This project has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 662287.

EJP-CONCERT

European Joint Programme for the Integration of Radiation Protection Research

H2020 – 662287

D 9.105 – An Indoor Positioning System (IPS) based on a developed camera network system and the multi-image acquisition computer system with the corresponding software

Lead Authors: Joan Aranda, Álvaro Zornoza, Mahmoud Abdelrahman, Pasquale Lombardo

Reviewer(s): Mercè Ginjaume, Filip Vanhavere

and CONCERT coordination team

Work package / Task	**WP9**	**T 9.6**	**ST9.6.1**	**STT 9.6.1.1**
Deliverable nature:	Report			
Dissemination level: (Confidentiality)	PU			
Contractual delivery date:	2018-12-31 (CONCERT M43)			
Actual delivery date:	2018-12-21 (CONCERT M43)			
Version:	1			
Total number of pages:	25			
Keywords:	Tracking system			
Approved by the coordinator:	M43			
Submitted to EC by the coordinator:	M43			

Table of Content

1 Introduction

The Indoor Positioning System (IPS) presented in the D9.103 deliverable [1] was successfully tested in simple workplaces within SCK•CEN and UPC, and in a catheterization laboratory (Cath-lab) of UZ Brussels (Vrije Universiteit Brussel), Liège University Hospital and Skåne University Hospital in Malmö. In these scenarios, the system showed to be reliable when tracking one person or even two people with few inter or self-occlusion. However, several limitations were identified that compromised its extension to a multi-user IPS, i.e. a system that would be able to correctly identify multiple individuals in the scene.

The main limitations can be enumerated as:

1. The range of the person tracking algorithm is restricted to about 4.5 meters;
2. The Field-Of-View of the Kinect v.2 depth camera is limited to 84 degrees horizontally and 54 degrees vertically;
3. The maximum number of tracked people is limited to six by Kinect;
4. There might be occlusions affecting the view of the tracked workers;
5. There might be a misidentification of the tracked workers;
6. The position of the joint coordinates can show some fluctuations and sporadic outliers.

These limitations have been addressed through two different strategies described in this deliverable. The first one is an improved version of that presented in the D9.103 deliverable using only one Kinect camera. The second one is based on a camera network system offering multi-view tracking.

Regarding first strategy, as the first three limitations are due to the intrinsic hardware features of the Kinect v.2, they have to be solved by optimizing the tracking setup. The Kinect sensor can track up to six people. The individuals in the scene should not be occluded with each other or objects in the scene (e.g. the ceiling shielding during fluoroscopy-guided procedures). Otherwise, the system may lose the tracking of this user. However, a correct placement of the Kinect can prevent the risk of occlusions, allowing a clear view of the part of the bodies of interest. With respect misidentification, the current tracking software does not provide an algorithm able to identify individuals on line, but they can be classified in an off-line process, after trajectories acquisition, with simple algorithms.

Some little fluctuations of ± 1.5 cm, known as jittering, have been identified in the joints' coordinates. The joint jittering is a well-known problem for the Kinect v.2, and it is due to the tracking algorithm and to the noise of the depth images. To reduce the jittering, we propose to sample the depth images with the highest frequency, i.e. 30 Hz. This high sampling rate allows us to filter the tracking data, thanks to which we can smoothen jittering and spikes.

The filtration algorithm will be applied to the output file from the Kinect Data Acquisition (KDA) [1] and it will provide a new file with the demanded constant sample rate of 1Hz, according to the dose calculation requirements.

As a second strategy, this deliverable will explore the use of multiple Kinect cameras capturing the scene from different and complementary viewpoints. The use of multiple cameras would allow to monitor larger areas and tracking from different perspectives should reduce the occlusion problem. This tracking software does not provide an algorithm able to identify individuals either. However, by using different viewpoints the misidentification problem caused by track interruptions due to occlusions from a given point of view will be highly reduced.

Regarding the calibration, the system must provide a procedure to calibrate easily the camera in the monitored workplace and get the geometrical transformation between the world and the camera coordinates. In the case of the system with multiple cameras, it is necessary to do this process for one of the cameras and perform an inter-calibration between the different cameras that belongs to the network. The development of an automatic calibration system is under progress to ease the set-up of the IPS.

2 Improved IPS based on a single Kinect camera

2.1 Filtering

2.1.1 Need for filtering

The IPS presented in the D9.103 deliverable uses the Microsoft Kinect SDK skeletal tracking system that provides with the joint positions of tracked persons' skeletons. These joint positions will be used in our application for staff-pose estimation and positioning of our phantoms in the workplace.

Just like any other measurement system, Kinect v.2 sensor is not free from characteristic noise, which affects the joint position estimation even when the tracking conditions are good enough. During our tracking experiments, the Kinect sensor exhibits a permanent precision error in joints' coordinates in the form of some little fluctuations of few centimetres (the exact dimension depends on distance), known as jittering. This joint jittering is a well-known problem for such tracking systems, and it is due to the tracking algorithm and to the noise of the depth images. Therefore, we can consider this noise as a relatively small white noise that is always present for all joints and caused by imprecision.

However, there exists another big source of noise mostly due to skeleton's joints occlusion or self-occlusion that generates lack of accuracy. Even during stationary pose, some joints often appear to be shifted in unrealistic manner. These sudden joints' coordinate changes are known as *spikes*.

Thus, an important step before using the raw joint data is to remove as much noise as possible from the input data by means of a noise reduction noise filter. Since the noises have different characteristics, different filtering techniques should be used for each of them, in order to achieve good results. In order to study the nature of the noises and their frequency, the evolutions of X, Y and Z coordinates of each joint were analysed, as well as their first and second derivatives.

As an example, Figure 1 shows the plot of time evolution for the z-coordinate of the spine base joint during a static pose. Below, the first derivative is plotted to raise the evidence of the noisy data. In this first derivative plot, some positive and negative peaks can be observed. These are the noisy values to be filtered.

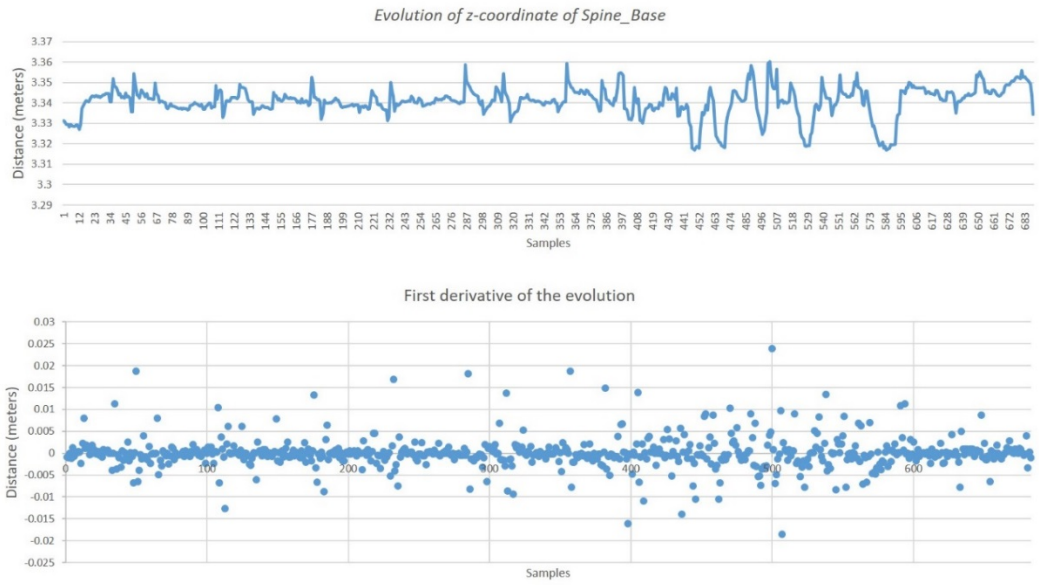

***Figure 1**. Evolution example of z-coordinate of the spine base (meters) for a static pose (30Hz). Above: Raw data; Below: Derivative of raw data, highlighting the jittering.*

In an IPS, any smoothing strategy will introduce a lag or delay in the output increasing system latency. Latency in our application can be defined as how much time it takes for filter output to catch up with the actual joint position when there is a movement in a joint. Thus, latency degrades the synchronization of the person tracking and can jeopardize the accuracy of our real time dose calculation. In general, the filtering delay depends on how quickly the input is changing, and hence, one cannot attribute a specific delay value to a given filter for all cases. However, this parameter will be explicitly considered when comparing filtering techniques.

