Markerless 3D Human Pose Tracking in the Wild with fusion of Multiple Depth Cameras: Comparative Experimental Study with Kinect 2 and 3

Jessica Colombel, David Daney, Vincent Bonnet, François Charpillet

Markerless 3D Human Pose Tracking in the Wild with fusion of Multiple Depth Cameras: Comparative Experimental Study with Kinect 2 and 3.

Jessica Colombel, David Daney, Vincent Bonnet, and François Charpillet

Abstract Human-robot interaction requires a robust estimate of human motion in real-time. This work presents a fusion algorithm for joint center positions tracking from multiple depth cameras to improve human motion analysis accuracy. The main contribution is the proposed algorithm based on body tracking measurements fusion with an extended Kalman filter and anthropomorphic constraints, independent of sensors. As an illustration of the use of this algorithm, this paper presents the direct comparison of joint center positions estimated with a reference stereophotogrammetric system and the ones estimated with the new Kinect 3 (Azure Kinect) sensor and its older version the Kinect 2 (Kinect for Windows). The experiment was made in two parts, one for each model of Kinect, by comparing raw and merging body tracking data of two sided Kinect with the proposed algorithm. The proposed approach improves body tracker data for Kinect 3 which has not the same characteristics as Kinect 2. This study shows also the importance of defining good heuristics to merge data depending on how the body tracking works. Thus, with proper heuristics, the joint center position estimates are improved by at least 14.6 %. Finally, we propose an additional comparison between Kinect 2 and Kinect 3 exhibiting the pros and cons of the two sensors.

Jessica Colombel
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
e-mail: jessica.colombel@inria.fr

David Daney
Inria Bordeaux Sud Ouest - IMS (UMR 5218), F-33405 Talence, France
e-mail: david.daney@inria.fr

Vincent Bonnet
Univ. Paris Est Creteil, LISSI, F-94400 Vitry, France
e-mail: bonnet.vincent@gmail.com

François Charpillet
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
e-mail: francois.charpillet@inria.fr

1 Introduction

Human movement observation is necessary in robotics as for other applications to analyze human behaviours or to control assistance systems. When the perceptive system is used in situ (e.g., at home or at work) the usability of the system is almost as important as the minimum metrological performance to ensure the validity of the results obtained.

Nowadays stereophotogrammetric systems such as Vicon, Qualysis or Optitrack systems are used as golden standard for position tracking for laboratory experiments. However, these systems are too expensive and difficult to install on site. Furthermore, these may be constraining for the observed human because of the markers' placement. These constraints prevent the natural observation of humans in their environments such as home or work and do not permit to handle real life situations.

The emergence of wearable sensors or depth cameras capable of measuring human movement has extended the scope of investigation for personal monitoring outside the laboratory. Numerous algorithms based on Machine Learning for vision have obtained appealing results on human motion observation [1, 2]. Nevertheless, the accuracy obtained and the reliability of skeleton tracker results are not always satisfactory for biomechanical analysis or to ensure safe interactions with assistance robots. However, the use of several devices and fusion techniques can deeply improve such systems with a good compromise between quality, usability and cost. In particular, it is interesting to favour data redundancy for visual systems that suffer from occlusions [3].

In this context a Kalman filter based on a human model is proposed to merge and filter skeleton data from multiple sensors in real time. The Kalman filter is well known for the fusion of data from sensors, even from different types of sensors (i.e. depth camera and wearable sensors as IMU) [4]. Especially for data from depth sensors, this type of algorithm can be also used with anthropometric constraints and a dynamical model for human motion to compensate for the lack of a model in the body tracking software [5]. One objective of this paper is to evaluate if this approach remains useful and meaningful for the body tracking software Kinect 3 (Azure Kinect). Technical improvements over the body tracking software Kinect 2 are expected and are promising given the improvements observed between the first two generations of Kinect and the technical characteristics of both Kinect (See Table 1) [6, 7].

In this paper we are comparing results of the proposed fusion algorithm on one experiment in two parts: one with Kinect 2 and the other with Kinect 3 body trackers (see Figure 1), with the Qualisys system as reference for the ground truth. The main contributions of this paper are:

- a skeleton fusion algorithm based on an Extended Kalman Filter (EKF) with human model independent of sensors;
- a comparison of skeleton tracking of Kinect 2 and Kinect 3 faced to a reference system.

	Kinect 2	Kinect 3 Azure			
		NFOV	NFOV binned	WFOV	WFOV binned
Depth Camera (pixel)	512×424	640x576	320x288	1024x1024	512x512
FPS	30	30		15	30
Min depth distance (m)	0.5	0.5		0.25	
Max depth distance (m)	4.5	3.86	5.46	2.21	2.88
Horizontal FOV (degree)	70	75		120	
Vertical Fov (degree)	60	65		120	
Skeleton joint define	25	26 (SDK v.9.2)			

Table 1: Table of technical characteristics of Kinect 2 and 3.

