Single View Motion Tracking by Depth and Silhouette Information

Daniel Grest[1], Volker Krüger[1] and Reinhard Koch[2]

[1] Aalborg University Copenhagen, Denmark
Aalborg Media Lab
[2] Christian-Albrechts-University Kiel, Germany
Multimedia Information Processing

Abstract. In this work[3]a combination of depth and silhouette information is presented to track the motion of a human from a single view. Depth data is acquired from a Photonic Mixer Device (PMD), which measures the time-of-flight of light. Correspondences between the silhouette of the projected model and the real image are established in a novel way, that can handle cluttered non-static backgrounds. Pose is estimated by Nonlinear Least Squares, which handles the underlying dynamics of the kinematic chain directly. Analytic Jacobians allow pose estimation with 5 FPS.

Keywords: optical motion capture, articulated objects, pose estimation, cue-integration

1 Introduction

Valid tracking of human motion from a single view is an important aspect in robotics, where research aims at motion recognition from data, that is collected from the robot's measuring devices. Additionally, the processing time should be at least near-to-real-time to make human-robot interaction possible. Both aspects are addressed in this work.

Motion capture and body pose estimation are also applied in motion analysis for sports and medical purposes. Motion capture products used in the film industry or for computer games are usually marker based to achieve high quality and fast processing. While the accuracy of markerless approaches is comparable to marker based systems [13, 4], the segmentation step makes strong restrictions to the capture environment, because these systems rely on segmentation of the person in the foreground, e.g. homogenous clothing and background, constant lighting, camera setups that cover a complete circular view on the person etc.

Our approach doesn't need explicit segmentation or homogenous clothing and gives reliable results even with non-static cluttered background. Additionally, motion can be accurately tracked even from a single view, because the underlying motion and body model is directly incorporated in the image processing step. We present here a combination of depth data and silhouette information, which extends the motion estimation from stereo data [7] with additional information from silhouette correspondences. Results are given for depth data from a novel measuring technique, called *Photonic Mixer Device (PMD)*, which gives a 64×48 depth image in real-time with 25FPS. The results show, that the characteristic of this depth data is not sufficient alone for valid tracking. However in

[3] Acknowledgment This work was partially funded by PACO-PLUS (IST-FP6-IP-027657) and by German Science Foundation project DFG-3DPoseMap.

combination with silhouette information, the accuracy is increased and motion can be successfully tracked over longer sequences.

Capturing human motion by pose estimation of an articulated object is done in many approaches and is motivated from inverse kinematic problems in robotics. Solving the estimation problem by optimization of an objective function is also very common [13, 8, 11]. Silhouette information is usually part of this function, that tries to minimize the difference between the model silhouette and the silhouette of the real person either by background segmentation [13, 11] or image gradient [12, 5]. In [2] a scaled orthographic projection approximates the full perspective camera model and in [13] the minimization of 2D image point distances is approximated by 3D-line-3D-point distances. A recent extensive survey on vision-based motion capture can be found in [9].

While some kind of template body model is common in most approaches, adaption of body part sizes of the template during the motion estimation is also possible [12], where depth and silhouette information were combined to estimate the size and pose of the upper body. In contrast to their approach, we estimate pose in near-to-real-time and minimize silhouette differences in the image plane rather than in 3D, which makes the estimation more accurate. The image processing with color histograms allows us to establish valid silhouette correspondences even with moving background, which in turn allows moving cameras. By combination with depth data from a PMD device motion can be tracked, which is not trackable from a single view with only one of these data types. Our method minimizes errors, where they are observed and makes no approximations to the motion or projection model. Additionally, it allows analytical derivations of the optimization function. This speeds up the calculation by more accuracy and less function evaluations than numerical derivatives. Therefore the approach is fast enough for real-time applications in the near future as we process images already with 5 frames per second on a standard PC Pentium IV 3 GHz.

2 Body and Motion Model

Depending on the kind of work different body models are used for the estimation process. The models range from simple stick figures over models consisting of scalable spheres (metaballs) [12] to linear blend skinned models [1]. We use models with motion capabilities as defined in the MPEG4 standard, with up to 180 DOF, an example model is shown in figure (1). The MPEG4 description allows to exchange body models easily and to reanimate other models with the captured motion data. The model for a specific person is obtained by silhouette fitting of a template model as described in [6].

