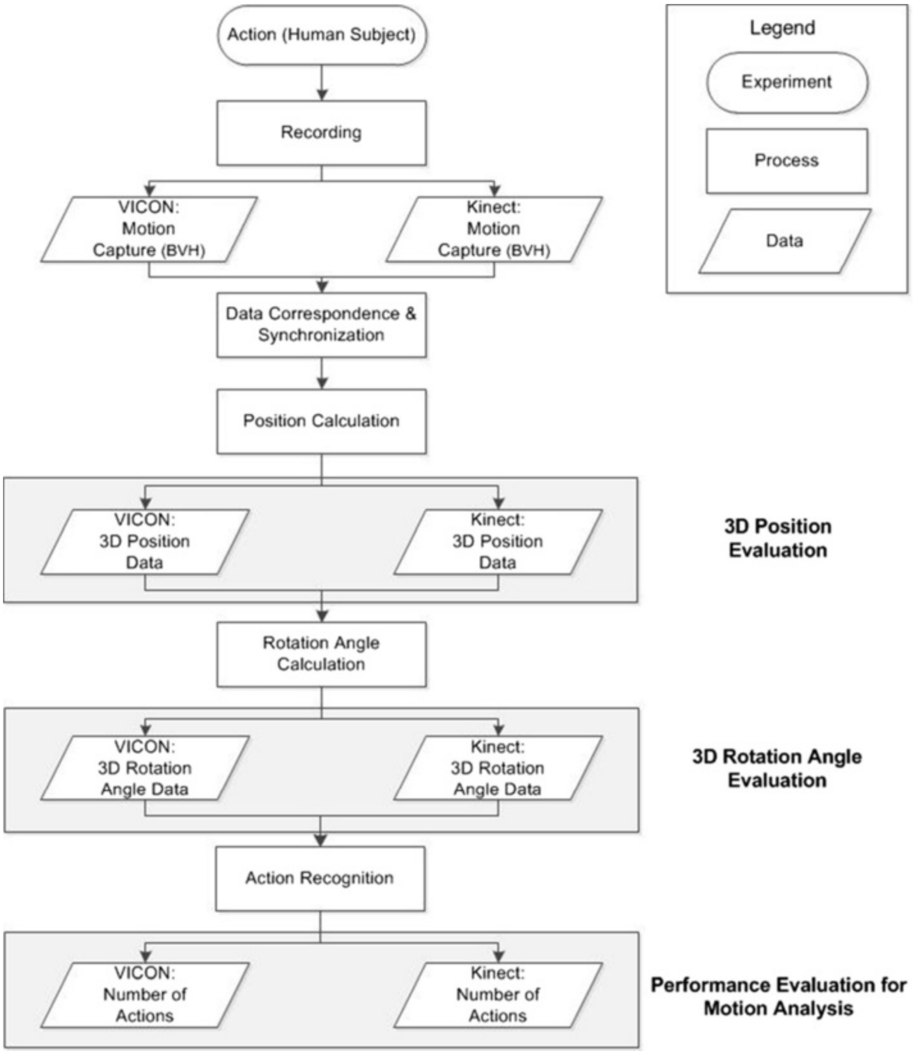
Action (Human Subject)
Recording
VICON: Motion Capture (BVH)
Kinect: Motion Capture (BVH)
Data Correspondence & Synchronization
Position Calculation
VICON: 3D Position Data
Kinect: 3D Position Data
3D Position Evaluation
Rotation Angle Calculation
VICON: 3D Rotation Angle Data
Kinect: 3D Rotation Angle Data
3D Rotation Angle Evaluation
Action Recognition
VICON: Number of Actions
Kinect: Number of Actions
Performance Evaluation for Motion Analysis
Legend
Experiment
Process
Data

Figure 1 Overview of evaluation processes.

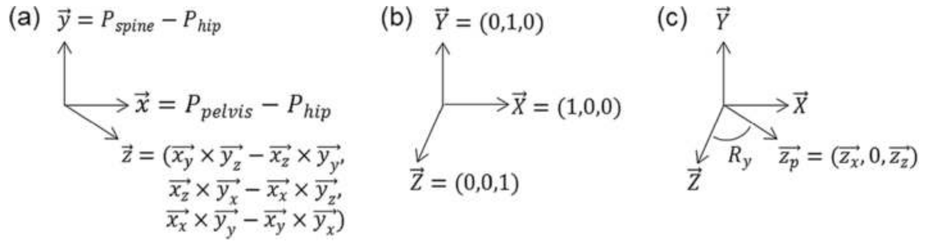
(a) $\vec{y} = P_{spine} - P_{hip}$
$\vec{x} = P_{pelvis} - P_{hip}$
$\vec{z} = (\overrightarrow{x_y} \times \overrightarrow{y_z} - \overrightarrow{x_z} \times \overrightarrow{y_y}, \overrightarrow{x_z} \times \overrightarrow{y_x} - \overrightarrow{x_x} \times \overrightarrow{y_z}, \overrightarrow{x_x} \times \overrightarrow{y_y} - \overrightarrow{x_y} \times \overrightarrow{y_x})$
(b) $\vec{Y} = (0,1,0)$
$\vec{X} = (1,0,0)$
$\vec{Z} = (0,0,1)$
(c) $\vec{Y}$
$\vec{X}$
R_y
$\overrightarrow{z_p} = (\overrightarrow{z_x}, 0, \overrightarrow{z_z})$
$\vec{Z}$

Figure 2 Y-axis rotation for data correspondences; (a) a local coordinate system of motion capture data, (b) a global coordinate system, and (c) Y-axis rotation between (a) and (b).

modeling of indoor environments, as well as for motion sensing. Khoshelham and Elberink (2012) report that the depth discrepancies between pairs of point clouds generated by a Kinect and a high-end laser scanner (i.e., Faro LS 880) are less than 3 cm for 84% of the point pairs, and that the point spacing in the depth direction (i.e., resolution) is about 2 mm, 2.5 cm, and 7 cm at the 1-, 3-, and 5-m distance. Rafibakhsh et al. (2012) also compare the accuracy and resolution of a Kinect with a laser scanner (i.e.. a Faro Focus3D scanner) and reveal that the average distance error between the point pairs is 3.49 cm, and that the resolution of the Kinect is about 4 times less than that of a laser scanner at 1.7- to 3.4-m distances from a sensor. Dutta (2012) measures the differences in distances between a Kinect and a VICON for a 0.1-m cube over a range of 1–3 m from a sensor, and the Root-Mean-Square Errors (RMSEs) are 6.5 mm in a horizontal direction, 5.7 mm in a vertical direction, and 10.9 mm in depth. On the other hand, Stoyanov et al. (2011) evaluate the accuracy of a Kinect in comparison with a laser scanner using the Three-Dimensional Normal Distributions Transform (3DNDT), which is a spatial representation accuracy evaluation technique, and conclude that the Kinect sensor performs well within 3.5-m distances. In Chow et al. (2012), a 3D reconstruction model of a mannequin is computed and compared with a laser scanner, and an RMSE of 11 mm is observed. In sum, previous studies reviewed herein conclude that the depth measurement and resolution of the Kinect are promising within a short range (e.g., 3 m), though not as accurate as those of a laser scanner, particularly in longer ranges.

