To appear in *Computer Methods in Biomechanics and Biomedical Engineering*
Vol. 00, No. 00, Month 20XX, 1–23

GUIDE

Stereo Vision System for Skeleton Joint Recognition in a Rehabilitation Context

Ana Clara Matos[a], Teresa Azevedo Terroso [bc*], Luis Corte-Real [ab] and Pedro Carvalho [b]

[a] *FEUP - Faculdade de Engenharia da Universidade do Porto, Portugal*; [b] *INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência, Portugal*; [c] *ESMAD, IPP - Escola Superior de Media Artes e Design, Instituto Politécnico do Porto, Portugal*

(*November 2017*)

The present demographic trends point to an increase in aged population and chronic diseases which symptoms can be alleviated through rehabilitation. The applicability of passive 3D reconstruction for motion tracking in a rehabilitation context was explored using a stereo camera. The camera was used to acquire depth and color information from which the 3D position of predefined joints was recovered based on: kinematic relationships, anthropometrically feasible lengths and temporal consistency. Finally, a set of quantitative measures were extracted to evaluate the performed rehabilitation exercises. Validation study using data provided by a marker based as ground-truth revealed that our proposal achieved errors within the range of state-of-the-art active markerless systems and visual evaluations done by physical therapists. The obtained results are promising and demonstrate that the developed methodology allows the analysis of human motion for a rehabilitation purpose.

Keywords: rehabilitation, stereo vision, human motion tracking, biomedical engineering

1. Introduction

The recent shift towards preventive, proactive and continuous care led to an increased need for intensive rehabilitation (Brennan and Barker 2008). Human motion tracking (HMT) for rehabilitation has been an active research topic (Zhou and Hu 2008). By providing adequate feedback and guidance, HMT systems could potentiate the proper performance of the rehabilitation exercises and increase the patient accountability and motivation. Also, the identification and correction of errors in the exercises by the clinician could enable the modification of the prescribed exercises and thus minimize unneeded trips to outpatient centers.

Currently, HMT systems are accomplished by using motion-sensor technologies. The gold standard are marker-based systems (Bonnechère, Jansen, et al. 2014), providing high accuracy, but are quite expensive. Furthermore, the markers placement needs to be performed by a specialist and is time consuming. Also, the analysis must be performed in specialized centers (Galna, Barry, et al. 2014). For the stated reasons, the lack of portability and easiness to use makes them unsuited for a home context.

The development of efficient, affordable, compact and easy to use three-dimensional (3D) acquisition sensors boosted their application for motion tracking in rehabilitation. The three main technologies currently used are: structured light, time-of-flight and stereo cameras (González-Ortega, Díaz-Pernas, et al. 2014). The most commonly used sensors are active (González-Ortega, Díaz-

*Corresponding author. Email: teresaterroso@esmad.ipp.pt

Pernas, et al. 2014) whereas the advantages of passive sensors, like higher quality depth images, less sensibility to illumination changes, portability and ease of use, have remained mostly unexplored.

This paper describes a proposal for the detection and tracking of the human body and its parts using a stereo camera, assessing its applicability in opposition to the commonly used active devices. Colour and depth information acquired by the camera were combined to obtain a 3D representation of the human body; this representation was then delivered to a skeleton tracking algorithm able to recognize a predefined set of skeleton joints. The used algorithm is based on the work developed by (Buys, Cagniart, et al. 2014) and was improved through the combination of biomechanical and temporal constraints. From the skeleton positions, clinically significant measures in the context of rehabilitation were extracted.

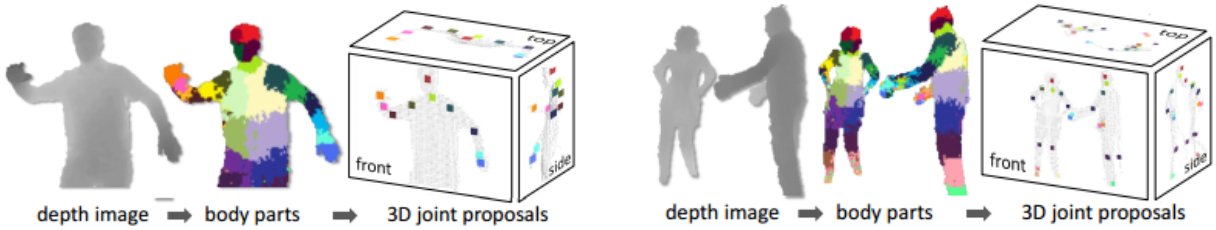
For accuracy evaluation and, given the absence of available datasets with annotated ground-truth for motion tracking with a stereo system, a dataset was also collected. The dataset contains the performance of three rehabilitation exercises by a male and a female. The database comprises color, depth and skeleton data acquired with the Microsoft® Kinect, stereo images acquired with the Bumblebee2 stereo camera and ground-truth data provided by a marker based system (Qualysis). According to the author knowledge this is the first created database that includes both depth, stereo and marker based information in a context of motion tracking evaluation for rehabilitation. The dataset is available by e-mail request to the corresponding author.

This paper is organized as follows. Section 2 presents some related work on 3D human reconstruction and motion tracking. In section 3 the proposed method for the analysis of human motion in a rehabilitation context is described in detail. Obtained results and correspondent discussion are presented in section 4. Main conclusions and some future improvements are provided in section 5.

2. Related Work

In rehabilitation, HMT systems should generate real-time data to dynamically represent the position changes of a human body (or portion of it). Visual tracking systems take advantage of optical sensors to improve the accuracy in pose estimation and can be divided into marker and marker free systems. Visual marker based systems (MBS) follow the human movement by using cameras and identifiers located on the human body. During body movement each body part performs its own motion trajectory with high degrees of freedom (DoFs) (Zhou and Hu 2008). MBS systems solve this problem by minimizing the ambiguity in the subject's movements. However, they are expensive and placing the markers can become a burdensome task. Moreover, data can be affected by noise due to the movement of skin under the markers or the marker itself. As well, one of the main problems of MBS is reproducibility due to the variation of marker placement between sessions and the identification of the standard bony landmarks. Also, image acquisition when using MBS is often limited to a specific laboratory setting which constitutes a major drawback in a rehabilitation application (Zhou and Hu 2008; Bonnechère, Jansen, et al. 2014; Yang, Christiansen, et al. 2014; Obdržálek, Kurillo, et al. 2012). Markerless systems (MLS) can overcome most of the aforementioned problems by relying on the information of the visual sensors and result in a less restrictive system. Nowadays cameras are relatively inexpensive, portable and non-obtrusive, which is a very important feature in a rehabilitation setup (Zhou and Hu 2008; Bonnechère, Jansen, et al. 2014).

Shotton et al. (Shotton, Fitzgibbon, et al. 2011) developed a human pose recognition method from a single depth image divided into two stages: body part labelling and 3D joint position estimation, Figure 1. First a dense probabilistic body part labelling was done using a segmented depth image. The labelling was accomplished using per pixel classification based on Randomized Decision Forests (RDF) trained using a large database of synthetic depth images. Then, mean-shift was used to find the spatial modes of each part distribution resulting in confidence-weighted proposals for the 3D locations of each skeletal joint.



(a) (b)

Figure 1.: Single (a) and multiple (b) human pose recognition, using a single depth image. From the depth image a per-pixel body distribution is inferred. Then, high-quality 3D proposals for the locations of each body joint are obtained by estimating local modes. From (Shotton, Fitzgibbon, et al. 2011).

Estimating the joint position using mean-shift has some drawbacks: the size and shape of the subject deeply influences the joint position estimation; the relative information obtained is related to the body surface, whereas joints are localized inside body parts. To surpass these limitations, a new 3D body pose recovery approach based on Principal Direction Analysis (PDA) of recognized human body parts from a series of depth images was developed (Dinh, Lim, et al. 2014), Figure 2. First, trained RDF were used to identify the human body parts within a synthetic training database. The recognized body parts were used to estimate the principal direction vectors using PDA. Finally, the 3D human body pose was recovered by mapping the directional vectors to each body part of the 3D model, that used a kinematic chain with predefined DoFs for each joint. Overall results were more robust and revealed that it was able to deal with sequences of unconstrained movements of persons with different sizes and shapes, Figure 3.

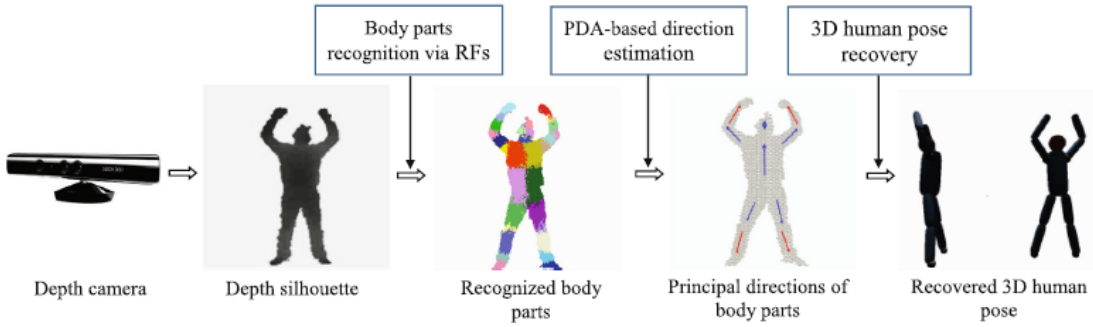


Figure 2.: Processing pipeline of the method proposed by Dinh et al. (Dinh, Lim, et al. 2014). Using as input a depth image without background the body parts are labelled and by applying PDA to the body parts the final 3D human body pose is recovered. From (Dinh, Lim, et al. 2014).