2.1.2 Types of Filters

Different filters have been tested before deciding upon which one to apply to our data [2]. The following ones are presented and discussed in this report:

- Exponential Smoothing
- Double Exponential Smoothing
- Median Filtering

Exponential Smoothing

An exponential smoothing filter, also known as an exponentially weighted moving average (EWMA), is a popular filter in many different fields due to its fastest computation and better performance compared with the simple averaging smoothing filter.

The exponential filter output estimation is given by:

$$\hat{X}_0 = X_0$$

$$\hat{X}_n = \alpha X_n + (1-\alpha)\hat{X}_{n-1}$$

Where $\{Xn\}$ represents the raw input data, sequence and α is called the smoothing factor with 0≤α ≤1.

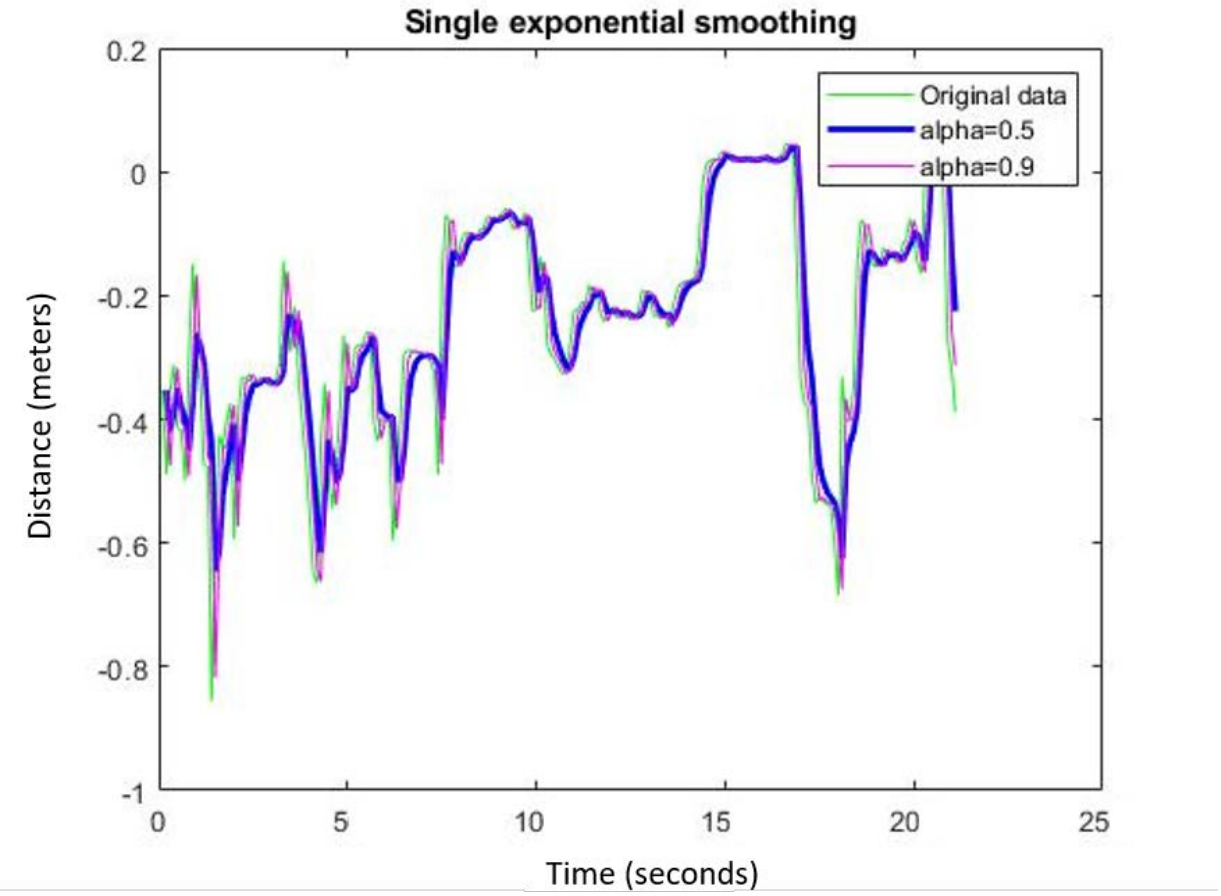

Figure 2. *Example of application of the single exponential filter to the raw data with two different values of alpha.*

Whereas in the simple moving average the past observations are weighted equally, exponential functions are used to assign exponentially decreasing weights over time. Higher values of α corresponds to larger weight on recent inputs, which results in less latency and less smoothing of the data.

Exponential filter output is determined by previous samples but it does not change depending on samples evolution or trend, so it does not matter if past values are increasing or decreasing. A variant of exponential filter, called a double exponential filter, addresses this limitation of exponential filters and is described in the next section.

Double Exponential Smoothing

The double exponential smoothing filter uses a second exponential filter (hence the name double exponential) to account for trends in input data. In the case of joint positioning, a trend can be thought of as the estimated velocity of the joint, calculated as the smoothed difference between the last two estimated joint positions ($Xn – Xn–1$).

Therefore, actual trend can be calculated by using an exponential smoothing filter as:

$$\text{Trend:} \qquad b_n = \gamma(\hat{X}_n - \hat{X}_{n-1}) + (1-\gamma)b_{n-1}$$

Where, γ acts as the smoothing factor used in exponential filtering of joint velocity estimation. The γ parameter controls how sensitive the trend is to recent changes in input. A large γ results in less latency in trend, but amplifies the negative effects of noisy input signals.

This trend is usually accounted in the joint position exponential filter in the next form:

$$\text{Filter output:} \quad \widehat{X}_n = \alpha X_n + (1-\alpha)(\widehat{X}_{n-1} + b_{n-1})$$

Including the trend in this way, helps to reduce the delay in exponential filtering output, especially when the joint velocity is constant. High sampling rates (in relation to joint speed) helps to satisfy this constraint.

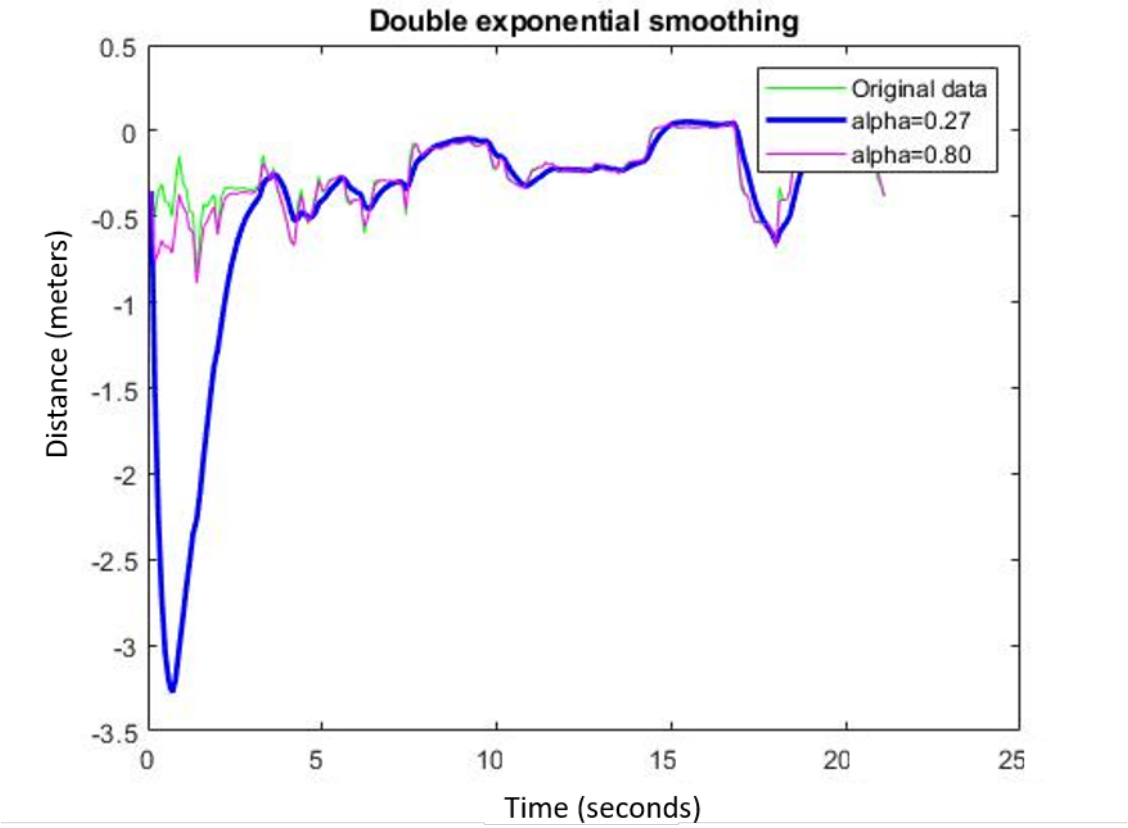

***Figure 3**. Example of application of the double exponential filter to the raw data with two different values of alpha.*

On the other hand, the trend factor bn can easily result in erroneous values when there are sudden joint movements or stops. In these cases, the trend term bn cannot be well estimated, which can result in overshoots and oscillations in filter output. However, by optimizing the trend smoothing factors these issues could be reduced.

