

Robust Estimation of Human Body Kinematics from Video

Aleš Ude

Kawato Dynamic Brain Project
Exploratory Research for Advanced Technology (ERATO)
Japan Science and Technology Corporation (JST)
2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

Abstract

This paper addresses the problem of estimating the human body motion from video. Its main contribution is the introduction of a new robust optimization framework that leads to reliable and accurate body tracking and posture recovery. The proposed approach is resistant to occlusions and demonstrates that it is possible to treat different problems arising in human motion analysis in a unified way without using many decision thresholds. The implemented system requires only a standard CCD camera and no special markers on the body. We present experimental results showing the reliability of the implemented tracker.

1 Introduction

The understanding of human actions and intentions by vision is essential for the design of intelligent robot systems capable of working in environments populated by humans. The key feature needed to equip a robot with such a capability is the ability to track and estimate the articulated body motion of a person. Other applications of human motion estimation include the creation of realistic computer animations by motion capture, virtual reality, medicine and sports (biomechanics). The main topic of this paper is the recovery of articulated body motions directly from image sequences without using any specialized magnetic or optical tracking devices. This is a very difficult problem because, firstly, only 2-D images are available and secondly, unlike when using an optical tracking device, the correspondences between image and body points are not known.

There has been a great spurt of interest in human motion analysis from image sequences in recent years. A significant part of this research dealt with the recovery of kinematic [3, 5, 4, 9, 10] and dynamic [15] motion parameters

from images captured by one or more cameras. Results presented in these papers demonstrate that it is possible to recover at least some of the human motion parameters from video. Theoretical results about the conditions under which the parameters of articulated motions can be estimated by vision [7, 11] support this experimental work. Approaches based on nonlinear least squares tracking techniques that minimize a region-based criterion function over the set of motion parameters have proven to be especially effective [3, 9]. Potentially, these approaches can treat the problems of image segmentation, feature tracking and motion estimation simultaneously. An example of a more traditional approach that decomposes the problem of human motion capture into a body tracking stage and a motion estimation stage is the work of Wren et al. [14, 15].

In this paper we propose to incorporate the nonlinear least squares tracking approach into a robust optimization framework. This makes the tracking process more resistant to occlusions and other sources of model violations. We begin by introducing a standard approach for the kinematic modeling of articulated structures. Using a kinematic model we relate the 3-D body motion directly to the image motion and show how to estimate the kinematic parameters by use of a robust optimization framework. We conclude the paper by presenting some experimental results illustrating the performance of our motion estimation scheme.

2 Kinematic and Geometric Modeling

To measure the motion of a human body by vision, we need to be able to relate the points on the body to the image points. A 3-D model of the human body is needed to accomplish this task. An extensive discussion of this topic can be found in [1]. For our purposes it is sufficient to



Figure 1: Computer graphics model of a human body

model the motion of a human body as an articulated motion of rigid body parts. At the moment we model the geometry of body parts by simple volumetric primitives like 3-D ellipsoids. However, we are currently working on the incorporation of a more accurate body model (see Fig. 1) into our motion estimation scheme. It will be clear from the discussion in this paper that the usage of accurate computer graphics models – apart from the fact that rendering of such models is more complex than rendering of simple volumetric primitives – does not require any changes in our motion estimation scheme.

Articulated motions can be represented by kinematic chains. There are many parameter systems that can be used to model kinematic chains. The most well-known one is the Denavit-Hartenberg parameterization, which is a de facto standard in robotics. An increasingly popular alternative are the twist coordinates [8], which were first used to model the human body motion in [3]. We also used twist coordinates to parameterize the articulated body motion in our experiments. We do not expect, however, that the choice of one or the other kinematic parameter system would have a great impact on the estimation of motion parameters in our framework.

