A Stereo Vision-based Approach to Marker-less Motion Capture for On-Site Kinematic Modeling of Construction Worker Tasks

Richmond Starbuck[1], JoonOh Seo[2], SangUk Han[3], and SangHyun Lee[4]

[1]Undergraduate Research Assistant, Tishman Construction Management Program, Department of Civil and Environmental Engineering, University of Michigan, 2350 Hayward Street, Ann Arbor, MI, 48109-2125; PH (734) 695-1324; email: rstarbuc@umich.edu
[2]PhD Student, Tishman Construction Management Program, Department of Civil and Environmental Engineering, University of Michigan, 2350 Hayward Street, Ann Arbor, MI, 48109-2125; PH (734) 763-8077; email: junoseo@umich.edu
[3]Assistant Professor, Department of Civil and Environmental Engineering, University of Alberta, 3-015 Markin/CNRL Natural Resources Engineering Facility, 9105 116th Street, Edmonton, Alberta, Canada, T6G 2W2; PH (780) 492-2722; FAX (780) 492-0249; email: sanguk@ualberta.edu
[4]Assistant Professor, Tishman Construction Management Program, Department of Civil and Environmental Engineering, University of Michigan, 2350 Hayward Street, Ann Arbor, MI, 48109-2125; PH (734) 764-9420; FAX (734) 764-4292; email: shdpm@umich.edu

ABSTRACT

Marker-less motion capture has been extensively studied in recent years as a means of evaluating productivity, safety, and workplace design for manual operations on-site. These technologies are ideal for circumstances in which traditional motion capture systems are ineffective due to the need for a laboratory setting and movement-inhibiting markers or sensors. However, many marker-less motion capture systems rely on RGB-D sensors that have limited range and susceptibility to interference from sunlight and ferromagnetic radiation, making them unsuitable for modeling worker actions on construction sites. To address this issue, we propose a marker-less motion capture approach utilizing optical images and depth data obtained from stereo vision cameras. Multiple camera lenses and triangulation algorithms generate depths maps similar to those produced by RGB-D sensors, while still utilizing an optical recording process unhindered by potentially harsh construction site conditions. These data are adapted for existing kinematic modeling systems (i.e. iPi Mocap Studio) for 3-D pose estimation. The experiments show that the proposed approach can provide data precision comparable to that of RGB-D-based systems with fewer operational constraints; thus, motion data can be collected where previously developed methods fail due to environmental or maneuverability restrictions. With the proposed approach, kinematic modeling of human movements can be carried out on construction sites without inhibiting the mobility of the recorded subject.

INTRODUCTION

Human actions are a major contributing factor to workplace concerns in construction. For example, 76.2% of a construction project's employees are directly or indirectly involved in production processes (BLS 2012). Additionally, 80%–90% of accidents in construction result from unsafe worker behaviors (Salminen and Tallberg 1996; Lingard and Rowlinson 2005), and awkward postures are one of the primary risk factors relating to construction workers' health issues (Everett 1999). As such, effective monitoring and evaluation of worker's postures and actions can enhance site safety and productivity. However, commonly employed worker appraisal techniques (e.g. direct human observation, controlled experiments, and surveys) may not provide quantitative data of real world worker actions necessary for motion analysis due to a lack of accurate kinematic measurements taken on-site.

Recently, marker-less motion capture-based systems have been proposed for on-site kinematic measurement in order to provide data necessary for automated unsafe action detection, biomechanical analysis, and operation analysis (Ray and Teizer 2012; Esorcia et al. 2012; Han et al. 2013a; Seo et al. 2013). Marker-less motion capture can allow for data to be collected outside of specialized recording studios and without movement-inhibiting markers or sensors. Unfortunately, many of these systems utilize infrared-based RGB-D sensors (e.g. Microsoft Kinect) that are severely impeded by sunlight and ferromagnetic radiation and have limited operating range (Weerasinghe et al. 2012; Han et al. 2013b). Because these systems cannot be used outdoors, they are ineffective for monitoring worker tasks on construction sites.

To address this issue, we propose a stereo vision-based approach to data acquisition for human pose estimation and skeletal tracking. The stereo cameras employed measure line-of-sight distance using two lenses with a narrow baseline in a self-contained unit. This allows for both optical and depth data to be collected with few environmental restrictions (e.g. outdoor environments) and limited field-of-view.