The accuracy of motion capture data obtained with the Kinect has also been investigated. In Livingston et al. (2012), human skeletons tracked by a Microsoft software development kit are evaluated based on the positions of body joints (e.g., arms and hands) along a meter stick, and the average error and standard deviation in this experiment are 5.6 mm and 8.1 mm, respectively. Fernandez-Baena et al. (2012) conduct an experiment associated with rehabilitation treatments to compare the accuracy between a Kinect—combined with Natural Interaction Technology for End-user (NITE)—and a VICON in terms of the rotation angles of knee, hip, and shoulder joints, defined as angles

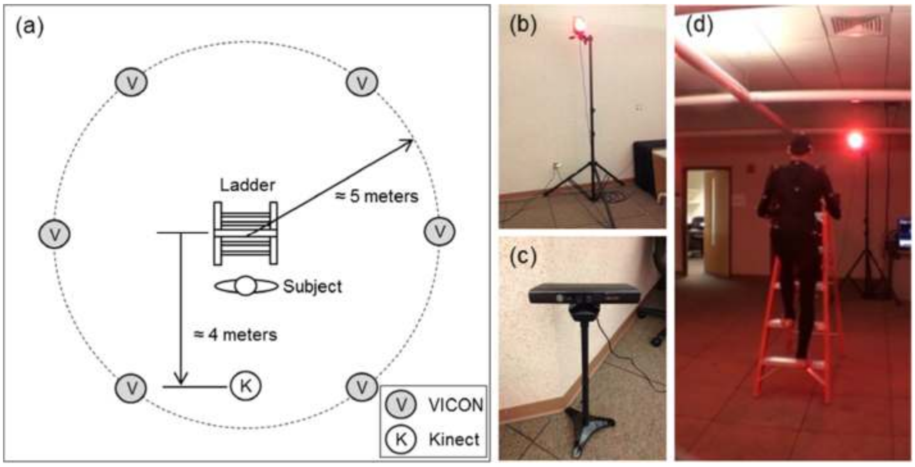

Figure 3 Experimental settings; (a) configurations of Kinect and VICON sensors, (b) a VICON sensor, (c) a Kinect, and (d) a human subject wearing a black suit and attaching reflective markers.

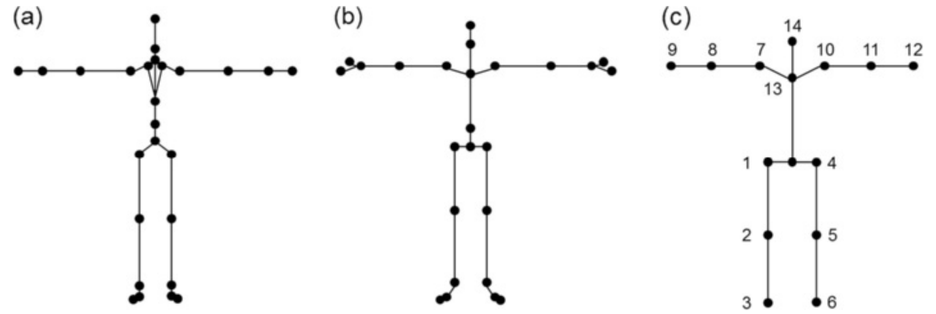

Figure 4 Human skeleton models; (a) a Kinect, (b) a VICON, and (c) a converted skeleton model (number: body joint ID).

between two vectors of body parts (e.g., one from knee to foot); the results show that the differences in rotation angles range from 6.78 to 8.98 degrees for a knee, from 5.53 to 9.92 degrees for a hip, and from 7 to 13 degrees for a shoulder. In the study of physical rehabilitation by Chang et al. (2012), the trajectories of the right hand, right elbow, and right shoulder that are tracked by a Kinect with OpenNI/NITE middleware are visually compared with those of marker-based OptiTrack motion capture system; the trajectories of a hand and an elbow are matched between two systems, while a shoulder is not accurately tracked by a Kinect system. To apply the Kinect to construction, however, further research efforts are required to address the following issues on the assessment of its motion capture performances: (1) the motions involved in construction activities need to be investigated, (2) the tracking results of full body joints need to be evaluated due to the characteristics of construction activities (i.e., manual work), and (3) the impact of the Kinect system's performances on action recognition needs to be studied for the analysis of construction worker monitoring and operation.

Methods

The objective of this paper is to assess the accuracy of Kinect motion capture data for the motion analysis of construction operations; Figure 1 illustrates an overview of evaluation processes comparing the outputs of VICON and Kinect motion capture systems. The evaluations are based on the error analysis of tracked 3D positions of full body joints, the 3D rotation angles at body joints used as a feature for motion classification, and the effect of the accuracy on action recognition. To compute the tracking errors, a VICON is used as the ground truth for motion tracking, and the iPi Motion Capture solution (http://ipisoft.com) is used with Kinect sensors to track the 3D positions of a human subject and extract 3D skeletons; the iPi Motion capture system estimates human poses mainly based on the depth measurements of a human body, and is thus less affected by a performer's appearance (e.g., special black suit and markers required by a VICON). In the experiment, human motions are thus simultaneously recorded with a Kinect and a VICON, and corresponding body joints of both systems—synchronized in time and space domains—are compared to compute the errors of Kinect outcomes. In addition, the ethics of this study including human subjects has been approved by the University of Michigan Institutional Review Board and the reference number is HUM00061888.

