IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Robust Upper Limb Kinematic Reconstruction Using a RGB-D Camera

Salvatore Maria Li Gioi, *Graduate Student Member, IEEE*, Giuseppe Loianno, *Member, IEEE*, and Francesca Cordella, *Member, IEEE*

***Abstract*—In this letter, we propose a new approach for human motion reconstruction based on Gaussian Mixture Probability Hypothesis Density (GM-PHD) Filter applied to human joint positions extracted from RGB-D camera (e.g. Kinect). Existing inference approaches require a proper association between measurements and joints, which cannot be maintained in case of the multi-tracking occlusion problem. The proposed GM-PHD recursively estimates the number and states of each group of targets. Furthermore, we embed kinematic constraints in the inference process to guarantee robustness to occlusions. We evaluate the accuracy of both the proposed approach and the default one obtained through a Kinect device by comparing them with a motion analysis system (i.e. Vicon optoelectronic system) even in presence of occlusions of one or more body joints. Experimental results show that the filter outperforms the solution provided by the baseline commercial solution approach available in the Kinect device by reducing the hand position and elbow flexion error of 55.8% and 36.3%, respectively. In addition, to evaluate the applicability of the approach in real-world applications, we employ it in a drone gesture-based context to remotely control a drone. The user is able to move the drone in a target position with a 100% success rate.**

***Index Terms*—Drone, GM-PHD, kinect one.**

I. Introduction

MOTION analysis is a domain of biomechanics that aims at collecting quantitative information about the mechanics of the musculoskeletal system during the execution of a motor task by means of certain measuring instruments, in order to describe and characterise the motor gesture for evaluative, diagnostic and improvement purposes [1]. Because of its excellent precision in reconstructing the subject's kinematic characteristics, marker-based optoelectronic systems are regarded as the gold standard in this field [2]. Despite their excellent efficacy, these devices have some limitations, including long experiment preparation times, structured acquisition locations and high costs [3]. As a result, there is a steady increase in the development of markerless approaches using RGB-D cameras [4], [5]. The Kinect One is a popular depth camera that incorporates skeleton tracking algorithms to retrieve the 3D positions of each body joint through Time of Flight (ToF), i.e. the distance is measured by calculating the phase shift distance of the modulated infrared light. However, it has lower levels of accuracy and precision than optoelectronic systems [6], and it is also prone to occlusion and misidentification of body joints, hindering image segmentation. In this letter, we introduce an innovative motion analysis technique Fig. 1 employing a GM-PHD filter, aimed at increasing the performance, applicability, and occlusion resistance of the Microsoft Kinect One RGB-D camera.

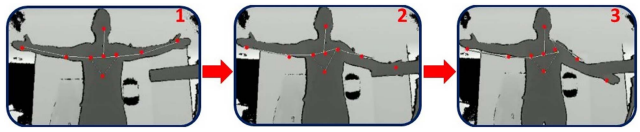

Fig. 1. GM-PHD filter (red dots) applied on the skeleton (white lines) for tracking even in presence of occlusions.

Several studies analyse the Kinect performances. In [3], the accuracy and repeatability of the Kinect is evaluated through a comparison with the Vicon optoelectronic system. Each healthy subject performs point-to-point and hand exploration movements, executed without occlusions. Using angular distance as a comparison metric, the obtained results show that the Kinect One algorithm tracks the majority of the upper limb degrees of freedom with an error of less than 10° for both movements. In [7], the accuracy of the reconstruction of joint position and limb lengths acquired from the Kinect V1 and V2 is compared with ground truth data from PhaseSpace's Impulse X2 optoelectronic motion capture device. The results reveal that the Kinect V2 reconstruction capabilities are more accurate than its predecessor, but it shows reduced performances compared to the Impulse X2 system with joint positions offsets ranging from 50 to 100 mm and standard deviations ranging from 10 to 50 mm. These results indicate that the Kinect One's kinematic reconstruction is not perfect, with flaws in accuracy and limb length variation during task performance. To ensure effective tracking of body joints Deep Learning (DL) and KF-based approaches have been proposed. The comparison of these two strategies [8] revealed several advantages of using KF with

Manuscript received 18 September 2023; accepted 18 February 2024. Date of publication 5 March 2024; date of current version 12 March 2024. This letter was recommended for publication by Associate Editor D. Sidib and Editor C. Cadena Lerma upon evaluation of the reviewers' comments. This work was supported by NSF CAREER Award under Grant 2145277, in part by DARPA YFA under Grant D22AP00156-00, in part by NSF CPS under Grant CNS-2121391, in part by Qualcomm Research, Nokia, and in part by NYU Wireless. *(Corresponding author: Salvatore Maria Li Gioi.)*

Salvatore Maria Li Gioi and Francesca Cordella are with the Research Unit of Advanced Robotics and Human-Centred Technologies, Department of Engineering, Università Campus Bio-Medico di Roma, 00128 Rome, Italy (e-mail: s.ligioi@unicampus.it; f.cordella@unicampus.it).

