Stereoscopic Estimation of Surface Movement from inter-frame Matched Skin Texture

Björn Holmberg*
Department of Information Technology
Division of Systems and Control
Uppsala University

Abstract

Marker-less human motion analysis is currently a hot topic in the research community. In this study three dimensional motion of a human limb is estimated using a large number of matched skin texture image patches. These two dimensional matches are triangulated and also matched to the next time frame. With these matched three dimensional points a Least Squares estimate of the rigid body motion obtained using standard methods. This motion estimate is subsequently compared to a marker based estimate acquired from a synchronized marker system.

The results show that this approach can be used for motion estimation but with less accurate results than the marker based system that is to be considered as the clinically used standard. However, the correlation surfaces indicate that the method has potential if for example subpixel correlation algorithms were to be employed.

1 Introduction

The motivation for this paper can be found in the well investigated problem of human motion analysis. Estimating human motion in a non-invasive way is basically trying to estimate the motion of skeletal segments that are completely encapsulated in soft tissue. This soft tissue effectively acts as a noise barrier making accurate measurement of, at least, small motion very hard. The standard systems used in clinical practice, to this date, are marker-based [11]. By marker-based is meant that reflective markers are attached to the subject under study and then the motion of these markers are captured and used for motion estimation. It has been shown [10, 5] that these marker-based methods have severe drawbacks in the form of sensitivity to soft tissue movement, marker placement etc.

Marker free motion analysis is a natural candidate to remove the problems associated with markers. The difficulties with marker-based motion estimation

*Corresponding author. e-mail: Bjorn.Holmberg@it.uu.se,
Phone number: +46 (0)184713391, Address: Box 337, 75105 Uppsala

are partly related to using too little information in the image material rather than the actual usage of markers. If an estimate of the actual skin surface motion, using e.g marker free methods, could be achieved in an accurate way then, models for the soft tissue motion [2] could be employed and hence reduce the effect on the estimated skeletal motion. Some promising attempts has been made using this approach [3] and the image quality in modern video systems make it plausible as a way forward.

In the present work the focus is on investigating if skin texture can be used as a base for estimating three-dimensional motion of human limbs. In previous work [9, 8] in our group it has been shown that two-dimensional motion estimation is possible. With these results as a starting point the next natural step is 3D analysis. Recent developments in the motion analysis scene with the new motion capture venture MaMoCa[1] sprung from the BioMotion Lab[2] at Stanford enhance the focus on this field.

The results reached here suggest that it is possible to use this type of input with the purpose to estimate human limb motion, but with much less accuracy than using marker based methods. There are only one "stereo head" camera setup and because of this the motion estimation is rather noise sensitive in the plane orthogonal to the main motion. This suggests that the method could benefit from some type of constraint e.g. contour data. A fusion between contour data an this type of point matching algorithm would be very interesting for the future.

The organization of this paper is: First the image material and capturing conditions are discussed, then the matching and triangulation methods producing the 3D estimates are described. Here, the rigid body transform estimation is also briefly discussed. In the last part the results are presented followed by a discussion of these.

2 Material

The capturing setup and the calibration of this is essential for three dimensional motion estimation. This is described below with some illustrations for easy understanding.

Figure 1 show three typical images from one of the two video cameras. The leftmost image contain an extracted patch that is later to be registered to the next image from the same camera. This patch is a good example of the 100 patches extracted in each matching step.

2.1 Capturing setup

The camera system used in the study was a four camera setup, as depicted in Figures 3 and 2. The cameras are four megapixel "Oqus® 500" high speed cameras produced by Qualisys® Inc. They can be configured for marker tracking or for high speed video. The cameras are daisy chained in a loop which also provides the synchronization. If configured for high speed video there is an on-camera memory used to store the video sequences. Image resolution for this

[1] www.mamoca.com

[2] www.stanford.edu/group/biomotion/

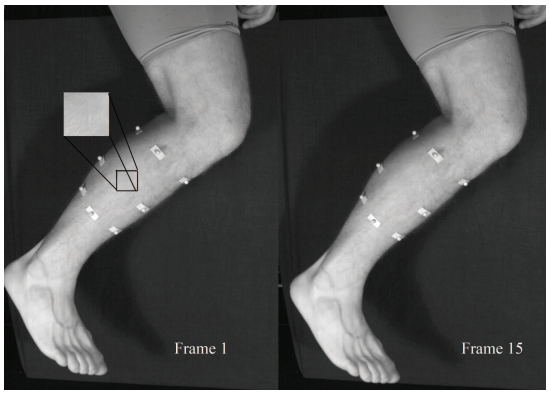

Figure 1: An example of typical images from one of the video cameras. In the leftmost image there is an example of an extracted skin texture patch that is to be registered to the next image in the series. In the rightmost image a white line surrounds the area where skin patches were extracted.

model is 2352 × 1728 pixels. The cameras can be run at a speed of 200 fps but in this case we used 100 fps. The exposure time was set to 1/2000 s.