More recently, developers of the Microsoft® Kinect skeletal tracker proposed two enhanced algorithms. Girshick et al. (Girshick, Shotton, et al. 2011) performed the regression directly on the raw depth information, instead of on the body part labelled intermediate stage. The algorithm was able to estimate the position of occluded joints and results revealed that it outperformed state-of-the-art implementations, such as the one of Shotton et al. (Shotton, Fitzgibbon, et al. 2011). Taylor et al. (Taylor, Shotton, et al. 2012) extended the initial machine learning approach by estimating correspondences directly between images pixels and a 3D mesh model, Figure 4. This was accomplished by employing a regression forest in an energy minimization approach. As an additional contribution the authors proposed a more realist evaluation metric (number of fully correct frames), instead of the mean average precision used by (Shotton, Fitzgibbon, et al. 2011) and (Girshick, Shotton, et al. 2011).

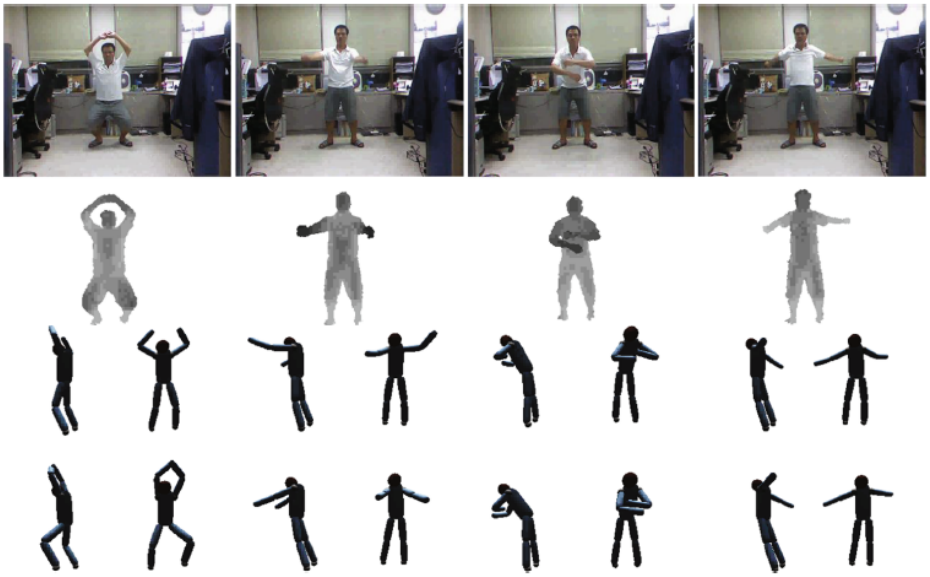


Figure 3.: Comparison between mean-shift and PDA algorithms for 3D human pose estimation. From top to bottom: RGB images, depth silhouettes, mean shift algorithm results and PDA algorithm results. From (Dinh, Lim, et al. 2014).

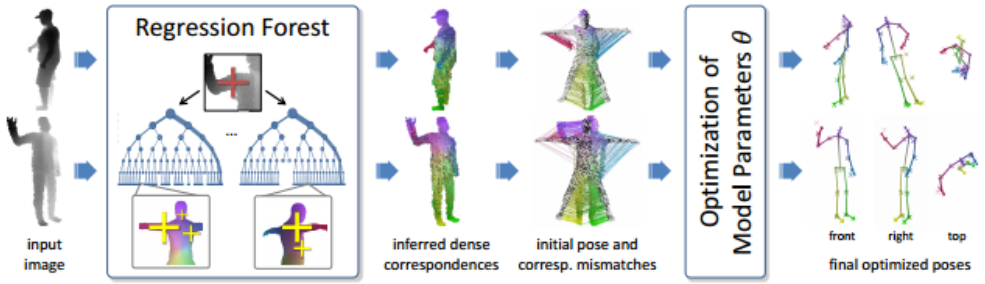


Figure 4.: Overview of the method proposed by Taylor et al. (Taylor, Shotton, et al. 2012). The correspondences are estimated directly between image pixels and a 3D mesh model. Without separate initialization or alternating minimization of pose and correspondence, a fast and reliable convergence to a good pose estimate can be obtained in a "single-shot". From (Taylor, Shotton, et al. 2012).

Zhou et al. (Zhou, Liu, et al. 2014) tried to overcome the difficulty of the method of Shotton et al. (Shotton, Fitzgibbon, et al. 2011) when dealing with severe occlusions. A probabilistic model based on Gaussian Process (GP) to reconstruct poses directly captured with a Microsoft® Kinect, Figure 5. Applying a GP based model allowed the use of a smaller training set. Results revealed that the system was able to deal with severe self-occlusion and outperformed the proposal of Shen et al. (Shen, Deng, et al. 2012).

3. Methodology

The developed system comprised two hierarchical stages, Figure 6. First, a stereo camera was used to acquire a 3D representation of the human body. Then, the obtained 3D representation was fed to the skeleton tracking system that recognized each skeleton joint of the given human body during the performance of a series of rehabilitation exercises.

(a) (b) (c)

Figure 5.: Postures from Microsoft® Kinect (left avatar) and their corresponding reconstructed poses (right avatar). The skeleton data presented in blue in the top two pictures was tracked from the Microsoft® Kinect results. (a) Rolling hands forward and backward; (b) bending leg; (c) Tai Chi motion. Adapted from (Zhou, Liu, et al. 2014).

Figure 6.: Overview of the general pipeline. Stage I: acquisition of a 3D scene representation using a stereo camera. Stage II: use of the acquired 3D representation to obtain the skeleton configuration of the subject.

3.1 *Image Acquisition*

Images were acquired using the stereo camera Bumblebee2 from Point Grey, placed at 2.7m to 3.5m from the subject performing a series of rehabilitation exercises, Table 1. From a biomedical perspective, these exercises were chosen since they are commonly used in a rehabilitation setting (Zhao, Espy, et al. 2014). From a computational perspective, they represent an increasing difficulty for a skeleton recognition system due to its growing complexity.

Table 1.: Rehabilitation exercises performed by the subjects during image acquisition.

Exercise	Description
1	Arm abduction and adduction in the coronal and sagittal planes.
2	Hip abduction and adduction in the coronal plane with the knee extended.
3	Toe touch: movement of the hands from the sides of the trunk in the direction of the toes.

3.2 *Human Body Reconstruction*

3.2.1 Point Cloud Generation

A 3D representation of the human body was obtained by combining the color and depth information provided by the stereo camera. To obtain a 3D model of the scene, stereo systems must deal with the correspondence and the reconstruction problems (Trucco and Verri 1998). Briefly, the first aims to determine the correspondent points in the different views, while the reconstruction problem uses

those correspondent points combined with the relative position between the views to obtain the 3D mapping of the scene.

First, the input images were normalized to reduce lighting differences and to enhance image texture. Correspondence between points in the two views were determined using the Semi-Global Block Matching (SGBM) algorithm (Hirschmuller 2008). The disparity computation was accomplished by a *winner takes it all* approach that was further refined by three post-processing steps: speckle filtering, consistency check and quadratic interpolation (Hirschmuller 2008).

After finding the correspondent points in the two views, the 3D scene was computed using triangulation, with the factory default calibration parameters, Table 2.

Table 2.: Camera calibration parameters of the Bumblebee2 stereo camera.

f_x	f_y	c_x	c_y
800.3968	800.3952	323.155	242.366

According to the pinhole camera model, the camera parameters can be summarized in the projection matrix that is used to estimate the world coordinates $P(X, Y, Z)$ from the pixel coordinates $p(u, v)$ (Jia, Yi, et al. 2012):

$$w \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{1}$$

where (f_x, f_y) are the pixel-related focal lengths representing the focal length f on the image coordinate system, (c_x, c_y) are the central points in pixel coordinates, s is the skew factor, w is the scaling factor, R is the rotation matrix and t is the translation vector. For the used stereo system, the pixels are considered to be squared and so the skew is zero. Also, the two views are parallel in the X axis, with a translation of 0.12 m (that corresponds to the stereo camera baseline), and with no rotation between then. The 3D point cloud of the scene was obtained by combining the disparity values with the camera's extrinsic and intrinsic parameters (Bradski and Kaehler 2008):

$$\begin{bmatrix} X/w \\ Y/w \\ Z/w \\ 1 \end{bmatrix} = Q \begin{bmatrix} u \\ v \\ d(u,v) \\ 1 \end{bmatrix} \Leftrightarrow \begin{bmatrix} X/w \\ Y/w \\ Z/w \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & c_x \\ 0 & 1 & 0 & -c_y \\ 0 & 0 & 0 & f \\ 0 & 0 & -1/t_1 & (c_x - c_x^{'})/t_1 \end{bmatrix} \begin{bmatrix} u \\ v \\ d(u,v) \\ 1 \end{bmatrix} \tag{2}$$

where $d(u, v)$ is the disparity value at the location (u, v) and Q is the perspective transformation matrix that represents the disparity-to-depth mapping.

3.2.2 Point Cloud Segmentation and Denoising

Due to the nature of the SGBM algorithm the obtained raw point cloud presented lateral noise, mainly located around the subject, resulting in noisy borders. To improve the raw point clouds, a segmentation stage was implemented, Figure 7. The subject position was estimated by applying the Otsu's binarization method on the disparity image (Figure 7a-b) (Otsu 1979). Since in a disparity image each pixel is inversely proportional to the distance from the camera, the histogram of the disparity image ideally presents two distinct peaks, for the subject and the background. With the Otsu's method the disparity value, located in the valley between those peaks, can be extracted. The presence of low texture, repetitive patterns, reflections, noise and occlusion can result in erroneous

disparities. To reduce them, an erosion morphological operation was applied to the binary image obtained from the Otsu's method (Figure 7b-c). The black black pixels of the binary image were marked as foregroundand black and the black ones as probable foreground. This information as provided as input data together with the RGB image (from the left image of the stereo vision system) to proceed with the segmentation by using the GrabCut method (Rother, Kolmogorov and Blake 2004). Briefly, this method is based on a color Gaussian Mixture Model and an iterative energy minimization is optimized to model the foreground and background. As a result, the binary mask of Figure 7d is obtained. However, as incorrectly labelling foreground pixels as background may occur, an hole filling methodology was applied. Also, assuming that the object of interest (in this case the subject) was the bigger blob, all the other smaller blobs were removed (Figure 7f). The final mask was used to obtain the refined RGB segmented subject presented in Figure 7h that was projected to 3D (Figure 7i).