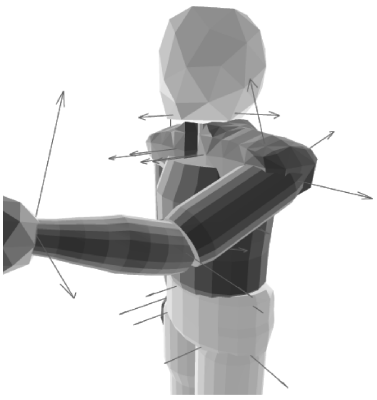

Fig. 1. The body model with rotation axes shown as arrows

The MPEG4 body model is a combination of kinematic chains. The motion of a point, e.g. on the hand, may therefore be expressed

as a concatenation of rotations [7]. As the rotation axes are known, e.g. the flexion of the elbow, the rotation has only one degree of freedom (DOF), i.e. the angle around that axis. In addition to the joint angles there are 6 DOF for the position and orientation of the object within the global world coordinate frame. For an articulated object with p joints the transformation may be written according to [7] as:

$$\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x}) = (\theta_x, \theta_y, \theta_z)^T + \left(R_x(\theta_\alpha) \circ R_y(\theta_\beta) \circ R_z(\theta_\gamma) \circ R_{\boldsymbol{\omega},\boldsymbol{q}}(\theta_1) \circ \cdots \circ R_{\boldsymbol{\omega},\boldsymbol{q}}(\theta_p)\right)(\boldsymbol{x})$$

where $(\theta_x, \theta_y, \theta_z)^T$ is the global translation, R_x, R_y, R_z are the rotations around the global x, y, z-axes with Euler angles α, β, γ and $R_{\boldsymbol{\omega},\boldsymbol{q}}(\theta_i), i \in \{1..p\}$ denotes the rotation around the known axis with angle θ_i. The axis is described by the normal vector $\boldsymbol{\omega_i}$ and the point $\boldsymbol{q_i}$ on the axis with closest distance to the origin.

The equation above gives the position of a point $\boldsymbol{x}$ on a specific segment of the body (e.g. the hand) with respect to joint angles $\boldsymbol{\theta}$ and an initial body pose.

The first derivatives of $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x})$ with respect to $\boldsymbol{\theta}$ give the Jacobian matrix $J_{ki} = \frac{\partial f_k}{\partial \theta_i}$. The Jacobian for the motion of the point $\boldsymbol{x}$ on an articulated object is

$$J = \begin{bmatrix} 1 & 0 & 0 & & & & & & & \\ 0 & 1 & 0 & \frac{\partial f}{\partial \theta_\alpha} & \frac{\partial f}{\partial \theta_\beta} & \frac{\partial f}{\partial \theta_\gamma} & \frac{\partial f}{\partial \theta_1} & \cdots & \frac{\partial f}{\partial \theta_p} \\ 0 & 0 & 1 & & & & & & \end{bmatrix} \quad , \tag{1}$$

with the simplified derivative at zero:

$$\left.\frac{\partial f}{\partial \theta_i}\right|_0 = \left.\frac{\partial R_{\boldsymbol{\omega},\boldsymbol{q}}(\theta_i)}{\partial \theta_i}\right|_0 = \boldsymbol{\omega}_i \times (\boldsymbol{x} - \boldsymbol{q}_i) = \boldsymbol{\omega}_i \times \boldsymbol{x} - \boldsymbol{\omega}_i \times \boldsymbol{p}_i \ . \tag{2}$$

Here $\boldsymbol{p}_i$ is an arbitrary point on the rotation axis. The term $\boldsymbol{\omega}_i \times \boldsymbol{p}_i$ is also called the momentum. The simplified derivative at zero is valid, if relative transforms in each iteration step of the *Nonlinear Least Squares* are calculated and if all axes and corresponding point pairs are given in world coordinates.