Regardless of the parameter system used to represent the human body motion, we can characterize the body motion by a mapping describing the forward kinematics of the underlying mechanical structure. The forward kinematics map specifies the position and orientation of the local frame attached to the body segment relative to the base frame. It is given by a series of matrix multiplications, parameterized by joint angles $\theta_1, \dots, \theta_n$. In homogeneous coordinates, we can calculate the position of a point attached to the k -th body segment after motion as follows

$$\mathbf{g}_k(\theta_1, \dots, \theta_k) \cdot \mathbf{x}^0, \quad (1)$$

where \mathbf{x}^0 is the position of a point attached to the k -th body segment given in the local body coordinate frame and \mathbf{g}_k is the mapping describing the forward kinematics of the first

k segments in the kinematic chain. For each fixed parameter set $\theta_1, \dots, \theta_k$, $\mathbf{g}_k(\theta_1, \dots, \theta_k)$ is a 4 by 4 matrix specifying a rigid body transformation in homogeneous coordinates. A complete model of the human body consists of a number of kinematic chains. However, consideration of this fact would only complicate the notation and does not influence the resulting estimation problems, and we neglect this fact in the rest of the paper.

Unlike in many problems in robotics, in which the position and the orientation of the base frame is fixed, we must consider the possibility that the base frame is moving. This motion can be caused either by human body motion or by camera motion. Hence the position of a body point after motion is given by

$$\mathbf{x} = \mathbf{g}(\mathbf{R}, \mathbf{d}) \cdot \mathbf{g}_k(\theta_1, \dots, \theta_k) \cdot \mathbf{x}^0, \quad (2)$$

where $\mathbf{g}(\mathbf{R}, \mathbf{d})$ is the homogeneous matrix corresponding to rotation \mathbf{R} and translation \mathbf{d} of the base coordinate frame with respect to the camera frame.

3 Relating human body appearance and motion to images

Given the body and the camera model as well as the current body posture, the relationship between the corresponding image and body points can be expressed by a combination of kinematic and camera transformations. We use the perspective projection to model the geometry of the camera transformation. This results in the following relationship between the image point \mathbf{u} and the corresponding world point \mathbf{x}

$$\begin{aligned} u &= Fx/z, \\ v &= Fy/z, \end{aligned} \quad (3)$$

where F is the focal length of the camera. We denote this mapping by \mathbf{f} , hence $\mathbf{u} = \mathbf{f}(F, \mathbf{x})$.

Let the coordinates of a body point in a local body coordinate frame be denoted by \mathbf{x}^0 . Assuming that this point belongs to the k -th body segment and that the configuration of the body at time t is given by $\mathbf{R}(t), \mathbf{d}(t), \theta_1(t), \dots, \theta_k(t)$, its image position $\mathbf{u}(t)$ after motion and projection can be calculated by a successive application of mappings (2) and (3)

$$\mathbf{u}(t) = \mathbf{f}[F(t), \mathbf{g}(\mathbf{R}(t), \mathbf{d}(t)) \cdot \mathbf{g}_k(\theta_1(t), \dots, \theta_k(t)) \cdot \mathbf{x}^0]. \quad (4)$$

While it is possible to model the body and the camera geometry accurately, it is very difficult to model the reflectance function of the body surface, lighting conditions and photometric properties of cameras. If these properties were modeled, then we could predict the intensity and

color of a body point projected onto the image plane and calculate the difference between the brightness of an image point \mathbf{u} and the predicted brightness of a corresponding body point \mathbf{x}^0

$$e(F(t), \mathbf{H}(t), \theta_1(t), \dots, \theta_n(t); \mathbf{x}^0) = I(\mathbf{u}, t) - \hat{I}(F(t), \mathbf{R}(t), \mathbf{d}(t), \theta_1(t), \dots, \theta_k(t); \mathbf{x}^0). \quad (5)$$

Here I denotes the measured, possibly filtered, image and \hat{I} denotes the predicted image of the body. Despite the difficulties involved with such a formulation, Rehg and Kanade [9] were able to predict the appearance of the human hand using a set of hand templates. They minimized the sum of squared differences using a simple gradient-based minimization algorithm to estimate the kinematic parameters of a hand. Assuming that a good initial approximation for the hand configuration is available (which is the case if the sampling rate is high enough), they were able to track the hand by sequentially minimizing their criterion function.