2.2 Calibration

The calibration of all four cameras in the camera system was performed with a vendor[3] specific fixed length "wand" algorithm. This calibration provides estimates of the distortions present in the camera which can be used to approximately linearize the cameras. This linearisation was applied to all data points retrieved in the images. The calibration parameters are the same as the ones described in the well known Camera Calibration Toolbox [4]. For reference Table 1 give the actual internal and external calibration parameters. The high speed video cameras were placed approximately 0.6 meters apart and approximately 1.7 m from a supporting plane used to guide the subject movement. The capturing setup is illustrated in Figure 2. The supporting plane is visible in the background of Figure 2. The function of the plane is to limit subject motion to planar motion.

The camera rotation matrices representing the camera axes orientation as compared to the lab reference coordinate system in Figure 2 are represented by matrices **R1** and **R2**.

$$\mathbf{R1} = \begin{bmatrix} 0.989 & 0.022 & 0.142 \\ 0.144 & -0.180 & -0.972 \\ 0.003 & 0.983 & -0.182 \end{bmatrix} \tag{1}$$

$$\mathbf{R2} = \begin{bmatrix} 0.978 & -0.025 & -0.204 \\ -0.205 & -0.189 & -0.960 \\ -0.014 & 0.981 & -0.190 \end{bmatrix} \tag{2}$$

The interpretation of these matrices are that the X axis of the cameras are in, roughly, the X direction of the lab frame. The camera Y axis is in the lab frame

[3] www.qualisys.com

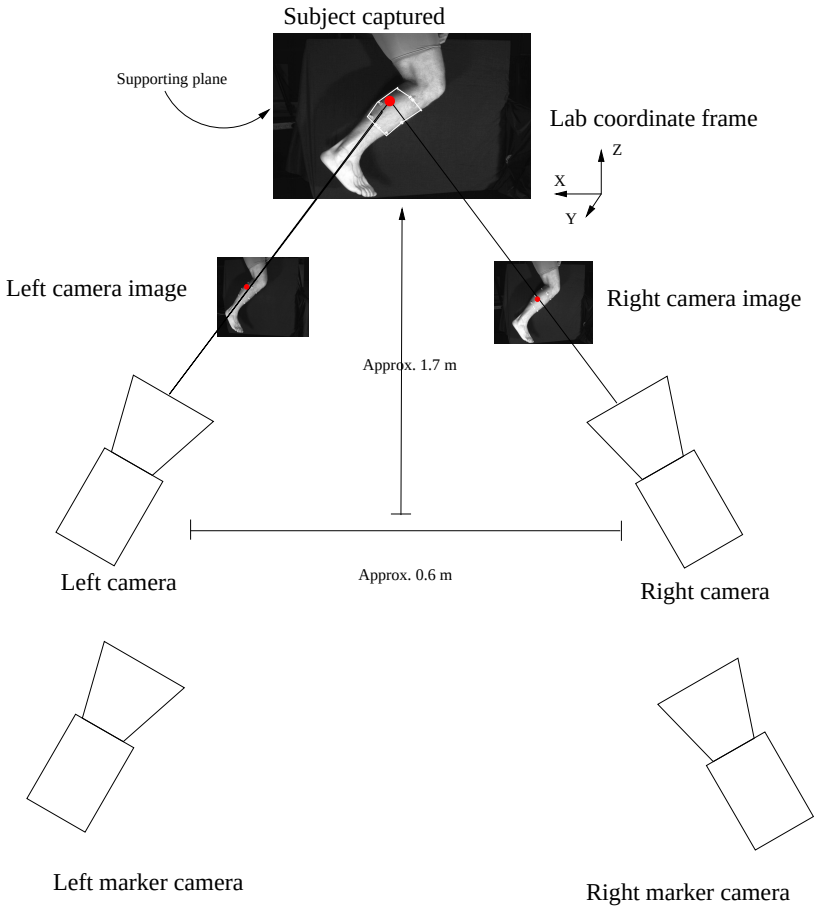

Figure 2: A graphical presentation of the capturing conditions is presented here. The two images in front of the cameras are the images of the subject in the cameras. The two front cameras are set in video capture mode as described before. The labels of the cameras match those used in Figure 3.

Z axis direction and the camera Z axis is looking down the negative lab frame Y axis.

The angular separation of the view axes in the cameras were approximately 40 degrees.