Double exponential smoothing was attempted with different values of damping factors. As seen in Figure 3, double exponential filter can introduce significant overshoots, mainly in the beginning of a track, when the trend is not already well evaluated. Their effect can be reduced with a convenient tune of smoothing parameters.

Median filter

The problem with lineal smoothing filters is that they take into account of all the previous samples with more or less weight. In fact, in the signal processing literature, the exponential smoothing is equivalent to a first-order Infinite Impulse Response or IIR filter. Therefore, given a spike or noise in the input data, this will be considered in future outputs, until its weight is reduced over time.

On the other hand, even if the PODIUM project only requires one set of joint positions per second, Kinect v2 can process and provide data up to 30 frames per second. Thus, to reduce the jittering, we could sample the depth images with a higher frequency, i.e. 30 Hz. This high sampling rate allows us to filter the tracking data using a non-causal filter, thanks to which we will smoothen better jittering and spikes, while maintaining the maximum time latency of one second demanded by PODIUM project.

The idea is to treat all the joint positions' samples in one-second period just as estimations of the actual joint coordinates. Then, from all the candidates (up to 30 depending on acquisition speed), the filter orders them, and select the median value, that is, the value separating the higher half from the lower half of the sample values. The benefit is two-fold: first, the selected value is a real input value obtained during the one second period (not an averaged one); secondly, it rejects the outliers (spikes) produced by the inferred joint values during occlusions, so they will not weight in future analysis. Additionally, this method allows fixing the 1Hz sample output even if the frequency of the input samples is adjusted at different frequencies or it oscillates due to problems with the acquisition system. The moving median filter is calculated over a jumping window of one-second length, thus it returns one median value per second of every joint coordinate.

The median filter can produce some discontinuities (discretization) in the signal, and they need to be smoothed out. So an extra filter, like a single exponential filter, is applied, the result of which can be shown in the example of Figure 4.

However, some inconveniencies of this method were observed. First, it introduces up to 1-second lag in the filter output. Second, the use of the median filter for each joint coordinate separately can generate a false joint position once combined, with X, Y, Z components belonging to the same 1 sec. interval but coming from different time stamps.

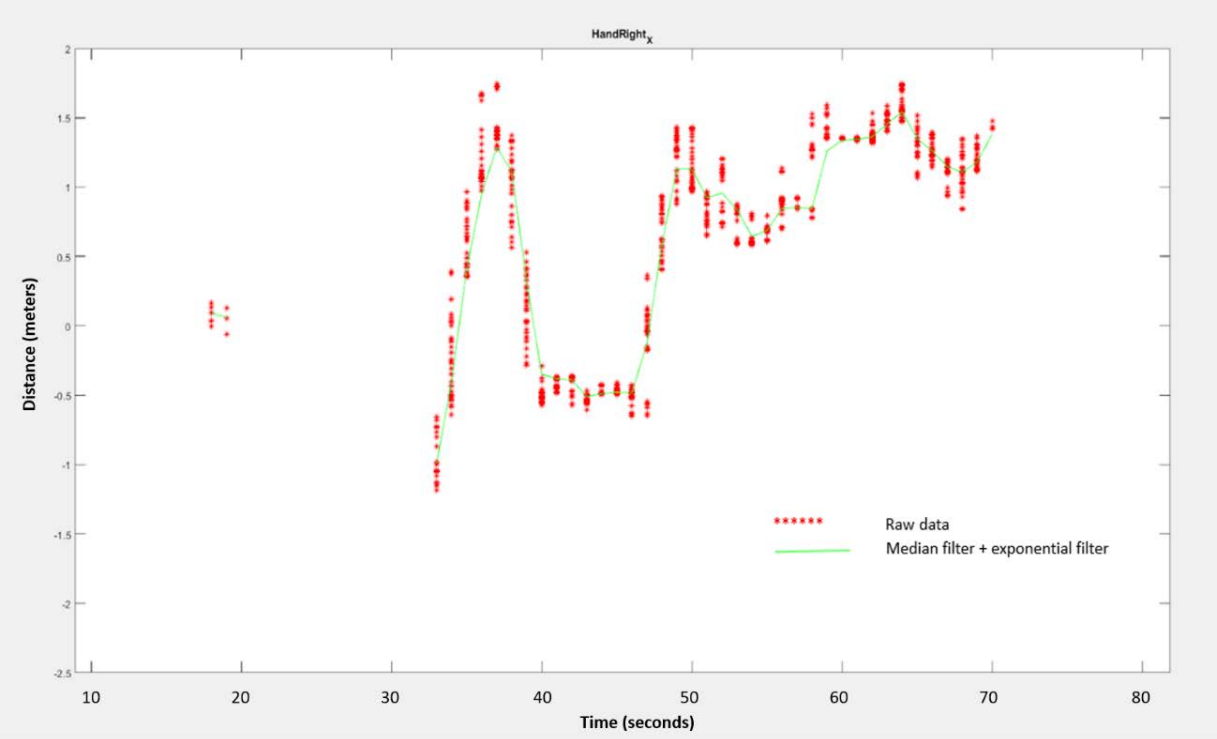

Figure 4. *Example of application of the median filter followed by a simple exponential smoothing filter. The red points represent the samples acquired in a 1-second interval; the green line is the filtered and smoothed data.*

2.1.3 Selected solution

We tested different filters for our application, including some not reported here, such as the Kalman filter, and we applied them using various permutations like applying median filter before and after exponential smoothing, different values of alpha, different window lengths for median filter, etc…

A good filtering solution is usually a combination of various filtering techniques, which may include applying a jitter removal filter to remove spike noise, a smoothing filter, and a forecasting filter to reduce latency, and then adjusting the outputs based on person kinematics and anatomy to avoid awkward cases caused by overshoot [3].

We implemented a filter architecture that integrate jitter reduction, as well as statistical smoothing based on Holt Double exponential method [4] as described in the scheme of Figure 5. This implementation was provided natively in later versions of Microsoft Kinect version 1 SDK, but was yet to be implemented to version 2 SDK. It is highly configurable using simple configuration options.

