MULTIPLE DEPTH CAMERAS CALIBRATION AND BODY VOLUME RECONSTRUCTION FOR GAIT ANALYSIS

Edouard Auvinet, Jean Meunier

IGB ,Université de Montréal
Montréal, Canada
—auvinet,meunier—@iro.umontreal.ca

Franck Multon

M2S - Université de Rennes 2
Rennes, France
franck.multon@uhb.fr

ABSTRACT

In the last decade, gait analysis has become one of the most active research topics in biomedical research engineering partly due to recent developpement of sensors and signal processing devices and more recently depth cameras. The latters can provide real-time distance measurements of moving objects. In this context, we present a new way to reconstruct body volume in motion using multiple active cameras from the depth maps they provide. A first contribution of this paper is a new and simple external camera calibration method based on several plane intersections observed with a low-cost depth camera which is experimentally validated. A second contribution consists in a body volume reconstruction method based on visual hull that is adapted and enhanced with the use of depth information. Preliminary results based on simulations are presented and compared with classical visual hull reconstruction. These results show that as little as three low-cost depth cameras can recover a more accurate 3D body shape than twenty regular cameras.

1. INTRODUCTION

Several works have been done on gait analysis for human identification, medicine and biomechanics, computer animation etc. [1]. In particular, significative changes occur in the walking pattern with aging and/or disease, increasing the risk of falls or other problems (reduced mobility, decreased balance, increased body stiffness, shorter step, slower pace, reduced motion of the limbs etc.) [2]. This risk may be decreased by early detection of gait abnormalities with clinical tests to assess possible degradation or change to prevent falls and/or diagnose a musculoskeletal disorder using a treadmill or a walkway. Different tests already exist but can not be used routinely on a large scale, due to their cost (e.g. the Vicon motion capture system [3] reaches a million dollars) or due to their lack of precision and accuracy. Recently, low-cost active cameras [4] have appeared on the marketplace. These cameras return 3D information as depth images. Briefly, such a camera projects an infrared (IR) dot pattern on the scene and observes it with an IR camera to produce a real time depth map by triangulation. This feature is particularly interesting for the development of a system to reconstruct the body volume during gait movement executed on a treadmill for instance. However some depth cameras working together are required for this purpose in order to have several points of view for a truly full 3D reconstruction. Therefore our aim is to show first, how to adapt existing methods to permit external camera calibration directly from a depth map which appears totally different than the usual color image of a scene. Secondly, we will demonstrate how this depth information can enhance results obtained with an existing shape-from-silhouette method named visual hull [5]. Results will be presented to demonstrate the superiority of a depth camera methodology with respect to the usual multi-camera setup for 3D body reconstruction.

Edouard Auvinet is a PhD student in both institutions

2. RELATED WORKS

2.1. Depth camera

The active cameras used in this project are based on the structured light concept developped in computer vision [6]. In this study, the depth camera uses a dot pattern projected on the scene with near infrared (IR) laser light which is observed by an infrared camera. Each dot has a unique neighborhood. Thanks to the unique signature of this neighborhood, each point can be easily localized in the picture. Then, knowing the geometrical relationship between the IR projector and camera the depth of this point can be easily computed by triangulation.

2.2. Calibration

Classical camera calibration is a well known domain. For instance, a widely used method is described in [7, 8]. It uses a checkerboard pattern, as a planar calibration object, which is simply moved in front of the camera. The feature points are the corners of the squares on the checkerboard. By fitting lines to this grid pattern and computing line intersections, subpixel precision can be achieved for those feature points. Unfortunately, this method cannot be used “as is“ for a depth camera. The problem is that printed details (checkerboard pattern) are not visible on the depth map and could not be measured. Some researchers [9] have proposed to use a similar grid pattern but with holes instead of black squares to create a depth pattern, but this does not work very well on hole edges

due to the absence of some IR dots (from the projected pattern) within a point neighborhood (some dots are on the planar calibration object but some are lost through the hole). Another work [10] has used a standard image accessible from the IR camera to calibrate the depth map assuming that they are perfectly registered (which is not necessarily the case). For our work, we prefer to directly calibrate the depth map to ensure its accurate calibration. This will be based on several plane intersections (see section 3.2). This paper proposes a new method for multiple depth camera calibration which was evaluated with real mesurements on a reference object (cylinder).