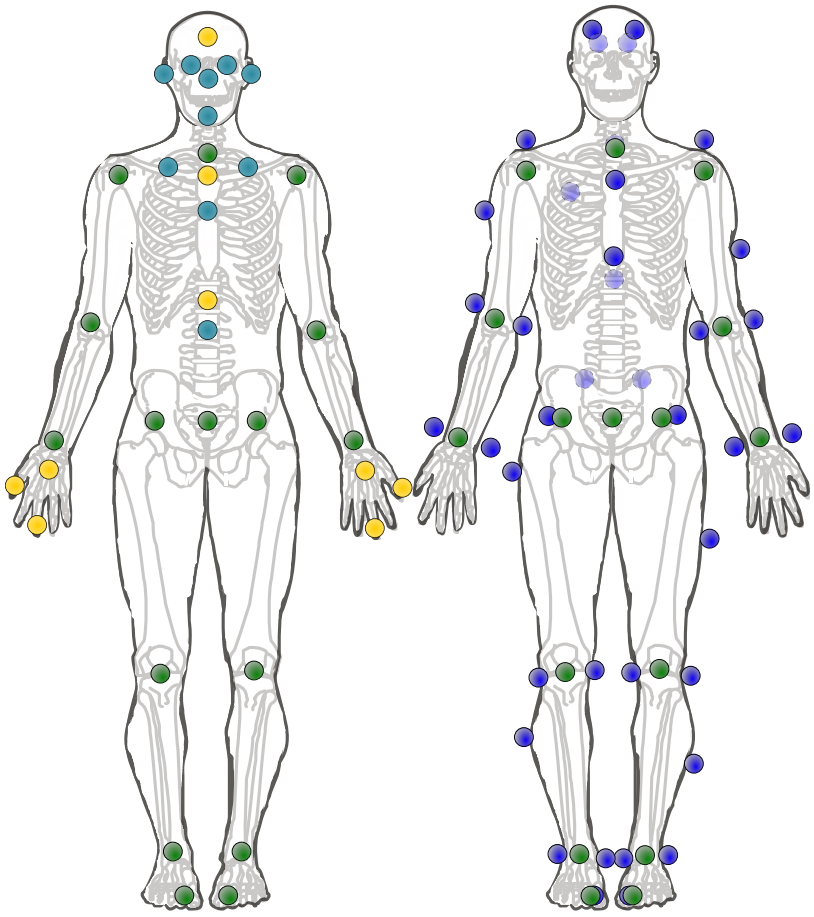
2 Human model and fusion methods

2.1 Joint center positions estimates

Both Kinect sensors provide skeleton estimate through the estimate of the 3D Joint Center Positions (JCPs). Kinect 2 estimates 25 JCPs including 3 joints for each hand, while Kinect 3 estimates the 26 JCPs including 5 joints for the head and 1 for each clavicle. Moreover, the joint center positions located on the spine are also different for each body tracker (namely the *SpineChest* and *SpineNaval* for Kinect 3 versus *SpineShoulder* and *SpineMid* for Kinect 2). Since both Kinects have different JCP estimates, we have selected 15 of them, called retained JCPs, located on the arms (except hands), legs and neck to be compared and analyzed. Figure 2a shows both skeleton models and the retained JCPs for the comparison. Other remaining joints were not considered in this analysis. The same 15 JCPs were also estimated using a reference stereophotogrammetric system and a set of 39 retroreflective-markers. These markers were set to match the Plug-In Gait full body template popularized by Vicon for the placement of retrofelective-markers [8]. Additional markers located on the lateral side of the knees, ankles and elbows were integrated to this template (see Figure 2b). Doing so reference JCPs were calculated using regression equations. For the elbows, wrists, knees and ankles it consisted to calculate the average position

Fig. 1 Pictures of the last two generations of Kinect : (Top) Kinect v2, (Bottom) Kinect Azure

between medial and lateral markers. The hip JCPs were estimated using the Bell's regression equation [9] based on the four pelvis markers. The shoulders JCPs were estimated by assuming an offset under the acromion markers along the vertical axis defined by the trunk segment [10]. Fig. 2b shows the location of the considered retroreflective markers and the estimated reference JCPs.



(a) (b)

Fig. 2: Location of (a) joint center positions of the Kinect 2 (yellow), and of the Kinect 3 (light blue). The green circles represent the 15 retained JCPs that stand for common JCPs to compare. (b) Retroreflective markers used for the reference stereophotogrammetric system (dark blue) and the estimated reference joint center position (green)

2.2 Extended Kalman Filter

The proposed EKF is based on a biomechanical model representing the human locomotor system. It is used to relate the JCPs estimated by the Kinect with the joint kinematics trough a kinematics model having fixed segment lengths. This kinematics model, composed of $N_\theta = 31$ revolute joints ($\boldsymbol{\theta}$) and $N_L = 12$ segments, was defined according to the recommendations of the International Society of Biomechanics [11, 12]. Modified Denavit-Hartenberg convention [13] was used to calculate the forward kinematics model that allows to express the position of each of the $N_J = 18$ estimated JCPs as a function of the joint angles $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$ and of segment lengths $\mathbf{L} \in \mathbb{R}^{N_l}$. This forward kinematics model is used as the measurement model h in the proposed Kalman filter. The EKF estimates the state vector $\mathbf{X}_k = [\boldsymbol{\theta}\ \mathbf{L}]^T$ while tracking the measurement vector $\mathbf{Z} \in \mathbb{R}^{3N_J}$ composed of N_J 3D coordinates of JCPs provided by each Kinect's skeleton data. The state and measurement vectors are modelled as follows at time k:

$$\begin{aligned} \mathbf{X}_k &= f(\mathbf{X}_{k-1}) + \mathbf{w}_{k-1}, \\ \mathbf{Z}_k &= h(\mathbf{X}_k) + \mathbf{v}_k, \end{aligned} \tag{1}$$

where f is the state model, $\mathbf{w}_k$ represents system noise defined by $p(\mathbf{w}) \sim \mathcal{N}(0, \mathbf{Q})$ with $\mathbf{Q}$ the model covariance noise matrix, and $\mathbf{v}_k$ represents the measurements noise defined by $p(\mathbf{v}) \sim \mathcal{N}(0, \mathbf{R})$ with $\mathbf{R}$ the measurement covariance noise matrix.

