LOW-COST ACCURATE SKELETON TRACKING BASED ON FUSION OF KINECT AND WEARABLE INERTIAL SENSORS

François Destelle, Amin Ahmadi,
Noel E. O'Connor, Kieran Moran [*]

Insight Centre for Data Analytics
Dublin City University
Dublin, Ireland

Anargyros Chatzitofis, Dimitrios Zarpalas,
Petros Daras

Information Technologies Institute
Centre for Research and Technology Hellas
Greece

ABSTRACT

In this paper, we present a novel multi-sensor fusion method to build a human skeleton. We propose to fuse the joint position information obtained from the popular Kinect sensor with more precise estimation of body segment orientations provided by a small number of wearable inertial sensors. The use of inertial sensors can help to address many of the well known limitations of the Kinect sensor. The precise calculation of joint angles potentially allows the quantification of movement errors in technique training, thus facilitating the use of the low-cost Kinect sensor for accurate biomechanical purposes e.g. the improved human skeleton could be used in visual feedback-guided motor learning, for example. We compare our system to the gold standard Vicon optical motion capture system, proving that the fused skeleton achieves a very high level of accuracy.

***Index Terms*—** Kinect, Inertial sensor, Motion capture, Skeleton tracking, Multi-sensor fusion

1. INTRODUCTION

The capture and analysis of human movements (e.g. walking, jumping, running) is common in a number of domains, including: sport science, musculoskeltal injury management, neural disease rehabilitation, clinical biomechanics and the gaming industry [1, 2]. The analysis of joint/body segment position, angles and angular velocities, requires highly accurate motion capture. Unfortunately, the more accurate motion capture systems tend to be expensive, whether camera based (e.g. Vicon, UK) or inertia sensor based (e.g. XSens, Holland). This places highly accurate motion capture outside the reach of most users.

To increase access to motion capture, researchers have explored the use of depth cameras, such as Microsoft Kinect, as low-cost alternatives. Kinect uses a infrared based active stereovision system to get a depth map of the observed scene [3]. While the Kinect sensor was designed to recognize gestures in gaming applications, it has the capacity to determine the position of the center of specific joints, using a fixed and rather simple human skeleton. This allows for both the provision of visual feedback on the body's motion (which is essential in motor learning in the above science domains), and the measurement of joint motion. However, Kinect has limitations in accurately measuring the latter, especially when the joint motions are not parallel to the Kinect's depth sensor and when parts of the body are occluded due to the body's orientation.

A second problem with the Kinect system is its low and varying sampling frequency (25 - 35Hz) [3], which cannot be determined by the user. In particular, when assessing joint angular velocities, small errors in joint angles are significantly magnified when differentiated [4]. In addition, it is very problematic to quantify body position or joint angle at individual key events (e.g. initial foot strike when running) when movements are fast and sampling frequencies are so low that they preclude both the identification of when the key events occur and the capture of the frame of data at that specific time. This is an important requirement in domains such as sport science and clinical biomechanics. A possible solution to the limitations of the Kinect system is to combine the Kinect based data with data from wireless inertial motion units (WIMUs) which can provide greater accuracy in the measurement of body segment angles and angular velocities, and also have much higher sampling frequencies (e.g. up to 512 Hz) at consistent rates [5]. WIMUs can incorporate tri-axial accelerometers and gyroscopes, to determine angular measures and facilitate an accurate identification of key events which involve impact (e.g. ground contact when jumping, striking a ball in tennis) and thanks to advances in memos technology, they are relatively low cost. The use of WIMUs alone, however, is limited because of significant challenges in determining accurate joint center position necessary in the provision of visual feedback on the body's motion. This provides the motivation for fusing information from Microsoft Kinect and multiple WIMUs.

[*] The work presented in this paper was supported by the European Commission under contract FP7-601170 RePlay.