2.1 Projection

If the point $\boldsymbol{x} = (x_x, x_y, x_z)^T$ is observed by a pin-hole camera and the camera coordinate system is in alignment with the world coordinate system, the camera projection may be written as:

$$p(\boldsymbol{x}) = \begin{pmatrix} s_x \frac{x_x}{x_z} + c_x \\ s_y \frac{x_y}{x_z} + c_y \end{pmatrix} \tag{3}$$

where s_x, s_y are the pixel scale (focal length) of the camera in x- and y-direction, and $(c_x, c_y)^T$ is the center of projection in camera coordinates.
We now combine $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x})$ and $p(\boldsymbol{x})$ by writing $\boldsymbol{g}(s_x, s_y, c_x, c_y, \boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x}))$. The partial derivatives of g can now be easily computed using the chain rule. The resulting Jacobian reads as follows:

$$\begin{aligned} J &= \begin{bmatrix} \frac{\partial g}{\partial s_x} & \frac{\partial g}{\partial s_y} & \frac{\partial g}{\partial c_x} & \frac{\partial g}{\partial c_y} & \frac{\partial g}{\partial \theta_x} & \frac{\partial g}{\partial \theta_y} & \frac{\partial g}{\partial \theta_z} & \frac{\partial g}{\partial \theta_\alpha} \cdots \frac{\partial g}{\partial \theta_p} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\boldsymbol{f}(\boldsymbol{\theta})_x}{\boldsymbol{f}(\boldsymbol{\theta})_z} & 0 & 1 & 0 & \frac{s_x}{\boldsymbol{f}(\boldsymbol{\theta})_z} & 0 & s_x \frac{-\boldsymbol{f}(\boldsymbol{\theta})_x}{(\boldsymbol{f}(\boldsymbol{\theta})_z)^2} & \frac{\partial g_x}{\partial \theta_\alpha} \cdots \frac{\partial g_x}{\partial \theta_p} \\ 0 & \frac{\boldsymbol{f}(\boldsymbol{\theta})_y}{\boldsymbol{f}(\boldsymbol{\theta})_z} & 0 & 1 & 0 & \frac{s_y}{\boldsymbol{f}(\boldsymbol{\theta})_z} & s_y \frac{-\boldsymbol{f}(\boldsymbol{\theta})_y}{(\boldsymbol{f}(\boldsymbol{\theta})_z)^2} & \frac{\partial g_y}{\partial \theta_\alpha} \cdots \frac{\partial g_y}{\partial \theta_p} \end{bmatrix} \end{aligned} \tag{4}$$

and

$$\frac{\partial g}{\partial \theta_i} = \begin{pmatrix} \frac{\partial \left(s_x \frac{f_x}{f_z} \right)}{\partial \theta_i} \\ \\ \frac{\partial \left(s_y \frac{f_y}{f_z} \right)}{\partial \theta_i} \end{pmatrix} = \begin{pmatrix} \frac{s_x \left(\frac{\partial f_x}{\partial \theta_i} \boldsymbol{f}(\boldsymbol{\theta})_z - \boldsymbol{f}(\boldsymbol{\theta})_x \frac{\partial f_z}{\partial \theta_i} \right)}{(\boldsymbol{f}(\boldsymbol{\theta})_z)^2} \\ \\ \frac{s_y \left(\frac{\partial f_y}{\partial \theta_i} \boldsymbol{f}(\boldsymbol{\theta})_z - \boldsymbol{f}(\boldsymbol{\theta})_y \frac{\partial f_z}{\partial \theta_i} \right)}{(\boldsymbol{f}(\boldsymbol{\theta})_z)^2} \end{pmatrix} \tag{5}$$

The partial derivatives $\frac{\partial f}{\partial \theta_i}, i \in \{\alpha, \beta, \gamma, 1, .., p\}$ are given in equation (1) and $\boldsymbol{f}(\boldsymbol{\theta}) = (f_x, f_y, f_z)^T$ is short for $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x})$. Note that $\boldsymbol{f}(\boldsymbol{\theta})$ simplifies to $\boldsymbol{x}$, if $\boldsymbol{\theta}$ is zero.