Giuseppe Loianno is with the Tandon School of Engineering, New York University, Brooklyn, NY 11201 USA (e-mail: loiannog@nyu.edu).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3373236, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3373236

respect to DL, such as the explainability, the reduced time needed for the training and the prediction phase, since the model parameters are specified and do not need to be learnt, less reconstruction error. Among the KF approaches proposed in literature, in [9] a constraint on limb lengths is included into the KF. Rather than filtering each joint's coordinates individually, coupled joints are simultaneously filtered such that the distance remains constant. According to the experimental results, using the constraint reduced limb length variations by 92% for the Kinect One and 94% for its predecessor. Another application of the KF is in [10], where an Extended Kalman Filter (EKF) was used to fuse data from Inertial Measurement Units (IMUs) and Kinect and estimate noise based on the context (i.e. fast, slow, and normal movements) in which the motion task was performed. Parameter estimation determines the actual values of the variances for each condition. Sensors data are then fused using the EKF to estimate the subject's joint angles. The algorithm outperforms the individual sensors, namely IMU and Kinect, and can automatically estimate variances based on the context. However, the presence of occlusions prevents the correct association between the traces and the measurements, which is required for the filter currently in use. To improve the robustness of the Kinect to occlusions, several existing approaches use the KF to fuse the information of the Kinect with the ones given by the IMUs [11], [12]. However, the introduction of IMUs not only increases the complexity of the system but also implies the use of wearable devices that can hinder natural movement behaviours. Another way to overcome occlusion problems is to fuse data from multiple Kinect sensors. In [13], information from multiple Kinects were combined to improve tracking accuracy by using a weighted measurement fusion technique based on the KF framework. The study compared the performance obtained by the proposed approach, during the execution of various tasks (i.e. running, crossing arms and legs, sitting on the chair, and walking around), with the ones obtained with a single Kinect and the ones obtained by averaging five skeleton poses. The OptiTrack motion-capture system was used to provide a set of ground truth trajectories. The results indicate that the authors' method outperformed traditional single-sensor and simple averaging methods. Although this method shows promising results, it requires the use of multiple cameras. Therefore a higher level of environmental structuring is required, making it challenging to be applied outside of laboratory settings.

In order to manage the joint occlusion problem and to have a completely not wearable system, the kinematic reconstruction of the human body using the Kinect One can be traced back to a multi-tracking object problem. Therefore, in this letter, we propose to employ a Gaussian Mixture Probability Hypothesis Density (GM-PHD) filter [14] to recursively estimate the number and states of a group of targets given a set of observations and it does not require individual observations to be connected with any trace. In the literature [15], it has been demonstrated that the filter outperforms the standard KF and has less computing cost than the Joint Probabilistic Data Association Filter (JPDAF) controlling swarms of drones [16]. This letter presents multiple contributions. First, we propose a new motion analysis approach based on a GM-PHD filter to increase the performance, applicability, and occlusion resistance of the Microsoft Kinect One RGB-D camera. Second, inspired by on past research findings [9], we include a kinematic constraint in the filter to keep the distances between joints constant during task execution, a condition that the Kinect One's skeleton tracking algorithm does not obey. Third, we evaluate the filter's performance with respect to the ones of the gold standard Vicon system to emphasise the benefits obtained with the filter with respect to the Kinect default algorithm solution as the occlusion level grows. Finally, in order to validate the filter adaptability in real-time scenarios, we propose a gesture-based drone control strategy and validate it through reaching target position tasks, with and without occlusions. The use of unmanned aerial vehicles (UAVs) in everyday life has become increasingly noticeable and is attracting more users who may not have specialised skills in operating them. Therefore, several drone control approaches aiming at reducing the physical and cognitive effort of the users have been proposed in literature, including the gesture-based approach. It can facilitate the communication between the user and the drone being an intuitive and effective mean to recognize user intention [17].