2.3. Volume reconstruction

As a markerless gait analysis method, visual hull [5] is an efficient method. This method uses multiple calibrated cameras and the shape-from-silhouette concept [11] to reconstruct a volume in a scene. In particular, it has been evaluated for human body reconstruction and gait cycle analysis by Mündemann in [12]. They showed that at least eight cameras were needed in a circular configuration to obtain a beginning of correct reconstruction. This was done using a numerical simulation of cameras placed arround a real body 3D model obtained with a laser system. Clinical tests were also made in [13].

The aim of this paper is to evaluate with the same kind of simulation method the benefits of using depth cameras (with adapted visual hull reconstruction) rather than normal cameras. Pictures of a real object (cylinder) reconstruction and a walking person on a treadmill made with three depth cameras will be presented to assess this method in a practical context.

3. METHOD

3.1. Time synchronisation and data acquisition

Each active camera is connected to a unique computer. All computers are synchronized with a time server using the NTP protocol [14] and data from the kinect sensor were obtained with OpenNI [15].

3.2. Calibration

For depth map calibration, the best primitive object is the plane. In fact, its equation can be robustly measured by integrating depth information of all points of each plane. Then, "virtual" calibration points needed for a classical camera calibration method [7, 8] could be obtained from the computed intersections of several triplets of measured planes.

3.2.1. Internal parameters

A projective camera model is typically composed of 4 parameters which are f the focal length, α the aspect ratio, c the optical axis coordinates in the image and $\mathbf{k}$ the radial distortion. Since the depth map is a computed image, it is reasonable to assume that the center of the picture is the intersection with the optical axis, and the aspect ratio is equal to 1. Furthemore, radial distortion is clearly negligeable in the depth map. The remaining parameter f, the focal length is given by the manufacturer library [15] and permits to correct the projective nature of the camera.

Thanks to those parameters, real coordinates of 3D points $[X, Y, Z]$ in the camera coordinate system can be computed from the depth map coordinates $[x, y, Z]$ using the following formula:

$$X = x\frac{Z}{f} \quad Y = y\frac{Z}{f}$$

3.2.2. External parameters

Once internal parameters are known for each camera, the spatial relationship between them is needed. Our method for external calibration only needs to move a large planar rectangle visible by at least two cameras simultaneously. Only one plane equation is measured by frame.

To compute the equation of a plane, several points belonging to the same plane are selected in the depth map by manually selecting the four corners of the plane. Using Singular Value Decomposition (SVD) one can find the parameters $[a, b, c, d]$ of the plane that best fits these points.

Then a "virtual" point is constructed by intersecting three planes from three different frames (with different plane positions) for each camera.

The intersection points needed for calibration are then simply the result of solving the following equation with an SVD :

$$\mathbf{Ax} = \mathbf{d}$$

where

$$\mathbf{A} = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \quad \mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}$$

where $\mathbf{n_i} = [a_i, b_i, c_i]$, d_i are the normal vector and the constant of each intersecting plane.

Due to the fact that all cameras are synchronized, there is an easy correspondence between an intersection point obtained for one camera and another camera. Then by selecting several triplets of planes in one camera, we are capable of generating a cloud of points with their matching points in all other cameras. Notice that we expect that 3D point assessment with our method will be less prone to errors if the planes are not too far from orthogonal to each others. Then only triplets of planes e, g, h which satisfy equation 1 are selected, where Th is a "minimum orthogonality" threshold.

$$|\mathbf{n}_e \cdot \mathbf{n}_g| + |\mathbf{n}_e \cdot \mathbf{n}_h| + |\mathbf{n}_g \cdot \mathbf{n}_h| < Th \qquad (1)$$

It becomes now possible to recover the translation and rotation between two cameras i and j, by finding the best registration transformation $\mathbf{R}_{i,j}$ and $\mathbf{T}_{i,j}$ of their corresponding clouds of points expressed in their respective camera 3D coordinate systems using Procrustes analysis [16] without scaling. Then all the two by two camera transformations are refined with a bundle adjustement method [17].