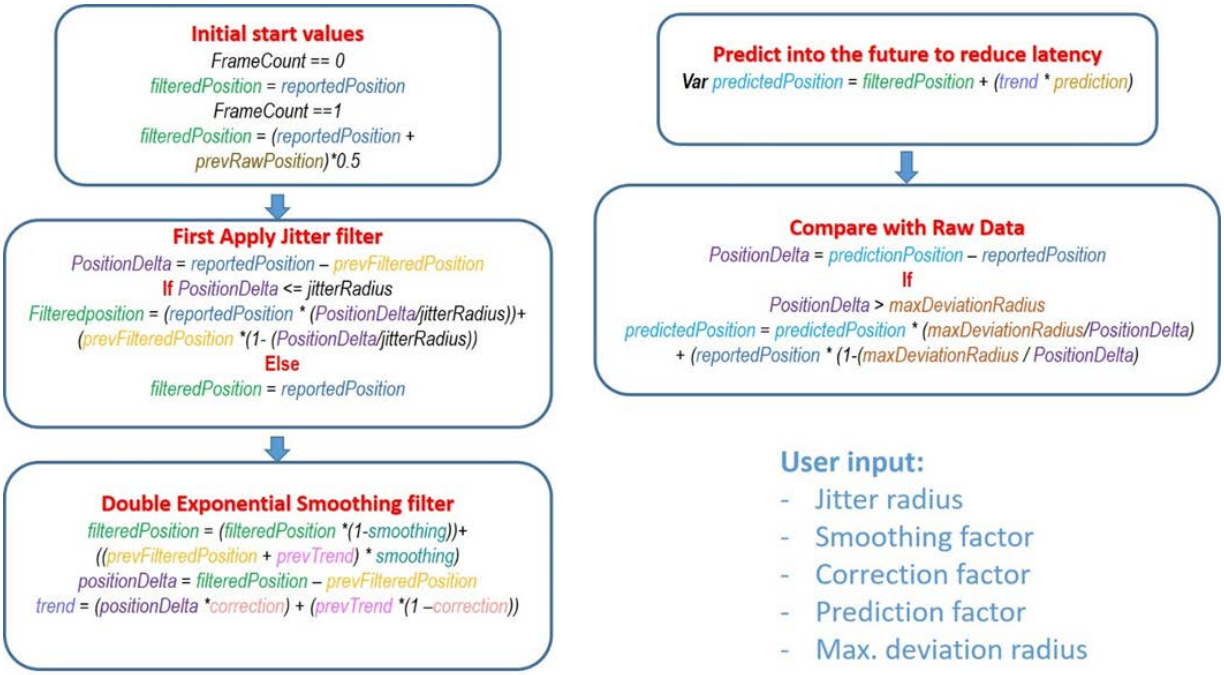

***Figure 5**. Scheme of jitter removal filter and Holt double exponential smoothing filter.*

The filter can be configured via five smoothing parameters:

1. **Smoothing** (*1-α; α* described in 2.1.2)
   Increasing the smoothing parameter value (*1-α*) leads to more highly smoothed skeleton position values being returned. It is the nature of smoothing that, as the smoothing value is increased, responsiveness to the raw data decreases. Thus, increased smoothing leads to increased latency in the returned skeleton values. Values must be in the range 0 through 1. Passing 0 causes the raw data to be returned.
2. **Correction** (gamma (γ); γ described in 2.1.2)
   Lower values are slower to correct towards the raw data and appear smoother, while higher values will correct toward the raw data more quickly. Values must be in the range 0 through 1.
3. **Prediction**
   The number of frames to predict into the future. Values must be greater than or equal to zero. Values greater than 0.5 will likely lead to overshooting when moving quickly. This effect can be damped by using small values of MaxDeviationRadius function.
4. **Jitter Radius**
   The radius in meters for jitter reduction. Any jitter below this radius will be smoothed. Lesser the changes in raw input data, higher the smoothing effect.
5. **Max. Deviation Radius**
   The maximum radius in meters that filtered positions are allowed to deviate from raw data. Filtered values that would be more than this radius from the raw data are clamped at this distance, in the direction of the filtered value. This last stage prevents from overshoots produced in some cases by double exponential filtering and prediction steps.

The configuration of the filtration algorithm is highly dependent on the application and the nature of movement. Experimentation is required on an application-by-application basis in order to provide the required level of filtering and smoothing for each application requirements. In fact, this configurable implementation was designed to enhance the user experience in video games; thus very low latency

configuration was set by default in the Microsoft SDK v1. On the other hand, for our application, we established that one frame per second rate is good enough to estimate doses to workers. Therefore, we set the filter parameters according to the following criterion:

- Applying jitter radius equivalent to the max. distance between two consecutive joints.
- Avoiding aggressive smoothing which may cause inaccurate joint positions.
- Avoiding lags so that the filter output can be synchronized with the acquisition time.

Table 1 shows the proposed filter parameters obtained by trial and error.

Proposed Filter Parameters	
Smoothing coefficient	0.8
Correction factor	0.1
Prediction coefficient	0.25
Jitter radius	0.5
Max. deviation radius	0.3

Table 1. Double exponential filter parameters used for the selected filtering solution

The advantage of the algorithm is that it will not introduce lags, which could cause a desynchronization in the acquisition time. In this way, we will be able to provide real-time tracking data as it is illustrated in Figure 6. Figure 7 shows an example of raw data of the hand from a real procedure once the proposed filter is applied.

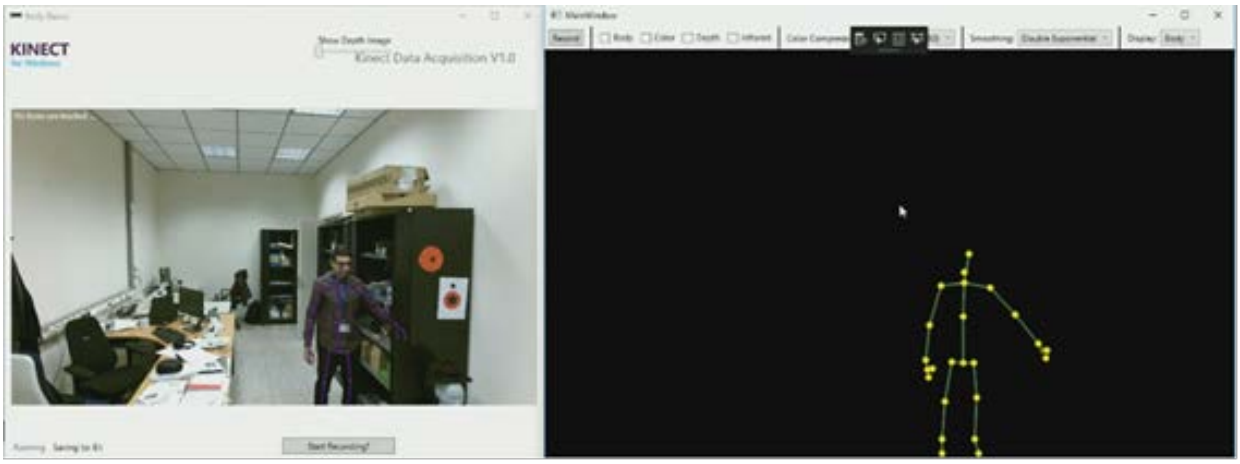

Figure 6*. Screenshots of the system.*

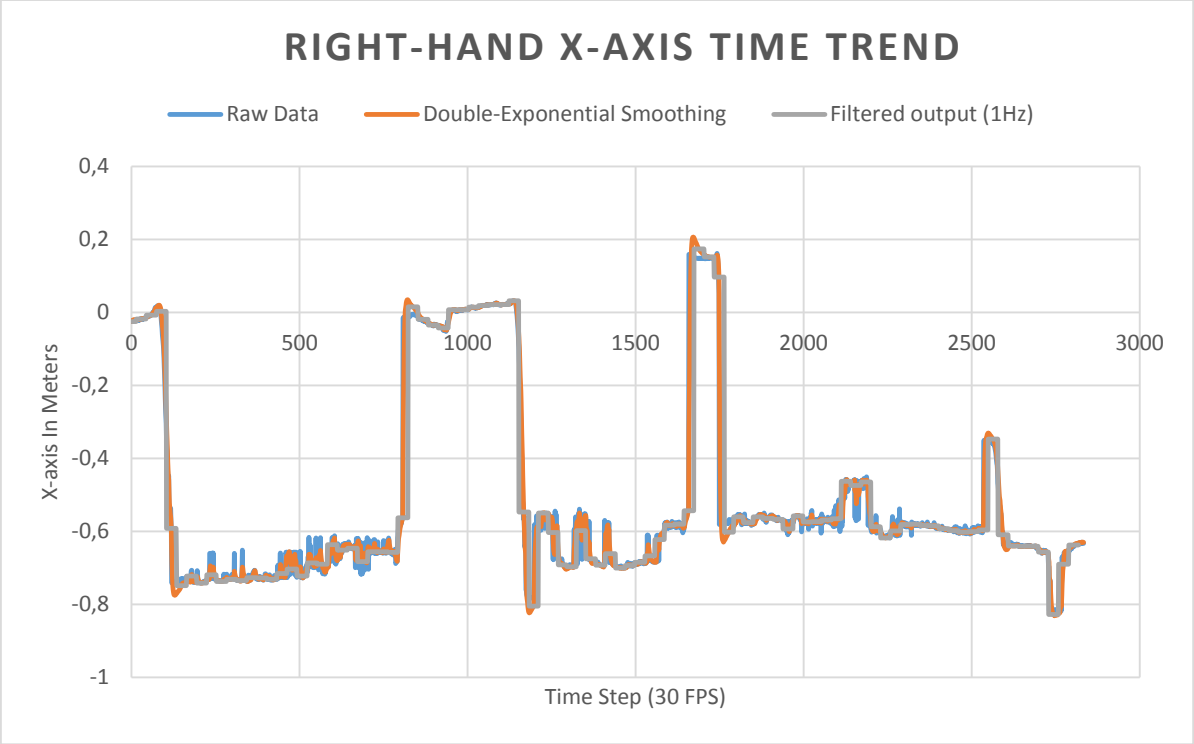

***Figure 7**. Example of application of the selected filtering solution to the right hand tracking during a real X-ray procedure*

2.2 Solution to occlusion problems and users' misidentification

As already addressed in deliverable D9.103 and in the introduction, an important limitation of a multi-user tracking with a single Kinect is that the current tracking software does not provide an algorithm for identifying individuals.