Data correspondence and synchronization

To compare the pose estimation results of a Kinect and a VICON, coordinate systems and data frames of both systems are matched through the rotation of coordinate systems and the synchronization of frames. For the spatial correspondence, local coordinate systems of both (i.e., coordinate systems defined by each system—an x-axis defined by the pelvis and a y-axis defined by the spine) are rotated about the y-axis into a global coordinate system (i.e., an absolute coordinate system newly defined for the coordinate system matching—a subject always faces the front) (Figure 2). In this experiment, a local coordinate system is defined based on the positions of a hip (i.e., P_{hip}), a spine (i.e., P_{spine}), and a pelvis (i.e.,

Table 1 Description of body parts and their joint IDs in Figure 4c

Body part ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Body part	Left upper leg	Left leg	Left foot	Right upper leg	Right leg	Right foot	Left arm
Body part ID	8	9	10	11	12	13	14
Body part	Left forearm	Left hand	Right arm	Right forearm	Right hand	Neck	Head

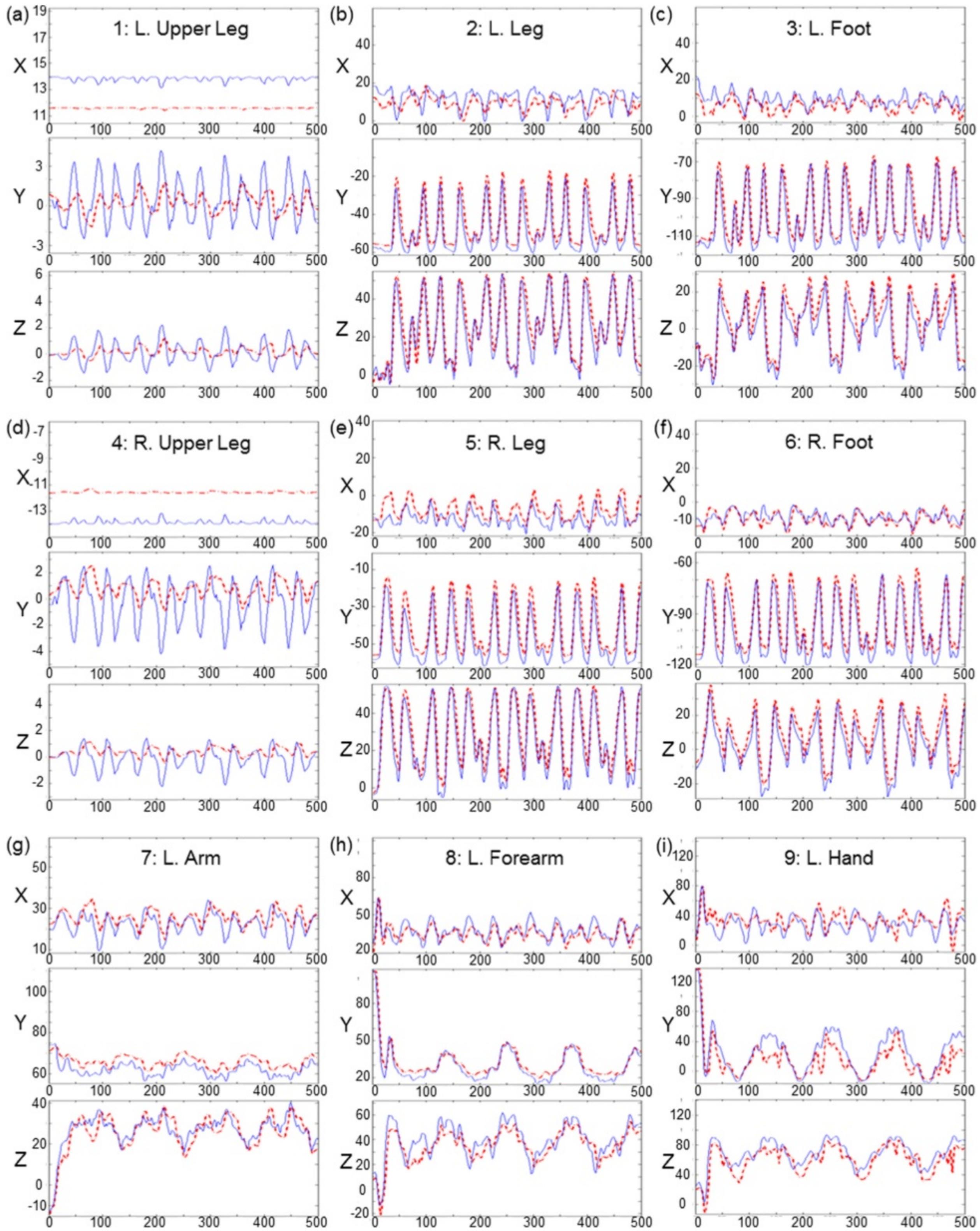

Figure 5 3D position trajectories of a Kinect and a VICON in x-, y-, and z-directions over the first 500 frames; (a) left upper leg, (b) left leg, (c) left foot, (d) right upper leg, (e) right leg, (f) right foot, (g) left arm, (h) left forearm, and (i) left hand.

P_{pelvis}) tracked by motion capture systems. The y-axis rotation angle, R_y, is calculated using Eq. (1):

$$R_y = \cos^{-1}\left(\overrightarrow{z_{p,x}} \times \vec{Z}_x + \overrightarrow{z_{p,y}} \times \vec{Z}_y + \overrightarrow{z_{p,z}} \times \vec{Z}_z\right) \quad (1)$$

where $\overrightarrow{z_{p,x}}, \overrightarrow{z_{p,y}}, \ and \ \overrightarrow{z_{p,z}}$ denote x, y, and z components of $\vec{z_p}$ in Figure 2c, and $\vec{Z}_x, \vec{Z}_y, \ and \ \vec{Z}_z$ denote x, y, and z components of $\vec{Z}$ in Figure 2b. Then, entire datasets of both systems are rotated using a rotation matrix obtained from R_y. In this manner, skeleton models of both systems face the front (i.e., z-axis), thus allowing for the comparison of skeletons in the same coordinate system regardless of viewpoints.

In the experiment, the synchronization of a pair of datasets is manually performed by identifying the same frame. For instance, we observe the frame in which a performer contacts a ladder's rung with a foot, and then we search for the exact frame among adjacent frames (e.g., 2 frames before and after the frame) by selecting the moment minimizing the distance between two datasets. In addition, the frame rates of the two systems are different (e.g., 120 frames per second for a VICON, and 30 frames per second for a Kinect). In the case of a VICON, thus 1 frame for every 4 is selected for the performance comparison. The accuracy is evaluated using RMSE in Eq. (2):

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{\left(x_{v,i} - x_{k,i}\right)^2}{n}} \quad (2)$$

where x_v denotes a VICON data value, x_k denotes a Kinect data value at each frame (i), and n is the total number of frames.