For pose estimation, the data collected from the stereo camera are converted into a format used by an existing kinematic modeling software solution designed for use with RGB-D sensors. The software uses these data to perform skeletal tracking and the results are outputted into a standard motion capture file format. Thus, the proposed approach allows two readily available but previously incompatible systems to be used for on-site worker monitoring without significant programming efforts, despite the technical difficulties associated with utilizing stereo camera data in a motion capture solution intended for use with RGB-D sensors (e.g. overcoming issues arising from different depth data densities and frame rates between the two systems).

In order to evaluate performance, the proposed method is compared to traditional RGB-D sensor-based approaches. Specifically, a motion sequence is recorded by a stereo camera and RGB-D sensor simultaneously and pose estimation is executed on both sets of data. An analysis of variance is performed on the kinematic models produced by each to evaluate the accuracy of the proposed approach relative to existing RGB-D-based marker-less motion capture systems.

STEREO VISION-BASED KINEMATIC MODELING

The proposed approach utilizes a stereo camera with built-in depth measuring capabilities for generating 3D point clouds. These point clouds are converted into a video format used in a commercially available pose estimation software system.

Depth Measurement. The stereo camera used is a Bumblebee XB3™ manufactured by Point Grey Technologies (ww2.ptgrey.com) as shown in Figure 1. Multiple camera lenses capture a scene from different positions, and correspondence algorithms determine the distance disparity between pixels in each of the two images. The result is a disparity map from which real world depth can be computed for any point within the camera's field-of-view.

Figure 1. Point Grey Technologies Bumblebee XB3 stereo camera.

Pose Estimation. For human skeletal tracking the proposed approach implements iPi Desktop Motion Capture, a commercially available marker-less motion capture solution (www.ipisoft.com) chosen for its high level of pose estimation accuracy relative to alternative systems. Additionally, iPi Desktop Motion Capture does not utilize Microsoft Kinect skeletal tracking features which cannot be modified as is necessary for the proposed approach due to hardware and software copyrights held by 3-D sensor manufacturer PrimeSense. Two software systems are provided: iPi Recorder, for interfacing with RGB-D sensors to record and compress depth data, and iPi Mocap Studio, which processes the data and performs skeletal tracking. Pose estimation is performed using a prior model with a deformable mesh skin and articulated skeleton which is posed to best match the recorded data. Temporal correspondences are employed as well, allowing for pose modeling continuity during limb occlusion using measured trajectories from prior and latter frames.

iPi Mocap Studio was not intended for use with stereo cameras and provides no built-in framework for modifying the system such that alternative data sources can be utilized. In order to implement the proposed approach, a stand-alone program was developed for generating disparity maps in the same uncompressed AVI format as those produced by iPi Recorder (v1.8). Samples of these files recorded with a Microsoft Kinect sensor were analyzed to determine the structure of the data format.

The data are stored in three channels (red, green, blue) for which integer values can be specified for each pixel. The red channel contains the raw Bayer pattern (image stream) from the RGB camera, while the blue and green channels store the disparity data recorded by the depth sensor. Each 11-bit disparity value is split between an 8-bit pixel in the blue channel and a 3-bit pixel in the green channel:

$$P_{\text{blue}} = \text{remainder}(D/256) \quad \text{(Equation 1)}$$

$$P_{\text{green}} = \text{floor}(D/256) \quad \text{(Equation 2)}$$

Where P is the integer stored in the respective channel and D is the disparity value.

These disparity values are later recombined and converted into depth values by Mocap Studio. The Bumblebee stereo camera and Kinect sensor each use different equations for converting these disparity values into real world depths:

$$\text{Kinect: } Z = 0.1236 \tan(1.1863 + D_{\text{kinect}}/2842.5) \quad \text{(Equation 3)}$$

$$\text{Bumblebee: } Z = fB/D_{\text{bumblebee}} \quad \text{(Equation 4)}$$

Where Z is distance from the sensor to the pixel (meters), D is the disparity value (pixels), f is camera focal length (pixels), and B is lens baseline (meters).

Setting Equations 3 and 4 equal to each other and solving for the Kinect disparity value yields:

$$D_{\text{kinect}} = 2842.5 \tan^{-1}(8.0906 fB/D_{\text{bumblebee}}) - 3372.06$$

This equation is used to convert the Bumblebee disparity values into those expected by Mocap Studio, which are written to the AVI file using Equations 1 and 2.

Data Output. For the proposed approach, Biovision Hierarchy (BVH) was chosen for kinematic data output. A hierarchical joint structure is defined where, for each joint, a parent joint and initial position relative to that parent is specified. The joint hierarchy defined by iPi Mocap Studio is shown in Figure 2 with key joints of interest labeled, which will be referenced in Section 3.