The proposed state model f is approximated by a linear form and denoted $\mathbf{F}$. It assumes that between two consecutive samples joint angles and segment lengths are constants, which is suitable for slow motion. Moreover, the Jacobian matrix $\mathbf{H}_k$ defined by $\frac{\partial h}{\partial \mathbf{X}_k}$ is used as the local linearization of the measurement model.

Thus, the equations of the extended Kalman filter are written as follows, with, at each time k the prediction phase first:

$$\mathbf{X}_k = \mathbf{F}\,\mathbf{X}_{k-1} \tag{2}$$

$$\mathbf{P}_k = \mathbf{F}\,\mathbf{P}_{k-1}\,\mathbf{F}^T + \mathbf{Q} \tag{3}$$

and then the update phase :

$$\mathbf{K}_k = \mathbf{P}_k\,\mathbf{H}^T\,(\mathbf{H}_k\,\mathbf{P}_k\,\mathbf{H}_k^T + \mathbf{R})^{-1} \tag{4}$$

$$\mathbf{X}_k = \mathbf{X}_k + \mathbf{K}_k(\mathbf{Z}_k - h(\mathbf{X}_k)) \tag{5}$$

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k\,\mathbf{H}_k)\,\mathbf{P}_k \tag{6}$$

with $\mathbf{P}_k$ the error co-variance matrix, $\mathbf{K}_k$ the Kalman's gain matrix and $\mathbf{I}$ the identity matrix with the good size.

2.3 Sensors data fusion

There are two methods for merging multiple sensors with Kalman as a function of the type of fused variables [14]. They are called *state-vector fusion* for the merge of the state vector $\mathbf{X}$ and *measurement fusion* for the measurement vector $\mathbf{Z}$. According to Gan et al [15], these two methods are functionally equivalent when the h_s measurement models of sensor s are identical for each sensors to merge. However state-vector fusion requires to augment the measurement vector while measurement fusion merges observation before filtering data with the Kalman filter. Thus for state-vector fusion the computational cost grows with the number of sensors. As the proposed algorithm tends to operate in real time with a variable number of sensors and the measurement model h_s are identical for all sensors, the measurement fusion was chosen. Thus the proposed method weights the measurements at each time k without increasing the matrix sizes as follows:

$$\mathbf{Z}_k = \left[\sum_{j=1}^{M} \mathbf{R}_{kj}^{-1}\right]^{-1} \sum_{j=1}^{M} \mathbf{R}_{kj}^{-1}\mathbf{Z}_{kj} \tag{7}$$

$$\mathbf{H}_k = \left[\sum_{j=1}^{M} \mathbf{R}_{kj}^{-1}\right]^{-1} \sum_{j=1}^{M} \mathbf{R}_{kj}^{-1}\mathbf{H}_{kj} \tag{8}$$

$$\mathbf{R}_k = \left[\sum_{j=1}^{M} \mathbf{R}_{kj}^{-1}\right]^{-1} \tag{9}$$

with $\mathbf{Z}_{ks}$ the measurement vector, $\mathbf{R}_{ks}$ is the covariance matrix of the measurement noise and $\mathbf{H}_{ks}$ the Jacobian of the measurement model of sensor s at time k, and M the number of sensors, here the number of Kinects.

2.3.1 Heuristics

The fusion of multiple sensors can be done without the knowledge of sensor confidence index. However, in the literature, several heuristics were used to weigh the measurements for merging multiple Kinect 1 or Kinect 2, such as point continuity with respect to velocity [16], segment length stability [3] or placement with respect to Kinect [17]. These heuristics are based on technical specifications of the manufacturer depending on the sensors.

So as to improve the fusion result, we propose a new set of heuristics inspired by those cited above that represents a confidence index in each sensor. Even though we merge several Kinect 2 on one side and several Kinect 3 on the other, they are based on the characteristics of the Kinect 2 only. These heuristics result in weights $w \in]0.1]$. Two of them are related to the placement of the body in relation to the

sensor:

$$w_{d_z} = \begin{cases} 0.1 & \text{if } d_z < 1.5 \\ 1 & \text{if } 1.5 \leq d_z < 3.5 \\ 0.5 & \text{if } 3.5 \leq d_z < 4.5 \\ 0.1 & \text{else.} \end{cases} \tag{10}$$

$$w_{\theta_y} = \frac{-2}{\pi} \cdot \theta_y + 1 \tag{11}$$

with d_z the depth distance of the spine base from the sensor and with θ_y the rotation of the body from the Kinect. The two following weights represent the confidence given in the sensor for the global body tracking trust (number of joints tracked over the total number of joints observed by the body tracking) and the time synchronisation to merge only closer frames (observations from different sensors must be close in time to be synchronized):

$$w_{skeleton} = \frac{N_{jt}}{N_T} \tag{12}$$

$$w_{time} = \frac{-0.75}{\Delta_{maxT}} \cdot \Delta T + 1 \tag{13}$$

with N_{jt} and N_T respectively the number of joints seen and the total number of joints of the camera as well as Δ_{maxT} and ΔT the maximum time between two accepted frames and the time between two frames, respectively. The number of joints seen are known thanks to the tracking state variable given by the body tracker to know if the joint is measured or estimated. These values are then used to weight the $\mathbf{R}_s$ matrix of each sensor s.