Ross A. Clark et al. presented in [3] a study about the precision of Kinect in a biomechanics or a clinical context, focusing on the movement of a subject's foot. It is noted that a Kinect is potentially able to achieve reliable gesture tracking of the subject's feet, especially if one combines the Kinect sensor with other modalities, i.e. another camera-based system. In [6], the authors explore the combined use of inertial sensors and Kinect for applications in rehabilitation robotics and assistive devices. The method was evaluated on experiments involving healthy subjects performing multiple degree-of-freedom tasks. As in the work presented in this paper, the author used Kinect as a first joint angle estimator as well as a visualization tool to give feedback to the patients in their rehabilitation process. But instead of considering the use of WIMUs for the analysis of human postures, this work aims to improve the WIMUs calibration using an online application based on Kinect captures. In [7], the authors present an application of the use of one Kinect to monitor and analyze post stroke patients during one specific activity: eating and drinking. The use of inertial-aware sensorized utensils can help this monitoring, introducing another source of information i.e. a fusion between the optical Kinect and inertial sensors. This study is however limited to the use of one specific inertial measurment unit and it is not designed to provide an end user feedback.

The aim of our work is to explore the benefit of combining the position-based information provided by Kinect with the orientation measures provided by the WIMUs sensors to determine an accurate skeleton representation of a subject along with measures of joint angle. The results are compared to a gold standard Vicon 3D motion analysis system.

2. CONSTRUCTION OF A FUSED SKELETON

2.1. Overview

In general, a Wireless/Wearable Inertial Measurement Unit, or WIMU, is an electronic device consisting of a microprocessor board, on-board accelerometers, gyroscopes and a wireless connection to transfer the captured data to a receiving client. WIMUs are capable of tracking rotational and translational movements and are often used in MoCap systems.

Although there are different technologies to monitor body orientation, wearable inertial sensors have the advantage of being self-contained in a way that measurement is independent of motion, environment and location. It is feasible to measure accurate orientation in three-dimensional space by utilizing tri-axial accelerometers, and gyroscopes and a proper filter. We have employed the filter described in [8] to minimise computational load and to operate at low sampling rates in order to reduce the hardware and software necessary for wearable inertial movement tracking. The mathematical derivation of the orientation estimation algorithm is described in the next section.

2.2. Computing orientation estimation from inertial sensors

In this paper, we use an algorithm which has been shown to provide effective performance at low computational expense. Utilizing such a technique, it is feasible to have a lightweight, inexpensive system capable of functioning over an extended period of time. The algorithm employs a quaternion representation of orientation and is not subject to the problematic singularities associated with Euler angles. The estimated orientation rate is defined in the following equations [8]:

$$\begin{cases} q_t = q_{t-1} + \dot{q}_t \Delta t \\ \dot{q}_t = \dot{q}_{\omega,t} - \beta \frac{\nabla f}{||\nabla f||} \end{cases}, \tag{1}$$

where

$$\begin{aligned} &\nabla f(q, E_g, S_a) = J^T(q, E_g) f(q, E_g, S_a) \\ &S_a = [0, a_x, a_y, a_z] \\ &E_g = [0, 0, 0, 1] \\ &q = [q_1, q_2, q_3, q_4] \end{aligned} \tag{2}$$

In this formulation, q_t and q_{t-1} are the orientations of the global frame relative to the sensor frame at time t and $t-1$ respectively. $\dot{q}_{\omega,t}$ is the rate of change of orientation measured by the gyroscopes. S_a is the acceleration in the x, y and z axes of the sensor frame, termed a_x, a_y, a_z respectively. The algorithm calculates the orientation q_t by integrating the estimated rate of change of orientation measured by the gyroscope. Then gyroscope measurement error, β, was removed in a direction based on accelerometer measurements. This algorithm uses a gradient descent optimization technique to measure only one solution for the sensor orientation by knowing the direction of the gravity in the Earth frame. The objective function f and its Jacobean J are defined by the following equations:

$$f(q, S_a) = \begin{bmatrix} 2(q_2 q_4 - q_1 q_3) - a_x \\ 2(q_1 q_2 + q_3 q_4) - a_y \\ 2(0.5 - q_2^2 - q_3^2) - a_z \end{bmatrix} \tag{3}$$

$$J(q) = \begin{bmatrix} -2q_3 & 2q_4 & -2q_1 & 2q_2 \\ 2q_2 & 2q_1 & 2q_4 & 2q_3 \\ 0 & -4q_2 & -4q_3 & 0 \end{bmatrix} \tag{4}$$

2.3. Initialization of the multi-modal sensor framework

As stated in section 2.2, each inertial sensor w is defined by a local coordinate system described by a quaternion q_w, thus a triple of orthonormal vectors (X_w, Y_w, Z_w). In order to join them in a common WIMU global coordinate system we initialize the sensors while they are fixed on a rigid plank, sharing the same orientation (X_W, Y_W, Z_W). From this initial configuration, we can evaluate a multiple WIMUs framework in a consistent manner.