Unfortunately, the necessity to predict the image brightness often forces us to make assumptions which are too restrictive in practice. An alternative approach was inspired by a region-based framework for the calculation of optical flow fields [2]. A standard assumption in the estimation of optical flow fields is brightness constancy. This assumption states that the image brightness of a world point does not change in an image. If a world point is projected onto the image point \mathbf{u} at time t and onto the image point $\mathbf{u} + \Delta\mathbf{u}$ at time $t + \Delta t$, then the brightness constancy can be expressed mathematically as

$$I(\mathbf{u}, t) = I(\mathbf{u} + \Delta\mathbf{u}, t + \Delta t), \quad (6)$$

where $I(\mathbf{u}, t)$ denotes the image brightness at pixel \mathbf{u} at time t .

Assuming that the change in the body configuration and in the focal length of the camera from t to $t + \Delta t$ is given by $(\Delta F, \Delta\mathbf{H}, \Delta\theta_1, \dots, \Delta\theta_n)$, we have the following relationship between the motion of the body point and the motion of the image point

$$\mathbf{u}(t) + \Delta\mathbf{u} = f[F(t) + \Delta F, \Delta\mathbf{H} \cdot \mathbf{g}(\mathbf{R}(t), \mathbf{d}(t)) \cdot \mathbf{g}_k(\theta_1(t) + \Delta\theta_1, \dots, \theta_k(t) + \Delta\theta_k) \cdot \mathbf{x}^0],$$

where $\Delta\mathbf{H}$ denotes the change in the body position and orientation specified by a 4 by 4 homogeneous matrix. Bregler and Malik [3] were the first to exchange the affine motion model, which is often used in the estimation of optical flow fields, with the above kinematic model (unlike us they used the orthographic projection to model the camera transformation). This leads to the following expression for the violation of the brightness constancy at projection of

point \mathbf{x}^0

$$e(\Delta F, \Delta\mathbf{H}, \Delta\theta_1, \dots, \Delta\theta_n; \mathbf{x}^0) = I(f[F(t), \mathbf{g}(\mathbf{R}(t), \mathbf{d}(t)) \cdot \mathbf{g}_k(\theta_1(t), \dots, \theta_k(t)) \cdot \mathbf{x}^0], t) - I(f[F(t) + \Delta F, \Delta\mathbf{H} \cdot \mathbf{g}(\mathbf{R}(t), \mathbf{d}(t)) \cdot \mathbf{g}_k(\theta_1(t) + \Delta\theta_1, \dots, \theta_k(t) + \Delta\theta_k) \cdot \mathbf{x}^0], t + \Delta t) \quad (7)$$

If the body configuration and the camera focal length at time t are known, then the body configuration and the camera focal length at time $t + \Delta t$ can be calculated by minimizing the violation of the brightness constancy between the two measurement time instants. This amounts to optimization of the following criterion¹

$$E(\Delta F, \Delta\mathbf{H}, \Delta\theta_1, \dots, \Delta\theta_n) = \sum_{k=1}^n \sum_{\mathbf{x}^0 \in \mathcal{B}_k(t)} e(\Delta F, \Delta\mathbf{H}, \Delta\theta_1, \dots, \Delta\theta_n; \mathbf{x}^0)^2, \quad (8)$$

where $\mathcal{B}_k(t)$ denotes the set of visible points which belong to the k -th body segment and projects onto the pixels of the captured image.

In practice $\mathcal{B}_k(t)$ can be determined by calculating the projection of a geometric model of the observed body from the given configuration $\mathbf{R}(t), \mathbf{d}(t), \theta_1(t), \dots, \theta_k(t)$ onto an image plane. At the moment, we approximate the body segments by 3-D ellipsoids. However, a model of an arbitrary complexity can be used as long as it is possible to render it onto an image plane from a given configuration. This is a significant advantage over approaches that do not allow the use of standard computer graphics models.