II. Methodology

A. Kinect One Skeleton Tracking Algorithm

The upper body joints reconstructed by the camera are: torso, head, shoulders, midpoint between shoulders, elbows, and hands. Before starting the detection, the program gives the subject three seconds to assume the T-pose (which consists of arranging the upper limbs in such a way that the posture assumed the form of the letter T), so that, on the first frame, the subject is in a suitable position for measuring the distances between body joints. After this phase, the detection and tracking of the subject's joints start. For each joint, the position in the image plane and the coordinate along the z-axis were extracted, which require to move from the image plane to the 3D space.

B. Gaussian Mixture PHD Filter

The Gaussian Mixture Probability Hypothesis Density (GM-PHD) filter [14] is a Bayesian filter that estimates the states of multiple objects from noisy observations at discrete time intervals. Traditional KF assumes a fixed number of objects with known identities, which may not perform well in complex scenarios such as occlusions. In contrast, Random Finite Sets (RFS) methods, such as the GM-PHD filter, offer a probabilistic description of the targets' states. The filter describes the uncertainty associated with the target states as a Gaussian mixture model. Each component represents a potential target with weights indicating the probability of the corresponding target's existence. Additionally, the filter updates a set of hypotheses, including the targets' birth and death processes, based on the likelihood of the measurements. These features are beneficial in situations where objects are temporarily hidden due to occlusions, as demonstrated in [15]. Hence, in this case, the measurements $(\mathbf{z})$ and states $(\mathbf{x})$ are represented as RFS to improve accuracy. Density functions in target space (body

joints) are used to represent the states, and the GM-PHD depicts the first moment of distribution on the RFS. At the instant n, the state of the target i denoted $\mathbf{x}_n^i$ is characterized as a single value m_n^i given by the weighted sum of the Gaussian components

$$m_n^i\left(\mathbf{x}_n^i\right) = \sum_{t=1}^{T_n} w_n^t \mathcal{N}(\mathbf{x}_n^i; \boldsymbol{\mu}_n^t, \mathbf{P}_n^t), \tag{1}$$

where T_n is the number of joints. The Gaussian components $\mathcal{N}$ for state $\mathbf{x}_n^i$ include a weight w_n^t, a mean value $\boldsymbol{\mu}_n^t$, and a covariance $\mathbf{P}_n^t$. The Gaussian components are propagated along the prediction and update steps based on the states of the preceding instant and observations.

1) Prediction Phase: Each target state is described as a Random Finite Sets, which is

$$m_{n|n-1}^i(\mathbf{x}_n^i) = \sum_{t=1}^{T_k} w_{n|n-1}^t \mathcal{N}(\mathbf{x}_n^i; \boldsymbol{\mu}_{n|n-1}^t, \mathbf{P}_{n|n-1}^t) + \gamma(\mathbf{x}_n^i). \tag{2}$$

As in [18], an adaptive agent birth model $\gamma(\mathbf{x}_n^i)$ particular to the PHD filter is considered, where the new Gaussian components are described by mean $\boldsymbol{\mu}_\gamma^i$ and covariance $\mathbf{P}_\gamma^i$. Instead, ω_γ^i is the weight associated to the the new Gaussian components.

2) Update Phase: The update of each status is obtained by

$$\begin{aligned} m_n(\mathbf{x}_n^i) &= (1 - p_d) m_{n|n-1}(\mathbf{x}_n^i) \\ &\quad + \sum_{j=1}^{Z_n} \sum_{t=1}^{T_n} w_n^{j,t}(\mathbf{z}_n^j) \mathcal{N}(\mathbf{x}_n^i; \boldsymbol{\mu}_{n|n}^t, \mathbf{P}_{n|n}^t), \end{aligned} \tag{3}$$

where p_d is the detection probability. The weight, mean, covariance and Kalman gain are updated as in [15].