Action recognition

To evaluate the impact of motion tracking accuracy on action recognition, this paper adopts the action detection framework presented in our previous work (Han et al. 2013). The framework consists of the dimension reduction of high-dimensional motion data, similarity measurements between a pair of motion data, and motion classification based on the measured similarity. First, dimension reduction is needed due to the high dimensions in motion data (e.g., 78), which hinder efficient and accurate action detection. Thus, we use Kernel Principal Component Analysis (Kernel PCA) (Schölkopf et al. 1998) to map motion data onto a 3D space, and then we compare the trajectories of datasets in the low-dimensional coordinate. In this space, a trajectory represents a sequential movement of postures (i.e., actions), and actions can be recognized by comparing the temporal patterns of transformed datasets. For the pattern recognition, temporal-spatial similarity between a pair of datasets is quantitatively measured using Dynamic Time Warping (DTW) (Okada and Hasegawa 2008). In this study, DTW measures Euclidean distances between datasets by warping the datasets in a time domain so as to compare datasets, even the sizes (i.e., durations) of which are different. For the performance evaluation, thus the similarity between a motion template (i.e., one trial of action datasets) and the entirety of the data is computed over all of the frames, and the behavior (e.g., fluctuation) of measured similarities is compared to investigate the effect of motion capture systems on the detection accuracy. Eventually, we perform the action detection that recognizes actions based on similarities by observing the ones with less similarity than a threshold (i.e., a classifier learned through classification); the detection results of Kinect and VICON datasets are compared in terms of accuracy (i.e., the fraction of correctly classified actions among all sample actions), precision (i.e., the fraction of correctly detected actions among detected ones), and recall (i.e., the fraction of correctly detected actions among ones that should be detected).

Experiment

To collect motion capture data, a lab experiment was conducted in the University of Michigan 3D Lab (Han et al. 2012); experimental configuration and scenes are illustrated in Figure 3. In this experiment, actions during

Table 2 3D position comparison (cm) of body joints between a Kinect and a VICON

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Mean
3D	2.7	8.3	8.8	3.1	10.0	9.5	6.8	9.1	24.3	6.8	12.4	21.7	19.0	7.7	10.7
(Std.)	(0.4)	(3.2)	(3.0)	(0.7)	(3.7)	(3.6)	(2.3)	(3.5)	(12.0)	(2.7)	(4.9)	(12.2)	(1.2)	(2.3)	(5.3)
X	2.3	5.2	4.5	2.3	5.9	2.9	4.5	4.4	11.3	4.1	8.7	14.7	4.1	4.8	5.7
(Std.)	(0.2)	(3.9)	(3.0)	(0.2)	(3.7)	(2.7)	(3.2)	(4.0)	(10.6)	(3.3)	(5.1)	(10.9)	(3.2)	(3.9)	(5.1)
Y	1.3	4.4	4.5	1.8	6.0	5.8	3.6	3.5	17.4	3.1	4.6	10.9	17.3	4.7	6.4
(Std.)	(1.3)	(3.0)	(3.7)	(1.3)	(3.1)	(4.1)	(1.5)	(3.1)	(13.7)	(1.5)	(2.5)	(10.4)	(1.0)	(1.0)	(5.1)
Z	0.6	4.8	6.0	0.9	5.4	7.0	3.7	7.1	12.6	4.4	7.6	11.7	6.6	3.8	5.9
(Std.)	(0.6)	(4.2)	(3.6)	(0.6)	(4.5)	(3.4)	(3.6)	(6.3)	(9.0)	(3.7)	(6.1)	(11.6)	(3.6)	(3.0)	(5.4)

(Unit: cm).