4 Making Gauss-Newton Iteration More Robust

A straightforward minimization of the least squares criterion function (8) leads to problems because it is not possible to determine the sets $\mathcal{B}_k(t)$ without errors in practice. Points that are falsely classified as body points and assigned to $\mathcal{B}_k(t)$ must be treated as outliers by the optimization algorithm. Moreover, since the change in body configuration is not known in advance, even some of the points from $\mathcal{B}_k(t)$ that really belong to the body will not be projected onto the body image at the next time instant $t + \Delta t$. Such points must also be treated as outliers. An EM-based segmentation algorithm was proposed to solve these problems [3]. However, the results from the optical

¹It is straightforward to include in the optimization criterion (8) residuals stemming from images taken by different cameras. The usage of color information is also possible and results in 3 residuals per pixel (each color channel contributes one residual).

flow literature suggest that a robust optimization approach might yield a better solution.

Robust estimators have proven to be effective for fitting model parameters to a data set in many computer vision problems [12]. Among the most useful robust estimators are the M-estimators (maximum likelihood type estimators), which were thoroughly studied by Huber [6] for the case of linear regression. A straightforward application of these estimators suggests that instead of minimizing the sum of squares (8), we should minimize a sum of less rapidly increasing functions of the residuals

$$E(\Delta F, \Delta \mathbf{H}, \Delta \theta_1, \dots, \Delta \theta_n) = \sum_{k=1}^n \sum_{\mathbf{x}^0 \in \mathcal{B}_k(t)} \rho(e(\Delta F, \Delta \mathbf{H}, \Delta \theta_1, \dots, \Delta \theta_n; \mathbf{x}^0)). \quad (9)$$

Many ρ functions with different properties have been proposed in the literature [2, 12]. Unfortunately, the minimization of the above criterion function is a very difficult problem. First of all, the number of residuals, which is equal to the number of pixels classified as part of the body image, is very large. For example, there were about 10,000 residuals per image when the videos shown in Fig. 2 and 3 were processed. Furthermore, unlike in the case of linear regression, where the residual function e is a linear function of parameters, here we have a nonlinear residual function. Finally, all useful ρ functions are nonconvex. All these factors make the direct minimization of (9) a very difficult problem and we in fact encountered convergence problems when we tried to minimize it directly using general optimization techniques available in Matlab.

For this reason we decided that instead of minimizing the robust criterion (9), we would rather try to calculate the minimum of criterion (8) in a robust way. Standard methods for nonlinear least squares problems are the Gauss-Newton and the Levenberg-Marquardt iteration. These two iterations are based on the computation of a vector-valued function $e = [e(\Delta F, \Delta \mathbf{H}, \Delta \theta_1, \dots, \Delta \theta_n; \mathbf{x}^0)]_{\mathbf{x}^0 \in \mathcal{B}_k(t)}$ and of its Jacobian matrix. The Jacobian matrix of the residuals can be calculated using the chain rule which enables us to combine the Jacobian of the underlying kinematic structure, the Jacobian of the camera transformation and the numerically calculated image gradient into the residual Jacobian. In this paper we consider only the Gauss-Newton iteration in which the modification for the current estimate of motion parameters is calculated by solving

$$\mathbf{J} \cdot [\Delta F, \Delta \mathbf{r}^T, \Delta \mathbf{d}^T, \Delta \theta_1, \dots, \Delta \theta_n]^T = -e. \quad (10)$$

Here the 4 by 4 homogeneous matrix specifying the modification for the position and orientation is given by $\Delta \mathbf{H} = \mathbf{g}(\exp(\Delta \mathbf{r}), \Delta \mathbf{d})$ (transformation \mathbf{g} is defined as in Eq.