These Jacobian allows full camera calibration from (at best) five 2D-3D correspondences or pose from 3 correspondences. An implementation of it with an extension to the *Levenberg-Marquardt* algorithm[3], which ensures an error decrease with each iteration, is available for public in our open-source C++ library[4].

3 Correspondences by Silhouette

To compensate the drift we add silhouette information to our estimation. This is achieved by calculating additional 2D-3D correspondences for the model silhouette and the silhouette of the real person. In contrast to [13] we don't utilize explicit segmentation of the images in fore- and background, but use the predicted model silhouette to search for corresponding points on the real silhouette. Previous work like [8] already took this approach by searching for a maximum grey value gradient in the image in the vicinity of the model silhouette. However we experienced that the gray value gradient alone gives often erroneous correspondences, especially if the background is heavily cluttered and the person wears textured clothes.

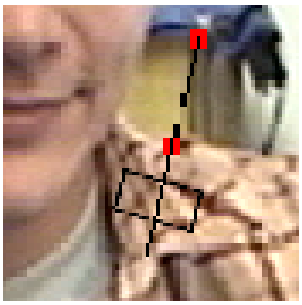

Fig. 2. Correspondence search along the normal.

Therefore we also take color information into account. As the initial pose is known, it is possible to calculate a color histogram for each body segment. We use the HSL color space to get more brightness invariance. This reference histogram is then compared with a histogram calculated over a small window on the searched normal. In figure 2 the normal is shown and the rectangular window, that are used for histogram and gradient calculation. The expectation is, that the histogram difference changes most rapidly on the point on

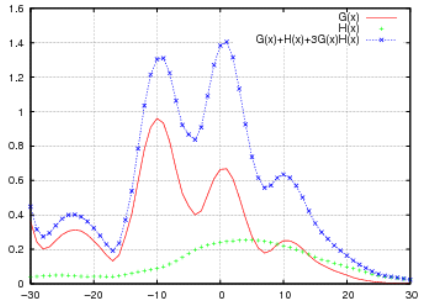

Fig. 3. The gradient (G(x)) and histogram (H(x)) values along the normal. Correct correspondence at 0.

[4] www.mip.informatik.uni-kiel.de/Software/software.html

the normal of the correct correspondence, where the border between person and background is. The type of combination function was chosen by analyzing the developing of gradient and histogram values over 15 normals in different images. The actual values of the combination were then evaluated experimentally by trying different values and counting the number of correct correspondences manually for about 100 silhouette points in 4 different images.

A rather difficult case is shown in figure 3, which shows a plot of the maximum search along the normal of figure 2. The grey value gradient $G(x)$ is shown as a solid line, the gradient of the histogram differences $H(x)$ as points and the combination with lines and points. As visible, the grey value gradient alone would give a wrong correspondence, while the combination yields the correct maximum at zero.

For parallel lines it isn't possible to measure the displacement in the direction of the lines (aperture problem). Therefore we use a formulation that minimizes the distance between the tangent at the model silhouette and the target silhouette point (normal displacement), resulting in a 3D-point-2D-line correspondence as visible in figure 4. For a single correspondence the minimization is

$$\min_{\boldsymbol{\theta}} \left[(\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{x}) - \boldsymbol{x'})^T \boldsymbol{n} - d\right]^2 \tag{6}$$

where $\boldsymbol{n}$ is the normal vector on the tangent line and d is the distance between both silhouettes. We compute d as $d = (\hat{\boldsymbol{x}}' - \boldsymbol{x}')\boldsymbol{n}$. The point on the image silhouette $\hat{\boldsymbol{x}}'$ is the closest point to $\boldsymbol{x}'$ in direction of the normal. In this formulation a motion of the point perpendicular to the normal will not change the error. We calculate the normal vector as the projected face normal of the triangle, which belongs to the point $\boldsymbol{x}'$.