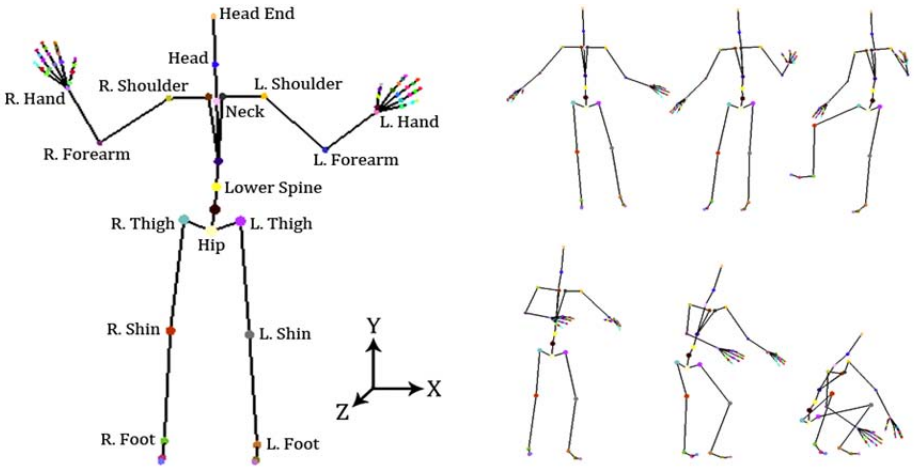

Figure 2. Skeletal hierarchy of BVH format with key joints labeled.

Motion data are stored as Euler rotations over each joint's local axes relative to an initial position implicitly defined by the joint hierarchy as well as translation of the root joint. Using these data and the initial orientation of each joint, the position of a joint at any frame of the animation can be computed using transformation matrices.

The coordinate system used in BVH files outputted by iPi Mocap Studio is right-handed with the Y-axis perpendicular to the ground plane. This convention will be maintained in the results section of this paper.

RESULTS

A controlled experiment was performed in order to evaluate the accuracy of the proposed method relative to existing RGB-D-based marker-less motion capture systems. A 21 second motion sequence was simultaneously captured using Microsoft Kinect with iPi Recorder and Bumblebee with the system described in Section 2.3. For this experiment, recording was performed indoors for the purpose of evaluating the proposed approach's accuracy relative to Kinect, which has been shown previously to be incapable of tracking human motions in outdoor environments (Weerasinghe et al. 2012; Han et al. 2013b). During recording, a subject performed motions that included rotations of all major body segments along all three local axes. To ensure that the system was capable of overcoming problems associated with marker-less motion capture, orientations that resulted in limbs being partially occluded from view of the sensors were included.

The depth data measured by each method were loaded into iPi Mocap Studio and skeletal tracking was performed as shown in Figure 3.

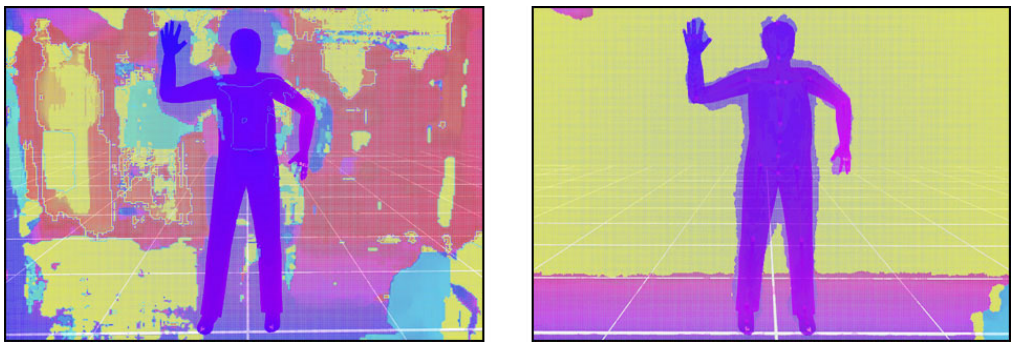

Figure 3. Pose estimation in iPi Mocap Studio using the proposed approach (left) and Microsoft Kinect (right).

For each of the BVH motion capture files outputted by Mocap Studio, positions of the joints of interest labeled in Figure 2 were computed using a transformation matrix from translation and rotation data contained in the files. The computed joint locations from each method were evaluated using the mean absolute error (MAE) in the X, Y, and Z dimensions. Additionally, the magnitudes of those errors were computed for each joint.

$$\mathrm{MAE}_{\mathrm{dim}} = \mathrm{sum}(s_{\mathrm{kinect,dim}} - s_{\mathrm{bumblebee,dim}}) \,/\, n$$

$$\| \mathrm{MAE} \| = \mathrm{sqrt}(\mathrm{MAE}_X^2 + \mathrm{MAE}_Y^2 + \mathrm{MAE}_Z^2)$$

Where s is the joint displacement in centimeters relative to an arbitrary fixed position and n is the number of frames in the animation.