(2))². Eq. (10) specifies an overconstrained system of linear equations. Instead of solving it in a least squares sense, we propose to calculate the solution by use of a robust estimator. This makes the underlying iteration more robust while, provided the robust estimator works well, its convergence properties are not altered. The main advantage of such an approach over direct minimization of the robust criterion (9) is that it calculates the optimal estimate by solving two well understood problems: nonlinear least squares optimization and robust linear regression. Many different estimators for the calculation of a robust linear regression estimate have been proposed in the literature. The most commonly used ones are the M-estimators and the LMS-estimator (least median of squares). We omit the details and relate the interested reader to the literature [6, 12].

To calculate the body configuration at time $t + \Delta t$, we first project the body model from the configuration calculated at time t onto the image taken at this time instant. Body points corresponding to pixels contained in the synthetic body image are assigned to the sets $\mathcal{B}_k(t)$ and the iteration is initialized by taking zero as an initial approximation for the change in the body configuration.

The body configuration is not known at the beginning of the tracking process, i.e. at $t = 0$. Currently, the tracking algorithm must be initialized manually by clicking with a mouse on the boundary of body segments which are then approximated by 2-D ellipses. Using this information and the body model we can calculate the initial parameters needed by the tracker.

It turns out that the estimation of focal length under perspective projection model often leads to convergence problems. However, making even a rough estimate for the focal length fixed throughout the optimization process results in a reliable convergence. Alternatively, the orthographic projection can be used.

5 Experiments and Future Work

In our experiments we utilized the Gauss-Newton iteration and M-estimators to track the human motion in a single video source. We made use of the Geman-McClure function (see Fig. 5), which was employed before for the calculation of optical flow fields [2], to specify the error function for the M-estimator. The Geman-McClure error function is defined by

$$\rho(x, \sigma) = \frac{x^2}{\sigma + x^2}. \quad (11)$$

²Note that the calculation of $\Delta \mathbf{H} = \mathbf{g}(\exp(\Delta \mathbf{r}), \Delta \mathbf{d})$ is more difficult than the calculation of other motion parameters because the set of all rotations is not a vector space. See [13] for the treatment of this problem.

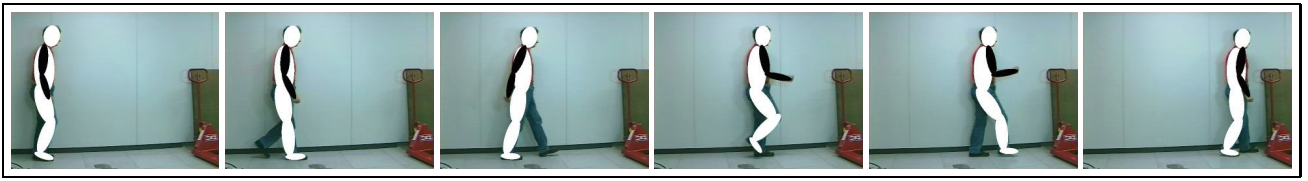


Figure 2: Projection of volumetric primitives building the body model from the estimated postures

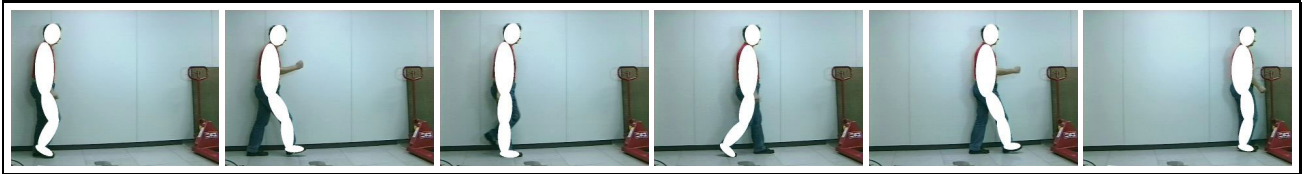


Figure 3: Tracking of a similar motion as above with an incomplete body model. Although the arm has not been modeled, the system was able to track the rest of the body reliably. This shows that our approach can handle a significant amount of occlusion.