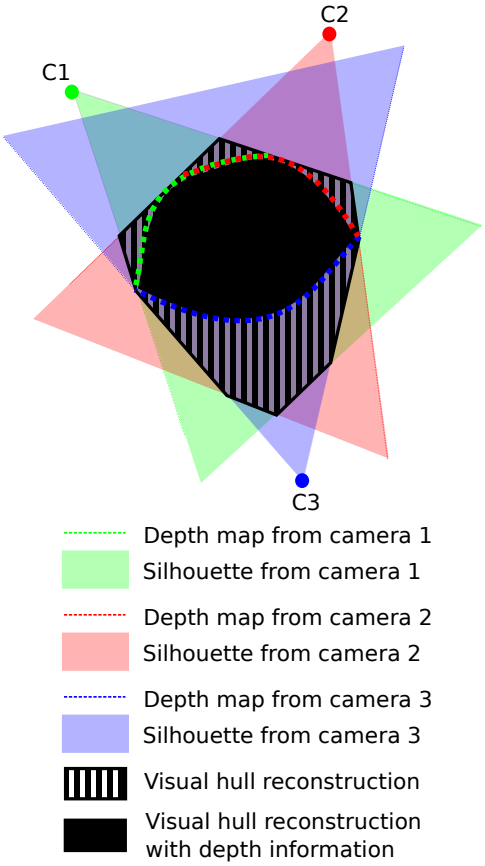

Fig. 1. Representation of the visual hull reconstructions with and without depth information.

3.3. Volume reconstruction

The volume reconstruction method is an extension of the visual hull method taking into account the depth map information.

All operations for the reconstruction are realized in a voxelized space. For each voxel the following operation is done. Let $\mathbf{V}$ be the voxel position in the real world. If for each camera i, the distance between $\mathbf{V}$ and $\mathbf{C_i}$, the optical center of this camera, is superior to the depth map value of its projection (within the object silhouette), the voxel is set to one, else it is null.

4. RESULTS

To evaluate the use of active cameras to reconstruct the body volume during gait movement, the following subparts describe the validation protocol for calibration and body volume reconstruction followed by the results obtained.

4.1. Evaluation protocol

4.1.1. Calibration

To validate the calibration results, we compared the reconstruction of a real reference object with its known size.

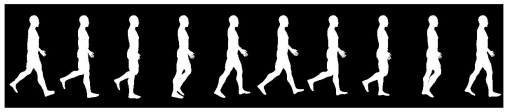

Fig. 2. Simulated walking person from 0% to 90% of the gait cycle with a 10% step viewed laterally.

The reference object was a cylinder with a mean radius of 0.2284 meter and height of 0.6 meter. The radius was measured from its perimeter. The reconstruction was tested for different minimum orthogonality threshold Th.

4.1.2. Body reconstruction

The correctness of our reconstruction method was also evaluated by comparing the reconstructed volume of a walking person to a ground truth. Since it is difficult to have a ground truth measured for each pose in good condition during the gait cycle, simulations were generated with a 3D human body model from the MakeHuman project [18] and animated with a reference gait cycle from Boulic [19]. The result of this simulation is presented in Figure 2.

Thanks to the OpenGL library, depth images and silhouettes could be computed efficiently for an unlimited number of cameras and with totally free positioning possibilities. We also simulated a simple visual hull method (i.e. without depth information) based on a voxel-based shape-from-silhouette algorithm [5].

The result presented here are differences in voxels between the ground truth and our reconstruction results. The ground truth was computed with a transformation of the reference model in the voxel space. Like [12], we used a voxel edge size of 10 mm to minimize voxelization artifacts. A sequence of one gait cycle was generated with a set of 30 frames. This to simulate a realistic situation (walk) of one gait cycle per second (1 Hz) at the frame rate (30 fps) of the depth camera (Kinect [10]). We used the same circular disposition as [12] but without the top camera.

4.2. Calibration results

The results presented in Table 1 show immediately that orthogonality of planes used to construct intersection points is important. In fact, the calibration became too weak for acceptable reconstruction when the sum of scalar products (equation 1) was higher than 1.0.

4.3. Reconstruction results

The reconstruction obtained with 64 simulated cameras during a simulated gait cycle (Figure 4) produced errors five times higher with classical visual hull compared to our method that used depth information. Moreover, our method dealt correctly with the mid-stance reconstruction deterioration already described for classical visual hull in

Th	Mean radius (meter)	Std. Dev. (meter)	error (meter)	nb points
0.2	0.2369	0.0144	0.009	72
0.4	0.2348	0.0139	0.006	948
0.6	0.2293	0.0135	-0.009	2226
0.8	0.2194	0.0147	-0.009	6714
1.0	0.2195	0.0153	-0.009	19098
1.2	-	-	-	80194
⋮	⋮	⋮	⋮	⋮
3.0	-	-	-	-

Table 1. Mean and standard deviation of reference object reconstructions with error with respect to the real mean radius (0.2284m). Different plane orthogonality threshold Th are presented. No results are reported when no reconstruction was possible. Three depth cameras were used in this experiement.