Table 1. Mean absolute errors of estimated joint positions (cm) recorded by Microsoft Kinect and Bumblebee XB3 in a controlled experiment.

Joint Name	MAE_X	MAE_Y	MAE_Z	|| MAE ||
Hip	3.245 cm	4.706 cm	1.358 cm	5.875 cm
Lower Spine	2.700 cm	4.588 cm	2.686 cm	5.962 cm
Neck	2.896 cm	4.334 cm	4.654 cm	6.988 cm
Head	2.953 cm	4.220 cm	2.674 cm	5.804 cm
Head End	3.593 cm	4.441 cm	1.733 cm	5.970 cm
L. Shoulder	2.853 cm	3.803 cm	4.670 cm	6.665 cm
R. Shoulder	3.389 cm	3.998 cm	5.211 cm	7.390 cm
L. Forearm	3.621 cm	2.971 cm	3.561 cm	5.884 cm
R. Forearm	3.212 cm	3.417 cm	3.280 cm	5.772 cm
L. Hand	4.831 cm	3.901 cm	5.615 cm	8.372 cm
R. Hand	5.317 cm	5.531 cm	6.305 cm	9.931 cm
L. Thigh	2.774 cm	5.115 cm	2.337 cm	6.295 cm
R. Thigh	3.283 cm	4.377 cm	2.971 cm	6.226 cm
L. Shin	2.211 cm	5.271 cm	2.372 cm	6.189 cm
R. Shin	3.090 cm	5.061 cm	1.246 cm	6.059 cm
L. Foot	1.177 cm	5.705 cm	3.884 cm	7.001 cm
R. Foot	3.701 cm	4.915 cm	7.502 cm	9.702 cm
Mean	3.226 cm	4.493 cm	3.651 cm	6.826 cm

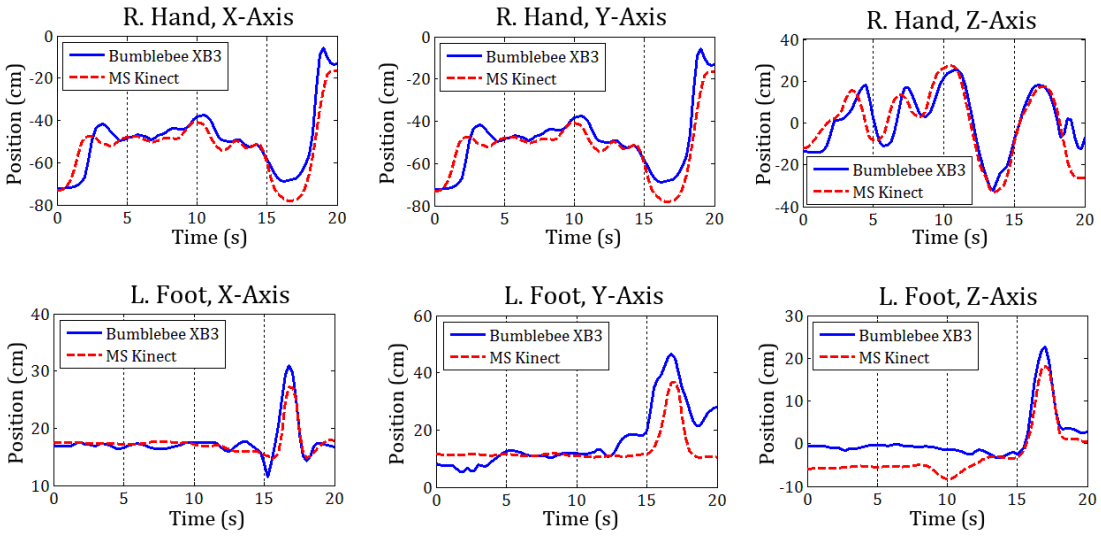

Figure 4. Graphs of right hand and left foot positions (cm) as a function of time (s) for Bumblebee XB3 (solid blue) and Microsoft Kinect (dotted red) over X/Y/Z axes.