3) Filter Application: The GM-PHD filter takes as input the positions, extracted from the Kinect One, of the nine joints in the image plane as well as the coordinate along the z-axis in the three-dimensional space, taking the Kinect reference frame. In the prediction phase, it is assumed that the system could be described by a constant acceleration model (4)

$$\mathbf{A} = \begin{bmatrix} 1 & dt & \frac{1}{2}dt^2 & 0 & 0 & 0 \\ 0 & 1 & dt & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & dt & \frac{1}{2}dt^2 \\ 0 & 0 & 0 & 0 & 1 & dt \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} u \\ \dot{u} \\ \ddot{u} \\ v \\ \dot{v} \\ \ddot{v} \end{bmatrix}. \tag{4}$$

The state of each joint consists of its position (u, v), velocity $(\dot{u}, \dot{v})$ and acceleration $(\ddot{u}, \ddot{v})$ in the image plane. On the first iteration, it is initialised by setting the positions equal to the measurements returned by the Kinect, instead the velocities and accelerations are set to zero. The $\mathbf{A}$ matrix contains the terms required to make the prediction under the constant acceleration model. The value assumed by the variable dt is equal to the inverse of the Kinect sampling frequency, which is 1/30 s. The $\mathbf{R}$ and $\mathbf{Q}$ matrices have been defined by means of diagonal matrices, since it is assumed that there is independence between the various process states. After the prediction phase is completed, the update phase is carried out

$$\hat{\mathbf{x}}_{n,n} = \hat{\mathbf{x}}_{n,n-1} + \mathbf{K}_n(\mathbf{Z}_n - \mathbf{H}\hat{\mathbf{x}}_{n,n-1}), \tag{5}$$

where $\mathbf{K}_n$ denotes the Kalman Gain, $\mathbf{Z}_n$ the measurement vector and $\mathbf{H}$ the observation matrix

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}. \tag{6}$$

The $\mathbf{H}$ matrix has the same shape for each joint and its values are necessary in order to extract the new joint coordinates in the image plane. After obtaining the filtered data, the Pinhole equations are used (7). The new positions (x, y, z) in the three-dimensional space, expressed in Kinect frame, can be determined using knowledge of the coordinate along the z-axis

$$x = \frac{(u - c_x)z}{f_x}, \quad y = \frac{(v + c_y)z}{f_y}. \tag{7}$$

In order to use (7), it is necessary to know the camera's intrinsic parameters, which are focal distance along the x-axis (f_x), focal distance along the y-axis (f_y), and the optical center coordinates (c_x, c_y). These parameters were determined using the NiTE library, which provides the 3D position and the image plane position for each joint. Knowing the positions of at least two joints in the two spaces, it is feasible to calculate the Kinect One's intrinsic parameters using a system consisting of four unknowns (intrinsic parameters) and four equations (Pinhole equations).

C. Kinematic Constrain Application

When the Kinect data is analysed, we can observe that the distance between the joints does not remain constant with and without occlusions. Because the limb lengths do not vary, a kinematic constraint is applied to the filter, allowing the subject's anatomical constraints to be respected during task execution. The anthropometric lengths, in pixels, of the subject's upper limbs are extracted at the first instant, implying that the subject must maintain a proper posture for the correct measurement of joint distances (e.g., T-Pose). Subsequently, once the prediction phase is completed, the constraint is applied. Knowing the filtered positions of the limb's proximal and distal joints, the new position of the latter can be calculated. Specifically, the new position of the hand (P_3), obtained after applying the kinematic constraint to the hand position (P_2) resulting from the prediction step, is dependent by the angular coefficient (m) and the intercept (q) of the line linking the elbow joint (P_1) and P_2. To obtain the new coordinates (u_3, v_3) of P_3, the following system should be resolved to ensure that the new point lies on the same line connecting P_1 and P_2, and is at a distance from the P_1 equal to the one measured in the first frame

$$\begin{cases} v_3 = mu_3 + q \\ l^2 = (u_3 - u_1)^2 + (v_3 - v_1)^2 \end{cases} \tag{8}$$

where l is the measured distance between P_1 and P_2 at the first frame, (u_1, v_1) and (u_3, v_3) are the image plane coordinates of the P_1 and P_2 joints, respectively. During the update process, the new position will be compared to the measurement extracted from the Kinect One.

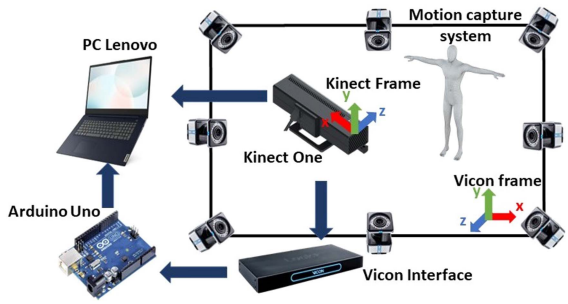

Fig. 2. Acquisition system.