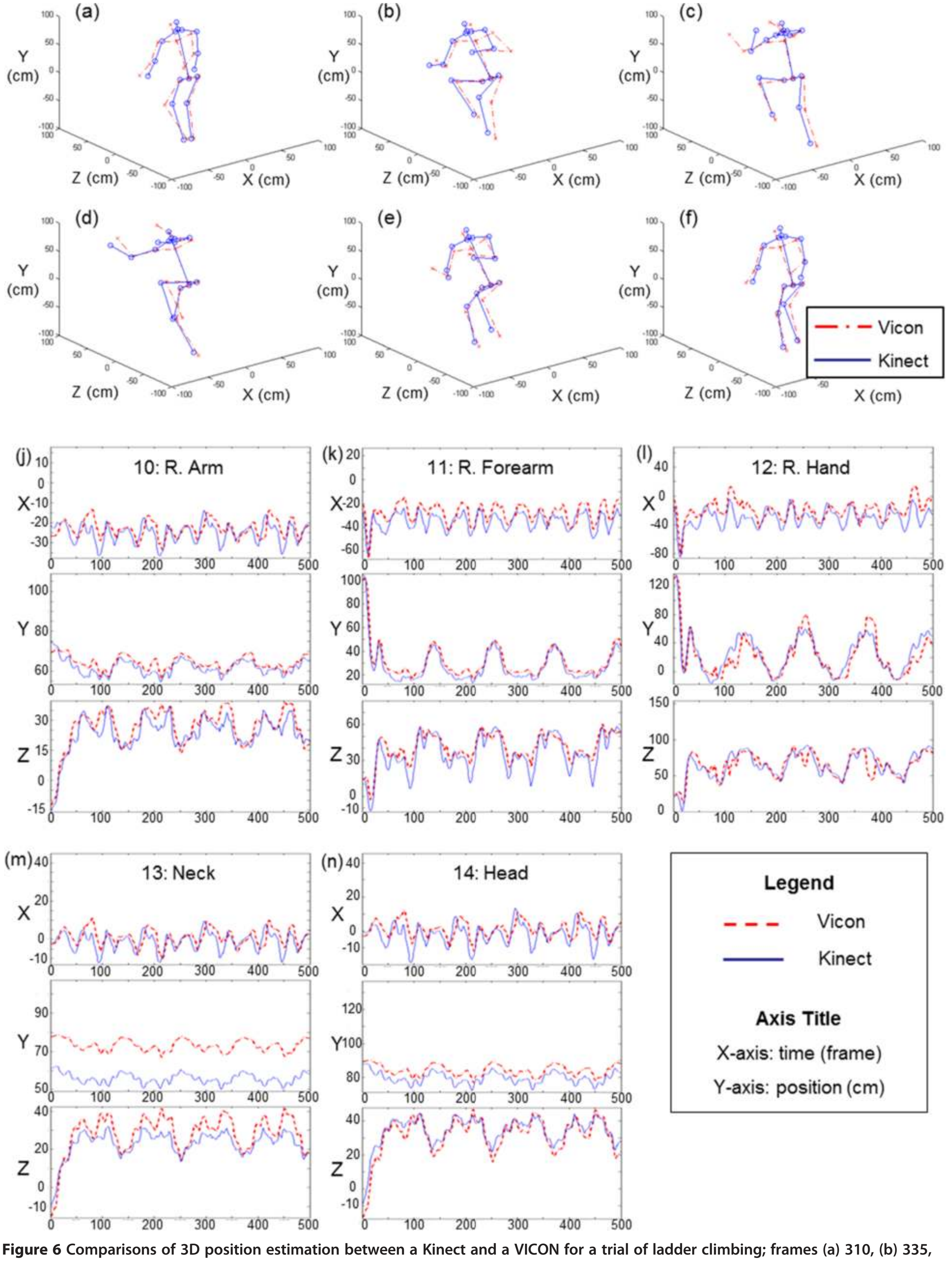

Figure 6 Comparisons of 3D position estimation between a Kinect and a VICON for a trial of ladder climbing; frames (a) 310, (b) 335, (c) 360, (d) 385, (e) 410, and (f) 435. 3D position trajectories of a Kinect and a VICON in the x-, y-, and z-directions over the first 500 frames; **(j)** right arm, **(k)** right forearm, **(l)** right hand, **(m)** neck, and **(n)** head (Note: graphs are scaled for each body part).

ladder climbing were recorded and analyzed; in construction, 16% of fatalities and 24.2% of injuries were caused by falls from a ladder in 2005 (CPWR 2008). 25 trials of each action (i.e., ascending and descending) taken by 1 subject were recorded with six 4-mega-pixel VICON sensors and a Kinect sensor. In total, 3,136 and 12,544 frames were collected with the Kinect and the VICON, respectively; and the datasets were synchronized for each system to have 3,136 frames for the comparison.

In this experiment, human skeleton models of the VICON and Kinect systems were slightly different in terms of the hierarchical structures of a human body; graphical illustrations of skeleton models extracted from each system are presented in Figure 4. Thus, for the comparison, corresponding body joints between the two systems are selected to convert the two models into the same form of a skeletal model (Figure 4c), and positions of such joints, as well as their rotation angles, are computed from motion capture data. For instance, motion capture data used in this study was in the Biovision Hierarchy (BVH) format (Meredith and Maddock 2001), in which a human posture at each frame is represented only with 3D Euler rotation angles. The BVH format also defines the 3D positions of body joints (i.e., translations) in an initial pose (e.g., T-pose as shown in Figure 4). This rotation and translation information forms a transformation matrix allowing for the computation of the 3D positions of all body joints in a global coordinate system (Meredith and Maddock 2001). To re-calculate Euler rotation angles (e.g., rotations in an order of x-, y-, and z-axes in this study) with respect to the converted skeleton model, an axis-angle between two body parts is first computed, a quaternion is defined with the axis-angle and axis vector, this quaternion forms a rotation matrix, and lastly a rotation angle is computed based on the rotation matrix (Han et al. 2012). Consequently, the 3D positions and rotation angles of each body part (Figure 4c) are compared to evaluate the tracking performances of the two systems; Table 1 describes body joint IDs corresponding to body parts in Figure 4c.

Results and discussion

To assess the performance of the Kinect as a motion capture system, we compare it with the VICON in terms of the results of: (1) 3D positions of body joints, (2) 3D rotation angles, and (3) motion detection for the datasets simultaneously collected through a lab experiment. Based on the error analysis, the applicability of the Kinect to the motion analysis of construction workers is discussed.

3D Position evaluation

To compare the 3D positions of body joints tracked by both systems, postures at each frame were iteratively rotated about the y-axis in a global coordinate system (Figure 2) over all of the temporally synchronized frames. Figure 5 visualizes skeleton models extracted from both systems at selected frames in the coordinate where two datasets are mapped. In this manner, the inspection of entire frames (i.e., animations) visually confirmed that the data correspondence and synchronization were successfully carried out for the two datasets. Through the visual investigation, we found that overall a Kinect model was closely matched with a VICON model, while hands and feet in particular were not exactly located in the same place.

For the quantitative assessment, RMSEs of body parts are computed over the entire frame using Eq. (2). Table 2 summarizes the RMSEs and standard deviations on distance differences in x-, y-, and z-directions, as well as in a 3D space; body part IDs refer to Figure 4 and Table 1. The temporal trajectories of the 3D positions of both systems in the first 500 frames are also presented in Figure 6. Compared with a VICON, a Kinect produces the discrepancy of 10.7 cm in a 3D coordinate, and no significant disparity in each direction was identified. The results show that the largest RMSEs are caused by the tracking of both hands (i.e., IDs 9 and 12) among body parts, and the large standard deviations of hands also indicate that the locations of such body parts are inconsistently estimated over the frames. Yet, Figures 6i and 6l imply that the patterns of a Kinect at large are still similar with those of a VICON. In addition, a large RMSE—the third greatest after that of the two hands—is found in a neck (i.e., ID 13). However, the standard deviation is relatively small, and most

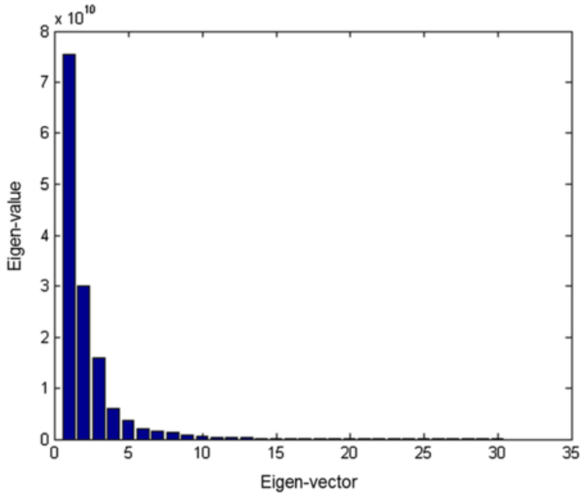

Figure 7 Eigen-decomposition for the internal dimensionality estimation.