3 Methods

This section deal with the methods for inter frame matching between images from the same camera and also between images from the two different cameras. The method can be summarized as follows: First extract N patches from image frame k in camera 1 and use correlation to match these N patches to a location in image frame k of camera 2. After triangulation this yields a set of three dimensional points. Next, match the initial patches from frame k in camera 1 to frame $k+1$ in the same camera. Use these new registration points and correlate

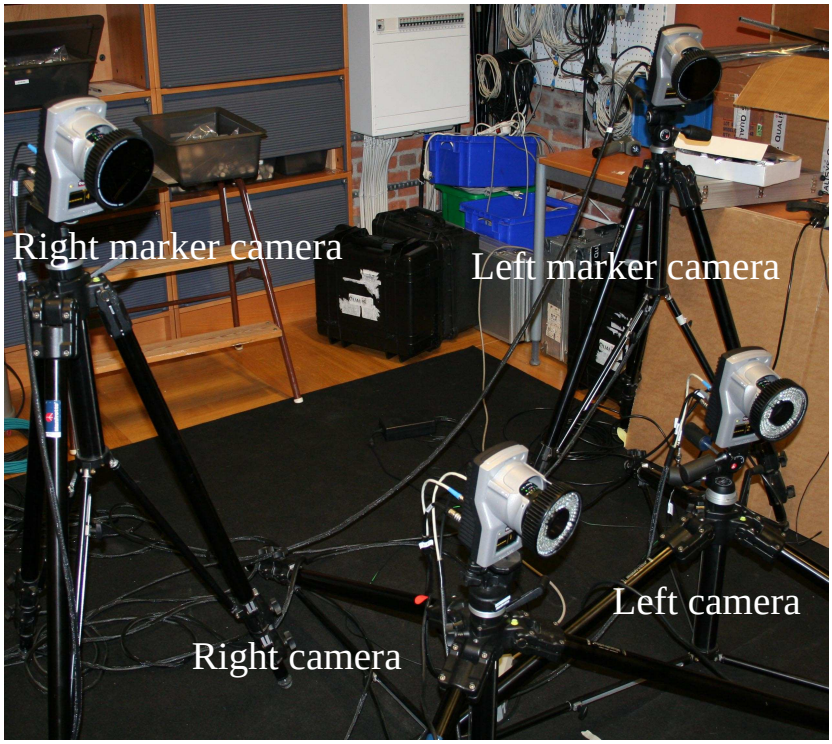

Figure 3: The real camera system as seen from the front is depicted. The two cameras in front are the ones set for video capture and the ones in the back for capturing the marker positions.

with image frame $k+1$ in camera 2 and get another set of three dimensional points in frame 2. Iterate this for the whole image sequence. Figure 1 show a typical image series in one camera. The patches are all extracted within the marker defined area, within the white lines in Figures 1 and 2, this to ensure that the results would not use other information than skin texture, and yield an unfair comparison with the marker-based motion estimate.

3.1 Matching through 2D correlation

The matching of the patches to some location in another image is determined by the two dimensional correlation value of that patch with the reference image. The two dimensional correlation value between two matrices A and B is calculated in a standard way [1] as

$$\mathbf{Corrval}(\mathbf{A},\mathbf{B}) = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum_m \sum_n (A_{mn} - \bar{A})^2 (\sum_m \sum_n (B_{mn} - \bar{B})^2))}} \qquad (3)$$

where $\bar{A}$ and $\bar{B}$ are the total matrix means. A_{mn} is the element on row m and column n in A.

Camera 1	**Position**	**X** -210(mm)	**Y** 1445(mm)	**Z** 425(mm)
fc	[3578 3578] pixels			
cc	[1232 872] pixels			
kc	[-9.72e-002]	[7.79e-002]	[-3.38e-004]	[9.23e-004]
alpha$_c$	0			
Camera 2	**Position**	**X** 413(mm)	**Y** 1429(mm)	**Z** 446(mm)
fc	[3546 3546] pixels			
cc	[1225 873] pixels			
kc	[-1.00e-001]	[8.56e-002]	[5.59e-005]	[3.12e-004]
alpha$_c$	0			

Table 1: Table of the camera center positions and the internal calibration parameters. The calibration parameters are described in detail in [4].

The size of the extracted image patches were chosen to be 100 by 100 pixels or approximately 50 by 50 mm [9] on the skin surface. The search area is also defined as a square of 100 by 100 pixels making the effective search area approximately 50 by 50 mm. There is a need here to establish an initial search position, or a center point for the search window. In this study this problem was solved by using the initial mean difference in image position of the marker centroids seen in both the left and right cameras. Our aim here is to evaluate whether skin texture can be used for matching and motion estimation in three dimensions, not to build an automatic system hence this simplification is no problem. An automatic initialization idea would be to use the limb contour to get a reasonable estimate for this image frame position difference.

3.2 Triangulation method

When the correlation method has been applied to the image material in camera one and two as described above there are a large number of matched two dimensional points. Starting with these set of two dimensional points and a good model of the cameras used it is possible to estimate the position of the three dimensional point corresponding to each pair of matched points. This is done using the method: *stereo_triangulation* implemented in [4]. This method basically calculates the rays representing each of these image points and then estimates the three dimensional position to be the intersection point of these rays. There is no imposing of epipolar constraints [7] here so the rays will generally not intersect exactly instead the mid point of the shortest straight line between the two rays is used as the estimate. Figure 4 show a graphical representation of the triangulation method.