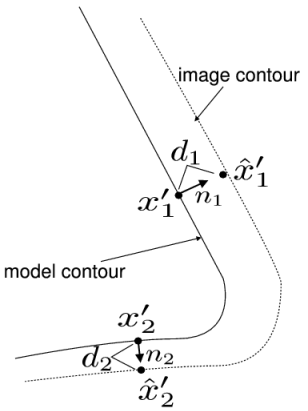

Fig. 4. Silhouette correspondences

For a set $\boldsymbol{X}$ with N points and projected image points $\boldsymbol{X'}$ the optimal solution is:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left[(\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{x_i}) - \boldsymbol{x'_i})^T \boldsymbol{n_i} - d_i\right]^2$$

This problem is known as *Nonlinear Least Squares* and can be solved by *Newton's Method* [3]. We use the *Gauss-Newton Method* [3], which doesn't require the second derivatives of $\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{x_i})$. The necessary Jacobian is given as:

$$J_{ik} = \left(\frac{\partial \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{x_i})}{\partial \theta_k}\right)^T \boldsymbol{n_i} \tag{7}$$

Note that each of these correspondences gives one row in the Jacobian.

The solution is found by iteratively solving the following equation:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - (J^T J)^{-1} J^T \left(G(\boldsymbol{\theta_t}, \boldsymbol{X}) - \hat{\boldsymbol{X}}'\right) \tag{8}$$

Here the Jacobian matrix J consists of all partial derivatives for all N points. The Jacobian for a single point is given in equation (4). In case of convergence the final solution $\hat{\boldsymbol{\theta}}$ is found.

4 Combining Multiple Cues

Integration of different vision cues into our parameter estimation problem is non trivial. Different cues like tracked edges or points give different information about the model parameters. Additionally, the measurement noise of different cues can vary dramatically.

In [5] both aspects are addressed by modeling different image cues, which are defined by regions. These regions are then propagated through the estimation by affine arithmetic. For example, tracked edges have a region that is elongated along the edge and less elongated perpendicular to it. These regions are combined into a generalized image force for each cue. The resulting region in parameter space is approximated by a Gaussian distribution. The Gaussians from each cue are then combined by a *Maximum Likelihood* Estimator and the result is integrated in a classical Euler integration procedure. The defined image regions are supposed to set hard limits on the possible displacements, however due to Gaussian approximation of the resulting parameter region the limits are softened. Therefore the approach becomes similar to a covariance based approach, where each image cue has an associated covariance matrix.

The approach taken in this work is different. The silhouette information is integrated by changing the objective function, such that the distance of the projected 3D-point to the 2D line is minimized. This is equivalent to a point-point distance with a covariance infinitely extended in direction along the edge. The different measurement noise of different cues is integrated in the estimation here by weighting each correspondence with a scalar. Weighting with a covariance matrix would be possible as well. However, for the different cues in this work the measurement noise is not exactly known and therefore covariance matrices are assumed to be diagonal and extended the same in all directions. Additionally it is assumed, that the measurement noise is the same for all measurements of one cue, resulting in one single scalar weight for each cue. In addition to the measurement noise, the weights reflect the different units of measurements, e.g. the measurement unit of 3D point positions from stereo images is meter, while the 2D measurement unit is pixels.

Let $X = \{\boldsymbol{x_0}, \boldsymbol{x_1}, .., \boldsymbol{x_k}\}$ be the set of model points and $Y = \{\boldsymbol{y}_1, .., \boldsymbol{y}_k\}$ be their corresponding 3D-points from the PMD camera found by nearest neighbor [7]. The correspondences for the silhouette information are built by the 3D-points $X = \{\boldsymbol{x_{k+1}}, .., \boldsymbol{x_{k+l}}\}$ and corresponding points on the image silhouette $X' = \{\boldsymbol{x'_0}, \boldsymbol{x'_1}, .., \boldsymbol{x'_l}\}$. Additionally assume that the pose of the person is known at that time, such that the projected body model aligns with the observed image as in the first image of figure 8. If the person now moves a little and an image I_{t+1} is taken, it is possible to capture the motion by estimating the relative joint angles of the body between the frames I_t and I_{t+1}. The pose estimation problem is to find the parameters $\hat{\boldsymbol{\theta}}$ that best fit the transformed and projected model points to the $k + l$ correspondences. This can be formulated as follows:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{k} w_i \ |\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{x_i}) - \boldsymbol{y_i}|^2 + \sum_{j=1}^{l} w_j \ \left[(\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{x_{k+j}}) - \boldsymbol{x'_i})^T \boldsymbol{n_j} - d_i\right]^2 \tag{9}$$

This problem is again a *Nonlinear Least Squares* and is solved with the *Gauss-Newton method* [3]. The necessary Jacobian is a row-wise combination of the Jacobians from equation (1) and (7).