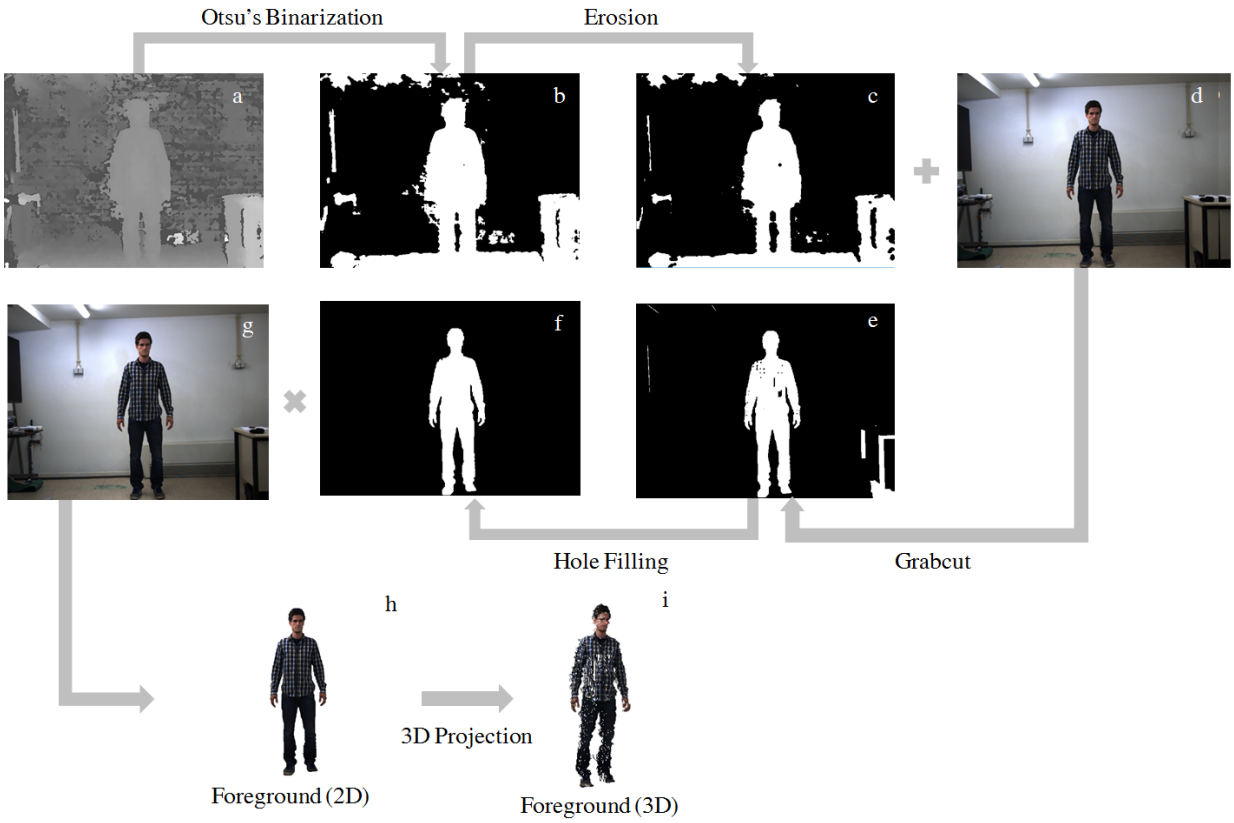


Figure 7.: Foreground segmentation pipeline.

In textureless backgrounds, the SGBM algorithm performs poorly on the background image areas. To improve the quality of the background point cloud a plane fitting method was followed using RANdom SAmple Consensus (RANSAC) supported by the calculation of surface normals (Rusu, Marton, et al. 2008). After the normals calculation, and considering that the goal was to estimate a plane, the RANSAC was implemented in the following way:

(1) the algorithm began by randomly selecting three points from the input cloud and calculating the correspondent plane parameters;
(2) according to a given threshold (empirically chosen as 0.15 for the wall plane and 0.10 for the floor plane), all the 3D points from the original cloud that belonged to the calculated plane were selected;
(3) steps 1 and 2 are repeated N times (here N was set to 100). In each iteration the previous

result was compared to the new one, by estimating the error of the inliers in relation to the model, and replaced if better. Or until a confidence of 99% was found.

Since the RANSAC algorithm was supported by the computation of normals a second threshold was considered in step 2. This threshold sets the relative weight (between 0 and 1) to give to the angular distance (0 to $\pi/2$) between point normals and the plane normal. Here, this threshold was empirically considered to be 0.1. The output of the RANSAC plane model fitting was a set of inlier plane points and plane parameters that represented each of the fitted planes. The output was used to produce the final background cloud by projecting the inlier points to the corresponding plane. Points of the initial raw cloud, that were not in the inlier plane points returned by RANSAC, were projected to the wall plane. For each point is done based on the following relation:

$$ax + by + cz + d = 0 \Leftrightarrow z = \frac{-d}{\frac{a(u-c_x)}{f_x} + \frac{b(v-c_y)}{f_y} + c} \tag{3}$$

where (u, v) are the 2D pixel coordinates, (x, y, z) the 3D point coordinates, a, b, c and d are the plane coefficients, f_x and f_y are the pixel-related focal lengths and (c_x, c_y) are the central points in pixel coordinates. The x and y coordinates are then calculated according to the following equations:

$$x = \frac{(u - c_x) * z}{f_x} \tag{4}$$

$$y = \frac{(v - c_y) * z}{f_y} \tag{5}$$

Before combining background and foreground into a single point cloud, foreground information was smoothed using a bilateral filter (Paris and Durand 2009), which removed small and weakly correlated differences between pixel values caused by noise, while preserving the edges. The filter was tested using different parameters, varying the spatial kernel (σ_s, from 1.0 to 15.0) and the range kernel (σ_r, from 0.01 to 5.0), independently. Results (not shown) were assessed through visual comparison and the parameters that resulted in an adequate compromise between smoothing without loss of detail were $\sigma_s = 5.0$ and $\sigma_r = 0.1$.

Finally, since the relative position of the 3D pixels is preserved during the foreground segmentation pipeline, the final point cloud can be obtained by merging the smoothed foreground with the refined background. The obtained final 3D point cloud served as input for the skeleton tracking system, Figure 8.

Figure 8.: Combination of the foreground, after the filtering process, with the plane fitted background.

3.3 *Human Pose Estimation and Motion Tracking*

In a rehabilitation setting, the markerless motion capture problem aims to extract clinically relevant information from the patient while he performs a set of prescribed exercises.

Our approach for skeleton tracking was based in (Buys, Cagniart, et al. 2014). The system developed by (Buys, Cagniart, et al. 2014) uses as input point clouds provided by the Microsoft® Kinect camera, but can be extended to use as input point clouds provided by other types of cameras, such as a stereo camera. Briefly, the pixel-wise body part labelling was accomplished by training a RDF classifier that is able to attribute body part labels to each image pixel. To extract a valid skeleton, the ensemble of pixel labels must be clustered into a smaller set of body part proposals. The clustering was accomplished by a breadth-first search over all the connected pixels with the same label within a given distance threshold in 3D.

3.3.1 Joint Positions Correction

Considering the implementation described above, the returned joints' position was not stable. Therefore, we propose several algorithms to revise and correct them. They were developed and implemented given the used kinematic tree, Figure 9. This orderly correction was performed since, in most cases, joint position was based on the previous joint in the kinematic chain. A joint position was corrected when its distance to its parent joint was not anthropometrically valid. Due to the variability of the human body (age, race, sex, among others) and also since the used information was dependent on the person height (that was automatically calculated and so it is prone to errors), a confidence level of ± 30% (empirically determined) was considered for the assessment of the distance between joints. Anthropometrically valid distances were obtained by the anthopometric data provided by (Drillis, Contini and Bluestein 1964) and are presented in Table 3. Despite being a standardized dataset, the used body segment lengths have an associated error (not documented by its authors) and for that reason the proposed confidence value was considered to be a good compromise. The lengths provided by Table 3 were used as guidance for the anthropometrically valid thresholds mentioned on the following correction and evaluation algorithms.

Table 3.: Anthropometrically feasible lengths between the body parts/joint locations and its children expressed as a percentage of total height (Drillis, Contini and Bluestein 1964). Example of interpretation: Lshoulder has two children, Larm and Lchest, which should be located at a distance of 0.160H and 0.095H meters, respectively. H – Height, L - Left, R - Right, B - Bottom, T - Top.

Father	**Child**				**Father**	**Child**			
	1st	**2nd**	**3rd**	**4th**		**1st**	**2nd**	**3rd**	**4th**
Lfoot	—	—	—	—	**Rhand**	—	—	—	—
Lleg	0.145	—	—	—	**Larm**	0.100	—	—	—
Lknee	0.123	—	—	—	**Lelbow**	0.073	—	—	—
Lthigh	0.123	—	—	—	**Lforearm**	0.137	—	—	—
Rfoot	—	—	—	—	**Lhand**	—	—	—	—
Rleg	0.145	—	—	—	**faceLB**	0.064	—	—	—
Rknee	0.123	—	—	—	**faceRB**	0.064	—	—	—
Rthigh	0.123	—	—	—	**faceLT**	—	—	—	—
Rhips	0.125	—	—	—	**faceRT**	—	—	—	—
Lhips	0.125	—	—	—	**Rchest**	0.210	—	—	—
Neck	0.080	0.080	0.085	0.085	**Lchest**	0.210	—	—	—
Rarm	0.100	—	—	—	**Lshoulder**	0.160	0.095	—	—
Relbow	0.073	—	—	—	**Rshoulder**	0.160	0.095	—	—
Rforearm	0.137	—	—	—					

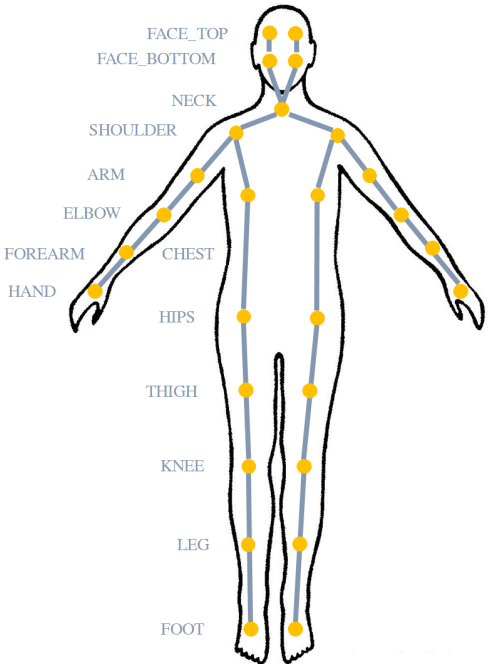


Figure 9.: Used kinematic skeleton model. The arrows indicate the order in which a parent is connected to its child. Adapted from Buys, Cagniart, et al. (2014).