Table 3 Rotation angle comparison (degree) at body joints between a Kinect and a VICON

ID	1	2	4	5	7	8	10	11	13	Mean
3D	5.1	5.6	6.2	8.2	13.9	34.2	18.9	49.0	4.4	16.2
(Std.)	(5.2)	(5.1)	(4.9)	(6.2)	(7.8)	(29.6)	(15.4)	(40.0)	(3.8)	(18.0)
X	6.6	7.3	6.7	8.1	12.1	31.2	18.7	38.6	6.0	15.1
(Std.)	(6.2)	(7.3)	(6.1)	(8.1)	(6.3)	(27.7)	(15.6)	(34.5)	(2.0)	(16.5)
Y	3.3	3.5	3.9	7.2	6.3	21.9	5.2	48.3	0.2	11.1
(Std.)	(3.2)	(2.9)	(2.7)	(5.1)	(5.5)	(19.6)	(5.2)	(33.8)	(0.2)	(13.5)
Z	5.6	5.9	7.9	9.2	23.4	49.5	32.9	60.0	7.0	22.4
(Std.)	(5.6)	(4.1)	(5.1)	(5.1)	(10.6)	(38.4)	(21.0)	(49.7)	(6.3)	(22.7)

(Unit: degree).

errors result from differences in a y-direction; this suggests that the tracking positions of a neck by the two systems are slightly different, as shown in Figure 6m. Next, relatively large RMSEs are caused by forearms (i.e., IDs 8 and 11), legs (i.e., IDs 2 and 5), and feet (i.e., IDs 3 and 6). As observed with hands, the trajectories of those body parts also similarly fluctuate over time with both a Kinect and a VICON (Figures 6h, 6k, 6b, 6e, 6c, and 6f).

The results show that the discrepancy between the Kinect motion capture system and a marker-based system is 10.7 cm on average in 3D positions. However, the estimated trajectories reveal that the Kinect sensor can still capture patterns of movements well, even with only one sensor. On the other hand, the use of one sensor may introduce the issue of occlusions. In the experiment, a Kinect sensor was positioned at the rear of a performer (Figure 3), and hence the performer's hands were frequently occluded by the performer, himself/herself. Also, forearms and legs, which were sometimes occluded as a performer climbed up and down a ladder, caused larger errors than other body parts. This implies that occlusions may have been a major source of errors in this experiment.

3D Rotation angle evaluation

In this experiment, rotation angles were the outcomes of both motion capture systems. However, other types (e.g., joint angles) of motion data—which can efficiently characterize human postures—can be obtained from motion capture systems and used as a feature for motion analysis. Taking into account that the selection of discriminating features significantly affects the classification performances (Mangai et al. 2010), we compared three data types in our previous study: rotation angles, joint angles (i.e., horizontal and vertical joint angles between a body part and x-y and x-z planes in a global coordinate system), and position vectors (i.e., normalized vectors of body parts) (Han, Lee, and Peña-Mora: Comparative study of motion features for similarity-based modeling and classification of unsafe actions in construction, submitted). The result reveals that, in the experiment, rotation angles outperformed the other two data types in applying the motion detection framework, which is also adopted in this paper. In this respect, 3D rotation angles used as inputs for motion analysis are compared to evaluate the accuracy of the Kinect and its impact on action recognition.

For the assessment, rotation angles were computed according to the converted skeleton model in Figure 4c. A rotation angle at a particular joint is defined as the angle rotating a vector of the joint (i.e., a vector from the joint to its child joint) from a corresponding vector in an initial pose. Thus, end-joints such as body part IDs 3, 6, 9, 12, and 14 are excluded from the comparison as not defined. Thus at the available joints, RMSEs of rotation angles in the x-, y-, and z-directions, as well as mean RMSEs of the three directions, were computed

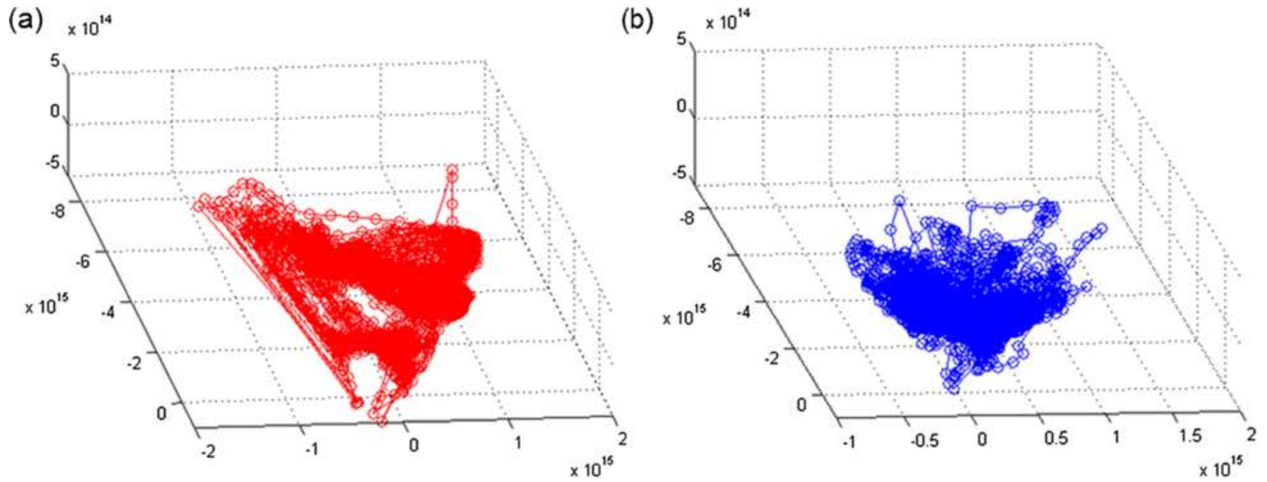

Figure 8 Dimension reduction results of (a) a VICON and (b) a Kinect (axis: eigen-vector; and unit: eigen-value).

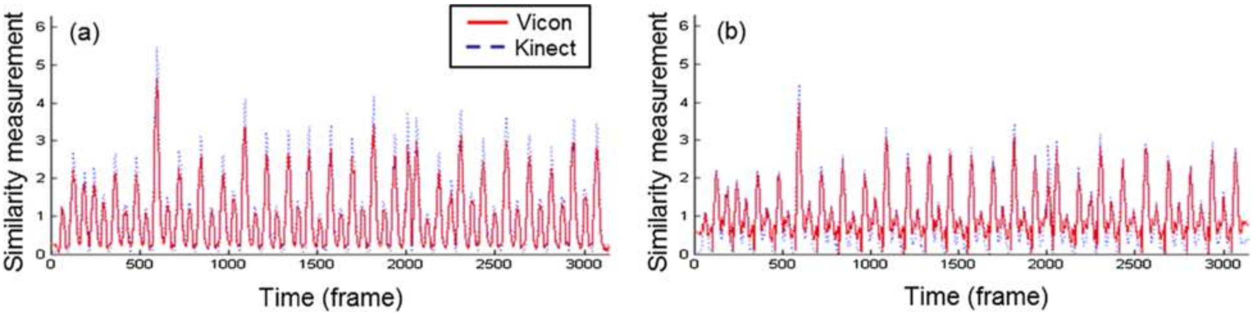

Figure 9 Comparison of similarity measurements between a VICON and a Kinect; (a) ascending, and (b) descending actions.