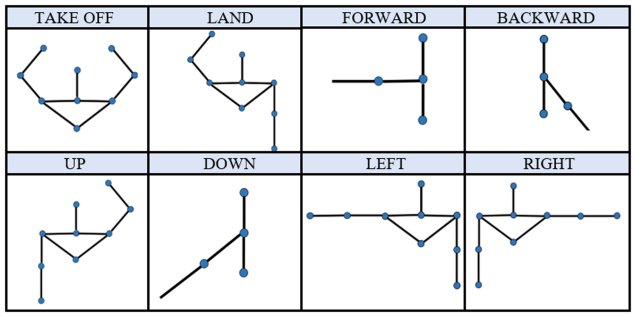

Fig. 3. Gestures associated to commands.

III. Experimental Setup

The experimental setup used to conduct the experiments consists of the RGB-D Kinect One camera, the Ryze Dji Tello drone, a Lenovo PC and the Vicon optoelectronic motion capture system Fig. 2. The framework was developed in Ros Noetic. According to Microsoft's recommendations [19], in our motion tracking settings, we place the device's height between 0.6 and 2 m, and the subject is asked to stand between 1.4 and 2 meters away. The upper body joints reconstructed by the camera are: torso, head, shoulders, midpoint between shoulders, elbows, and hands. As Ubuntu does not have an official Kinect SDK, we used the NiTE library, which is part of the OpenNI framework. This library supports both Kinect V2 and Kinect One. Therefore, there is no discernible difference between using Kinect One or Kinect V2 in Ubuntu as they are supported by the same unofficial library. The NiTE library provides, for each joint, both the position in the image plane and the position in three-dimensional space with respect to the Kinect frame Fig. 2.

IV. Experimental Protocol

The experimental protocol consists of two experiments to validate the filter and two to understand the applicability of remote control of the drone. In accordance with ethical guidelines and regulations governing research involving human subjects, this study has been determined to be exempt from obtaining approval from a relevant review board since the research poses minimal risk to the participants primarily involving observations and not involving any invasive or potentially harmful procedures.

Drone control algorithm: The proposed algorithm uses the filter output to determine the command chosen by the subject. Fig. 3 depicts the commands assigned to each gesture. The algorithm uses the three-dimensional positions of the joints, in the Kinect frame, as input and analyses their relative positions to determine the subject's pose.

1) First Experimental Test

The aim is to compare the Kinect One and the GM-PHD with constrain joints position and angle results to those obtained with a gold standard system (i.e., Vicon Vero 2.2 optoelectronic system). Four healthy subjects, aged between 26 ± 2, participated in the experiment. They were placed in the Vicon's acquisition volume and in front of the Kinect, and asked to abduct the right shoulder three times with different levels of occlusion of different portions of the right arm: firstly the right hand was occluded, then the right elbow and hand Fig. 4.

Fig. 4. Elbow-hand occlusion (red).

Fig. 5. Marker placement according to the ULM.

In this test, an Arduino Uno electronic board is used to synchronize the Vicon and Kinect acquisitions, operating as a bridge for the two systems' communication: it connects with the Vicon via an analogue pin and with the computer via the USB port Fig. 2. To reconstruct the joint positions through the Vicon system, 15 markers are placed on the upper body of the subjects Fig. 5 according to the Upper Limb Model (ULM) for Vicon Nexus software. This protocol is chosen since it computes the position of the same joints retrieved by the Kinect One. To compare the positions from the Vicon with those from the Kinect, a common reference system was defined: three markers were placed on the Kinect to define the Kinect frame. The positions of the joints were expressed in the Kinect frame by using rotation matrices Fig. 2.

To validate the constraint imposed, the standard deviations of the shoulder-elbow and elbow-hand lengths of the limb subjected to occlusion were compared in 3D space for each subject. Moreover, the filter performance were assessed by comparing

Fig. 6. Setup of the second experiment: the drone is in the red box and the target position is highlight with a red X.

the elbow flexion and hand position error of the filter with the ones given by the Kinect One.

2) Second Experimental Test

To assess the efficacy of the kinematic constrain in keeping limb lengths constant during movement, the same acquisitions from the first experiment were processed with a without the kinematic constraint. To validate the imposed constraint, we compared the standard deviations in 3D space of the shoulder-elbow and elbow-hand lengths of the occluded limb for each subject.