Elbow Calculation and Correction

Our proposal for the elbow calculation stage (Algorithm 1) was always considered as a mandatory joint correction step, even when the elbow blob was found, since, due to its the small dimensions, most times its position was not accurate, Figure 10. On the top of the calculation, a correction (Algorithm 2) was introduced to ensure that the arm to elbow and elbow to forearm distances were anthropometrically valid.

Algorithm 1 Elbow Calculation

if elbow blob missing **then**
  Get point P from the arm blob with the maximum distance from shoulder
  Fuse arm with forearm blob. Remove all points outside a search square of 5cm around P point
  Elbow position = centroid of resulting blob
else if elbow blob is found **then**
  Initial elbow position = centroid of elbow blob
  Fuse arm, forearm and elbow blobs.
  Select all points on fused blob within a 5cm search square around the initial elbow position proposal.
  Elbow position = centroid of resulting blob
end if

Algorithm 2 Elbow Correction

while arm to elbow and elbow to forearm distances are invalid **do**
  if elbow to forearm distance > threshold OR arm to elbow distance < threshold **then**
    add lower points (belonging to the arm-forearm blob) to blob used for centroid calculation
  else if elbow to forearm distance < threshold OR arm to elbow distance > threshold **then**
    add upper points (belonging to arm-forearm blob) to blob used for centroid calculation
  end if
end while

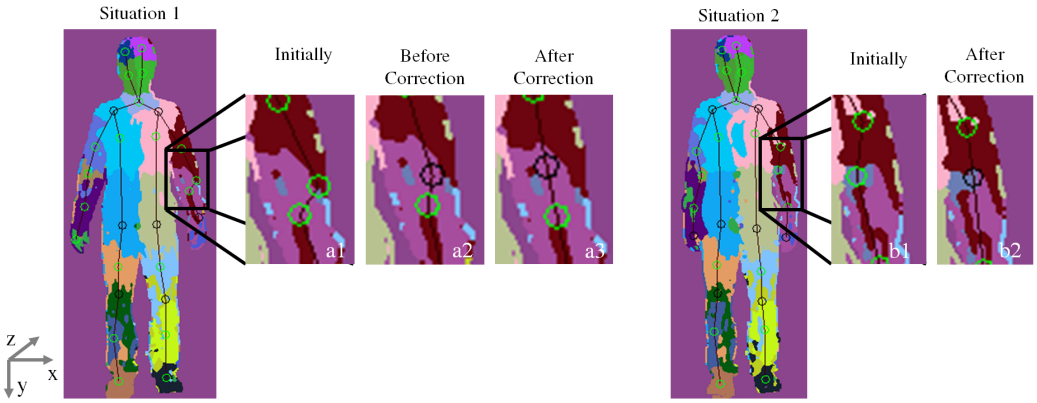


Figure 10.: Situation 1: elbow position correction when the elbow blob is missing. Situation 2: elbow position correction when the elbow blob is found. Black circles represent the inferred joints.

Hand Calculation and Correction

Due to the size of the hand blob, its labelling was noisy and often not found or misplaced, particularly when the hands were closed or sideways. To prevent this, an evaluation followed by a correction were developed and enforced, Figure 11. The evaluation (Algorithm 3) helped to remove hand positions that were not valid in relation to previous joints (in the kinematic chain), such as the elbow and the forearm. The correction (Algorithm 4) provided a joint position proposal based on kinematic relationships when the initial proposal was not valid.

Algorithm 3 Hand Evaluation

if Criteria 1: forearm to hand distance >= threshold OR Criteria 2: given a hand candidate (by adding elbow-forearm vector to forearm position), the distance between hand candidate and hand proposal > 20cm **then return** hand position proposal not accepted
end if

Algorithm 4 Hand Correction

Fuse forearm and hand blobs (if hand position rejected based on Criteria 2, only forearm blob is considered).
Remove all points above forearm center.
Get point P with maximum distance from forearm center within the new forearm blob.
Select all points around P within a 10cm search square.
Hand position = centroid of resulting blob.

Thigh Correction

Thigh correction was introduced when the hip to thigh distance was above or under an anthropometrically valid threshold (Algorithm 5). This correction helped to prevent situations in which the left and right thighs were not parallel in the Y dimension in a standing position with static legs, due to an incorrect labelling, Figure 12.

Algorithm 5 Thigh Correction

Set limit as double (Limit 1) or half (Limit 2) of the hip to thigh standard distance.
Remove all points of thigh blob that are above Limit 1 or under Limit 2.
Thigh position = centroid of resulting blob.

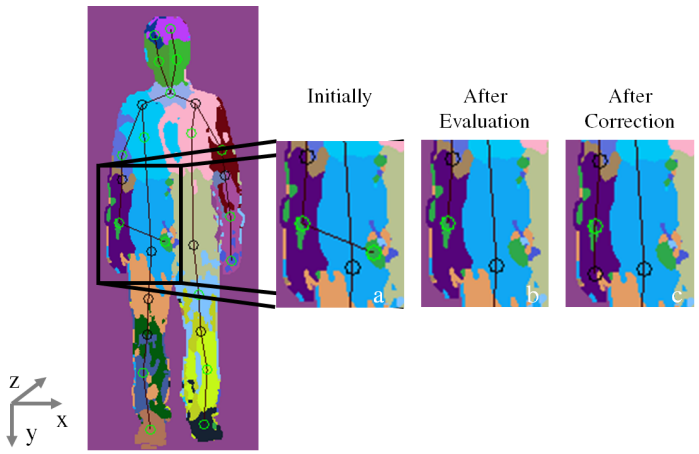


Figure 11.: Hand position evaluation and correction. Black circles represent the inferred joints.

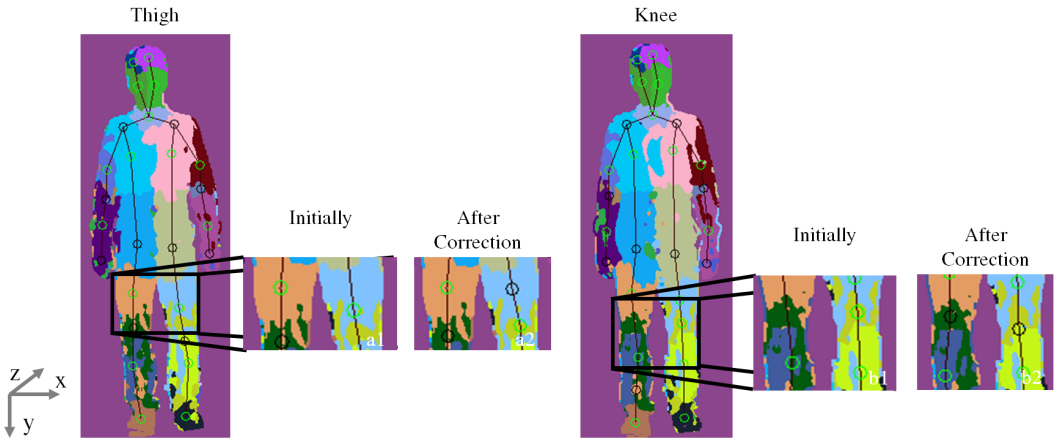


Figure 12.: Thigh and knee positions before and after the correction. Black circles represent the inferred joints. After the correction, thighs and knees positions are parallel to each other, like expected in a standing position.

Knee Correction

Due to the small size of the knee blob, its position was often severely misplaced or at least, its location tended to be noisy. Therefore a correction algorithm was implemented (Algorithm 6).

Algorithm 6 Knee Correction

if thigh to knee distance $>$ threshold **then**
  Knee candidate K = thigh center + thigh to knee standard distance
  Remove all points from thigh blob above thigh center and below K.
  Get point P with maximum distance from thigh center within the thigh blob.
  Fuse thigh, knee and leg blobs.
  Using entire leg blob, select all points around P within a 5cm search square.
  Knee position = centroid of resulting blob.
end if

Leg Correction

The leg correction was enforced in the case of position misplacement, due to an incorrect labelling. Two situations were covered (Algorithm 7): when the foot blob was correctly placed (Situation 1) the leg position was recovered based on the knee and foot positions; when the obtained foot blob was not anthropometric valid (situation 2), the leg position was determined based on the knee and leg positions. Figure 13 illustrates the two aforementioned situations and the corrected result.