through all of the frames, and the results are presented in Table 3 (extended from Han et al. 2012). Compared to a VICON, overall the mean difference of 16.2 degrees and the standard deviation of 18.0 degrees were observed for the average of the three directions. For each direction, large errors occur in an order of z-, x-, and y-axis rotations (the z-axis has the largest error). In particular, the largest errors of up to 49 degrees were caused by forearms (i.e., IDs 8 and 11), which define the hand position. This implies that position errors at forearms and hands, which determine the rotation angles at forearms, can heavily magnify errors of rotation angles as combined. This phenomenon also explains the large errors of arms (i.e., IDs 7 and 10). The position errors of arms and forearms were not relatively large, but the combination of errors produces the second largest errors of rotation angles at the arms. Except for forearms and arms, the rotation angle errors of other body parts were less than 10 degrees.

Performance evaluation for motion analysis

To evaluate the performance of the Kinect for motion analysis, we applied a motion detection method (Han et al. 2013) to motion capture datasets from a Kinect and a VICON, and compared the results of detection based on conventional measures of classification performances (i.e., accuracy, precision, and recall). For motion analysis, dimensionalities of motion datasets were first reduced using kernel PCA. To determine the dimension to be reduced, eigen-decomposition was performed for the estimation of internal dimensionality in the datasets. As shown in Figure 7, the first three eigen-vectors have large eigen-values, which means that most information can be represented with three dimensions. In this regard, motion datasets were transformed onto a 3-dimensional coordinate; Figure 8 illustrates the distributions of each dataset in the low-dimensional space. In this space, a data point represents posture information at one frame, and hence the trajectories describe actions as changing postures over time. In Figure 8, the minimum and maximum values of each system are slightly different; for example, the ranges of x, y, and z values of Vicon are [−8.5*10^{15}, 1.1*10^{15}], [−1.0*10^{15}, 1.6*10^{15}], and [−6.0*10^{14}, 4.3*10^{14}], while the ranges of x, y, and z values of Kinect are [−4.9*10^{15}, 0.9*10^{15}], [−0.4*10^{15}, 1.5*10^{15}], and [−3.5*10^{14}, 3.2*10^{14}], respectively. However, the results indicate that the motion data captured by both systems could be mapped onto the same space. More importantly, despite large errors associated with body parts (e.g., arms and forearms) in rotation angles, the result of mapping (i.e., the transformation of high-dimensional data onto a low-dimensional space) reveals that the patterns of motion data can be preserved though dimension reduction; action detection is based on the comparison of patterns (i.e., trajectories) in a 3D space.

To compare the trajectories between actions, temporal-spatial similarities are measured using the DTW. In this experiment, one trial of datasets among 25 for each ascending and descending action was used as a motion template to compare its similarity with testing data and detect similar actions when the similarity is higher—or the distance is smaller—than a threshold. To avoid a

Table 4 Detection error comparison

Data source	Action type	# of actions in data	# of correctly detected actions		# of incorrectly detected actions		Acc. (%)	Prec. (%)	Rec. (%)
			Template (TP)	Other action (TN)	Not detected (FN)	Mis-detected (FP)			
Vicon	Ascending	25	25	25	0	0	100	100	100
	Descending	25	25	25	0	0	100	100	100
Kinect	Ascending	25	24	25	1	0	98	100	96
	Descending	25	25	25	0	0	100	100	100

biased assessment, a motion template from a Kinect and a VICON were compared with the same testing dataset (i.e., an entire frame of VICON data) for the detection; for instance, consistent errors (e.g., constantly estimating locations of a hand at a wrong but similar place) caused by a Kinect over the frames can positively affect the detection accuracy. The similarities measured over all of the frames are illustrated in Figure 9. Notwithstanding errors in Kinect data, the fluctuations of both datasets behave similarly over time. This result suggests that the errors of a Kinect system have not significantly affected the motion analysis in this experiment. Detection results (Table 4) also show that the accuracy, precision, and recall of a Kinect system are 98%, 100%, and 96%, respectively; only one trial among 25 was not detected.

Conclusions

This paper evaluates the performance of an RGB-D sensor (e.g., Kinect sensor) as a motion capture system based on the accuracy in estimated 3D positions and computed rotation angles, and the sensor's impact on action recognition. We conducted an experiment to collect motion capture data for 25 trials of ladder climbing actions, and we analyzed the accuracy on the datasets to identify the sources of errors. In the experiment, a 3D position RMSE and standard deviation were 10.7 cm and 5.3 cm, compared with a VICON. In the case of rotation angles, the RMSE and standard deviation were 16.2 degrees and 18.0 degrees, respectively. The rotation angles were used for motion detection, and the results show that among 25 trials, only 1 case of an ascending action was incorrectly detected (i.e., accuracies of 98% and 100% for ascending and descending actions, respectively). The experimental study implies that the inaccuracy of the Kinect motion capture system, particularly on occluded body parts, did not have a considerable effect on action recognition. However, the Kinect system produces large errors in estimating the positions of body parts, which can even increase errors as converted into rotation angles. The relatively lower accuracy of the Kinect system than that of marker-based systems can thus limit its application to construction; for example, the Kinect system may not be suitable for applications requiring high accuracy such as hand-related ergonomic analysis. Moreover, further investigation of Kinect performance evaluation on various actions (e.g., walking, running, lifting and carrying an object, and slipping) in construction operations is required for the thorough review of the feasibility of a Kinect for construction applications (e.g., productivity and safety). In addition, occlusions by a performer or other moving objects might be common in construction; thus a single Kinect motion capture system may potentially produce noise in a field setting. In this respect, further investigation on the use of multiple Kinect sensors is required to collect reliable motion information on a jobsite.

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

SH carried out the motion analysis studies, participated the sequence alignment and drafted the manuscript. MA undertook the experiments to collect data using a motion capture system. SL and FP directed the entire processes of this study, provided suggestions on each procedure in the data collection and analysis, and reviewed the manuscript. All authors read and approved the final manuscript.

Author details

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [2]Department of Computer Science, University of Michigan, Ann Arbor, MI 48109, USA. [3]Department of Civil & Environmental Engineering, University of Michigan, Ann Arbor, MI 48109, USA. [4]Department of Civil Engineering and Engineering Mechanics, Earth and Environmental Engineering, and Computer Science, Columbia University, New York, NY 10027, USA.