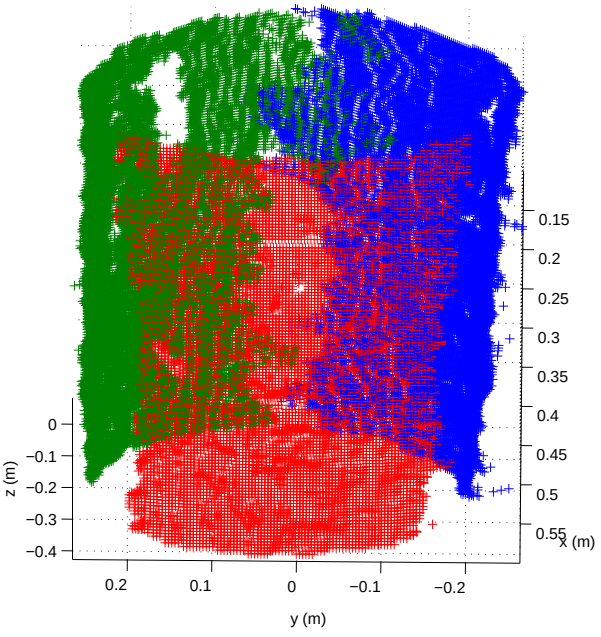

Fig. 3. Reconstructed points of the reference cylinder. Each color corresponds to one of the depth camera.

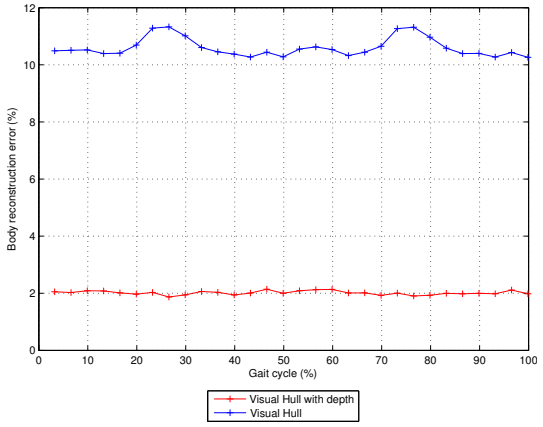

Fig. 4. Evolution of the body volumic reconstrution error during the gait cycle.

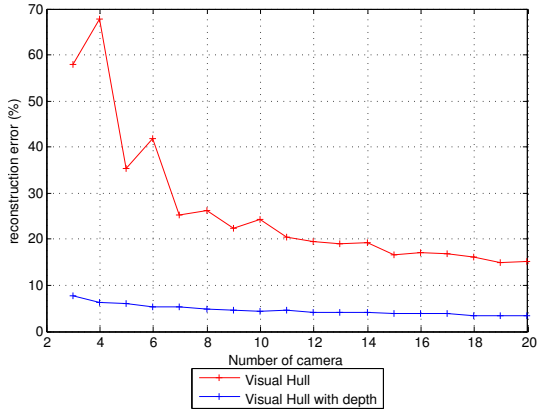

Fig. 5. Volumic reconstruction error versus number of cameras.

[11] and that typically takes the form of the two increases of error at 25% and 75% of the gait cycle.

The variation of the reconstruction error versus the number of cameras (Figure 5) showed that errors with visual hull using depth was low even with a three-camera configuration with a 8.5% error. This number was divided by two when using a 20-camera configuration. The classical visual hull method became better as the number of cameras increased but remained largely above 15% even with 20 cameras. The error standard deviation (for the different gait poses) decreased with the number of cameras (Figure 6) in part because both methods became less sensitive to the mid-stance peaks when the number of cameras increased. In all cases, visual hull with depth produced a much lower variability.

5. DISCUSSION AND FUTUR WORKS

The results presented for calibration and reconstruction are now discussed in more details.