3.3 Motion estimation

The method chosen for estimation of the three dimensional motion of the triangulated points is described in [12]. The method uses the singular value decomposition and the assumption of rigid segment motion to estimate a transformation matrix

$$\mathbf{T} = [R|t] \tag{4}$$

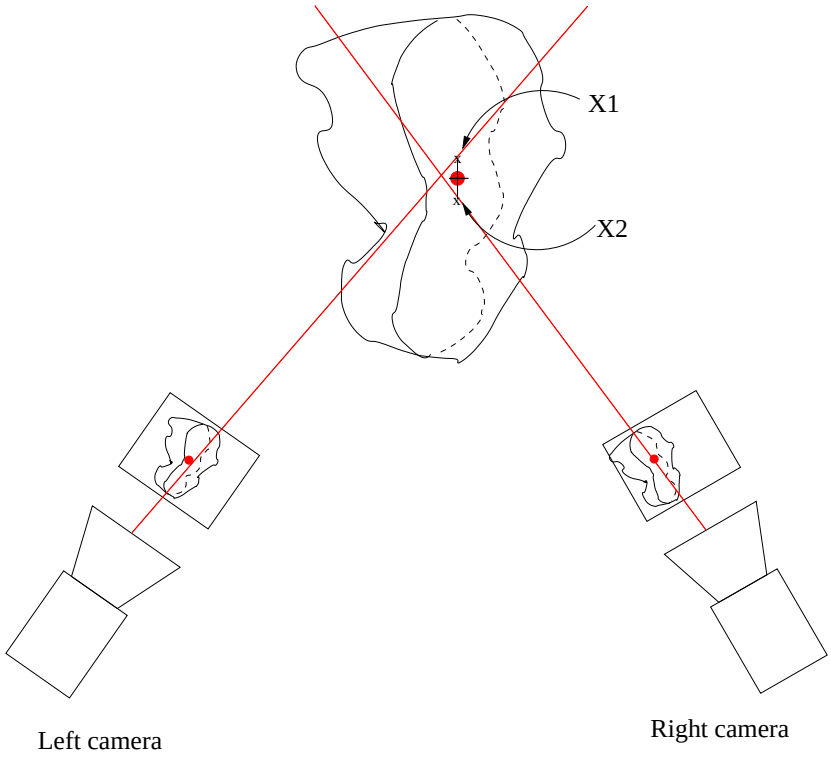

Figure 4: The triangulation method depicted schematically. The three dimensional object is seen in both cameras. Correlation matching has yielded two matched points and the three dimensional rays associated with those are projected forward. The resulting three dimensional point estimate is seen as the middle point between where these two rays are closest to each other.

where R is a proper rotation matrix and t is a translation vector. The transformation T takes triangulated points in frame k to points in frame $k+1$ in a way that minimizes the least squares error.

3.4 Removing outliers

Because of the rigid body assumption in the transform estimation method [12] outliers can be quite a problem and affect the solution to a large degree. An attempt was made to remove such outliers by simply calculating the maximum inter-frame displacement on any of the markers and then not allowing any triangulated points with larger displacement in the motion estimation. This method would not be usable in a fully marker-free system but as stated previously the aim here is to investigate whether skin texture can be used for this purpose, and not to design a automatic system marker-free system.

4 Results

The results from the matching of the image patches as described in Section 3 are exemplified in the first three figures in this section. Figure 5 show the registration of an image patch to the image it was taken from, i.e the image noise is the same in the two images. Here a very sharp correlation peak is the result, indicating that there is a well defined match between the image patch and the reference image. This correlation plot is shown as a reference for interpretation of the other correlation plots, since this is basically as good as it can get for the current image data.

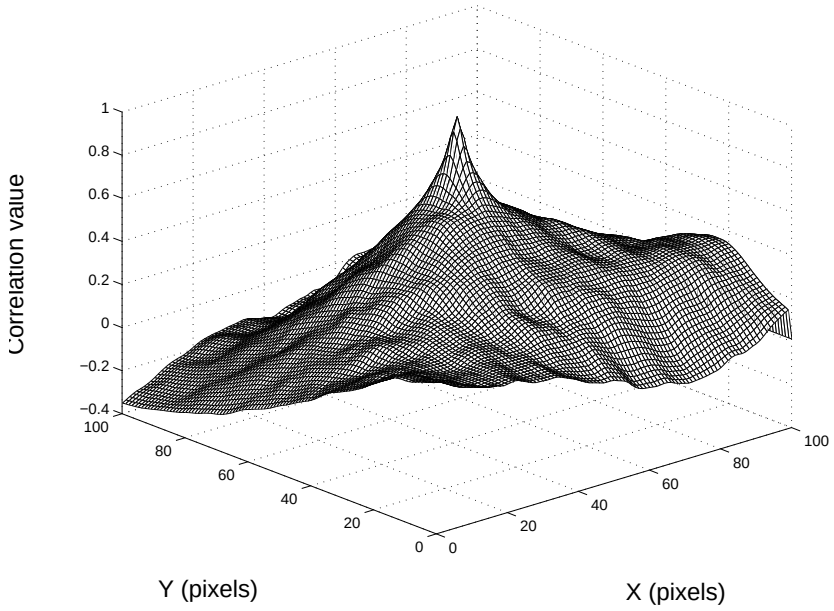

Figure 5: This Figure show a , 100 by 100, floating patch correlated over a 200 by 200 reference image that contain the actual patch location. There is a very clear peak at the "true" position. This means that there is a well defined "best" match at a certain position.