3) Third Experimental Test

Using the gesture-based approach, we evaluate the efficacy of the control strategy in absence of occlusions. One healthy subject is asked to control with gestures the drone to reach a predefined target point, outlined with a red X in Fig. 6. Only one subject is involved in this second and in the third experimental tests because the purpose is only to understand the applicability in a real-time scenario of the developed joint position reconstruction strategy and not to evaluate the drone controllability by different subjects. The success rate and the average completion time were used as metrics to evaluate the gesture-based approach.

4) Fourth Experimental Test

In this test, the efficacy of the joint position reconstruction strategy is evaluated during a real-time control of a drone in presence of the occlusion of the right hand. One healthy subject is tasked to control the drone, for three times, by means of the gesture-based approach defined in Section IV, to lift off and advance to the right. The success rate is evaluated to assess the gesture-based approach.

V. Results and Discussions

1) First Experimental Test: Analysing Table I, it is feasible to discover that employing the kinematic constraint. For hand occlusion, with the filter a reduction respect to the raw data in the standard deviations of the shoulder-elbow and elbow-hand lengths of 44.4% and 59.1%, respectively, is found. For the complex occlusion of the elbow-hand, a reduction of the standard deviations of the shoulder-elbow and elbow-hand lengths by 41.5% and 69.7%, respectively, is obtained. The qualitative and quantitative results achieved with the GM-PHD filter are presented below. In the first scenario (hand occlusion), it is clear from Fig. 7(a) that the filter output (red dots) better identifies location than the Kinect skeleton tracking (blue lines).

TABLE I
Mean and Standard Deviations for Each Subject

		Hand Occlusion		Elbow-Hand Occlusion	
	#Subj.	S.-E. std[mm]	E.-H. std[mm]	S.-E. std[mm]	E.-H. std[mm]
Kinect	1	20.5	16.4	35.4	37.9
Filter		8.3	9.6	15.0	17.0
Kinect	2	12.6	19.5	32.0	38.5
Filter		7.5	7.3	23.00	8.9
Kinect	3	24.1	23.6	29.7	49.9
Filter		13.8	9.4	20.7	12.7
Kinect	4	23.7	39.1	38.0	46.7
Filter		15.5	8.1	19.2	11.7

S.-E. stands for "Shoulder-Elbow" and E.-H for "Elbow-Hand".

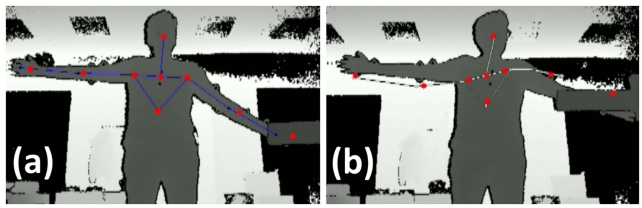

Fig. 7. (a) Hand occlusion. (b) Elbow−Hand occlusion.

TABLE II
Hand Occlusion, Mean Errors

#Subj.	Kinect Elbow Flex. Error[°]	Filter Elbow Flex. Error[°]	Kinect Hand Pos. Error[mm]	Filter Hand Pos. Error[mm]
1	6.1 ± 7.7	5.9 ± 7.2	50.7 ± 29.4	27.7 ± 18.3
2	2.3 ± 4.3	3.5 ± 5.4	14.8 ± 24.9	11.8 ± 11.6
3	1.7 ± 9.7	3.1 ± 6.7	23.3 ± 41.9	6.9 ± 17.0
4	6.7 ± 19.3	1.4 ± 4.1	32.2 ±57.1	21.2 ± 23.8

Quantitative results were obtained by comparing the hand position error and elbow flexion angle of the occluded limb with Vicon data for each subject Table II.

Concerning the hand position, it was feasible to see that, after removing the offset due to the incorrect measurement of the limb length by the Kinect, the filter reduced the average error by 43.7% and the standard deviation by 52.3%. In contrast, because the elbow joint is not involved in the occlusion, the filter performs the same as the Kinect in terms of elbow flexion angle. The performance of the filter is even more evident in experimental trials with both elbow and hand occlusion, because the Kinect completely misses the limb subject to occlusion, whilst the filter does not Fig. 7(b). By increasing the degree of occlusion, Table III shows that the hand position error and standard deviations in the filtered data are reduced by 55.8% and 57.4%, respectively. In this example, improvements in elbow flexion angle were discovered, with a reduction in mean and standard deviation of 36.3% and 70.2%, respectively.