Recognizing the different users in different locations is considered as an important and challenging task. Marker-less identification methods can be coarsely grouped, according to the used data, into three classes: RGB appearance-based identification, depth-based identification and skeleton-based identification. Most of the existing algorithms rely on RGB-based appearance features, for examples, references [5] and [6]. Nevertheless, RGB sources suffer always from pose changes, occlusions, clothes change (using of PPE, glasses,…etc.) and illumination variations. This hypothesis represents a very strong restriction, since it constraints identification methods to be applied under a limited temporal range (reasonably, in the order of minutes). In fact, the people appearance is subject to dramatic changes in camera views due to changes in the angle view, the body pose, lighting conditions and in the background cluttering caused by objects in the FOV. In our considered applications, in interventional radiology and in selected neutron workplaces, the use of RGB appearance-based identification might be a very challenging task. For instance, the same person can have different images depending on the different clothes he/she wears or the illumination conditions. Moreover, there are concerns about privacy and security of obtaining colour images of individuals operating in restricted workplaces.

On the other hand, there is a growing interest in depth-based and skeleton-based approaches. Indeed, this information is robust to variations of illumination, scale and rotation [7]. Depth-based

identification methods aim to create human signature from depth images or videos in the absence of RGB information [8]. While, in skeleton-based identification a small number of joint positions can effectively represents human motion.

One skeleton-based identification method is based on Gait Analysis that identifies people through the analysis of the way of walk. However, this analysis of walk is limited to medical, sports and surveillance applications.

Barbosa et al. [9] proposed a person signature that exploits skeleton-based features while exploiting soft-biometric features using a set of ratios of joint distances. Similarly, the study of Murano et al. [10] proposed skeletal tracker to determine joints as key points. They extract 2D and 3D descriptors in order to compute signatures of people. By comparing signatures of each target in the test frame with those in training frames, the best match is selected as a result. Those approaches differ depending on the required input data, feature quality and selection, or the need for a training dataset. The basic principle that skeleton-based identification approaches share is finding a skeleton-based person signature that can be unique to each individual. Some proposed skeleton-based person signatures are:

- Euclidean distance between floor and head
- Ratio between torso and legs
- Height estimate
- Euclidean distance between floor and neck
- Euclidean distance between neck and left shoulder
- Euclidean distance between neck and right shoulder
- Euclidean distance between torso center and right shoulder
- Geodesic distance between torso center and left shoulder
- Geodesic distance between torso center and left hip
- Geodesic distance between torso center and right hip

2.2.1 Geometrical filtration of single Kinect data for Skeleton Identification

Despite the fact that the skeleton tracking of the Kinect SDK is giving natively a body ID to different users in the scene, this body ID is not consistent over the tracking sequence. The body ID can change when users are overlapping (full body occlusions) or when a user exits the view. Although, performing multiple view skeleton tracking can help retrieving consistent body ID per user along a whole sequence, we investigated different techniques to improve skeleton identification in our single camera IPS. Mainly, it was decided that one camera IPS solution would be easy and convenient to be installed in interventional radiology workplaces.

Facial recognition is particularly difficult in interventional radiology where monitored workers are usually wearing protective glasses and caps. To overcome this problem, we developed an identification algorithm using skeleton joints data based on the specific geometrical configuration found in interventional radiology rooms.

In the case of interventional radiology, the algorithm makes use of skeleton tracking to correlate the relative position of monitored workers to the camera location. In fact, during our tests we observed that the main operator (the doctor) is usually working in a specific space region on the patient side, while the first assistant is positioned close to the main operator but in a separate region compared to the doctor' one. Overlapping between these two regions is unlikely when the x-ray beam is on.

After each procedure, histories of different body joints of different users can be clustered per user. Hence, the skeleton sequences are modelled as trajectories on 2D plans. This helps identifying monitored users based on their first known geometrical configuration. Skeleton-based person signature can also be used in suitable scenarios. For examples, worker's height can be used as complementary information that helps to identify different users when applicable. Giving the fact that in interventional radiology workplaces, we see only the upper part of the body, it is difficult to rely on a single skeleton-based person signature to identify different users, for example, using the Euclidean distance between feet and neck. Thus, different techniques can be used for different geometrical configurations and for different body size of monitored workers. For now, this procedure is done offline after the procedure is finished.

On the other hand, in neutron workplaces, RGB-based appearance features could be used to identify the monitored workers whenever they do not have to use PPE equipment that cover or partially cover the face.

Table 2 summarizes the different options for personal identification in PODIUM applications.

Method	Workplace	Requirement
RGB-based appearance Facial recognition	Neutron workplaces where workers do not use masks or glasses	- Face frontal view - Good illumination
Single-View Skeleton-based	Interventional Radiology	- Distinctive geometrical configuration
Skeleton-based person signature	Interventional Radiology	- Distinctive feature among tracked users
Multi-view Skeleton-based ID	Neutron workplaces Interventional Radiology	- Multiple cameras

Table 2. Personal identification methodology when using a 1 single Kinect

3 Multi-user IPS with multi-view approach

3.1 Introduction

Because of the reported occlusions and Field-of-View (FOV) limitations, a second tracking approach based on a multi camera solution is also proposed. We have developed a software to acquire the skeleton data from different view points so that the identification is correctly performed thanks to the data fusion from different sensors. At the moment of writing this deliverable, the software is a beta version and its verification is in progress. The goal of this new tracking approach is to avoid the previously mentioned problems of occlusions and limited field of view.

Prior to the development of the software for multi-view approach, existing literature was reviewed to explore several approaches. The most reliable way is to use marker-based motion capture systems [11]. These systems show great results in terms of accuracy (normally less than 1mm), but they are very expensive and require the users to wear many markers. However, in PODIUM, we aim at finding a marker-less solution and thus this approach was discarded.

Now, two possibilities can be considered about skeleton obtention. The first possibility is to use the skeletons provided by this Microsoft Kinect SDK. The second possibility is to use OpenPTrack [13]. In

this software they compute each single view by using convolutional neural networks (CNNs) for 2D pose estimation (as they do in OpenPose [12]) and extending the resulting skeletons to 3D by means of the sensor depth.

OpenPTrack would require the use of GPU and the payment of a license (OpenPose) making the solution more complex and more expensive. Therefore, we discarded the second option and we decided to proceed developing a software that performs the body estimation using the skeletal data coming from individual Kinect sensors (using the Microsoft Kinect SDK) connected through a network.

3.2 Description of the two Kinect system

3.2.1 Architecture: hardware needs

The proposed multi-user IPS with multi-view approach is based on the use of multiple Kinect sensors, which require two separate computers and an adapted software consisting of two Windows Presentation Foundation (WPF) applications developed in C#. The first application is the master (or admin) (figure 8).

Labels 1 and 2 in Figure 8 represent the two columns of data (cameras 1 and 2 respectively). In label 3, we can see the RGB images coming from cameras and in label 4 we can see the depth image. In label 5, we can see some data about how many skeletons is identifying each camera. In label 6. we can see logs of the programme, in this case messages indicate that data is being received. In label 8, we can see the number of fused skeletons and its position and in label 7, we can see a 3D plot of the fused data.

Figure 8. *View of the master application.*

This software manages the connection between the cameras, the calibration and the fusion of the data coming from the multiple cameras. The second application is the slave (figure 9). This software handles the connection with an individual Kinect sensor and sends both the skeletons and the images to the master software (when required by the master). Note that a computer can only handle one Kinect sensor due to the restrictions of the SDK 2.0 provided by Microsoft.

In label 1 of Figure 9, we can see the RGB frame, in label 2 the depth frame and in label 3 the recognized skeletons in the scene. In label 4, we can see some logs about how many individuals are being recognizes and in label 5, we can see several buttons to control the program.