2.4 Calibration

Calibration of visual sensors is specially of importance when merging data coming from multiple sensors. In this study, solely the skeleton data coming from several Kinect sensors are merged. Skeleton data consist in a labelled point cloud representing the whole body at each time. The objective of the calibration process is to align these point clouds. Haralick et al. [18] proposed a method to estimate a transformation matrix from 2 paired sets of $j = 1, \ldots, N$ 3D points: $\mathrm{Z}_1 = [\mathrm{z}_{1,1}^T, \ldots, \mathrm{z}_{1,N}^T]^T$ and $\mathrm{Z}_2 = [\mathrm{z}_{2,1}^T, \ldots, \mathrm{z}_{2,N}^T]^T$, with $\dim(\mathrm{z}_{s,j}) = 3 \times 1$, $s = 1, 2$.

$$\text{Let } \mathbf{D} = \frac{1}{\sum_{j=1}^{N} w_j} \sum_{j=1}^{N} w_j (\mathbf{z}_{2,j} - \bar{\mathbf{z}_2})(\mathbf{z}_{1,j} - \bar{\mathbf{z}_1})^T \tag{14}$$

$$\text{with} \quad \bar{\mathbf{z}}_1 = \frac{\sum_{j=1}^{N} w_j \mathbf{z}_{1,j}}{\sum_{j=1}^{N} w_j} \text{ and } \bar{\mathbf{z}}_2 = \frac{\sum_{j=1}^{N} w_j \mathbf{z}_{2,j}}{\sum_{j=1}^{N} w_j}$$

$$\text{Then} \quad \{\mathbf{U}, \mathbf{S}, \mathbf{V}\} = \text{SVD}(\mathbf{D}) \tag{15}$$

$$\text{such that} \quad \mathbf{D} = \mathbf{U}.\mathbf{S}.\mathbf{V}^T$$

$$\mathbf{C} = \mathbf{U}.\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(U.V^T) \end{pmatrix}.\mathbf{V}^T \tag{16}$$

$$\mathbf{t} = \bar{\mathbf{z}}_1 - \mathbf{C}\bar{\mathbf{z}}_2 \tag{17}$$

where $\mathbf{Z}_{s=1,2}$ are the measurement vectors composed of N 3D JCPs for each sensors $s = 1, 2$ and w_k are weights of point k (0 or 1 based on the tracking state variable given by the body tracking SDK). The rotation matrix $\mathbf{C}$ and the translation vector $\mathbf{t}$ represent the positioning of the sensors $s = 2$ with respect to $s = 1$.

A singular value decomposition (Eq. 15) is used to solve this orthogonal Procrustes problem (Eq. 16). To identify the 6 parameters of the transformation matrix in a robust way, the information provided by the point clouds is cumulated on several frames. The identification continues until the updating of the parameters converges to constant values. Unlike Kinect 2 which has a tracking state variable to give a quality index on each joint, Kinect 3 does not have this type of confidence index. Thus, we propose to use a median type method to select points for correspondence. For each frame, distance between each joint to the spine base are calculated for both sensor and compared. The 50% of points with a fewer error distance between sensor are taken ($w_j = 1$).

Once all the transformation matrices between cameras are calculated, the fused body can be express in a common base. This method is useful when it is needed to estimate the absolute position of the human body in the environment.

It should be noted that, to facilitate the correspondence between the three systems (multiple Kinect 2 and Kinect 3 and reference systems) and avoid error due to calibration, the common base chosen for this work is the skeleton itself. The base is defined by the spine base as the origin, the hip right-hip left axis for the x-axis, the spine axis for the y-axis and the z-axis orthogonal to the other two axes. To insure an orthonormal base, the Gram-Schmidt method was used.

3 Experiments and results

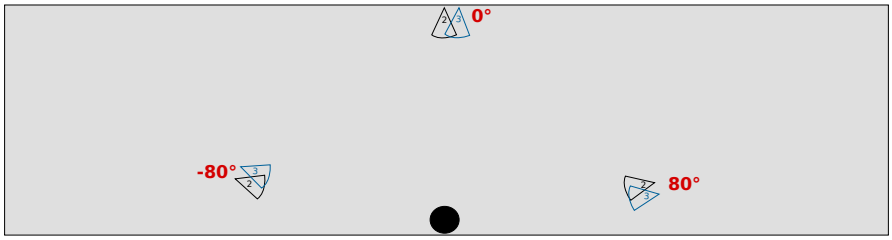
3.1 Experimental setup and protocol

Four tasks requiring whole body motions were considered in this study: squat with lateral extensions of arms, stepping in position with both legs, body tilt, and waving arms. Each task was performed with ten consecutive repetitions for each movement. All trials started and finished in a resting position: standing, arms along the body. These tasks were chosen to cover most degrees of freedom for arms, legs and trunk.

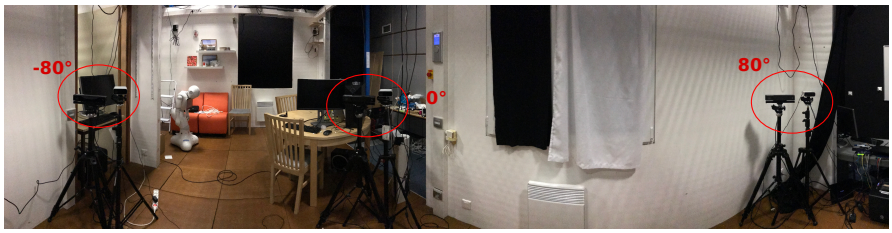
Whole-body motion was simultaneously collected by a reference stereophotogrammetric system consisting of 8 infrared cameras (Qualisys, 39 markers, 30 Hz) and by three Microsoft Kinect 2 (SDK, 30 Hz) and three Kinect Azure (NFOV, SDK v. 9.2, 30 Hz). The Kinect sensors were located in a triangle fashion with two Kinects in front (0°) of the participant, two on his right side (80°) and two on his left (-80°) side. Fig. 3 shows an overview of the experimental setup. Each Kinect was connected to a Windows 10 computer and all data were streamed on a ROS-based software to be elaborated on-line. Kinect 3 were synchronized with 3.5 mm audio cables in a daisy chain configuration. Due to interference between Kinect 3 and the reference system, the master cable has been split to trigger the reference system with the signal of the Kinect 3 master. Moreover, Kinect 3 body tracking SDK is badly disturbed by the motion capture retro-reflective markers. To minimise theses disturbances miniature markers of 2.5 mm of diameter were used.