Algorithm 7 Leg Correction

if Situation1: knee to foot distance is valid **then**
  Leg candidate L = mean point of knee to foot vector
  if leg proposal is located above a set search distance of L **then**
    leg proposal is replaced by L
  else
    leg proposal remains unchanged
  end if
else if Situation2: foot blob is not valid **then**
  Get point P with maximum distance from knee within leg blob or knee blob (if leg blob is not valid)
  Select all the points around L within a 10cm square
  Foot position = centroid of resulting blob
  Given foot and knee position, leg = mean point of knee to foot vector
end if

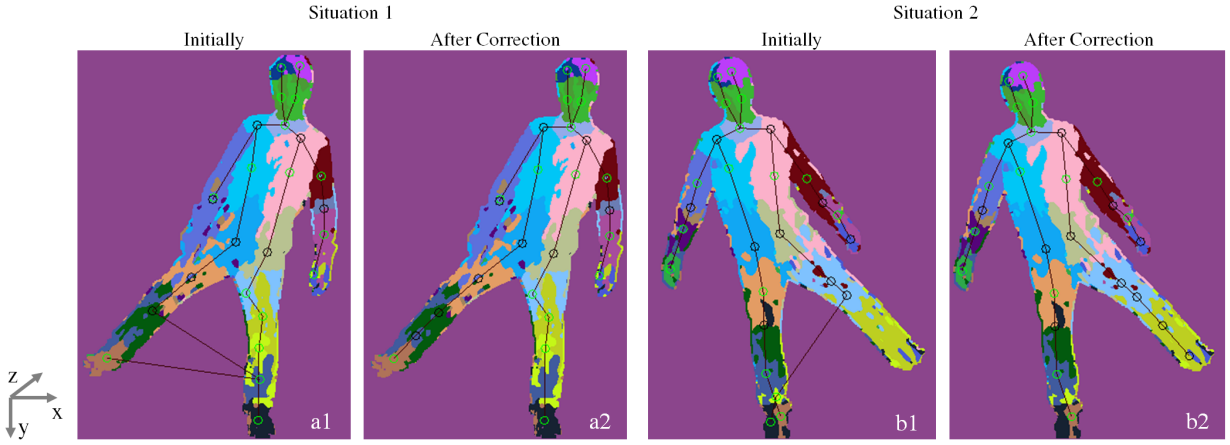


Figure 13.: Leg position before and after the correction for situations 1 and 2. Black circles represent the inferred joints.

3.3.2 Joint Positions Tracking

Given the absence of temporal constraints, results were prone to instability and jitter. To improve performance, individual Kalman filters were applied to the 27 retrieved joints. The state vector was considered to be the true 3D coordinates of the joints and their respective velocities, denoted as $x, y, z, \dot{x}, \dot{y}, \dot{z}$ (without the discrete time subscript t). As stated by the process model, the state

at time t evolved from the prior state at time $t-1$ according to:

$$X_t = AX_{t-1} + w_{t-1} \Leftrightarrow \begin{pmatrix} x_t \\ y_t \\ z_t \\ \dot{x}_t \\ \dot{y}_t \\ \dot{z}_t \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \triangle t & 0 & 0 \\ 0 & 1 & 0 & 0 & \triangle t & 0 \\ 0 & 0 & 1 & 0 & 0 & \triangle t \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \\ \dot{x}_{t-1} \\ \dot{y}_{t-1} \\ \dot{z}_{t-1} \end{pmatrix} + w_{t-1} \tag{6}$$

where A is the state transition matrix and w represents the normal distributed process noise with covariance Q. $\triangle t$ is the time step in seconds that was updated according to the time stamp of the acquired image frames.

The measurement model, that is composed by the 3D coordinates of each joint, relates the current state to the measurement Z with the matrix H. v is the normal distributed measurement noise with covariance R:

$$Z_t = HX_t + v_t \Leftrightarrow \begin{pmatrix} x_t \\ y_t \\ z_t \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \\ z_t \\ \dot{x}_t \\ \dot{y}_t \\ \dot{z}_t \end{pmatrix} + v_t \tag{7}$$

To avoid feeding the filter with incorrect measurements, a validation step was introduced to evaluate all joints in relation to their parents and discards those that were considered invalid, Figure 14. The validation was based on the anthropometric distance between joints, as described before. When a joint was considered invalid all the ones that follow on the kinematic chain were also set to invalid and the Kalman filter was updated by the prediction.

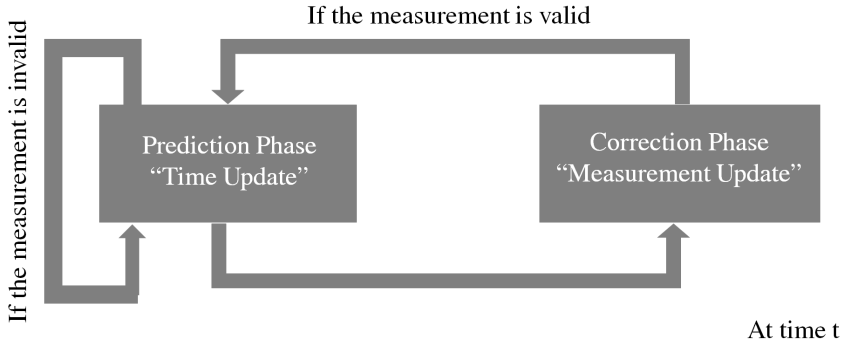


Figure 14.: Block diagram describing the adopted Kalman filtering methodology.

The filter was initialized with the estimates of the skeleton joint positions with zero initial velocities. To initialize the error covariance of the Kalman filter state vector, the joints were divided in three types, considering the amount of movement. The upper (except the hands) and the lower members were considered dynamic, since its position varied the most during the course of the exercises. The hands were considered to be a special case because they can move faster than the other joints. The torso and head joints were considered static, since their position remained mostly the same through the entire exercise duration. For this reason a higher value was attributed to the error covariance of the state vector for the hands, an intermediate for the dynamic joints and a lower one to the static joints. The error covariance of the state vector attributed to the three mentioned types of joints is a diagonal matrix with all entries equal to 0.0001, 0.000017, 0.0000029 respectively for

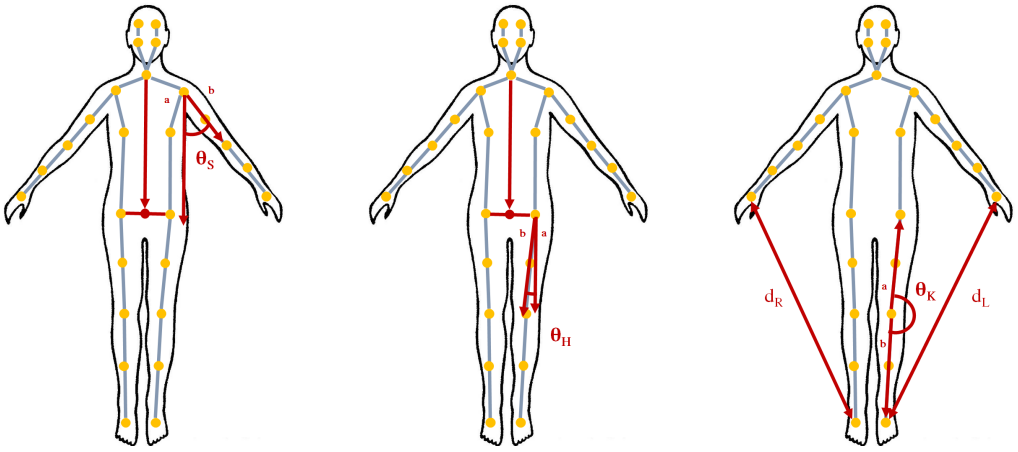
the hands, dynamic joints and static joints. The process noise covariance (Q) is considered to be a diagonal matrix with all entries equal to 0.005. The measurement noise covariance (R) is considered to be a diagonal matrix with all entries equal to 0.05. The mentioned values are empirically determined.

3.3.3 Range of Motion Calculation

From a medical perspective, it is important to provide quantitative and ready to use evaluation data from the obtained joints positions. The proposed rehabilitation exercises were evaluated by considering the Range Of Motion (ROM) of the main involved articulation, Figure 15. The aforementioned angles were calculated according to the following equation:

$$\theta = \arccos\left(\frac{x_a x_b + y_a y_b + z_a z_b}{(\sqrt{x_a^2 + y_a^2 + z_a^2}) \cdot (\sqrt{x_b^2 + y_b^2 + z_b^2})}\right) \tag{8}$$

where $a = (x_a, y_a, z_a)$ and $b = (x_b, y_b, z_b)$ are two vectors that form the angle of interest. For the shoulder angle calculation, a is the neck to hip center vector and b is the shoulder to elbow vector. For the hip angle calculation, a is the neck to hip center vector and b is the hip to knee vector. For the knee angle, a is the knee to hip vector and b is the knee to foot vector.



(a) Shoulder Angle (b) Hip Angle (c) Hand to Foot Distance

Figure 15.: ROM measurement of the main involved articulations of the proposed rehabilitation exercises.

3.3.4 System Validation

The accuracy of the proposed skeleton tracking system was evaluated using as ground-truth a MBS from Qualisys; it uses 12 cameras and tracks the spatial trajectories of the reflective markers on the subjects. Markers were placed in order to mimic the positions of the joints in the skeleton model used by the developed system. For that, 27 reflective markers were placed on predefined body positions as suggested in (Soltani and Vilas-Boas 2016). Two healthy subjects were instructed to perform the exercises described in Section 3.1. For each subject and exercise, 3 trials were performed.

Comparisons were performed based on the ROM of the main articulations involved in the proposed exercises. Based on the data provided by the MBS, the ground-truth ROM for each of the discrete timestamps of the markerless system was calculated using a spline interpolation.

4. Results and Discussion

The impact of the proposed segmentation and denoising methodology to enhance the 3D point clouds of the first stage of the pipeline is presented in Figure 16. As shown, the quality of the final point clouds was significantly improved. Besides solving the problem related with the lack of texture, the proposed plane fitting and segmentation approach projected all the points that do not belonged to the human body into the wall plane, improving the subject position determination in the subsequent skeleton tracking stage. In the presence of a more complicated background the human motion tracking task could have been hampered. With a cluttered environment, the segmentation stage would probably not be as efficient as with the images presented. Also, during the body part labelling stage objects in the background could be incorrectly identified as possible human bodies increasing the number of false positive detections.