Figure 6 show the correlation of a patch taken from image k in camera 1 and correlated to a reference image in the same camera at time $k+1$, i.e the next frame. This means that the image noise realization is not the same and the result is also a slightly less peaky correlation surface. The results however indicate the there is a clear registration point match to one specific pixel location.

The most interesting part of these image patch matchings is where an image patch taken from image k in camera 1 is to be registered to a reference image taken from image k in camera 2 as seen in Figure 7. Comparing this correlation surface to the ones previously seen it is evident that the peak is much more rounded and hence give a less accurate registration or matching. It is however clear that there is a well defined match to be made in this 100 by 100 pixel search area and even though the peak is less sharp than the previous ones it is still rather well defined.

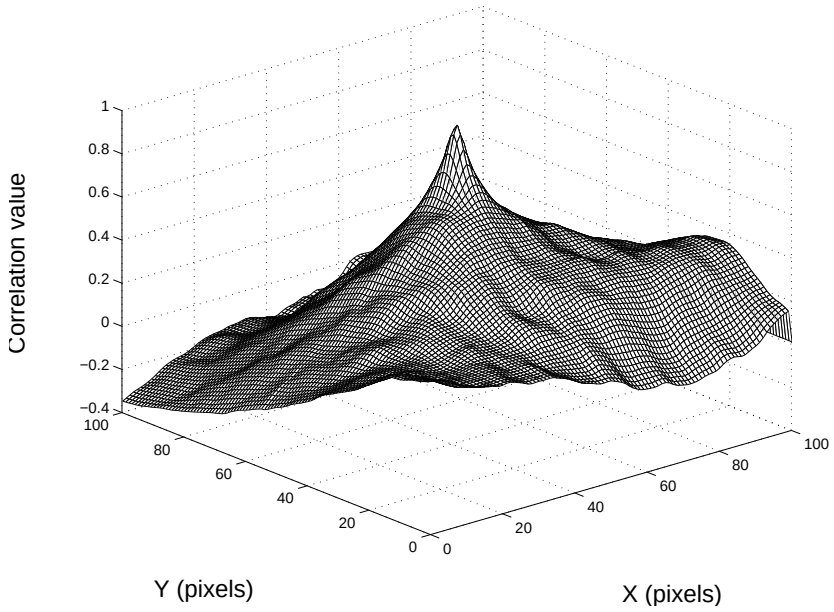

Figure 6: This Figure show the same thing as in 5 with the difference that the reference is the next image in that same camera. Also here there is a very clear correlation peak, a little less peaky but very clear. This basically indicate that it is possible to match images patches from the same camera in different frames.

The correlation surface plots shown here are examples randomly taken from one of the correlations used in the motion estimation. This means that the floating image patches are taken from within the marker defined area as shown in Figures 1 and 2. There is of course variations in the quality of these correlation based matchings that depend on different parameters such as lighting and local texture of the skin.

In Figure 8 a plot of the three dimensional triangulated correlation points is shown. These points are shown as the point cloud of stars in the middle. The points in frame k are shown in green and the points in frame $k+1$ are shown in blue. It is clear from Figure 8 and 9 that the triangulated points mostly fall inside the marker defined skin area. In Figure 9 where this point cloud is seen from the top it is clear that some of the points are estimated to lay at positions in front of and behind the marker defined surface.

Figures 10 and 11 show the measured markers position as green circles. These figures also show the marker position from frame 1 transformed throughout the whole image sequence using first the marker based transform as blue circles, and then using the texture based transform as red stars. Figure 10 show these plots as seen from the cameras. It is apparent that the texture based approach yield an overestimate of the motion in this dimension. But the shape of the motion is very similar to that achieved using the marker based approach.