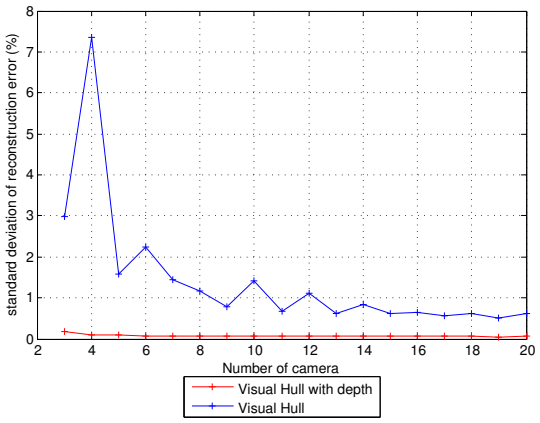

Fig. 6. Standard deviation of volumic reconstruction error versus the number of cameras.

5.1. Calibration

With a reconstruction error near half a centimeter, we believe that the calibration process using planes as primitive objects for external calibration is a valid methodology. Even if the depth (Z) error of the sensor is 0.01 m and 0.003 m for X and Y [20], the reconstruction error is less than 0.006 m with 3 depth cameras in the best case (second line, Table 1). This is due to the integration of several thousand of reconstructed points in the computation of the radius of the reference cylinder. However, even if the accuracy of the reconstruction is good, the operational spatial range of the depth camera is limited to distances between 0.8 m to 3.5 m. So the spatial volume where the reconstruction should be accurate is pretty small. This is why, with the current setup, gait analysis should be conducted on a treadmill instead of a walkway that would require a much more complex setup with more depth cameras. Moreover, the precision actually obtained with the Kinect camera is not uniform within the measured volume. It decreases with depth because of the integer quantification of the disparity (inverse of Z) returned by the Kinect. In future works, we plan to study the reconstruction error as a function of the camera configuration and the position of the reference object in the measured volume (although, as a rule of thumb, it seems reasonable to place it in the center of the camera configuration as we did here).

5.2. Volume reconstruction

The simulation analysis did not take into account all possible sources of error like projected pattern interference between active cameras on one hand and segmentation errors on the other hand. However it permitted to compare two different methodology concepts: visual hull reconstruction with and without depth information. Moreover, as shown in Figures 7 and 8, pattern interference did not introduce so many troubles on preliminary tests on real body reconstruction.

An important point to address is the positioning of the active cameras. For instance, with a 3-camera setup for gait analysis, 2 cameras (out of 3) should be placed in front looking at the subject. This is due to a particularity of the body positioning during the walk. During the gait cycle, as the arms move forward, they create some occlusions of the frontal part of the body. Therefore 2 cameras should be placed in front of the subject in order to manage this problem and retrieve a maximum of data. A 2-camera configuration (e.g. front-back, or left-right) would not give enough details for a proper arm reconstruction due to these occlusion problems. In the future, we intend to use an articulated dummy to validate precisely the reconstruction of the human body with real data.

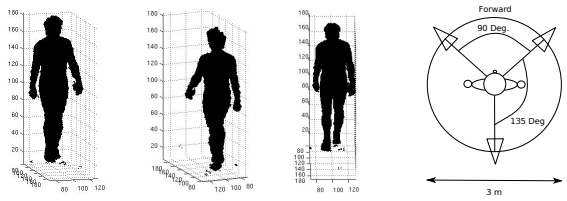

Fig. 7. Near frontal, isometric and rear views of a real human body reconstruction and the position of the 3 kinect cameras used for reconstruction.

6. CONCLUSION

These preliminary results showed that a network of depth cameras is definitively a good sensor system for markerless human body motion reconstruction, especially for gait analysis. In particular, it was not affected by the mid-stance problem reported in the literature for the classical visual hull methodology and the number of camera needed for an acceptable reconstruction is much lower (down to only 3). Our proposed method is therefore very affordable compared to any other gait analysis methods.

7. REFERENCES

[1] T.N. Tan R. Chellapa M.S. Nixon, *Human identification based on gait*, Springer, 2006.

[2] Raja I. Salgado, Stephen R. Lord, Frederick Ehrlich, Nabil Janji, and Abdur Rahman, "Predictors of falling in elderly hospital patients," *Archives of Gerontology and Geriatrics*, vol. 38, no. 3, pp. 213–219, 2004.

[3] "Vicon," http://www.vicon.com.

[4] "Kinect," http://www.xbox.com/kinect.

[5] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 150–162, 1994.