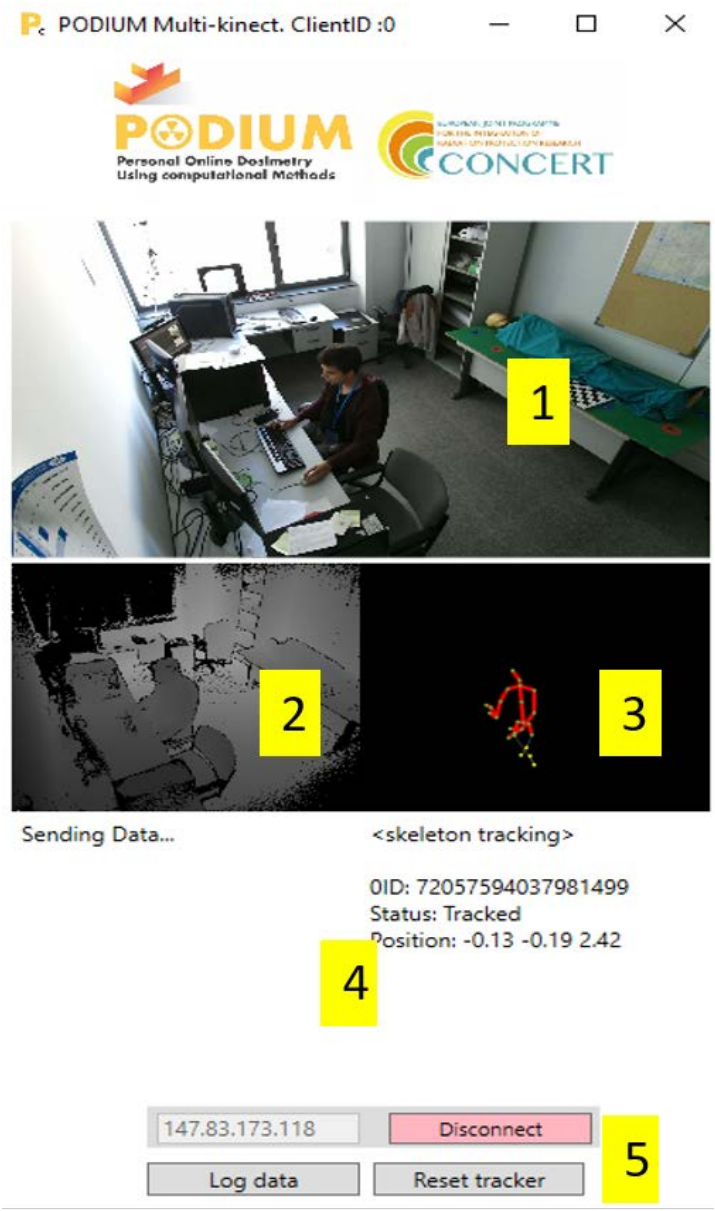

Figure 9. View of the slave application.

Figure 10 shows a simplified diagram of the architecture when having a solution with two cameras. The first computer runs both the master node and one of the slaves that it is attached to the first kinect. The second computer will run the second slave, controlling the second Kinect. Computers communicate with each other thanks to a connection based on TCP-IP protocol.

The slaves do not send the RGB, depth frames to the master node automatically since this transfer demands a high bandwidth, and it would reduce the frequency of skeleton data. They can be obtained under request of the master at any time. They are needed for test and in the calibration process.

With our system, we offer two alternatives for connecting the computers controlling the Kinect(s). The first possibility consists on connecting directly both computers with an Ethernet cable. The second one relies on a LAN network present in the area and connecting both computers to the same LAN. Then, we only need to input the respective IP addresses in order to connect the slave to the master.

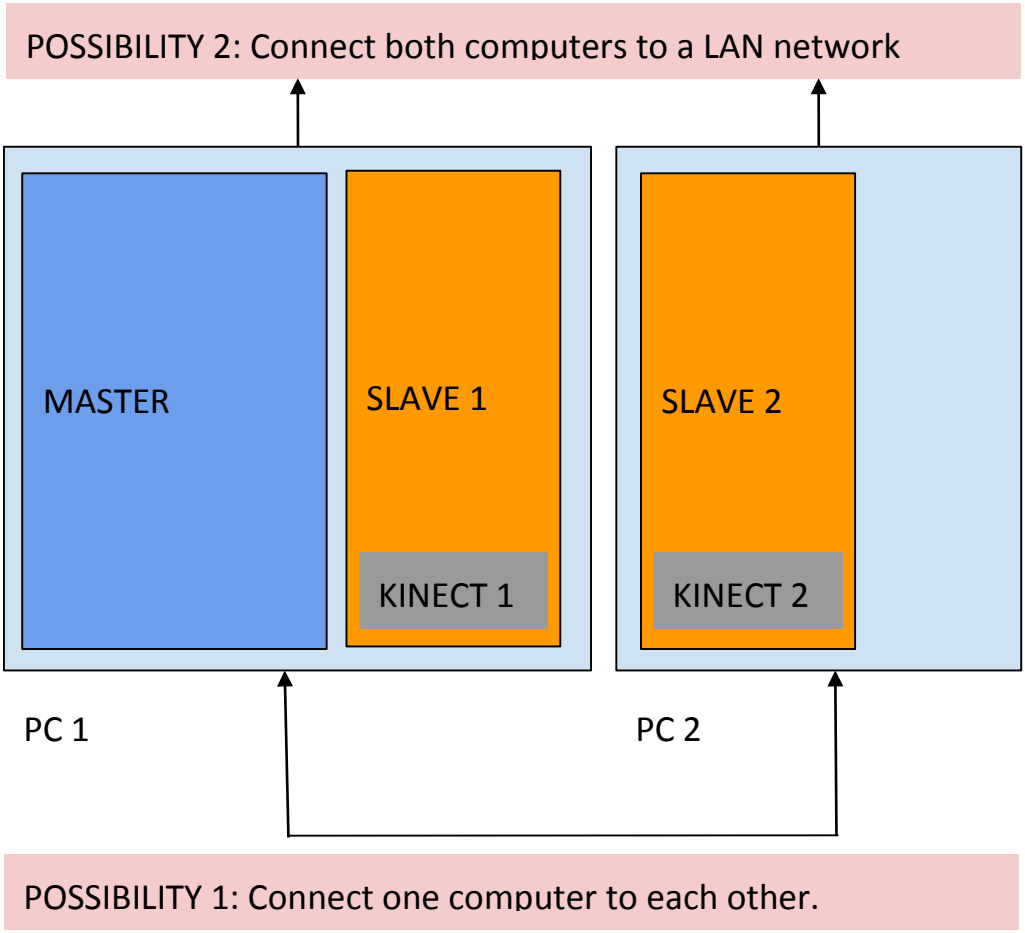

Figure 10. Architecture for two camera-two computer solution.

3.2.2 Calibration

The calibration of the system consists of two different procedures. First of all, the cameras that belong to the network should be calibrated with each other, i.e. finding the geometrical transformation between the multiple cameras. The second process consists in getting the position of one of the cameras with respect to the world applying the same method described in [1].

Regarding the calibration between cameras, we use a known pattern consisting of a red circle printed in both sides of a piece of a white paper (figure 11). This is seen and easy to identify by both cameras. This object is placed in at least four different positions and the X, Y and Z coordinates of the central point with respect to each camera are obtained. Once we obtain these points, we run the algorithm presented in [14] to find the optimal rotation and translation (6 degrees of freedom) between two sets of 3D points trying to minimize the least square error.

Figure 11. Camera calibration: both cameras must see the selected pattern

This is a first solution adopted for this beta software. Other possibilities are the use of a QR pattern or calibrate the camera with the Kinect skeleton own data as it is proposed in [15].

3.2.3 Software description

The software consists of two Windows Presentation Foundation (WPF) applications developed in C# language. As it is shown in the architecture, we need to execute one master software and as many slaves in separated computers as the cameras that we use. Note that the first slave can be run along the master software in the same computer. Even if the presented solution is designed to support many cameras, in the framework of PODIUM only two cameras are used. In fact, two cameras should be sufficient to monitor most of PODIUM applications. Furthermore, the use of only two cameras should make the set-up of the IPS easier.

The fusion of the data coming from the multiple cameras is the key feature of this software. The master node is in charge of fusing the different information that is receiving from the single-view detectors in the network. One of the common limitations in multi-camera motion capture systems is the need of having synchronized cameras. Moreover, off-the-shelves RGB-D sensors, such as the Microsoft Kinect v2, do not have the possibility to trigger the image acquisition. In order to overcome this limitation, our solution merges the different data streams controlling they belong to the same specified time intervals (few tens of milliseconds). The two slaves send skeleton data as soon as they have a new frame available. The server receives this data and updates the corresponding buffer that store the skeleton data of the different Kinect sensor.