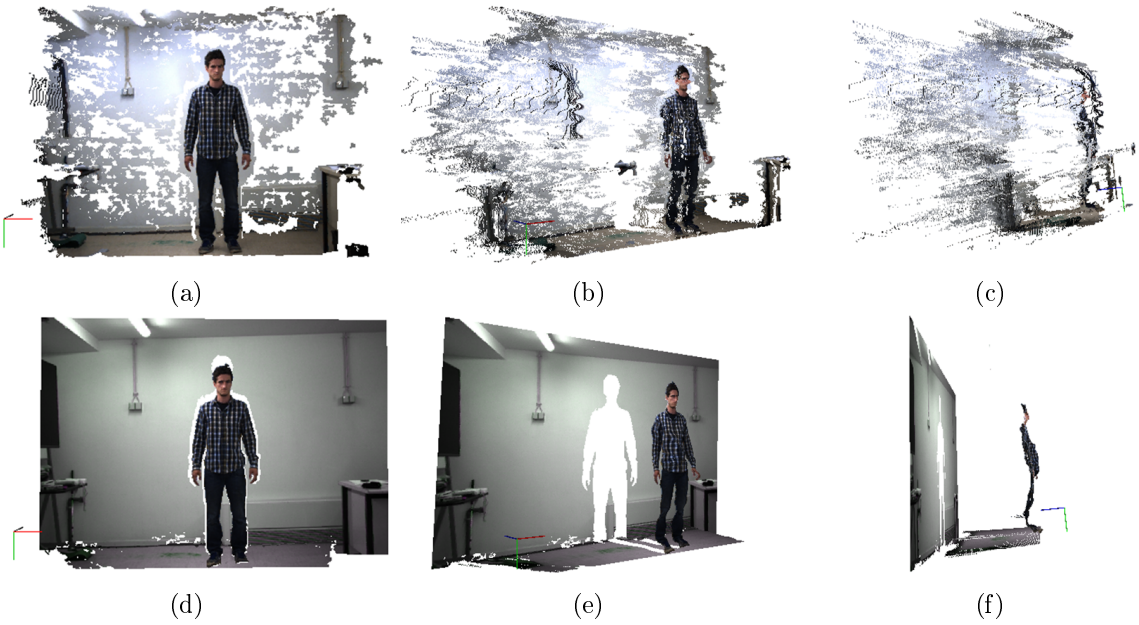


Figure 16.: Obtained 3D point clouds (a-c) before and (d-f) after segmentation and denoising. (a,d) Frontal view, (b,e) diagonal view and (c,f) lateral view.

The human body reconstruction accuracy was not determined. The quality of the obtained point clouds was only assessed through visual inspection. Nevertheless, it is expected that the accuracy of the reconstructed point clouds would influence the determination of the skeleton joints on the skeleton position estimation stage. The skeleton positions were obtained from the pixel wise body part labelling performed on the final point clouds. For that reason, if the quality of the final points clouds were to be very low that would have a negative impact on the body part labelling stage and hence on the accuracy of the obtained skeleton joints' position.

The sequences of images used present the subject in frontal position facing the camera. During the rehabilitation exercises the subject's direction remains unchanged. If the images were to include situations in which the subject's direction changes, namely during turnaround movements, the skeleton joints detection would be significantly compromised. During turnaround movements some parts of the subject's body would be occluded by others. Occluded joints position cannot be directly determined by the labelling stage and needs to be inferred by the position of non-occluded joints. If most of the joints are occluded the human pose cannot be correctly estimated.

The consistency in skeleton joints detection was assessed considering the kinematic and length relationships between the returned skeleton joints positions. The proposed correction algorithms

were evaluated considering the percentage of invalid joints (I):

$$I_i = \frac{1}{F}\sum_{f=0}^{F}\delta_{inv}, i = 0, ..., 26 \tag{9}$$

where i represents each one of the 27 retrieved joints, F is the total number of frames and $\delta_{inv} = 1$, if the correspondent joint position was considered to be invalid. A joint was considered to be invalid if the location proposed by the algorithm was not possible considering the anthropometrically feasible lengths presented on Table 3.

The average percentage of invalid joints for the evaluated exercises is presented in Figure 17. An overall decrease of invalid joints is visible after correction. Neck, shoulders and chest are the most stable joints. consistent with the fact that all skeleton proposals were built from the neck, with shoulders and chest being the child and grandchild joints, respectively. The hands position was also considerable improved, with an overall decrement of invalid retrieved joints from 52% to 13%.

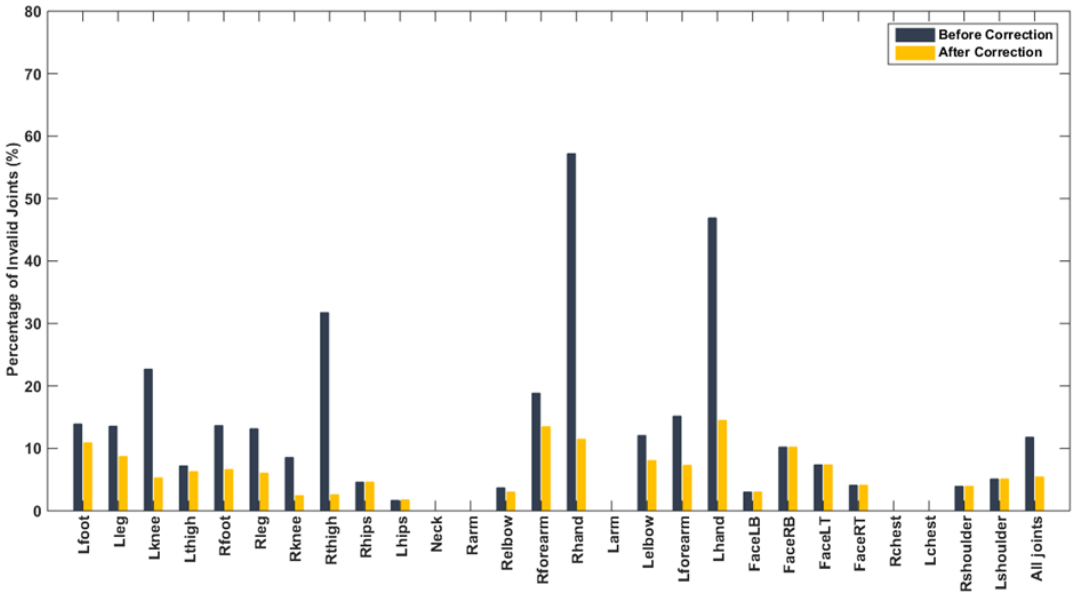


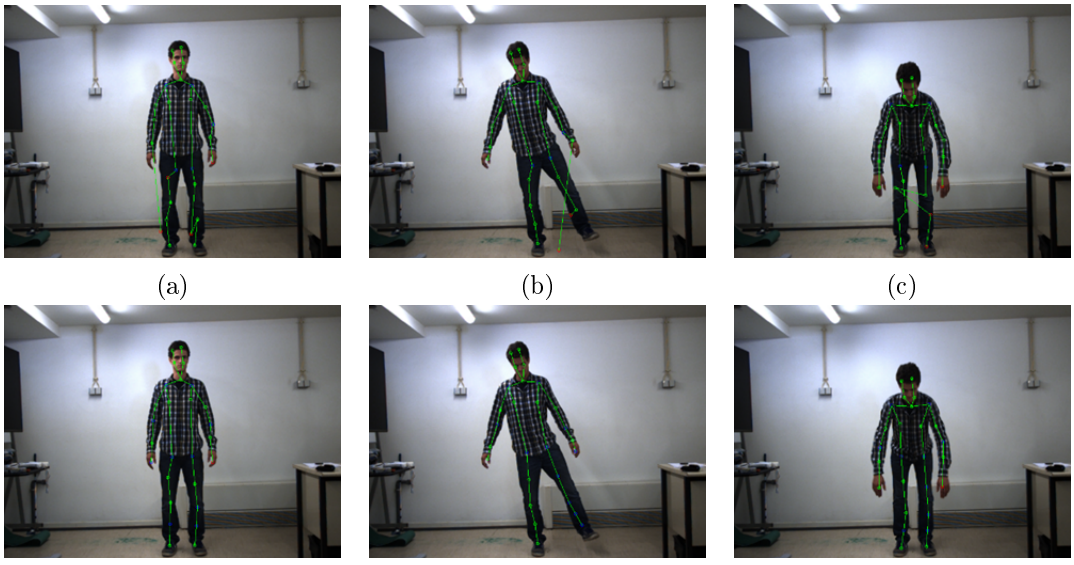
Figure 17.: Average of the percentage of invalid joints before (blue) and after (yellow) the implementation of the new correction algorithms for the all the image sequences.

The impact of the joint position correction stage on the overall consistency of the obtained skeletons is shown in Figure 18. As proved by the previous quantitative analysis, the implemented corrections improved the overall quality of the returned skeletons.

Figure 19 compares the trajectories (X, Y and Z world coordinates) of the raw and Kalman filter estimated data for selected joints, obtained during the joints position tracking stage described in subsection 3.3.2. The Kalman filter outputted smoother estimates whereas the unfiltered joint positions had more jitter. Also, the filter accurately predicted the joints position in the absence of measurement information.