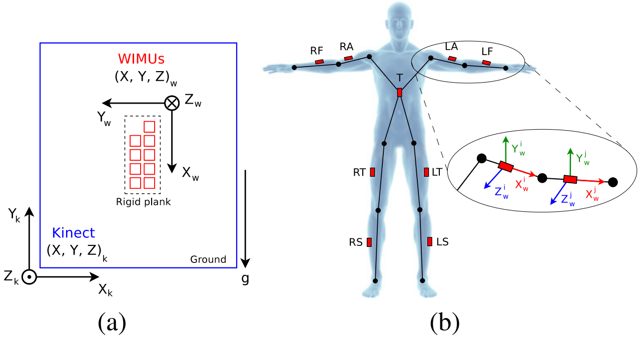

Fig. 1. Configuration of our testing platform. During the initialization step (a) the spatial coordinate system of each inertial sensor are the same and this spatial reference is aligned with the Kinect one. (b) Each WIMU is fixed on the subject's bones: X_w is fixed along the bone toward the smaller joint. The black skeleton is a representation of our fused skeleton.

The inertial sensors and Kinect do not share the same spatial reference coordinate system. As a second initialization frame, we map these two spatial systems together in order to compare consistently the Kinect rotational joints (X_K, Y_K, Z_K) and the WIMUs estimate orientation (X_w, Y_w, Z_w), see Figure 1(a). These two spatial configurations are linked by a known fixed rotation. We can then consider in our computation a unique global coordinate system embedding the multiple WIMUs and the Kinect sensor.

To apply this sensor framework to a real experimentation process, we need to map each inertial sensor to a specific bone of the subject. Furthermore, we need to identify what kind of rotation was performed over time from the local frame of each WIMU. We depict in Figure 1(b) the standard configuration we designed to tackle these two issues. On the one hand, the multiple WIMU framework is designed to be linked with the Kinect skeleton system. As a consequence, our nine inertial sensors are fixed to the subject's forearms, arms, thighs, shanks and finally to the chest. These correspond respectively to the fused skeleton joints R/LF, R/LA, R/LT, R/LT and T. On the other hand, to identify the three different kinds of rotation (flexion-extension, abduction-adduction, pronation-supination), each inertial sensor is fixed according to the scheme depicted in Figure 1(b): each sensor is fixed on the side of each limb. Each local sensor axis X_w^i is aligned with the associated bone, oriented toward the ground while in a standing pose. The Z_w axis is pointing toward the interior of this bone. As a consequence, the local rotations along the axes (X_w, Y_w, Z_w) describe the pronation-supination, the abduction-adduction and the flexion-extension of each limb relative to their associated joint respectively. The sensor attached to the torso is oriented as X_{torso} is pointing toward the ground and Z_{torso} is directed toward the back of the subject. Our experiments, as well as the global synchronization issue, are analyzed in section 3.

2.4. Skeleton fusion

We build our fused Kinect / WIMUs skeleton using three separate information sources given by each modality. The Kinect sensor provides the initial joint positions of our skeleton, as well as the global positioning of the subject's body over time. The WIMUs provide the orientation information we need to animate each bone of our fused skeleton over time.

Firstly, we consider a reference skeleton provided by the Kinect sensor and the associated skeleton extraction algorithm. This reference skeleton is the starting point of our fused skeleton synthesis method and is built from a reference frame captured by the Kinect. We need this reference skeleton to be as accurate as possible, in the sense that the Kinect algorithm produces a stable result. In this work, the reference frame is selected manually from a sequence where the subject stands still in front of the Kinect sensor. This step could be achieved automatically by measuring the relative stability of the results produced by the skeleton extraction algorithm over time. At this point, the fused skeleton is similar to the Kinect skeleton. The more carefully the reference frame is chosen, the more accurate the result will be.