Another difficulty is performing the fusion of the data continuously. First, it needs to transform all the skeletons, provided by the multiple Kinects, to a unique reference frame (using the calibration data). Second, we associate the skeletons that belong to the same person. To do that, we match those skeletons whose heads are near each other, in a distance below than a certain threshold. This produces a list of sets of skeletons. Then, we calculate the weighted average of all the joints of the input skeletons that belong to the same set. This weighted average takes into account the distance to the skeleton from each camera (the further the object is located from the camera, the worse confidence in the Z coordinate) and whether the joint is being tracked or inferred by the corresponding Kinect sensor.

For future improvements of the software, other approaches can be considered. For instance, in [13] they use Kalman filter for the data association algorithm. In [15] they compute and improve the fused skeleton taking into account the constraints of bone-lengths.

3.3 Evaluation of the methods.

3.3.1 Preliminary tests

As it was mentioned before, this software is proposed with the aim of solving occlusions and field of view limitations. In order to show the behaviour of the system in front of those situations, an experimental room has been used.

The room dimensions are 5.2 meters length by 3.2 meters wide (area is 16.64 m^2). Two cameras were placed in two corners of the room at height of 2 and 2.5 meters respectively, forming an angle of 45º with respect to the wall and an angle of 90º between each other. This camera configuration allows an easy calibration. Moreover, a table was placed in the middle of the room with a dummy in order to simulate the conditions of the PODIUM requirements (figure 12 illustrates the experimental set up and shows the camera location).

As can be seen in the figure, the scene is well covered by the field of view of the two cameras. We can distinguish three different areas. Both cameras cover zone 1 at the same time. Calibration should be performed in this area. Zones 2 are areas covered by only one of the cameras.

This distribution allows two or more individuals moving around the scene and it is used to test the performance of the system.

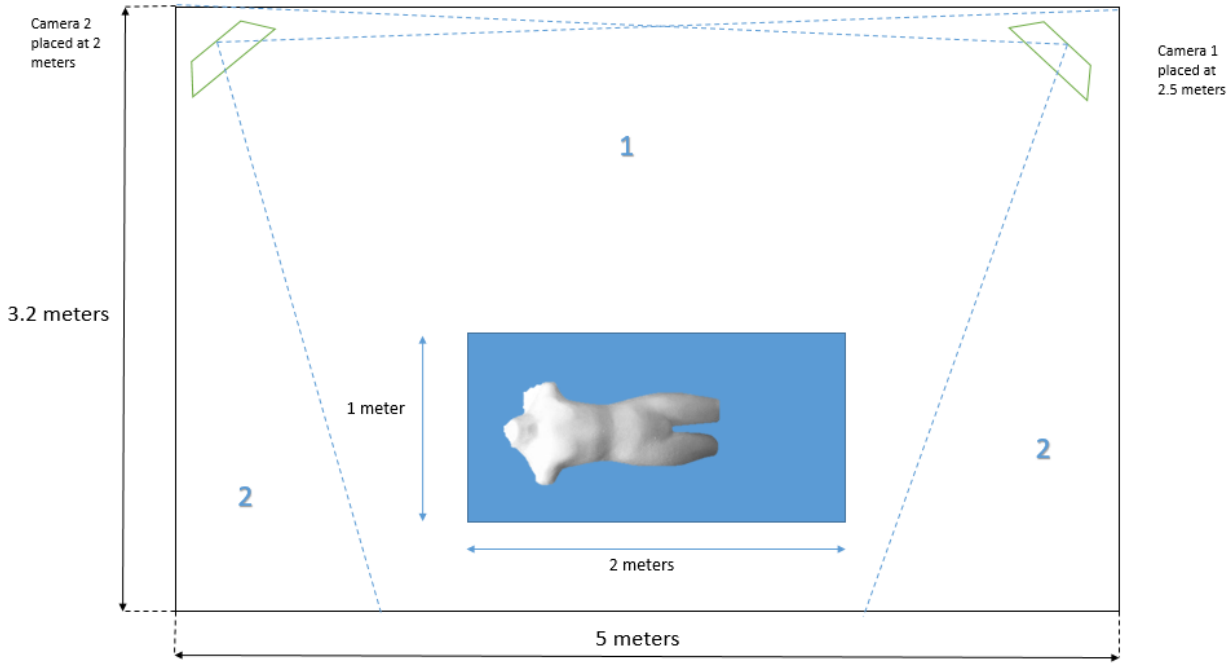

Figure 12. Experimental set up for the two-camera tracking

The video where the test is performed is available under request. In the video, one can see the master node and the two client nodes. First, the movement of the two individuals with partial occlusions in one of the cameras was recorded. Then, the performance of the system was evaluated.

Regarding occlusions, in Figure 13 we can see a trial with two computers and two Kinect sensors. In the left part of the figure, we can see the master node and one of the slaves. In the right part of the figure, we can see the other node. In this instant of the video we can see that in camera 2, one of the individuals is totally occluded by the other individual (label 4). However, the camera 1 is able to see both individuals (label 5) and indeed the fusion of the data produces an output of two skeletons (label 3).

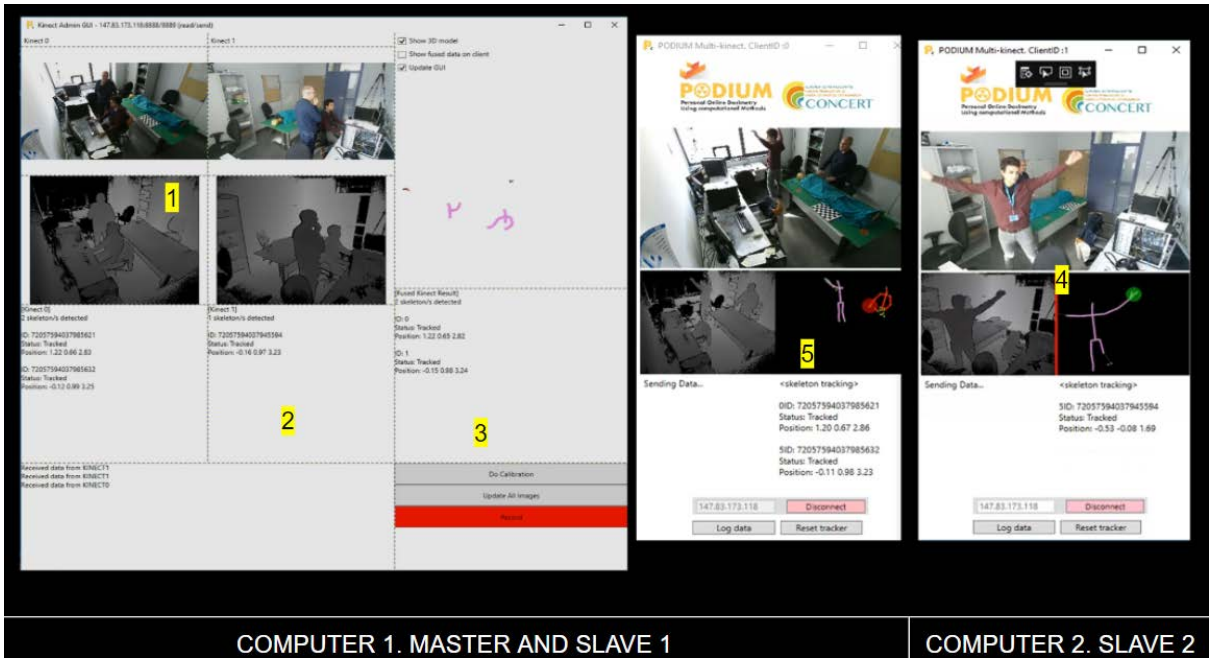

Figure 13. Evaluation of the software when there is occlusion in one of the cameras

In addition, it is shown that the field of view is enlarged thanks to the second camera (Figure 14). One of the cameras does not see the two individuals. However, the other camera is able to see them and the fused output is correct. Thus, by placing the cameras in a good position, the field of view can be extended. Such an advantage is especially important for large workplaces, such as some neutron facilities.