The smoothness effect was quantified using the smoothness measure (S) from (Larsen, Hauberg and Pedersen 2011), calculated as the average deviation of all joints J over all frames F:

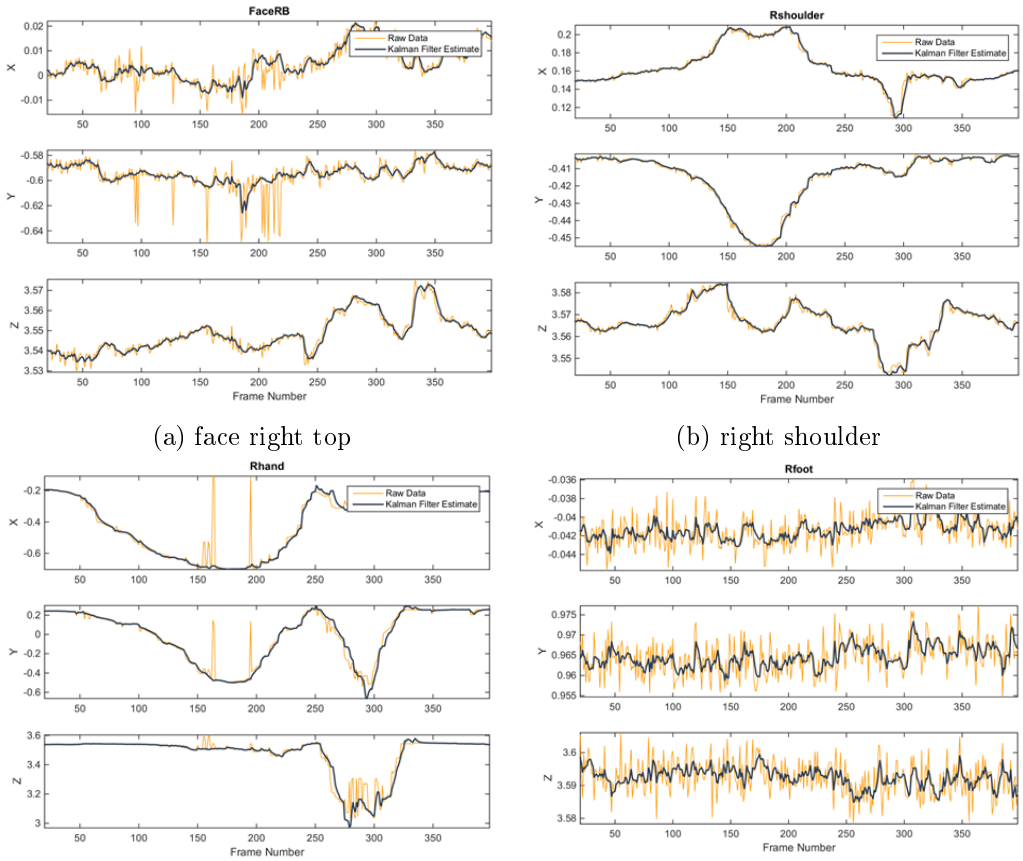
$$S(x_{1:T}) = \frac{1}{FJ}\sum_{f=0}^{F}\sum_{j=0}^{J} \parallel a_{f,j} - a_{f-1,j} \parallel \tag{10}$$



(a) (b) (c)

(d) (e) (f)

Figure 18.: Impact of the correction stage in the returned skeleton joint positions: (top) before and (bottom) after the correction. Blue circles are the inferred joints red circles are the invalid joints.



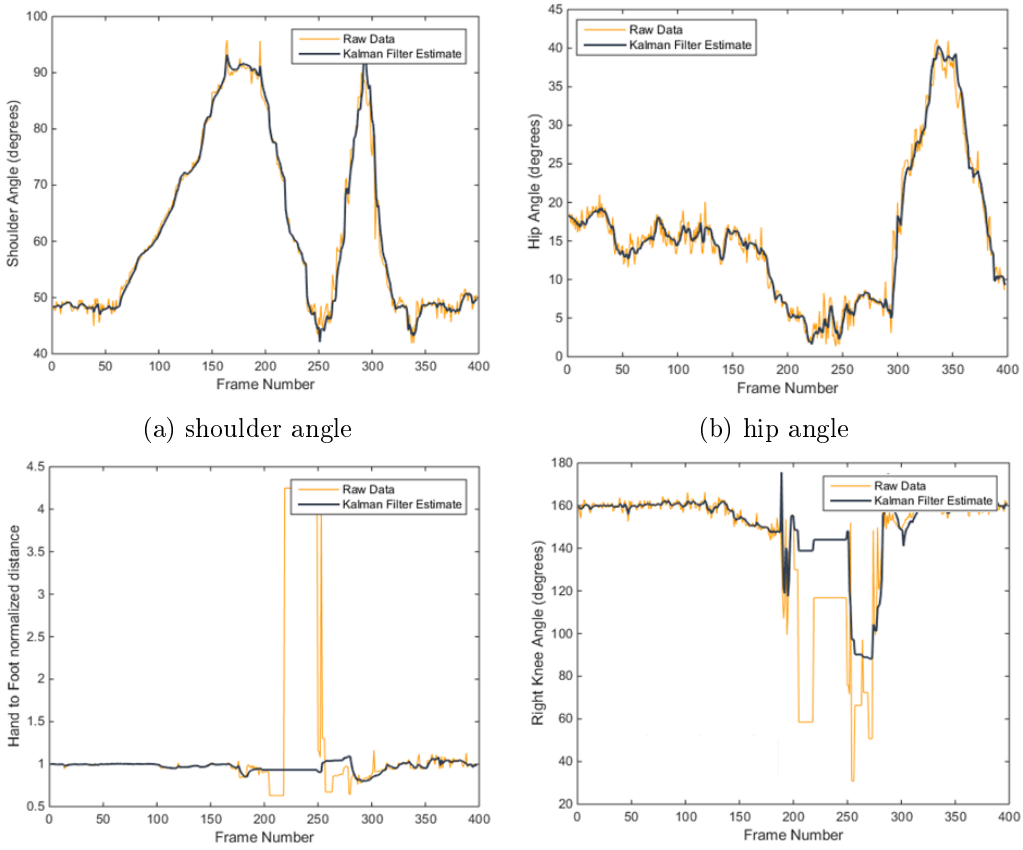
(a) face right top (b) right shoulder

(c) right hand (d) right foot

Figure 19.: Comparison of the Kalman filter estimate (blue) and the raw data (yellow), in meters, of selected joints trajectories during exercise 1. When the skeleton tracking system was not able to return a joint position, the coordinates were marked as -1.

where $a_{f,j}$ is the absolute position of joint j on frame f. Results of the smoothness measure for the joints positions obtained before and after the tracking stage showed, as would be expected, that the Kalman filter decreased the joint deviation by 35%.

The skeleton tracking system was evaluated considering proposed rehabilitation exercises (Figure 20). The system performance degrades with the increasing complexity of the evaluated exercises. Also, as expected, the kalman filter estimates present a smoother trajectory in comparison to the raw data. For the soulder and hip angles the skeleton tracking system was able to estimate the trend of the movement. However, for the third sequence movement the obtained results were not reliable. During the toe touch exercise the subjects were told to approximate the hands to the toe as close as possible, maintaining the knees extended. The exercise performance was evaluated by calculating the normalized hand to foot distance. This distance should be close to 1 in the beginning of the exercise, approximate to 0 as the hands touch the feet and then back to 1 as the subject returns to the initial position. The knee extension was assessed considering the knee angle, that should be close to 180^{o} during the entire exercise. The systems incapability to return stable results for this movement was due to severe occlusion of the hip and knee joints as the hands began to approximate the feet.



(a) shoulder angle

(b) hip angle

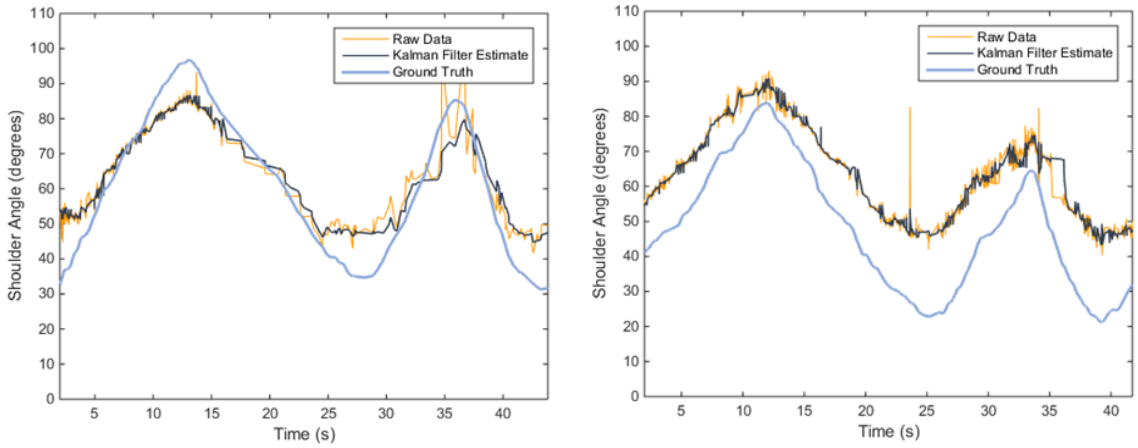
(c) hand to toe distance

(d) knee angle

Figure 20.: Samples of the range of motion evaluation for the three exercises. Results are present for both the raw data (yellow) and the Kalman filter estimate (dark blue).

As described in subsection 3.3.3, performance of the proposed markerless skeleton tracking system was evaluated from a medical perspective considering the ROM. Joint angle trajectories of the shoulder angle during the arm abduction and adduction are shown in Figure 21. Trajectories before and after the tracking stage and ground-truth present an evident correlation, since they are time synchronized and follow the same pattern. Nevertheless, it can be observed that the calculation of the shoulder angle in the sagittal plane (second peak) is more irregular than the one in the coronal

plane (first peak). When the shoulder angle is closer to 90°, the subject's hands may occlude the shoulders and the elbows which can contribute to the observed behavior.



(a) Left Shoulder Angle (b) Right Shoulder Angle

Figure 21.: Range of motion evaluation for the left and right shoulder angles during exercise 1.

Accuracy was evaluated considering the mean error, ME (Fernandez-Baena, Susin and Lligadas 2012):

$$ME = \frac{1}{F}\sum_{f=1}^{F}|ST_f - O_f| \tag{11}$$

where F is the number of frames, ST_f and O_f are the joint angles provided by the skeleton tracking system and the optical motion capture in frame f, respectively.

Figure 22 presents the mean errors for the shoulder angle calculation during the abduction and adduction of the left and right arm. Error was higher for the right shoulder angle of the female subject. This situation might be explained by an erroneous marker placement, since as shown in Figure 21b, the ground-truth trajectory presented a smaller correlation with the trajectories obtained from the tracking system and the observed behavior was consistent for the three trials. Excluding this case, the error was similar for both subjects. However, considering the small number of observations further comparisons using a larger number of subjects should be done in order to better support this observation. The implemented system performance was within the range of other results obtained using state-of-the-art active markerless systems. In (Fernandez-Baena, Susin and Lligadas 2012), a similar comparison was performed using the OpenNI, obtaining a range of errors in the shoulder angle calculation from 7° to 13°. When visually assessing the ROM, the physical therapist evaluation normally reports an error of 10° (Fernandez-Baena, Susin and Lligadas 2012), which is within the range of the overall error obtained by the proposed system ($9.42° \pm 5.02°$, mean of all trials).