3.2 Performance analysis

Kinect 2 has shown satisfactory results when the human body is facing the sensor without occlusion. Thus, to highlight the contribution of the proposed fusion algorithm, only the side sensors were considered (-80°and 80°), with heuristics and without heuristics, now referred to as Fusion H and Fusion !H, respectively. This was done for both versions of the Kinect sensor. The ability of the proposed fusion algorithm was then assessed by calculating the accuracy through the Euclidean distance between filtered JCPs and those obtained from the reference stereophotogrammetric system (see section II.A). The accuracy of the raw data gathered from the three sensor positions was also investigated for both Kinects. This was done to better understand the limitations of their specific body tracking software. Moreover, the mean and standard deviation of segment lengths were calculated on the raw and filtered data to attest to the consideration of anthropomorphic constraints by reducing the variability on segment lengths.



(a)



(b)

Fig. 3: (a) Schematic top view of the experimental setup: black circle is the human placement, black camera represents a Kinect 2 and blue camera a Kinect 3. (b) Panorama showing the location of the Kinect sensors facing the subject.

3.3 Results

Joint center position estimate accuracy assessment

First, the accuracy of the JCP estimates is evaluated by calculating the Euclidean distance between the JCPs obtained from the reference system and the JCPs estimated with the Kinect sensors. This accuracy is an absolute error in position. Table 2 reports mean and standard deviation distance for each joint. This is calculated for both types of Kinects when using raw data with different sensor's positions (0°, -80° and 80°) and when using merged data with and without heuristics.

As expected for Kinect 2, the Fusion H shows a better accuracy with a slightly shorter average distance (106 ± 41 mm) than Fusion !H (114 ± 36 mm). Fusion H improves on average of 14.6% for left (-80) and 28.9% for right (80) Kinect 2 results. When Fusion !H is used the improvement is of 8.0% and 23.4% for the left and right sensor, respectively. Globally, fusion algorithms tend to improve each joint accuracy and seem to be slightly equivalent. However the left elbow and wrist present a large difference between Fusion H and Fusion !H with both 121 mm for Fusion H, and 201 and 183 mm for Fusion !H, respectively. For these two joints, the results are particularly different between the right and left sensors, more than twice the mean and standard deviation. The difference on these two joints leads to the global difference between placements. It can be explain by the fact that the motion was not

exactly symmetrical and the self-occlusion affected more these joints for the right sensor than the other.

Unlike its previous version, Kinect 3 shows better results when using Fusion !H (84 ± 36 mm) than when using Fusion H (99 ± 206 mm). Both side Kinect 3 results are improved by Fusion !H with 21.4% and 16.8% for the left and right Kinect 3, respectively. The improvement of raw data by Fusion H are less good with 7.6% for the left sensor and only 2.2% for the right sensor. Finally, the accuracy of the feet position obtained when Fusion H algorithm is used decreases greatly. This can be observed with the very high standard deviations of the feet (1187 mm and 1259 mm).

Segment length variation

There is a consensus in the community stating that the body tracking of Kinect 2 is not anthropometry-based, resulting in segment length variations. Thus, numerous researchers tried to prevent segment length variations to improve human body tracking [19, 20]. The proposed Extended Kalman Filter used the process matrix covariance $\mathbf{Q}$ to force the length converge to a constant value. Fig. 4 represents the dispersion over all trials of the segment length. It shows that segment lengths of Kinect 2 have greater dispersion and more outliers than Kinect 3 for raw data. Moreover it illustrates the constraints on segment length achieved with the proposed Kalman filter. The segment length stability improves significantly with the Kalman filter. This is visible on the comparison between raw and filtered data for the right lower leg. There is also a strong reduction of the outliers. Table 3 confirmed these observations with an average standard deviation obtained for Kinect 2 of 19 mm instead of 8 mm or 9 mm when filtering.

It is interesting to notice that the Kalman filter also slightly improves the accuracy of the average absolute error of segment length estimate (37 mm for raw data instead of 35 mm for Fusion H). On the contrary, results obtained with the Kinect 3 show that Fusion !H degrades length estimate with an absolute error of 20 mm instead of 17 mm without filtering. Finally, Fusion H has terrible results on standard deviation, particularly for feet.

Additional results

These experimentations with both sensors raised some additional results that highlighted the different characteristics between Kinect 3 and Kinect 2. As mentionned in section 3.2 and confirmed by the results in Table 2, the central Kinect 2 shows much higher accuracy than those located on the side (82 ± 36 mm versus 124 ± 54 and 149 ± 71 mm). It is interesting to note that the Kinect 3 presents opposite results. The accuracy is lower for the central Kinect (121 ± 85 mm) than for the ones located on the side (107 ± 59 and 101 ± 53 mm).

Table 2: Joint center positions estimate obtained with the Kinect 2 and the Kinect 3 for raw (0°, -80° and 80°) and filtered data from the fusion algorithm with (Fusion H) and without heuristics (Fusion !H). Accuracy as absolute error is reported as Mean ± Standard Deviation [mm] over all the analyzed trials. Results in bold present the best accuracy for each joint between Fusion algorithms and raw data (except the facing camera (0°)).