[6] Daniel Scharstein and Richard Szeliski, "High-accuracy stereo depth maps using structured light,"

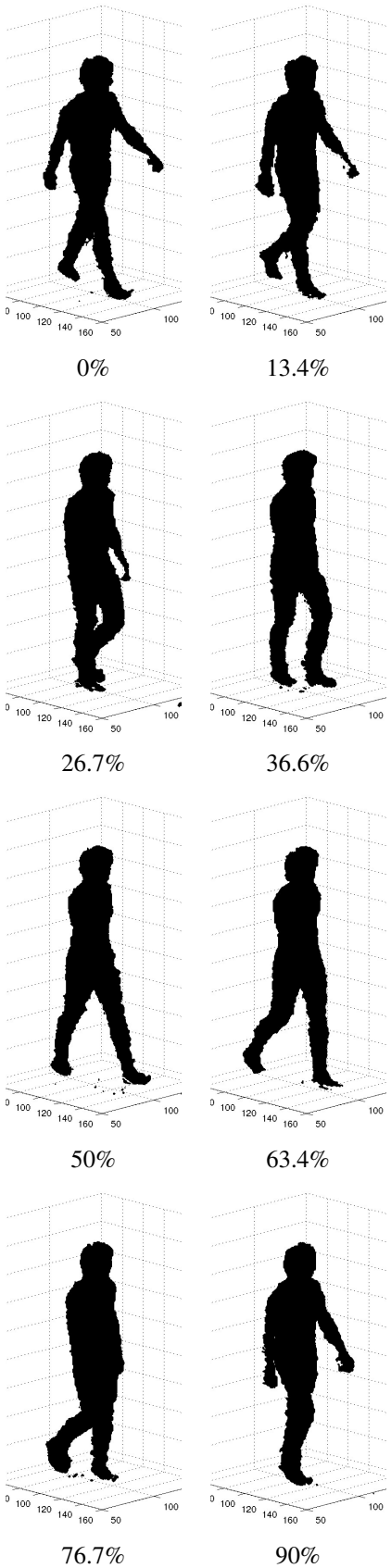

Fig. 8. Isometric view of the reconstructed volume for a real gait sequence on a treadmill starting with the right heel strike (corresponding to 0% of the gait cycle).

in *IEEE Computer Society Conference On Computer Vision And Pattern Recognition*, 2003, pp. 195–202.

[7] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[8] Zhengyou Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1330–1334, 2000.

[9] William Morris Stephane Magnenat Ivan Dryanovski, "Kinect calibration spec," *http://www.ros.org/wiki/kinect*, 2010.

[10] N. Burrus, "Kinect calibration spec," *http://nicolas.burrus.name*, 2010.

[11] B.G. Baumgard, *Geometric Modeling for Computer Vision*, Ph.D. thesis, University of Standfort, 1974.

[12] Lars Mundermann, Stefano Corazza, Ajit M. Chaudhari, Eugene J. Alexander, and Thomas P. Andriacchi, "Most favorable camera configuration for a shape-from-silhouette markerless motion capture system for biomechanical analysis," 2005, vol. 5665, pp. 278–287, SPIE.

[13] Lars Mundermann, Stefano Corazza, and Thomas Andriacchi, "The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications," *Journal of NeuroEngineering and Rehabilitation*, vol. 3, no. 1, pp. 6, 2006.

[14] D. Mills, U. Delaware, J. Martin, J. Burbank, and W. Kasch, "Network time protocol version 4: Protocol and algorithms specification - rfc 5905," Tech. Rep., Internet Engineering Task Force (IETF), June 2010.

[15] "Openni library," http://openni.org/.

[16] David G. Kendall, "A survey of the statistical theory of shape.," *Statistical Science*, vol. 4, no. 2, pp. 87–99, 1989.

[17] Bill Triggs, Philip Mclauchlan, Richard Hartley, and Andrew Fitzgibbon, "Bundle adjustment – a modern synthesis," in *Vision Algorithms: Theory and Practice, LNCS*. 2000, pp. 298–375, Springer Verlag.

[18] "Make human project - http://www.makehuman.org," .

[19] Ronan Boulic, Nadia Magnenat-thalmann, and Daniel Thalmann, "A global human walking model with real-time kinematic personification," *The Visual Computer*, vol. 6, pp. 344–358, 1990.

[20] "Prime sense kinect description," http://www.primesense.com/?p=514.