Looking at Figure 11 it seems that the rather noisy triangulated position es-

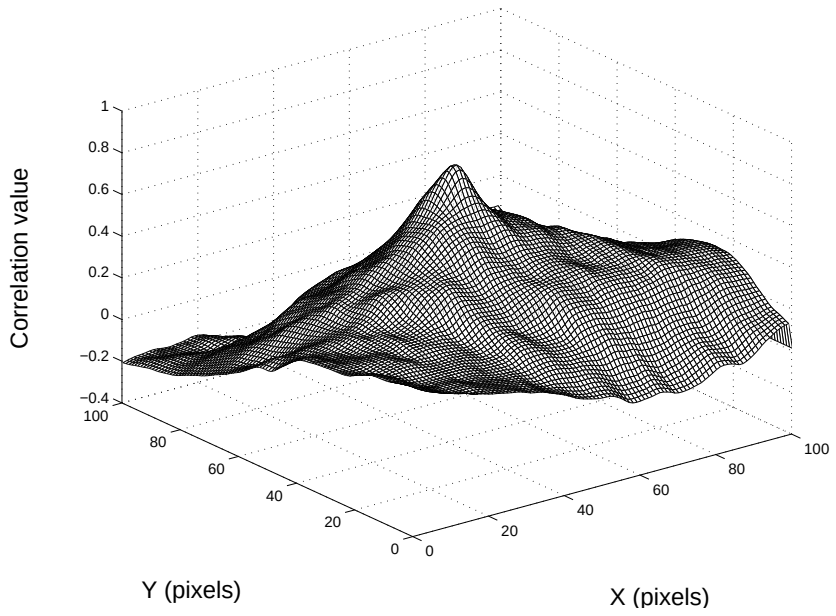

Figure 7: This Figure show the floating patch taken from the left camera and searched for in the right camera reference image, for the same time frame k. Here as well there is a clear peak but with a much more rounded shape, i.e. the estimate is not as accurate but still well defined on a pixel level. This correlation show that there is just a few pixels wide area possible in the matching.

timates affect the motion estimate more in the dimension with the least motion, i.e in the plane normal to the marker defined plane. The shape of the motion curves are much more erratic in this case.

The average euclidian distance between the measured markers and the marker positions from the first frame propagated using the marker and texture based transform is shown in Figure 12. This figure show that the average euclidian distance to the measurements is consistently much larger when using the texture based approach. If Figure 10 and 11 are considered again it is apparent that this larger error comes from both an overestimation of the motion in the marker plane and also from an erratic motion in the plane orthogonal to the motion plane.

5 Conclusions and future work

The work presented here does today not present an alternative to marker based motion estimation. The performance is simply not good enough, yet. This does not mean that there is not a potential in the usage of skin texture for estimation of human motion. The correlation plots presented in Figure 5,6 and especially 7 show that clear matchings are possible between cameras with this large angular separation. The correlation used here is discrete, i.e only integer pixel steps are allowed. Marker center estimates in modern systems take advantage of the

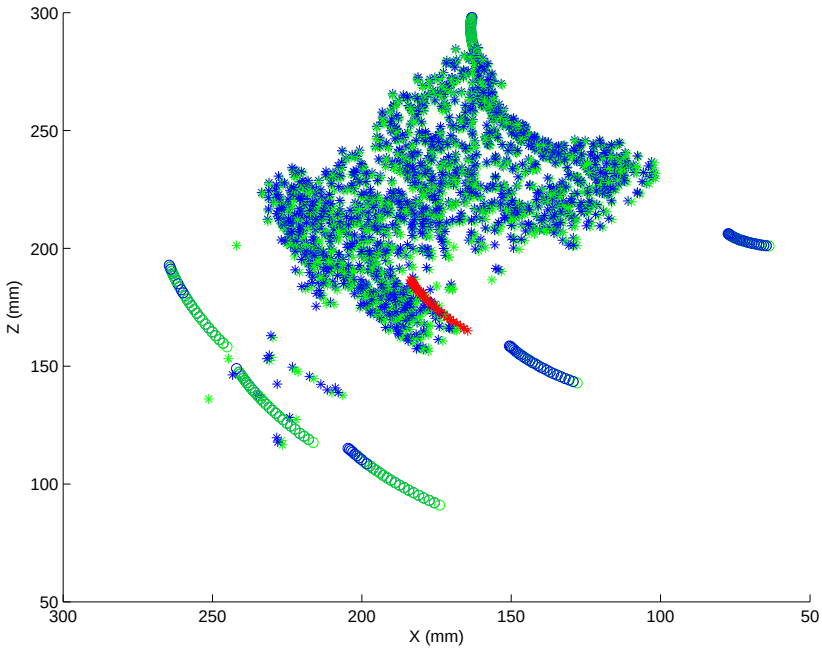

Figure 8: The point cloud of triangulated points seen from the side. The starred points are triangulated texture points, and the markers are represented by circles. Markers in frame k is displayed in green and markers in frame k+1 in blue. The same goes for the texture points. It is clear that the some triangulated points should be considered outliers since they are not located within the marker defined area depicted in Figure 2.

whole gray scale image and hence yield very high accuracy. If the correlation method was instead used in combination with rotations, scalings and also sub pixel displacements the results might be enhanced.

The main potential that usage of this skin texture based estimates has is to enhance the type of contour based methods dominating this field today. Estimation of internal/external rotation could benefit much from a less sensitive measure than contour data.

Many interesting possibilities open up in the wake of this work and a lot of work also has to be done to investigate if texture matching is a viable alternative or complement to the marker-free methods used today. Future interesting work would be:

- Usage of aligned cameras, i.e without angular separation and also less translational separation. This would yield less projective distortion between the images and hence make correlation easier and also faster due to smaller search area.

- Using more cameras of the type used here, capturing all sides of the subject under study. This would yield data usable for visual hull construction and then also for enhancement using the texture surface exemplified in Figures 8 and 9.