DISCUSSION

As indicated in Table 1, the mean absolute error in position for the 21 seconds of recorded motion averaged over all joints of interest was 6.826 cm. Figure 4 shows that overall trajectories were accurately captured even for hands and feet, which produced the largest errors. Although these results are promising, there are several ways in which measurement accuracy may be improved upon that will be discussed.

The greatest measurement errors occurred in the vertical direction, the most likely reason for which is that Mocap Studio incorrectly estimated the orientation of the ground plane used for posing the prior model along the Y-axis. These estimations are based partially upon the field-of-view angle for the camera, which is assumed to be a Kinect sensor, and were erroneous due to the Bumblebee having a wider viewing angle. Implementing pose estimation techniques designed for use with stereo cameras may eliminate problems resulting from incorrectly assumed camera specifications.

Additionally, the use of multiple stereo cameras may provide better results. In previous experiments, it was hypothesized that utilizing multiple RGB-D sensors with overlapping fields-of-view would increase pose estimation accuracy (Han et al. 2012; Rafibakhsh et al. 2012). However, this approach produced poorer results than those achieved by a single sensor. This may have been due to the use of infrared in RGB-D sensors, where field-of-view overlaps could result in infrared beams emitted by one sensor to be received by another. Stereo cameras do not suffer from this potential limitation of multiple RGB-D sensors due to the use of an optical sensor mechanism; thus, multiple sensors may increase pose estimation accuracy in stereo camera-based methods by reducing occlusions and providing greater data density.

CONCLUSIONS

We have proposed a method for on-site recording and kinematic modeling of construction worker tasks to provide real world quantitative data necessary for motion analysis using stereo cameras. The depth data obtained from these cameras are adapted for existing pose estimation systems. Our initial results have shown that the measurement accuracy of the proposed method is comparable to that of traditional RGB-D-based systems. Thus, this paper opens up new opportunities to collect motion data even in outdoor environments, such as construction sites, where previously developed RGB-D-based methods have failed.

We plan to expand upon this work by designing pose estimation techniques specifically for stereo camera data and by employing multiple stereo cameras simultaneously. The proposed method will be implemented on construction sites to determine the practical limitations of modeling worker tasks and evaluate the efficacy of the system under harsh environmental conditions. Additionally, we will evaluate the accuracy of these methods relative to traditional marker-based motion capture systems and investigate the impact that measurement errors in marker-less systems have on motion analysis.

ACKNOWLEDGEMENT

The work presented in this paper was supported financially with a National Science Foundation Award (No. CMMI-1161123).

REFERENCES

Bureau of Labor Statistics (2012). "May 2011 occupational employment and wage estimates: National sector NAICS industry-specific estimates." Occupational Employment Statistics, U.S. Department of Labor, Washington, DC <http://www.bls.gov/oes/oes_dl.htm> (Dec 13 2012).

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

Everett, J. G. (1999). "Overexertion injuries in construction." Journal of construction engineering and management, 125(2), 109-114.

Han, S, Achar, M, Lee, S, Peña-Mora, F (2012). Automated 3D human skeleton extraction using range cameras for safety action sampling. Taipei, Taiwan: National Taiwan University.

Han, S., Lee, S., and Peña-Mora, F. (2013a). Vision-based Detection of Unsafe Actions of a Construction Worker: A Case Study of Ladder Climbing. Journal of Computing in Civil Engineering, ASCE, Vol. 27, No. 6, 635–644.

Han, S., Achar, M., Lee, S., and Peña-Mora, F. (2013b). Empirical Assessment of an RGB-D Sensor on Motion Capture and Action Recognition for Construction Worker Monitoring. Visualization in Engineering, Springer, 1:6.

Lingard, H., and Rowlinson, S. (2004). *Occupational health and safety in construction project management*. Taylor & Francis.

Rafibakhsh, N., Gong, J., Siddiqui, M., Gordon, C., and Lee, H. (2012) Analysis of XBOX Kinect sensor data for use on construction sites: Depth accuracy and sensor interference assessment. Construction Research Congress 2012: pp. 848-857.

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

Salminen, S., and Tallberg, T. (1996). "Human errors in fatal and serious occupational accidents in Finland." Ergonomics, 39(7), 980-988.

Seo, J., Han, S., Lee, S., and Armstrong, T. J. (2013). Motion Data-driven Unsafe Pose Identification through Biomechanical Analysis. Computing in Civil Engineering (2013), 693-700.

Weerasinghe, I. P. T., Ruwanpura, J. Y., Boyd, J. E., and Habib, A. F. (2012). Application of Microsoft Kinect sensor for tracking construction workers. Proceeding of 2012 Construction Research Congress (CRC), West Lafayette, IN, 858–867.