Received: 18 March 2013 Accepted: 6 May 2013
Published: 9 July 2013

References

Chang, CY, Lange, B, Zhang, M, Koenig, S, Requejo, P, Somboon, N, Sawchuk, AA, & Rizzo, AA. (2012). *Towards pervasive physical rehabilitation using Microsoft Kinect* (pp. 159–162). San Diego, CA: 2012 6th international conference on pervasive computing technologies for healthcare (pervasiveHealth).
May 21–24, 2012.

Cheng, T, Migliaccio, GC, Teizer, J, & Gatti, UC. (2013). Data fusion of real-time location sensing and physiological status monitoring for ergonomics analysis of construction workers. *Journal of Computing in Civil Engineering, 27*(3), 320–335.

Chow, J, Ang, K, Lichti, D, & Teskey, W. (2012). *Performance analysis of a low-cost triangulation-based 3D camera: Microsoft Kinect system*. Melbourne, Austrailia: The XXII Congress of the International Society for Photogrammetry and Remote Sensing.

CPWP – The Center for Construction Research and Training. (2008). *The construction chart book: the U.S. construction industry and its workers*. Washington, D.C: CPWP.

Dutta, T. (2012). Evaluation of the Kinect sensor for 3-D kinematic measurement in the workplace. *Applied Ergonomics, 43*, 645–649.

Escorcia, V, Dávila, MA, Golparvar-Fard, M, & Niebles, JC. (2012). *Automated vision-based recognition of construction worker actions for building interior construction operations using RGBD cameras*. West Lafayette, Indiana: Proceeding of 2012 Construction Research Congress (CRC). May 21–23, 2012.

Fernández-Baena, A, Susín, A, & Lligadas, X. (2012). *Biomechanical validation of upper-body and lower-body joint movements of Kinect motion capture data for rehabilitation treatments* (pp. 656–661). Bucharest: 2012 4th international conference on intelligent networking and collaborative systems. Sep.19–21, 2012.

Golparvar-Fard, M, Bohn, J, Teizer, J, Savarese, S, & Peña-Mora, F. (2011). Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques. *Automation in Construction, 20*(8), 1143–1155.

Gong, J, & Caldas, CH. (2011). *Learning and classifying motions of construction workers and equipment using bag of video feature words and Bayesian learning methods* (pp. 274–281). Miami, FL: Proceeding of 2011 ASCE International Workshop on Computing in Civil Engineering. June 19–22, 2011.

Gonsalves, R, & Teizer, J. (2009). *Human motion analysis using 3D range imaging technology*. Austin, Texas: 26th International Symposium on Automation and Robotics in Construction (ISARC). June 24–27, 2009.

Han, S, Achar, M, Lee, S, & Peña-Mora, F. (2012). *Automated 3D human skeleton extraction using range cameras for safety action sampling* (12th international conference on construction applications of virtual reality (CONVR)). Taipei, Taiwan: National Taiwan University.

Han, S, Lee, S, & Peña-Mora, F. (2013). Vision-based detection of unsafe actions of a construction worker: a case study of ladder climbing. *ASCE Journal of Computing in Civil Engineering*. in press.

Johnson, A, & Sackett, R. (1998). Direct systematic observation of behavior. In HR Bernard (Ed.), *A handbook of methods in cultural anthropology* (pp. 301–331). Walnut Creek, California: AltaMira Press.

Khoshelhan, K. (2011). Accuracy analysis of Kinect depth data. In DD Lichti & AF Habib (Eds.), *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (pp. 133–138). Calgary, Canada: Proceeding of ISPRS workshop laser scanning 2011 (Volume XXXVIII-5/W12th ed.). Aug. 29–31, 2011.

Khoshelham, K, & Elberink, SO. (2012). Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors, 12*(2), 1437–1454.

Livingston, MA, Sebastian, J, Ai, Z, & Decker, JW. (2012). *Performance measurements for the Microsoft Kinect skeleton* (pp. 119–120). Costa Mesa, CA: 2012 IEEE in Virtual Reality Workshops (VR). March 4–8, 2012.

Mangai, UG, Samanta, S, Das, S, & Chowdhury, PR. (2010). A survey of decision fusion and feature fusion strategies for pattern classification. *The Institution of Electronics and Telecommunication Engineers (IETE) Technical Review, 27*(4), 293–307.

Meredith, M, & Maddock, S. (2001). *Motion capture file formats explained*. University of Sheffielf: Department of Computer Science. http://www.dcs.shef.ac.uk/intranet/research/public/resmes/CS0111.pdf. Accessed 8 May 2013.

Okada, S, & Hasegawa, O. (2008). *Motion recognition based on dynamic-time warping method with self-organizing incremental neural network*. Tampa, FL: Proceeding of 19th International Conference on Pattern Recognition (ICPR).

Peddi, A, Huan, L, Bai, Y, & Kim, S. (2009). *Development of human pose analyzing algorithms for the determination of construction productivity in real-time* (pp. 11–20). Seattle, WA: Proceedings of the 2009 Construction Research Congress, Building a Sustainable Future. April 5–7, 2009.

Rafibakhsh, N, Gong, J, Siddiqui, MK, Gordon, C, & Lee, HF. (2012). *Analysis of XBOX Kinect sensor data for use on construction sites: depth accuracy and sensor inference assessment* (pp. 848–857). West Lafayette, IN: Proceeding of 2012 Construction Research Congress (CRC).

Ray, SJ, & Teizer, J. (2012). Real-time construction worker posture analysis for ergonomics training. *Advanced Engineering Informatics, 26*, 439–455.

Schölkopf, B, Smola, A, & Müller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*, 1299–1319.

Stone, E, & Skubic, M. (2011). Evaluation of an inexpensive depth camera for in-home gait assessment. *Journal of Ambient Intelligence and Smart Environments, 3*, 349–361.

Stoyanov, T, Louloudi, A, Andreasson, H, & Lilienthal, AJ. (2011). *Comparative evaluation of range sensor accuracy in indoor environments*. Orebro, Sweden: European Conference on Mobile Robots (ECMR). Sep. 7–10, 2011.

Weerasinghe, IPT, Ruwanpura, JY, Boyd, JE, & Habib, AF. (2012). *Application of Microsoft kinect sensor for tracking construction workers* (pp. 858–867). West Lafayette, IN: Proceeding of 2012 Construction Research Congress (CRC).

doi:10.1186/2213-7459-1-6
Cite this article as: Han *et al.*: **Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring.** *Visualization in Engineering* 2013 **1**:6.