To get the initial pose, the user has to position the model manually in a near vicinity to the correct image position. After a few ICP iterations, the initial correct pose is found.

4.1 Arm Tracking from Silhouette and PMD-Data

A *Photonic Mixer Device* (PMD) is able to measure the distance to scene objects in its field of view. Similar to laser range scanners it is based on the time-of-flight of light. In contrast to the rather expensive laser range scanners, which usually give only one line of distances at a time. A PMD device gives distance values for a complete volume at a time. The construction and working principle is similar to conventional cameras. The time of flight is measured by phase differences between modulated emitted light and received light. To become more invariant to scene illumination and less disturbing, infrared light is used. More details can be found in [10].

Fig. 5. Setup used for the arm tracking. PMD camera on the top with IR-LEDs next to it.

In figure 5 the setup used in the experiments is shown. On the top one sees the PMD camera with Infrared-LEDs next to it. On the bottom is a conventional cameras installed. The PMD-depth image is best visualized with a view on the

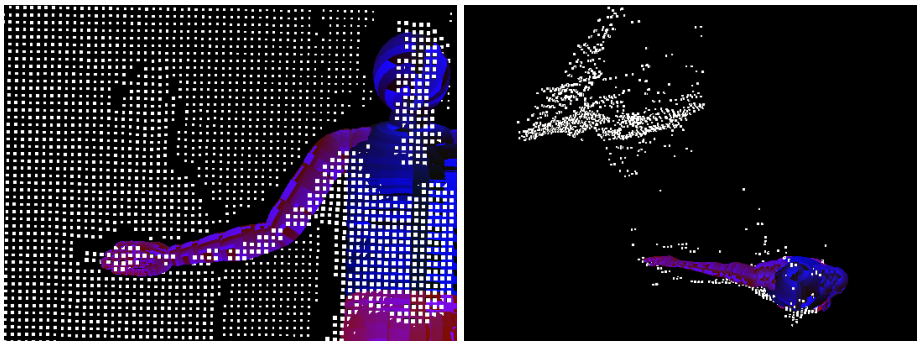

Fig. 6. Two views on the depth points, variance filtered

resulting 3D-scene points as shown in figure 6. Where one sees two views on the same point cloud from different angles. The depth image has been altered to eliminate erroneous depth values between fore- and background. To reduce

the influence of in-between-points and that of outliers, a variance filter is run on the depth image, that calculates the variance within a 3×3 window and sets all pixels with a deviation larger than a threshold to zero. Typical values are in between 0.1m and 0.25m.

5 Results

The field of view of the PMD with 20 degrees is rather small and could not be exchanged with other lenses, because the lens has a special daylight filter. Therefore the compromise between a large visible scene volume and low outlier rate is taken, which is at approx. 3m distance to the camera. In this distance the motion of one arm is completely visible. The motion in the following sequence is estimated from 3D-point-3D-point correspondences and 3D-2D-line correspondences established from silhouette information. The motion of shoulder and elbow as well as global translation and rotation were estimated, all together 10 DOF.

The motion of the right arm could be successfully tracked over the whole length of different sequences. Even though the background is non-static and cluttered (a person is walking around in the background) silhouette correspondences are accurate as visible in figure 7. This is achieved by the combination of grey value gradient and color histograms. An example sequence of 670 frames which was recorded with 7 FPS is shown in figure 8. Depicted is the image of the conventional camera superimposed with the estimated model pose. When the arms are moving in front of the body, there is not enough silhouette visible for a valid single view tracking from silhouette data alone. In that case the tracking relies on the depth data and becomes less accurate. The accuracy of the estimation is limited by the accuracy of the fitted model, which does not reveal the exact person's shape in the shoulder region.