Secondly, for each subsequent frame captured by the two sensory modalities, we consider one specific joint captured by the Kinect skeletonization algorithm, and the rotational data provided by the WIMUs. The aim of this specific Kinect skeleton joint is to track the global displacement of the subject's body over time, as the WIMUs cannot provide this information easily. For stability and simplicity purposes, we choose to consider the *torso* joint of the Kinect skeleton. As a consequence, the location of the central joint T (see Figure 1(b) for the sensor labels) of our fused skeleton is updated with respect to the displacement of this Kinect joint.

Finally, our fused skeleton is built from the reference skeleton. For each set of data captured by the WIMUs, each bone of our fused skeleton is rotated according to this rotational information in a hierarchical manner. It should be noted that the fused skeleton bones may not be aligned with the X_w axis of their associated inertial sensor: this case can only happen if the Kinect skeleton is perfectly aligned with the local orientation of each inertial sensor. Consider an inertial sensor $\mathcal{W}_t : \{q_t\}$ associated with a fused skeleton bone $\mathcal{B}_t \in \mathbb{R}^3$ constructed at a time t. We aim to rotate $\mathcal{B}_t$ according to the subsequent rotational information provided by $\mathcal{W}_{t+1}$. Let the quaternion Δq_{t+1} be the rotational offset occurring from t to $t+1$:

$$\Delta q_{t+1} = q_t^* \otimes q_{t+1} \tag{5}$$

The $\otimes$ denotes the quaternion product and $*$ denotes the quaternion conjugate The resulting rotated bone $\mathcal{B}_{t+1}$ can then be expressed by

$$\mathcal{B}_{t+1} = M^{\Delta q_{t+1}} \mathcal{B}_t \ , \tag{6}$$

where $M^{\Delta q_{t+1}}$ is the rotation matrix induced by Δq_{t+1}. The four bones linked to the fused skeleton joint T (see Figure 1(b)) are rotated using (6). This first process defines a new position for the starting and the ending points of our fused skeleton bones RA, LA, RT and LT (arms and thighs). In a hierarchical way, this displacement implies respectively new positions for the bones RF, LF, RS and LS (forearms and shanks). From this point, the bones RA, LA, RT and LT are rotated using the same method (6), inducing new hierarchical position changes. Then the bones RF, LF, RS and LS are rotated, our fused skeleton $\mathcal{B}^b_{t_{i+1}}, b \in [1, .., 12]$ is then finally complete from a time t_i to a subsequent one t_{i+1}.

3. RESULTS

3.1. Data Collection

To evaluate the proposed technique, data was captured using nine x-IMU wearable inertial sensors from X-IO Technologies, recording the data at 256 frames per second. The location of the sensor on each body segment was chosen to avoid large muscles, as soft tissue deformations due to muscle contractions and foot-ground impacts may negatively affect the accuracy of joint orientation estimates. In addition, a Kinect depth sensor was also employed to record the movements. In order to have a ground truth reference, the Vicon motion-capturing system using the standard Plug-in Gait model was also used. Reflective markers were placed on the body corresponding to Vicon's standard Plug-in Gait model. Twelve cameras were used to record the data at 250 frames per second. The subject was asked to perform a series of different actions with five trials for each gesture. However in this paper, only the knee and elbow flexion-extension are reported for a subject standing on their left leg while flexing and extending their right knee (simulated kicking). Since each sensor recorded data independently, a physical event was required to synchronize all inertial sensors together. This was achieved by instructing the subject to perform five vertical jumps, ensuring large acceleration spikes would occur simultaneously on each device, that would be clearly visible in the accelerometer stream.

3.2. Accuracy Evaluation

The Vicon data gathered provides orientation information, which serves as the ground truth of this evaluation procedure. Tracking of the Kinect skeleton was performed using OpenNI2, NiTE2 that computes positions and orientations of 13 human skeleton joints, see Figure 1(b).