Joint name	Kinect 2					Kinect 3				
	0°	-80°	80°	Fusion H	Fusion !H	0°	-80°	80°	Fusion H	Fusion !H
Neck	63 ± 29	111 ± 27	106 ± 33	101 ± 17	**97 ± 19**	67 ± 34	52 ± 29	42 ± 26	59 ± 256	**37 ± 18**
Shoulder R	42 ± 28	69 ± 31	59 ± 36	43 ± 25	**37 ± 23**	55 ± 36	51 ± 32	42 ± 24	38 ± 31	**34 ± 14**
Elbow R	54 ± 30	135 ± 107	124 ± 55	107 ± 42	**103 ± 37**	69 ± 54	63 ± 55	63 ± 38	49 ± 22	**40 ± 19**
Wrist R	53 ± 32	186 ± 162	132 ± 87	110 ± 53	**102 ± 49**	123 ± 125	124 ± 146	104 ± 93	80 ± 54	**61 ± 40**
Shoulder L	43 ± 29	63 ± 39	93 ± 57	**62 ± 24**	76 ± 38	65 ± 29	47 ± 30	48 ± 26	52 ± 36	**38 ± 17**
Elbow L	55 ± 28	**94 ± 45**	270 ± 165	121 ± 70	201 ± 54	67 ± 43	60 ± 36	91 ± 79	65 ± 25	**50 ± 16**
Wrist L	73 ± 57	115 ± 76	315 ± 176	**121 ± 85**	183 ± 48	116 ± 108	105 ± 104	185 ± 149	127 ± 60	**87 ± 31**
Hip R	33 ± 3	35 ± 7	34 ± 8	22 ± 6	**18 ± 6**	23 ± 2	20 ± 1	**19 ± 2**	25 ± 8	23 ± 3
Knee R	119 ± 35	122 ± 41	143 ± 53	115 ± 52	**111 ± 48**	155 ± 121	126 ± 51	**101 ± 51**	122 ± 81	110 ± 60
Ankle R	125 ± 56	176 ± 58	214 ± 94	**148 ± 32**	150 ± 30	193 ± 142	163 ± 72	131 ± 66	127 ± 82	**122 ± 57**
Foot R	155 ± 50	239 ± 63	217 ± 80	190±42	**185 ± 33**	230 ± 166	203 ± 91	168 ± 86	178 ± 1187	**142 ± 75**
Hip L	36 ± 2	41 ± 10	38 ± 8	22 ± 7	**17 ± 6**	32 ± 3	28 ± 2	**27 ± 2**	30 ± 9	**27 ± 2**
Knee L	124 ± 40	149 ± 43	141 ± 77	122 ± 54	**121 ± 56**	155 ± 138	125 ± 71	**109 ± 56**	124 ± 80	112 ± 62
Ankle L	131 ± 59	185 ± 58	178 ± 80	**162 ± 46**	166 ± 44	215 ± 164	183 ± 89	156 ± 60	159 ± 80	**154 ± 69**
Foot L	148 ± 57	193 ± 48	244 ± 78	**191 ± 54**	199 ± 45	249 ± 169	223 ± 108	197 ± 74	220 ± 1259	**183 ± 80**
Total	82 ± 36	124 ± 54	149 ± 71	**106 ± 41**	114 ± 36	121 ± 85	107 ± 59	101 ± 53	99 ± 206	**84 ± 36**

Table 3: Segment length estimation obtained by Kinect 2 and Kinect 3 for raw data (independently of the placement) and filtered data from fusion algorithm with (Fusion H) and without heuristics (Fusion !H) comparing to segment length reference. Accuracy as absolute error of segment length of interest has been reported as Mean ± Standard Deviation [mm] over all the analyzed trials.

Joint name	Kinect 2			Kinect 3		
	raw	Fusion H	Fusion !H	raw	Fusion H	Fusion !H
Clavicle R	13 ± 15	9 ± 4	**8 ± 4**	5 ± 4	10 ± 36	**3 ± 2**
Upper Arm R	**33 ± 18**	53 ± 8	58 ± 9	**9 ± 7**	19 ± 25	19 ± 5
Arm R	14 ± 12	**7 ± 6**	9 ± 7	**8 ± 6**	20 ± 24	23 ± 8
Clavicle L	16 ± 19	9 ± 4	**8 ± 4**	8 ± 5	10 ± 35	**3 ± 2**
Upper Arm L	**31 ± 16**	52 ± 8	57 ± 9	**10 ± 8**	18 ± 25	18 ± 5
Arm L	14 ± 14	**7 ± 6**	9 ± 7	**9 ± 7**	20 ± 24	23 ± 8
Hip R	18 ± 10	4 ± 2	**3 ± 2**	**21 ± 3**	40 ± 73	23 ± 1
Upper Leg R	85 ± 35	**83 ± 11**	87 ± 12	**42 ± 15**	66 ± 67	57 ± 5
Lower Leg R	41 ± 34	**30 ± 14**	31 ± 11	22 ± 12	31 ± 84	**9 ± 3**
Foot R	**53 ± 13**	59 ± 14	62 ± 10	**5 ± 4**	100 ± 4275	9 ± 4
Hip L	25 ± 10	4 ± 2	**3 ± 2**	30 ± 3	40 ± 73	**23 ± 1**
Upper Leg L	94 ± 31	**82 ± 11**	86 ± 12	**40 ± 15**	66 ± 67	56 ± 5
Lower Leg L	**24 ± 22**	29 ± 14	30 ± 10	20 ± 12	32 ± 84	**10 ± 3**
Foot L	**53 ± 14**	57 ± 9	62 ± 9	15 ± 7	108 ± 4650	**7 ± 4**
Total	37 ± 19	**35 ± 8**	37 ± 9	**17 ± 8**	41 ± 691	20 ± 4