Experiments with depth data alone showed, that the estimation is less accurate and during the 670 frame sequence tracking was lost for 50 frames. The

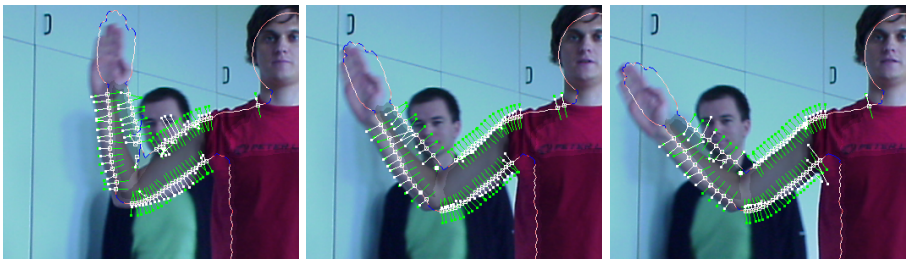

Fig. 7. Silhouette correspondences are accurate though the background is very dynamic.

depicted body model has 90000 points and 86000 triangles and processing time was about 1.5 seconds per frame. For this type of motion however a less detailed model is sufficient. In our experiments with a model consisting of a 10000 points and approx. 3500 triangles the processing time was about 5FPS on a standard PC Pentium IV 3GHz.