For the evaluation of the proposed methodology, we chose to compare the joints angle of the knees and elbows, given their biomechanical importance. Typically a joint rotation is defined as the orientation of a distal segment with respect to the proximal segment. In order to measure body joint angles, the orientation of the two wearable inertial sensors attached on the distal and proximal segments were calculated using the described fusion algorithm. Then a technique based on leg and hand segment movements was used to calibrate and align the reference frame of the two inertial sensors [9, 10]. For instance, this can be applied to the upper arm and forearm segments to calculate elbow joint angles. This is described by the following equation:

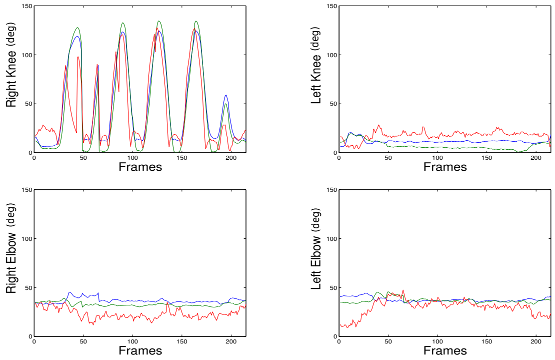

Fig. 2. Plots of four joint angles (deg) during the right knee flexion-extension. We compare the Kinect skeleton (red curves) and the WIMU orientations (blue curves) to the Vicon system (green curves) as a ground truth reference.

$$q_{elbow} = q^*_{upperarm} \otimes q_{forearm} \tag{7}$$

where $q_{upperarm}$ and $q_{forearm}$ are the quaternion representation of the orientation of the upper arm and forearm respectively.

Figure 2 shows the plots of four joint angles during the action of the right knee flexion-extension: the sujects both knees and elbows. We are comparing both the Kinect skeleton and the fused skeleton against the Vicon ground truth skeleton. These plots clearly shown that the fused skeleton produces joint angles that are much closer to the Vicon derived angles. One can see that the Kinect's knee angle behaves abnormally when the corresponding leg stretches during the knee flexion-extension. This occurs because Kinect allows the knee joint to bend in any direction, as depicted in the skeleton's left leg in Figure 3. Another observation is that for the joints that do not participate in one particular action (e.g. left knee during right knee flexion-extension) Kinect generates unreliable joint angles, which is not the case for the proposed fused scheme. Moreover, in all the plots it can be seen that the Kinect joint angles produce large fluctuations (i.e. greater noise) than the angles of the proposed method. Results from the fused scheme show smaller errors, with a relatively consistent offset to the Vicon data. The offset we can observe is due to the misalignement of the WIMUs along the subject's bones. Future research could focus on resolving this misalignment.

Further to the joint angle plots, Table 1 depicts the root mean squared error values (RMSE) and the normalized cross correlation measure (NCC) of each of the four joints by comparison with Vicon during two specific gestures. The actions chosen in this table are the right and the left knee flexion-

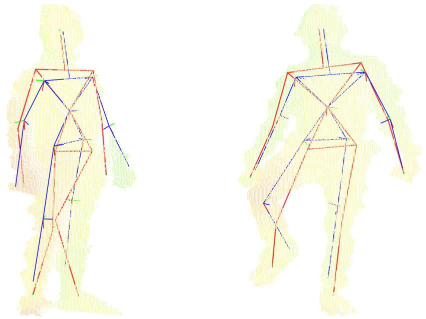

Fig. 3. The Kinect (with red color) and the proposed (blue) skeleton drawn over the Kinect's reconstructed depth map. Note the Kinect's inability to capture the left hand and leg on the left image and the right leg in the right image.

Joint angle	Left knee flexion		Right knee flexion	
	RMSE	NCC	RMSE	NCC
Kinect L-Elbow	16.73 $^\circ$	0.13	9.93 $^\circ$	0.61
Fusion L-Elbow	**14.19** $^\circ$	**0.70**	**3.81** $^\circ$	**0.85**
Kinect R-Elbow	12.06 $^\circ$	0.41	10.34 $^\circ$	0.56
Fusion R-Elbow	**6.97** $^\circ$	**0.89**	**5.12** $^\circ$	**0.84**
Kinect L-Knee	29.51 $^\circ$	-0.63	26.94 $^\circ$	-0.02
Fusion L-Knee	**6.79** $^\circ$	**0.73**	**8.98** $^\circ$	**0.50**
Kinect R-Knee	9.82 $^\circ$	0.82	12.96 $^\circ$	0.80
Fusion R-Knee	**4.10** $^\circ$	**0.99**	**5.86** $^\circ$	**0.99**