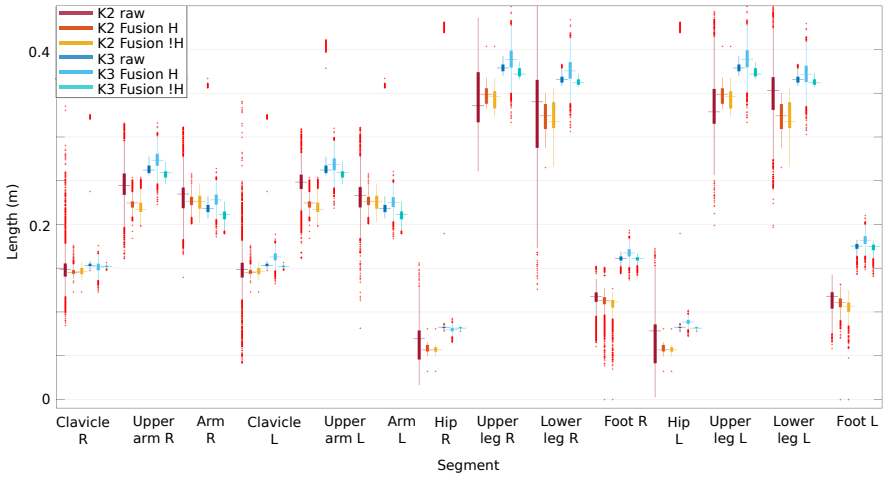


Fig. 4: Boxplot representing the dispersion of length on all trials for Kinect 2 and Kinect 3 with raw data and filtered data. Raw Kinect 2 is dark red, Fusion H Kinect 2 is orange, Fusion !H Kinect 2 is yellow, raw Kinect 3 is blue, Fusion H Kinect 3 is light blue and Fusion !H Kinect 3 teal.

4 Discussion and limits

The proposed fusion algorithms show improved accuracy on both Kinect 2 and Kinect 3. The differences due to the choice of heuristics were expected since the two Kinects do not present the same technical characteristics, e.g. the field of view and the depth resolution. Moreover, Kinect 2 was designed to view people facing the sensor for video games, but Kinect 3 is presented as an industrial sensor with a body tracking algorithm designed for multiple purposes, including back and front recognition. Several other differences as directly-included time synchronization for Kinect 3 and the absence of a tracking state variable for each joint makes at least two of the heuristics unnecessary. It can be concluded that the proposed algorithm improves body tracking but can still be enhanced by using sensor specific heuristics.

As expected, thanks to the use of the Kalman filter and anthropomorphic constraints, the length estimates vary less, although their estimation may be less accurate. The proposed approach is really interesting for Kinect 2 but seems to be less useful for Kinect 3. Unlike Kinect 2, Kinect 3 promises to provide an anatomically consistent skeleton tracking.

Despite these improvements on Kinect 3, it does not appear to greatly improve the accuracy of skeletal observation. The results obtained with the central Kinect 3 are quite surprising given that they are particularly poor compared to the ones obtained with the side sensors. However, even if the body tracker improves, the fusion will still be necessary to counter the effect of occlusion. These experiments gave additional results on the new Kinect 3 such as interference with stereophotogrammetric systems and reflective markers. The use of miniature markers of 2.5 mm of diameter reduces the effect of interference with the Kinect 3. In addition, the synchronization between the sensors constrains the Qualisys system to work at 30 Hz instead of 300 Hz. Naembadi et al. studied this question with Kinect 2, but Kinect 3 seems to be much more sensitive to interference [21]. Another point to highlight is the computational cost for running Kinect 3. It requires a good GPU (NVIDIA GEFORCE GTX 1070 or better) to work at 30 Hz. These results could question the portability of the sensor for embedded systems. It is interesting to notice that, if the portability of the system is critical and that the visual system as depth camera is not adapted to the application (i.e. the person had to be observed when moving in large area), other approach can be used [22].

5 Conclusion

This paper has presented a fusion algorithm based on an Extended Kalman Filter for merging skeleton data obtained from multiple depth cameras. The fusion algorithm was tested with the new Kinect 3 and with the Kinect 2 with regard to a reference provided by a stereophotogrammetric system. The results showed that this algorithm is robust to a change of body tracker software and overall improves joint center positions estimate. When the proposed fusion algorithm is used, the JCP estimate

accuracy improved by at least 14.6 %. The results prove that with or without fusion heuristics, the algorithm improves accuracy for each type of sensor. As shown for Kinect 2, a good definition of the heuristics allows even better results. However, the heuristics chosen to improve the fusion algorithm of multiple Kinect 2 are not suitable when using Kinect 3 sensors. This is due to differences between both sensors such as time synchronization, the lack of tracking confidence index number in Kinect 3 or to the fact that Kinect 3 is able to recognize back-front positioning. In addition to this last point, which is very important for usability, the improvement of segment length estimates is the main improvement obtained when the proposed algorithm is applied to Kinect 3 rather than Kinect 2. These two improvements open the door to a better observation of humans in their environment with the possibility of observing them at 360°.

Acknowledgement

This work has been partly funded by the CPER IT2MP (Contrat Plan État Région, Innovations Technologiques, Modélisation Médecine Personnalisée) and FEDER (Fonds européen de développement régional).