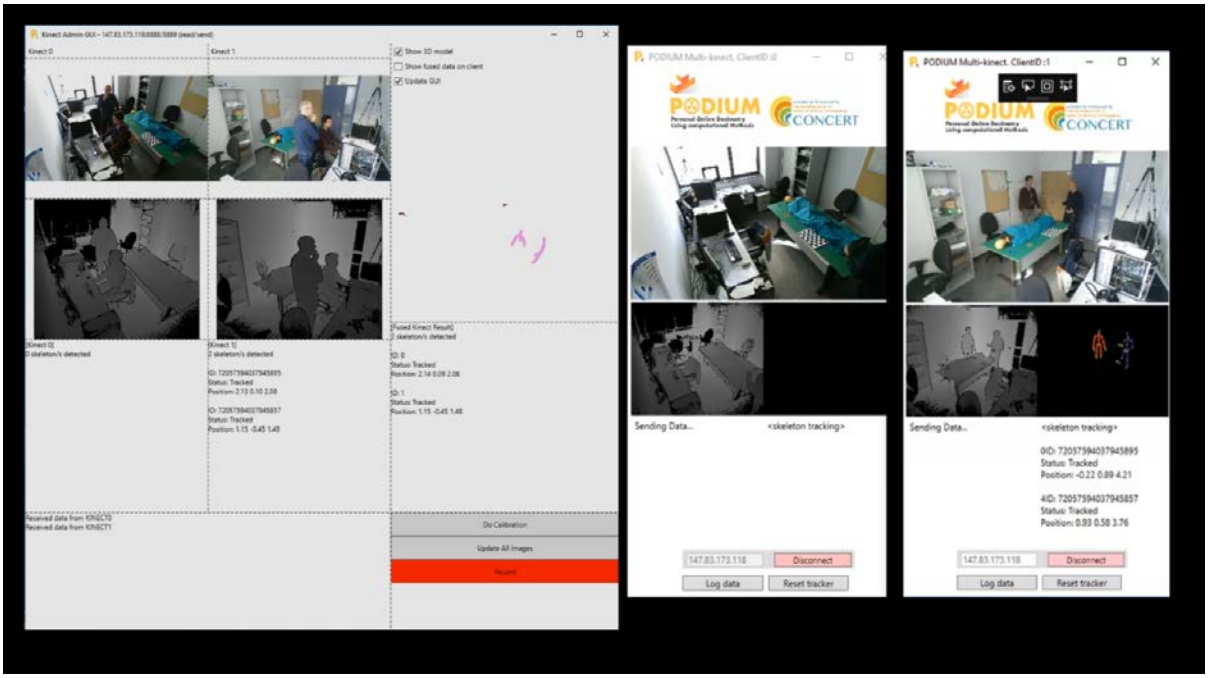

Figure 14. Evaluation of the software, showing extension of the field of view

3.3.2 Work in progress

The multiple camera software was tested successfully and shows promising results. It is planned to complete its verification in the next months, and in particular to test it against the improved KDA single Kinect software described in paragraph 2. The main foreseen steps are:

1. **Continue the software improvement and verification**

We would like to restructure the graphical user interface to make it easier and friendlier. Moreover, improvements in the communication between the different computers and the concurrency of processes are expected for early 2019. As indicated in the paragraph on calibration, a new calibration method based on the skeletons coming from the Kinect will be developed. The fusion part will be reviewed too in order to get more realistic skeletons according to physical bones length. Finally, in the next months (February-March), it is foreseen the integration of this software in the PODIUM application developed under WP3 (online dosimetry application).

2. **Comparison 1 Kinect / 2 Kinect, examples of realistic movements in UPC lab**

In the next months (February-March), it is planned to set up a new test where the tracking approaches will run in parallel and both output will be compared. This will be done first under controlled conditions in a UPC lab. The set up includes a first camera and one computer in which both software: KDA and multi-camera are run. This camera will be located in the best position of the scene to warrant good KDA output (as good as possible). Another camera with another computer will be added and it will be connected to the multi camera software (that is already running in the first computer along the KDA software). In this way we will be able to benchmark and compare both systems, looking in which situations double view system overcome the KDA and when its extra cost is justified.

3. **Verification in realistic workplaces**

Some tests will be performed before March 2019 in St James Hospital, to compare both techniques. Depending on the results, the final programming of realistic tests will be agreed.
In collaboration with WP5, partners some tests will be programmed in neutron workplaces, the need of the two cameras is foreseen when larger surfaces have to be monitored.

4 Conclusions

We have set up two indoor position systems (IPS) to track monitored people.

The first system is based on the use of a Kinect 2.0 depth sensor camera with an adapted software to reduce some of the limitations highlighted in D9.103 deliverable. Mainly the new proposal reduces jittering by incorporating a filtering algorithm based on based on Holt Double exponential method as described in 2.1.3, and improves the identification of workers by introducing several specific methodologies, depending on the workplace. In interventional radiology, at the moment an algorithm using skeleton joints data is applied. This one Kinect system has been selected to be used in the foreseen measurement campaigns within WP4 and WP5, because of its ease of use and installation, as well as its price, and previous experience from the preliminary tests in hospitals performed during the first year project.

The second system is a two-camera solution based on the use of two Kinect 2.0 depth sensor cameras and an adapted software that is capable to fuse the images of the two cameras and thus reduce occlusion problems and increase the field of view of the cameras. The development of this system is still in progress and will be compared to the first one for some workplaces, in particular, in those cases

where a longer range of person tracking is required or where occlusions are not satisfactorily solved by the proposed improved single camera solution. This solution is more expensive since it needs as many computers as cameras. Further, this extra hardware can compromise their installation in some interventional rooms with low free space. It is planned to complete the development of this multi-camera option and to test it both in lab and in realistic workplaces, so that by the end of the project we can recommend the best solution depending on the needs.

5 REFERENCES

[1] D9.103 –An IPS based on an infrared reflection time-of-flight sensor camera together with the corresponding software. Podium (2018) In: https://podium-concerth2020.eu/deliverables/.

[2] George E.P. Box [et al.]. Time series analysis: forecasting and control. Wiley, cop, ISBN: 9781118675021. (2016)

[3] Skeletal Joint Smoothing White Paper, Kinect for Windows 1.5, 1.6, 1.7, 1.8, by Mehran Azimi, Advanced Technology Group

[4] Holt, C.C. (1957). Forecasting trends and seasonals by exponentially weighted averages, Carnegie Institute of Technology, Pittsburgh ONR memorandum no. 52.

[5] Alavi, A., Yang, Y., Harandi, M., Sanderson, C.: Multi-shot person re-identification via relational stein divergence. In: Image Processing (ICIP), 2013 20th IEEE International Conference on. pp. 3542{3546. IEEE (2013).

[6] Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 144{151 (2014)

[7] Han, F., Reily, B., Ho_, W., Zhang, H.: Space-time representation of people based on 3d skeletal data: A review. Computer Vision and Image Understanding (2017).

[8] Haque, A., Alahi, A., Fei-Fei, L.: Recurrent attention models for depth-based person identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1229{1238 (2016).

[9] Barbosa, I., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: Computer Vision{ECCV 2012. Workshops and Demonstrations. pp. 433{442. Springer (2012).

[10] Munaro, M., Ghidoni, S., Dizmen, D.T., Menegatti, E.: A feature-based approach to people identification using skeleton keypoints. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on. pp. 5644{5651. IEEE (2014).

[11] Ceseracciu, Elena & Sawacha, Zimi & Cobelli, Claudio. Comparison of Marker-less and Marker-Based Motion Capture Technologies through Simultaneous Data Collection during Gait: Proof of Concept. PloS one. 9. e87640. 10.1371/journal.pone.0087640. (2014).

[12] Cao, Zhe & Simon, Tomas & Wei, Shih-En & Sheikh, Yaser. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. 1302-1310. 10.1109/CVPR.2017.143. (2017).

[13] Carraro, Marco & Munaro, Matteo & Burke, Jeff & Menegatti, Emanuele. Real-time marker-less multi-person 3D pose estimation in RGB-Depth camera networks. (2017).

[14] Besl, Paul & McKay, H.D.. A method for registration of 3-D shapes. IEEE Trans Pattern Anal Mach Intell. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 14. 239-256. 10.1109/34.121791. (1992).

[15] Yeung, Kwok-Yun & Kwok, Tsz Ho & Wang, Charlie. Improved Skeleton Tracking by Duplex Kinects: A Practical Approach for Real-Time Applications. Journal of Computing and Information Science in Engineering. 13. (2013).