Despite the higher accuracy provided by marker based systems (Windolf, Götzen and Morlock 2008), it should be noted that other factors may affect their accuracy and hence influence the comparison. These factors are related to the marker placement and soft tissue artifacts (Taylor, Ehrig, et al. 2005). Also, it is possible that slightly discrepancies of the orientation of the stereo camera in relation to the Qualysis system may have introduced additional error (Galna, Barry, et al. 2014).

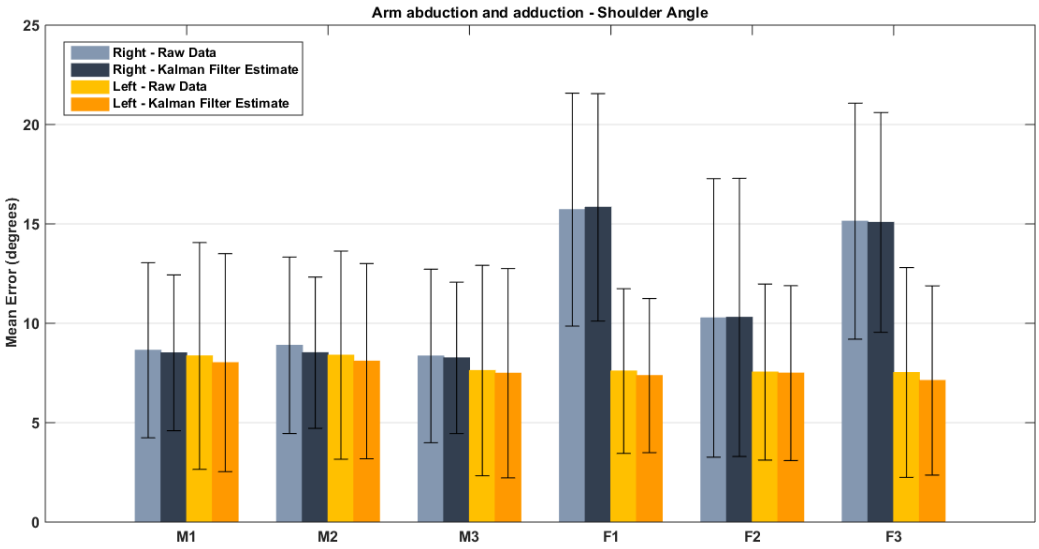


Figure 22.: Mean error (and standard deviation) of the shoulder angle calculation during exercise 1, for all three trials of both male (M) and female (F) subjects.

5. Conclusions and Future Work

Results reveal that the markerless system using a stereo camera reached errors within the range of other state-of-the-art active markerless systems and the average error reported by physical therapists. It is noteworthy that in the context of rehabilitation an extreme accuracy is not needed. In fact, many of the considered exercises are evaluated based on repetitions and in comparison to a motion pattern. Also, automatic systems remove the subjectivity inherent to the visual assessment done by a therapist. The obtained results were promising and proved that a passive sensor, such as a stereo camera, can be used in the context of motion tracking in rehabilitation.

The provided skeleton tracking system is device agnostic and so it can potentially be used with any type of RGB-D information, independently of the acquisition sensor. This is not the case for most of the commonly used motion tracking open-source algorithms. The developed system could be used to explore a wider variety of acquisition sensors.

Despite the potential of the obtained results, further developments can still be pointed. The used ground truth annotated model from which the RDF classifier was learned could be refined, for example by adding labels to areas of interest that were not contemplated by the current model. The training data could be adapted to better suit the purpose of rehabilitation. A dataset of users performing rehabilitation exercises could be used for training, which would considerably improve the recognition task. Many of the rehabilitation exercises used in the clinical practice are aided by the presence of common objects such as chairs, balls and elastics. However, the presence of this objects hampers the task of body recognition. The inclusion of some of this objects in the training dataset could overcome this limitation. Furthermore, a wider variety of subjects should be used, varying for example the gender, age, height, weight, hair and clothing, including as well subjects with amputations and physical handicaps. The proposed tracking methodology was implemented considering each joint individually; given the articulated nature of the human skeleton, the tracking outcome would substantially benefit from incorporating the kinematic relationship between each joint in the used measurement model.

Despite revealing that the system is able to reach errors within the range of state-of-the-art markerless systems and lower than the visual evaluation done by a physical therapist, a thorough validation study remains necessary. For that, the system's performance should be evaluated with a larger population, increasing the variability in sex, age and physical ability. Also, other clinically relevant joints should be considered for evaluation. This will allow a more consistent evaluation of

the proposed system.

6. Acknowledgements

The first author would like to thank LABIOMEP for the availability in providing the location for the acquisition of the marker based software used in the collected dataset.

References

Bonnechère B, Jansen B, Salvia P, Bouzahouene H, Omelina L, Moiseev F, Sholukha V, Cornelis J, Rooze M, Jan SVS. 2014. Validity and reliability of the kinect within functional assessment activities: Comparison with standard stereophotogrammetry. Gait & Posture. 39(1):593 – 598.

Bradski GR, Kaehler A. 2008. Learning opencv, 1st edition. 1st ed. O'Reilly Media, Inc.

Brennan DM, Barker LM. 2008. Human factors in the development and implementation of telerehabilitation systems. Journal of Telemedicine and Telecare. 14(2):55–58.

Buys K, Cagniart C, Baksheev A, Laet TD, Schutter JD, Pantofaru C. 2014. An adaptable system for rgb-d based human body detection and pose estimation. Journal of Visual Communication and Image Representation. 25(1):39 – 52.

Dinh D, Lim M, Thang N, Lee S, Kim T. 2014. Real-time 3d human pose recovery from a single depth image using principal direction analysis. Applied Intelligence. 41(2):473–486.

Drillis R, Contini R, Bluestein M. 1964. Body segment parameters; a survey of measurements and techniques. Artif Limbs. 8:44–66.

Fernandez-Baena A, Susin A, Lligadas X. 2012. Biomechanical validation of upper-body and lower-body joint movements of kinect motion capture data for rehabilitation treatments. In: Intelligent Networking and Collaborative Systems (INCoS), 2012 4th International Conference on; Sept. p. 656–661.

Galna B, Barry G, Jackson D, Mhiripiri D, Olivier P, Rochester L. 2014. Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson's disease. Gait & Posture. 39(4):1062 – 1068.

Girshick R, Shotton J, Kohli P, Criminisi A, Fitzgibbon A. 2011. Efficient regression of general-activity human poses from depth images. In: ICCV; October. IEEE.

González-Ortega D, Díaz-Pernas F, Martínez-Zarzuela M, Antón-Rodríguez M. 2014. A kinect-based system for cognitive rehabilitation exercises monitoring. Computer Methods and Programs in Biomedicine. 113(2):620 – 631.

Hirschmuller H. 2008. Stereo processing by semiglobal matching and mutual information. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 30(2):328–341.

Jia W, Yi W, Saniie J, Oruklu E. 2012. 3d image reconstruction and human body tracking using stereo vision and kinect technology. In: Electro/Information Technology (EIT), 2012 IEEE International Conference on. p. 1–4.

Larsen ABL, Hauberg S, Pedersen K. 2011. Unscented kalman filtering for articulated human tracking. In: Image analysis. Lecture notes in computer science; vol. 6688. Springer Berlin Heidelberg; p. 228–237.

Obdržálek , Kurillo G, Han J, Abresch T, Bajcsy R. 2012. Real-time human pose detection and tracking for tele-rehabilitation in virtual reality. Stud Health Technol Inform. 173:320–324.

Otsu N. 1979. A threshold selection method from gray-level histograms. IEEE Transactions on Systems,Man and Cybernetics. (9 (1)):62–66.

Paris S, Durand F. 2009. A fast approximation of the bilateral filter using a signal processing approach. International Journal of Computer Vision. 81(1):24–52.

Rother C, Kolmogorov V, Blake A. 2004. Grabcut -interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics (SIGGRAPH).

Rusu RB, Marton ZC, Blodow N, Dolha M, Beetz M. 2008. Towards 3d point cloud based object maps for household environments. Robotics and Autonomous Systems. 56(11):927 – 941.

Shen W, Deng K, Bai X, Leyvand T, Guo B, Tu Z. 2012. Exemplar-based human action pose correction and tagging. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on; June. p. 1784–1791.

Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A. 2011. Real-time human pose recognition in parts from a single depth image. In: CVPR; June. IEEE.

Soltani P, Vilas-Boas J. 2016. Muscle activation during exergame playing. IGI Global.

Taylor J, Shotton J, Sharp T, Fitzgibbon A. 2012. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In: Proc. CVPR; June. IEEE.

Taylor WR, Ehrig RM, Duda GN, Schell H, Seebeck P, Heller MO. 2005. On the influence of soft tissue coverage in the determination of bone kinematics using skin markers. Journal of Orthopaedic Research. 23(4):726–734.

Trucco E, Verri A. 1998. Introductory techniques for 3-d computer vision. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Windolf M, Götzen N, Morlock M. 2008. Systematic accuracy and precision analysis of video motion capturing systems—exemplified on the vicon-460 system. Journal of Biomechanics. 41(12):2776 – 2780.

Yang SX, Christiansen MS, Larsen PK, Alkjær T, Moeslund TB, Simonsen EB, Lynnerup N. 2014. Markerless motion capture systems for tracking of persons in forensic biomechanics: an overview. Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization. 2(1):46–65.

Zhao W, Espy D, Reinthal M, Feng H. 2014. A feasibility study of using a single kinect sensor for rehabilitation exercises monitoring: A rule based approach. In: Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on; Dec. p. 1–8.

Zhou H, Hu H. 2008. Human motion tracking for rehabilitation—a survey. Biomedical Signal Processing and Control. 3(1):1 – 18.

Zhou L, Liu Z, Leung H, Shum HPH. 2014. Posture reconstruction using kinect with a probabilistic model. In: Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology. ACM; p. 117–125. VRST '14.