Table 1. The RMSE values of the chosen joint angles against the Vicon system and their normalized cross correlation measure NCC. We are measuring two different movements: left and right knee flexion-extension. Each gesture is performed while the subject stand still in front of the Kinect sensor, see an illustration in Fig.3 (right).

extension. RMSE values generated by our proposed method are lower than those from Kinect, and normalized cross correlation measures imply that our fused skeleton is far more accurate than the Kinect one referring to the Vicon skeleton. Figure 3 depicts two snapshots of the extracted skeleton. The skeletons are drawn over the Kinect foreground surface to enable a natural evaluation of the produced joints' positions and angles. As can be seen, the Kinect skeleton is not fully aligned with its depth map and results in large errors especially in the knee. The whole captured sequence that depicts the extracted skeleton is available on our website.

4. CONCLUSION

In this paper we have presented a multi-sensor fusion approach to improving the skeleton provided by the popular Kinect sensor. Whilst Kinect, designed as a games controller, provides an important low-cost approach to motion capture and measurement, the accuracy obtained is not sufficient for many biomechanical applications. For this reason, we introduce a second data modality, corresponding to multiple wireless inertial measurement units and present a framework that allows the efficient fusion of these complementary data sources. The results show that the proposed approach can obtain more accurate joint angle measurements, approaching those of very expensive gold standard optical capture systems.

REFERENCES

[1] Enda F Whyte, Kieran Moran, Conor P Shortt, and Brendan Marshall, "The influence of reduced hamstring length on patellofemoral joint stress during squatting in healthy male adults," *Gait & posture*, vol. 31, no. 1, pp. 47–51, 2010.

[2] Carole M Van Camp and Lynda B Hayes, "Assessing and increasing physical activity," *Journal of applied behavior analysis*, vol. 45, no. 4, pp. 871–875, 2012.

[3] Ross A. Clark, Yong-Hao Pua, Karine Fortin, Callan Ritchie, Kate E. Webster, Linda Denehy, and Adam L. Bryant, "Validity of the Microsoft Kinect for assessment of postural control," *Gait & Posture*, vol. 36, no. 3, pp. 372–377, July 2012.

[4] David A Winter, *Biomechanics and motor control of human movement*, John Wiley & Sons, 2009.

[5] Marc Gowing, Amin Ahmadi, François Destelle, David S Monaghan, Noel E OConnor, and Kieran Moran, "Kinect vs. low-cost inertial sensing for gesture recognition," in *MultiMedia Modeling*. Springer, 2014, pp. 484–495.

[6] Antonio Padilha Lanari Bo, Mitsuhiro Hayashibe, and Philippe Poignet, "Joint angle estimation in rehabilitation with inertial sensors and its integration with kinect.," *Conf Proc IEEE Eng Med Biol Soc*, vol. 2011, pp. 3479–83, 2011.

[7] H. M. Hondori, M. Khademi, and Cristina V Lopes, "Monitoring intake gestures using sensor fusion (microsoft kinect and inertial sensors) for smart home telerehab setting," in *IEEE HIC 2012 Engineering in Medicine and Biology Society Conference on Healthcare Innovation*, Houston, TX, Nov 7-9 2012.

[8] Sebastian OH Madgwick, Andrew JL Harrison, and Ravi Vaidyanathan, "Estimation of imu and marg orientation using a gradient descent algorithm," in *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–7.

[9] J Favre, BM Jolles, R Aissaoui, and K Aminian, "Ambulatory measurement of 3d knee joint angle," *Journal of biomechanics*, vol. 41, no. 5, pp. 1029–1035, 2008.

[10] Amin Ahmadi, David D Rowlands, and Daniel A James, "Development of inertial and novel marker-based techniques and analysis for upper arm rotational velocity measurements in tennis," *Sports Engineering*, vol. 12, no. 4, pp. 179–188, 2010.