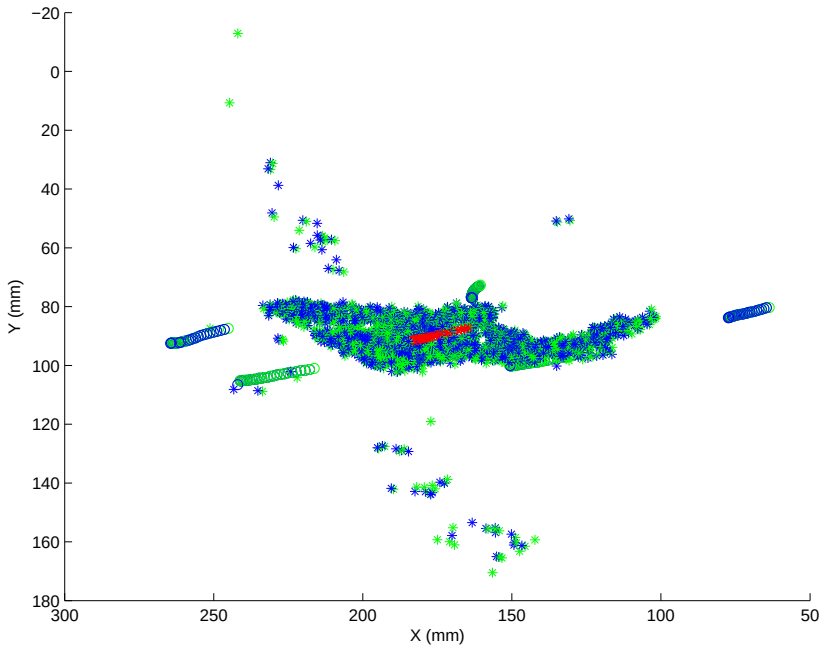

Figure 9: The point cloud of triangulated points seen from above. Starred points are triangulated texture and circles are markers, points in frame k are displayed in green and points in frame k+1 in blue.

- Building a particle filter [6] implementation with both contour and texture points as measurements.
- Estimating the error in the skin texture matching by using some type of 3D scanner that can estimate a detailed surface with high accuracy. Using such a scanner synchronized with a stereo head setup as the one used here would give a good indication of both bias and variance of the stereo based surface estimate.
- In this work 100 skin patches were used as a base for the motion estimation. It is clear that this is not the limit of the number of patches possible. It would be interesting to see how much improvement increasing the number of patches with one order of magnitude would yield.

6 Acknowledgment

We want to acknowledge Qualisys® corporation for providing us with the opportunity to use their cameras and also assisting with the capturing process.

References

[1] Matlab. www.themathworks.com.

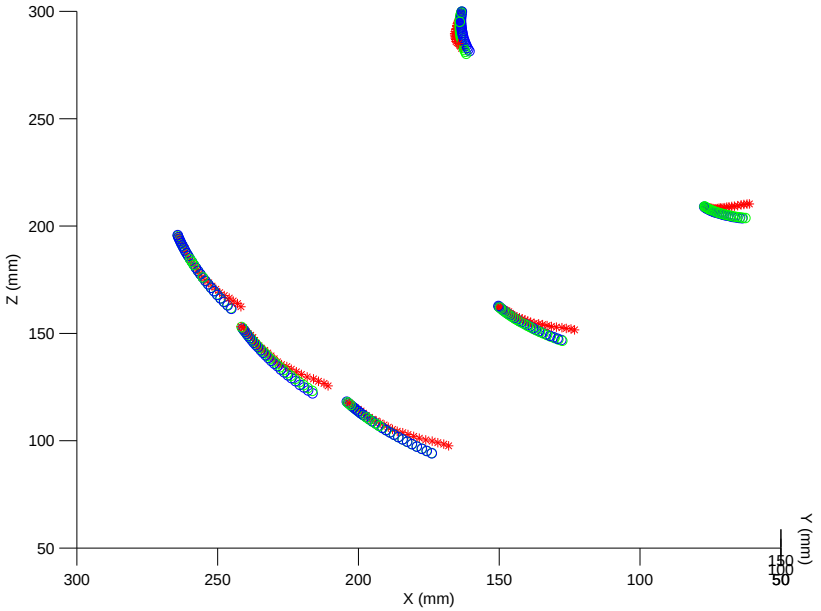

Figure 10: The markers from the first frame transformed with the marker-based, blue, and texture-based, red, transform. The measured markers are also shown, in green, for reference

[2] Alexander, E., and Andriacchi, T. Correcting for deformation in skin-based marker systems. *Journal of Biomechanics 34*, 7 (2001), 729–732.

[3] Alexander, E. J., Andriacchi, T. P., and Bregler, C. Estimation of skeletal kinematics through high feature density video based motion capture. In *Eigth International Symposium on the 3D Analysis of Human Movement* (2004).

[4] Bouguet, J.-Y. Camera calibration toolbox for matlab. www.vision.caltech.edu/bouguetj/calib_doc.