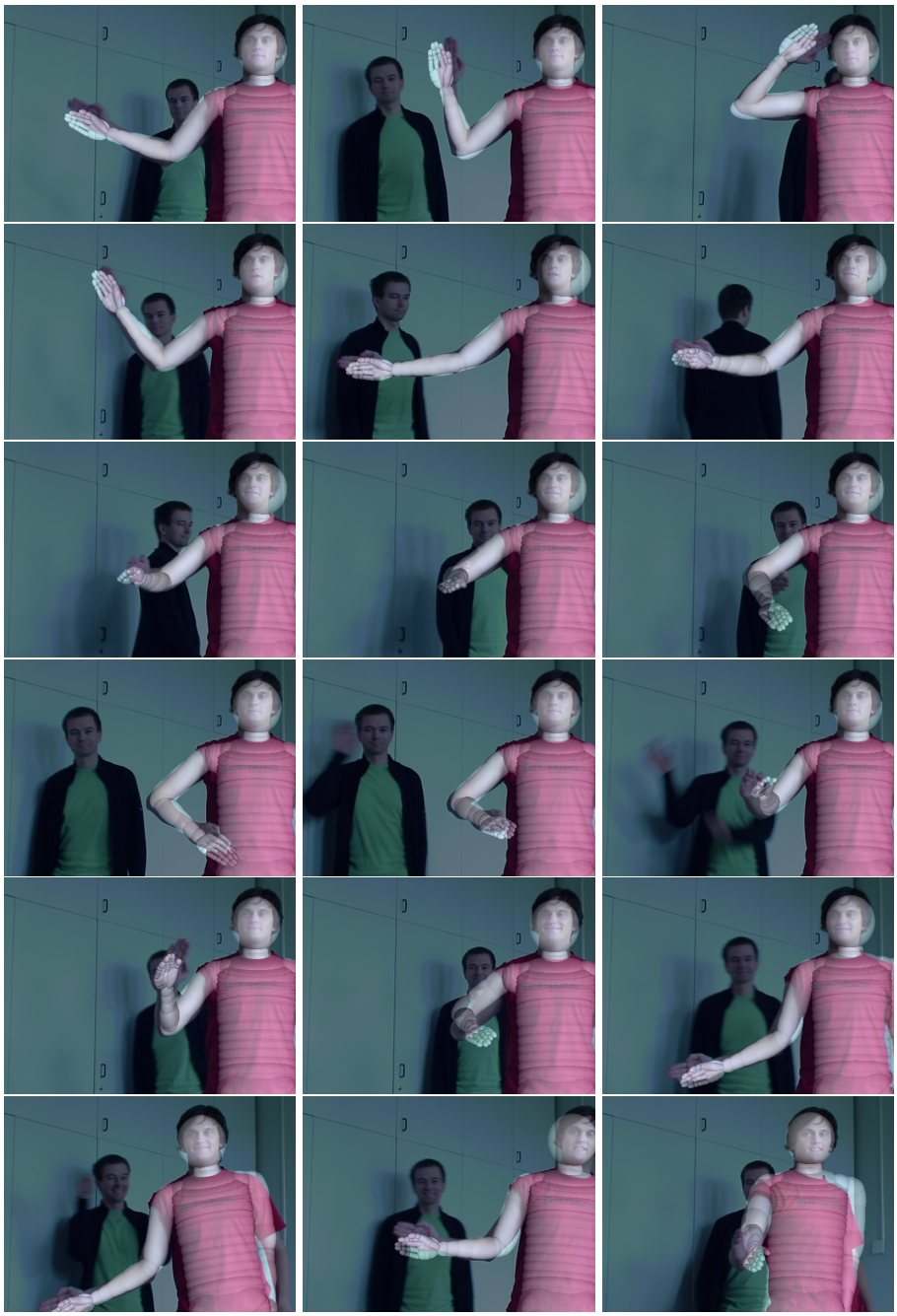

Fig. 8. Result sequence with dynamic background. The estimated model pose is overlayed on the original camera image. Ten DOF were estimated.

6 Conclusions

We showed how estimation of human motion can be derived from point transformations of an articulated object. Our approach uses a full perspective camera model and minimizes errors where they are observed, i.e. in the image plane. The combination of depth and silhouette information by color histograms and gradients allows to establish correct correspondences in spite of non-static background and people wearing normal clothing. Therefore the approach allows moving cameras as well. Ongoing research analyzes the quality of depth information of the PMD and stereo algorithms. We expect the depth data from stereo to be less accurate, but also exhibit less outliers than the PMD. Open problems are the necessary known initial pose and the need of a fitted body model, because the accuracy of the fitted model is a lower bound on the accuracy of the estimation.

References

1. M. Bray, E. Koller-Meier, P. Mueller, L. Van Gool, and N. N. Schraudolph. 3D Hand Tracking by Rapid Stochastic Gradient Descent Using a Skinning Model. In *CVMP*. IEE, March 2004.
2. C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proceeding IEEE CVPR*, pages 8–15, 1998.
3. Edwin K.P. Chong and Stanislaw H. Zak. *An Introduction to Optimization, Second Edition*, chapter 9. Wiley, 2001.
4. Lars Mündermann et al. Validation Of A Markerless Motion Capture System For The Calculation Of Lower Extremity Kinematics. In *Proc. American Society of Biomechanics*, Cleveland, USA, 2005.
5. S. Goldenstein, C. Vogler, and D. Metaxas. Statistical Cue Integration in DAG Deformable Models. *PAMI*, 25(7):801–813, 2003.
6. D. Grest, D. Herzog, and R. Koch. Human Model Fitting from Monocular Posture Images. In *Proc. of VMV*, Nov. 2005.
7. D. Grest, J. Woetzel, and R. Koch. Nonlinear Body Pose Estimation from Depth Images. In *Proc. of DAGM*, Vienna, Sept. 2005.
8. I. Kakadiaris and D. Metaxas. Model-Based Estimation of 3D Human Motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22(12), 2000.
9. T. B. Moeslund, A. Hilton, and V. Krüger. A Survey of Advances in Vision-Based Human Motion Capture Analysis. *Journal of CVIU*, 2006.
10. T. Möller, H. Kraft, J. Frey, M. Albrecht, and R. Lange. Robust 3D Measurement with PMD Sensors. In *IEEE PacRim*, 2005.
11. M. Niskanen, E. Boyer, and R. Horaud. Articulated motion capture from 3-D points and normals. In *CVMP*, London, 2005.
12. Ralf Plaenkers and Pascal Fua. Model-Based Silhouette Extraction for Accurate People Tracking. In *Proc. of ECCV*, pages 325–339. Springer-Verlag, 2002.
13. B. Rosenhahn, U. Kersting, D. Smith, J. Gurney, T. Brox, and R. Klette. A System for Marker-Less Human Motion Estimation . In W. Kropatsch, editor, *DAGM*, Wien, Austria, Sept. 2005.