TABLE III
OCCLUSION OF THE ELBOW AND OF THE HAND, MEAN ELBOW FLEXION AND HAND POSITION ERROR

#Subj.	Kinect Elbow Flex. Error[°]	Filter Elbow Flex. Error[°]	Kinect Hand Pos. Error[mm]	Filter Hand Pos. Error[mm]
1	8.7 ± 16.2	7.4 ± 8.8	57.3 ± 57.6	33.5 ± 35.8
2	10.1 ± 44.3	9.6 ± 11.5	60.5 ± 82.9	25.3 ± 25.7
3	9.6 ± 42.8	3.5 ± 6.8	50.1 ± 88.8	16.1 ± 26.0
4	22.3±63.5	8.5 ± 14.6	64.2 ± 84.1	28.4 ± 40.4

TABLE IV
PERCENTAGES OF STANDARD DEVIATION REDUCTION

	Hand Occlusion		Elbow-Hand Occlusion	
#Subj.	S.-E.[%]	E.-H.[%]	S.-E.[%]	E.-H.[%]
1	36.3	56.7	41.5	42.0
2	45.5	65.0	29.3	53.4
3	31.0	63.5	32.5	49.0
4	28.4	26.5	26.8	46.6

Fig. 8. First trial result.

As seen from the results,[1] even though the occlusion, the filter maximum hand error position is still lower than the raw maximum joint error position [7]. Moreover, this new motion analysis approach does not need IMUs, so it is suitable for home rehabilitation and remote drone control tasks.

2) Second Experimental Test: To emphasise the behaviour of the kinematic constraint in the filter, the mean percentage of the standard deviation reduction of the limb lengths for each subject was calculated Table IV. As evident, the application of the kinematic filter allowed to reduce the limb length standard deviation almost of the 30% in each subject.

3) Third Experimental Test: In the first experimental test of the drone control, a success rate of 100% was achieved with an average time of 94 s.

4) Fourth Experimental Test: Even though the obstruction induced by the obstacle is present, the predicted results show a success rate of 100% in this testing case. The Kinect completely loses hand detection in the first trial, which is visible since the Kinect One's skeleton tracking is not present (blue lines). Conversely, the filter can determine the hand's location (red dot) Fig. 8. The filter is always able to predict the hand position for the second and third repetitions Fig. 9, but significant reconstruction errors are found in the elbow tracking. Concerning the hand position, because the Kinect error is significant, the filter is not able to associate the new estimate with one of the measurements and retained the position calculated at the previous instant in the output.

[1]The multimedia material: https://youtu.be/4UAGFdxn9mw

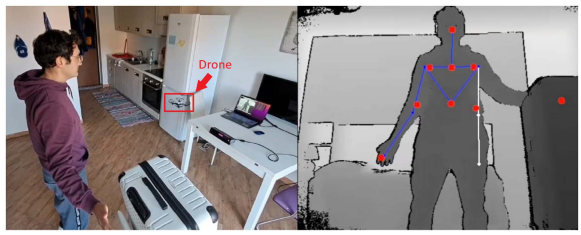

Fig. 9. Second trial result.

5) Statistical Analysis: Once the data is obtained, and because the distribution of the data is not Gaussian, a Wilcoxon test is used to verify that there is a statistically significant difference between the outputs of the Kinect One and those of the GM-PHD with kinematic constraint. Particularly for the occlusion of the hand, p-values of less than 0.05% (p-value $= 0.0039$) are observed for the standard deviations of limb lengths in the 3D space, as well as the hand position error. However, in the elbow flexion angle, no significant difference was identified (p-value > 0.05). This analysis was also applied in the case of elbow-hand occlusion, yielding p-values of less than 0.05%. This significant difference, in this case, was also obtained for the elbow flexion angle as well, because, unlike the previous scenario, elbow occlusion is present. This result demonstrated that, in the presence of occlusions, the filter performance is better than the ones of the Kinect One.

VI. CONCLUSION AND FUTURE WORK

In this letter, we proposed a new inference approach for human motion reconstruction tailored to upper-limb kinematic reconstruction. We show how it can robustify the Kinect One detections especially with respect to occlusions. Finally, we also show how to use these outputs of filter for precise gesture control of a drone. The filter achieves the desired results, according to the analysis of the experimental results; in presence of occlusions, the filter reduced the elbow flexion angle average error and standard deviation by 36.3% and 70.2%, respectively and 100% success rate in the drone control case. Future developments of this work are the use in the filter prediction phase of a model other than the constant acceleration model in order to improve the obtained results. In addition, we would also like to perform extended user case studies of the remote control modality of the drone in order to understand the usability and intuitiveness of the approach.