[5] Cappozzo, A., Catani, F., Leardini, A., Benedetti, M. G., and Croce, U. D. Position and orientation in space of bones during movement: experimental artefacts. *Clinical Biomechanics 11*, 2 (1996), 90–100.

[6] Gordon, N., Salmond, D., and Smith, A. Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing 140*, 2 (1993), 107–113.

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

[8] Holmberg, B., and Lanshammar, H. Possibilities in using skin texture based image registration for human movement. In *Proceedings of the "Ninth International Symposium on the 3-D Analysis of Human Movement"* (2006).

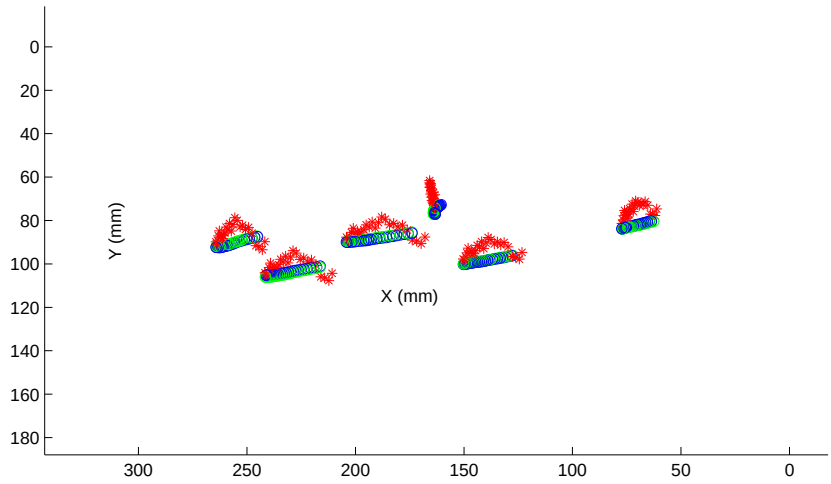

Figure 11: The markers from the first frame transformed with the marker-based, blue, and texture-based, red, transform. The measured markers are also shown, in green, for reference

[9] Holmberg, B., Nordin, B., Bengtsson, E., and Lanshammar, H. On the plausibility of using skin texture as virtual markers in the human analysis contect, a 2d study. Tech. Rep. 2008-001, Uppsala University, 2008.

[10] Leardini, A., Cappozzo, A., Catani, F., Toksvig-Larsen, S., Petitto, A., and Sforza, V. Validation of a functional method for the estimation of hip joint centre location. *Journal of Biomechanics 32*, 1 (1999), 99–103.

[11] Moeslund, T. B., Hilton, A., and Krüger, V. A survey of advances in vison-based human capture and analysis. *Computer Vision and Image Understanding 104* (2006), 90–126.

[12] Söderkvist, I., and Wedin, P. Determining the movements of the skeleton using well-configured markers. *Journal of Biomechanics 26* (1993), 1473–1477.

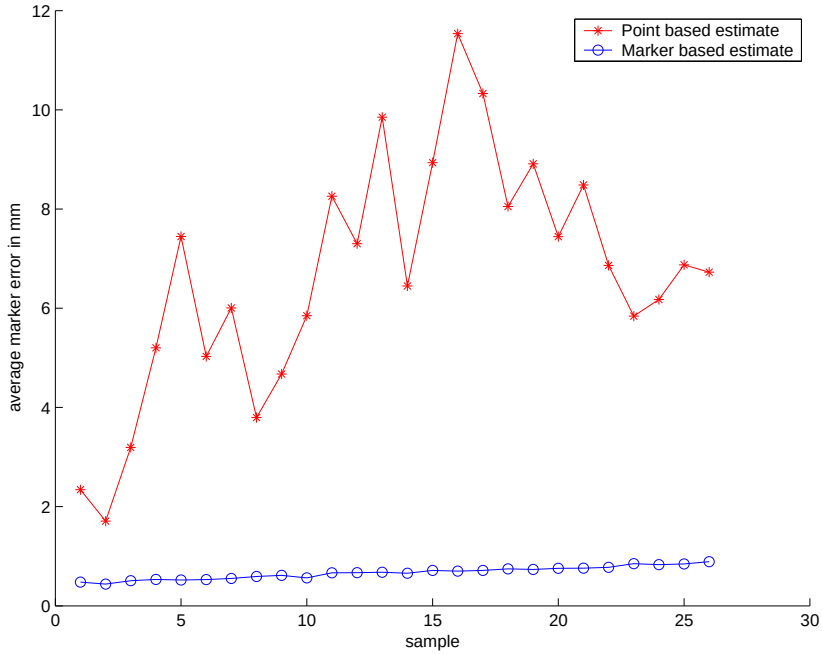

Figure 12: The difference in mm between the measured marker points and the marker points from frame 1 transformed with first the marker-based transformation and then the texture-based one. The texture based one show higher levels of difference than the marker-based one.