References

1. Microsoft, "Kinect - Windows app development." [Online]. Available: https://developer.microsoft.com/en-us/windows/kinect
2. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *arXiv:1812.08008 [cs]*, Dec. 2018.
3. S. Moon, Y. Park, D. W. Ko, and I. H. Suh, "Multiple Kinect Sensor Fusion for Human Skeleton Tracking Using Kalman Filtering," *International Journal of Advanced Robotic Systems*, vol. 13, no. 2, p. 65, Mar. 2016.
4. S. Feng and R. Murray-Smith, "Fusing Kinect sensor and inertial sensors with multi-rate Kalman filter," in *IET Conference on Data Fusion Target Tracking 2014: Algorithms and Applications (DF TT 2014)*, Apr. 2014, pp. 1–8.
5. J. Colombel, V. Bonnet, D. Daney, R. Dumas, A. Seilles, and F. Charpillet, "Physically Consistent Whole-Body Kinematics Assessment Based on an RGB-D Sensor. Application to Simple Rehabilitation Exercises," *Sensors*, vol. 20, no. 10, p. 2848, Jan. 2020, number: 10 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/20/10/2848
6. Q. Wang, G. Kurillo, F. Ofli, and R. Bajcsy, "Evaluation of Pose Tracking Accuracy in the First and Second Generations of Microsoft Kinect," in *2015 International Conference on Healthcare Informatics*, Oct. 2015, pp. 380–389.
7. D. Pagliari and L. Pinto, "Calibration of Kinect for Xbox One and Comparison between the Two Generations of Microsoft Sensors," *Sensors*, vol. 15, no. 11, pp. 27 569–27 589, Nov. 2015.
8. R. B. Davis, S. Õunpuu, D. Tyburski, and J. R. Gage, "A gait analysis data collection and reduction technique," *Human Movement Science*, vol. 10, no. 5, pp. 575–587, Oct. 1991.
9. A. L. Bell, R. A. Brand, and D. R. Pedersen, "Prediction of hip joint centre location from external landmarks," *Human Movement Science*, vol. 8, no. 1, pp. 3–16, Feb. 1989.
10. G. Rab, K. Petuskey, and A. Bagley, "A method for determination of upper extremity kinematics," *Gait & Posture*, vol. 15, no. 2, pp. 113–119, Apr. 2002.
11. G. Wu, S. Siegler, P. Allard, C. Kirtley, A. Leardini, D. Rosenbaum, M. Whittle, D. D. D'Lima, L. Cristofolini, H. Witte, O. Schmid, I. Stokes, and Standardization and Terminology Committee of the International Society of Biomechanics, "ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion–part I: ankle, hip, and spine. International Society of Biomechanics," *J Biomech*, vol. 35, no. 4, pp. 543–548, Apr. 2002.
12. G. Wu, F. C. T. van der Helm, H. E. J. (DirkJan) Veeger, M. Makhsous, P. Van Roy, C. Anglin, J. Nagels, A. R. Karduna, K. McQuade, X. Wang, F. W. Werner, and B. Buchholz, "ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—Part II: shoulder, elbow, wrist and hand," *Journal of Biomechanics*, vol. 38, no. 5, pp. 981–992, May 2005.
13. W. Khalil and D. Creusot, "SYMORO+: A system for the symbolic modelling of robots," *Robotica*, vol. 15, no. 2, pp. 153–161, Mar. 1997.
14. J. A. Roecker and C. D. McGillem, "Comparison of two-sensor tracking methods based on state vector fusion and measurement fusion," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 24, no. 4, pp. 447–449, Jul. 1988.
15. Q. Gan and C. J. Harris, "Comparison of two measurement fusion methods for Kalman-filter-based multisensor data fusion," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 1, pp. 273–279, Jan. 2001.
16. K.-Y. Yeung, T. H. Kwok, and C. Wang, "Improved Skeleton Tracking by Duplex Kinects: A Practical Approach for Real-Time Applications," *Journal of Computing and Information Science in Engineering*, vol. 13, p. 041007, Oct. 2013.
17. M. M. Otto, P. Agethen, F. Geiselhart, M. Rietzler, F. Gaisbauer, and E. Rukzio, "Presenting a Holistic Framework for Scalable, Marker-less Motion Capturing: Skeletal Tracking Performance Analysis, Sensor Fusion Algorithms and Usage in Automotive Industry," *Journal of Virtual Reality and Broadcasting*, vol. 13(2016), no. 3, Feb. 2017.

18. R. M. Haralick, H. Joo, C. Lee, X. Zhuang, V. G. Vaidya, and M. B. Kim, "Pose estimation from corresponding point data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 6, pp. 1426–1446, Nov. 1989.
19. S. R. Tripathy, K. Chakravarty, and A. Sinha, "Constrained Particle Filter for Improving Kinect Based Measurements," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep. 2018, pp. 306–310.
20. S. Skals, K. P. Rasmussen, K. M. Bendtsen, J. Yang, and M. S. Andersen, "A musculoskeletal model driven by dual Microsoft Kinect Sensor data," *Multibody Syst Dyn*, vol. 41, no. 4, pp. 297–316, Dec. 2017.
21. M. Naeemabadi, B. Dinesen, O. K. Andersen, and J. Hansen, "Investigating the impact of a motion capture system on Microsoft Kinect v2 recordings: A caution for using the technologies together," *PLOS ONE*, vol. 13, no. 9, p. e0204052, Sep. 2018.
22. O. Banos, A. Calatroni, M. Damas, H. Pomares, I. Rojas, H. Sagha, J. del R. Millán, G. Troster, R. Chavarriaga, and D. Roggen, "Kinect=IMU? Learning MIMO Signal Mappings to Automatically Translate Activity Recognition Systems across Sensor Modalities," in *2012 16th International Symposium on Wearable Computers*, Jun. 2012, pp. 92–99, iSSN: 2376-8541.