REFERENCES

[1] A. Cappozzo, U. Della Croce, A. Leardini, and L. Chiari, "Human movement analysis using stereophotogrammetry: Part 1: Theoretical background," *Gait Posture*, vol. 21, no. 2, pp. 186–196, 2005.

[2] M. Yahya, J. A. Shah, K. A. Kadir, Z. M. Yusof, S. Khan, and A. Warsi, "Motion capture sensing techniques used in human upper limb motion: A review," *Sensor Rev.*, vol. 39, no. 4, pp. 504–511, 2019.

[3] A. Scano, R. M. Mira, P. Cerveri, L. Molinari Tosatti, and M. Sacco, "Analysis of upper-limb and trunk kinematic variability: Accuracy and reliability of an RGB-D sensor," *Multimodal Technol. Interaction*, vol. 4, no. 2, 2020, Art. no. 14.

[4] S. H. Lee et al., "Measurement of shoulder range of motion in patients with adhesive capsulitis using a kinect," *PLoS One*, vol. 10, no. 6, 2015, Art. no. e0129398.

[5] G. Kurillo, A. Chen, R. Bajcsy, and J. J. Han, "Evaluation of upper extremity reachable workspace using kinect camera," *Technol. Health Care*, vol. 21, no. 6, pp. 641–656, 2013.

[6] L. Cai et al., "Validity and reliability of upper limb functional assessment using the microsoft kinect V2 sensor," *Appl. Bionics Biomech.*, vol. 2019, 2019, Art. no. 7175240.

[7] Q. Wang, G. Kurillo, F. Ofli, and R. Bajcsy, "Evaluation of pose tracking accuracy in the first and second generations of microsoft kinect," in *Proc. Int. Conf. Healthcare Inform.*, 2015, pp. 380–389.

[8] A. P. Yunus, N. C. Shirai, K. Morita, and T. Wakabayashi, "Comparison of RNN-LSTM and kalman filter based time series human motion prediction," *J. Phys.: Conf. Ser.*, vol. 2319, no. 1, 2022, Art. no. 012034.

[9] S. R. Tripathy, K. Chakravarty, A. Sinha, D. Chatterjee, and S. K. Saha, "Constrained Kalman filter for improving kinect based measurements," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2017, pp. 1–4.

[10] A. Akbari, X. Thomas, and R. Jafari, "Automatic noise estimation and context-enhanced data fusion of IMU and kinect for human motion measurement," in *Proc. IEEE 14th Int. Conf. Wearable Implantable Body Sensor Netw.*, 2017, pp. 178–182.

[11] J. Chen, H. Zhu, Z. Zeng, J. Liang, and Y. Guan, "Motion tracking of both hands with occasional mutual occlusion using RGB-D camera and IMU," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2017, pp. 848–853.

[12] Y. Tian, X. Meng, D. Tao, D. Liu, and C. Feng, "Upper limb motion tracking with the integration of IMU and kinect," *Neurocomputing*, vol. 159, pp. 207–218, 2015.

[13] S. Moon, Y. Park, D. W. Ko, and I. H. Suh, "Multiple kinect sensor fusion for human skeleton tracking using Kalman filtering," *Int. J. Adv. Robot. Syst.*, vol. 13, no. 2, pp. 65–74, 2016.

[14] B.-N. Vo and W.-K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, Nov. 2006.

[15] R. Ge, M. Lee, V. Radhakrishnan, Y. Zhou, G. Li, and G. Loianno, "Vision-based relative detection and tracking for teams of micro aerial vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 380–387.

[16] M. Pavliv, F. Schiano, C. Reardon, D. Floreano, and G. Loianno, "Tracking and relative localization of drone swarms with a vision-based headset," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 1455–1462, Apr. 2021.

[17] J. R. Cauchard, J. L. E, K. Y. Zhai, and J. A. Landay, "Drone & me: An exploration into natural human-drone interaction," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 361–365.

[18] F. Schilling, F. Schiano, and D. Floreano, "Vision-based drone flocking in outdoor environments," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 2954–2961, Apr. 2021.

[19] "Setup tips for your kinect sensor and play," *Microsoft*, 2013. Accessed: Jul. 05, 2023. [Online]. Available: https://support.xbox.com/en-US/help/hardware-network/kinect/kinect-sensor-setup-tips