Figure 4: Arm tracking in a 12 seconds long video

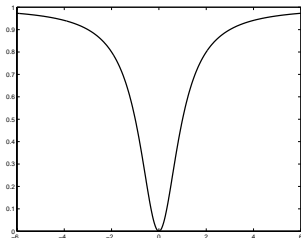


Figure 5: Geman-McClure error function

As its influence function $\psi(x) = \rho'(x)$ tends to 0 as $|x|$ tends to infinity, points that violate the model constraints significantly have only little influence on the estimated parameters. Typical convergence of the resulting iteration is shown in Tab. 1.

We tested our approach by processing several real body motions. Results from three different experiments are shown in Fig. 2, 3 and 4 (please check the web site <http://www.erato.atr.co.jp/~ude/> to see the QuickTime movies of the estimated motions). These motions were recorded on a video tape and then digitized using a standard video system. The digitized image sequences contained 30 images per second of video. We use different

Table 1: Convergence of the robustized Gauss-Newton iteration

Norm of the iteration step
3.342622e+01
5.367020e+00
1.602283e-01
1.229976e-03
1.182431e-04
1.182889e-05
1.158693e-06

image sizes: 240×320 and 240×640 were used in the presented experiments. The images were smoothed by a Gaussian filter before being processed by the tracker. Although the quality of such images is rather low, we could successfully process the presented walking sequences as well as the arm motion sequence. The trajectories of the estimated angles of the arm motion from Fig. 4 are shown in Fig. 6. Careful analysis of the video (see the web site)

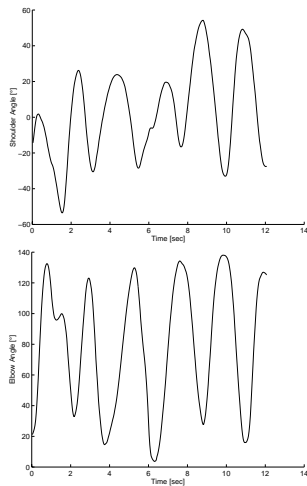


Figure 6: Motion parameters (shoulder and elbow angle) estimated by processing the motion from Fig. 4

shows that the measured motion parameters make sense.

Our future work will concentrate on learning texture maps of the observed surfaces. This will make the proposed approach less sensitive to the accuracy of the initial approximation for the body configuration and will also prevent the model from drifting from its image. In the current implementation such a drift can occur when there is a significant error in the estimated body configuration which consequently introduces a significant error into a body model used to estimate the body configuration at the next measurement time instant.

Acknowledgment: Aleš Ude is on leave from the Department of Automatics, Biocybernetics and Robotics, Jožef Stefan Institute, Ljubljana, Slovenia.

References

- [1] N. I. Badler, C. B. Phillips, and B. L. Weber. *Simulating Humans: Computer Graphics Animation and Control*. Oxford University Press, Oxford, 1993.
- [2] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow-fields. *Computer Vision and Image Understanding*, 63(1):75–104, January 1996.
- [3] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Santa Barbara, California, June 1998.
- [4] D. M. Gavrilu and L. S. Davis. 3-D model-based tracking of humans in action: a multiview approach. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pages 73–80, San Francisco, California, June 1996.
- [5] L. Goncalves, E. Di Bernardo, and P. Perona. Monocular tracking of the human arm in 3D. In *Proc. 5th Int. Conf. Computer Vision*, pages 764–770, Cambridge, Massachusetts, June 1995.
- [6] P. J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.
- [7] D. D. Morris and J. M. Rehg. Singularity analysis for articulated object tracking. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, Santa Barbara, California, June 1998.
- [8] R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Boca Raton, New York, 1994.
- [9] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. 5th Int. Conf. Computer Vision*, pages 612–617, Cambridge, Massachusetts, June 1995.
- [10] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, 59(1):94–115, January 1994.
- [11] R. Sharma and S. Hutchinson. Motion perceptibility and its application for active vision-based servo control. *IEEE Trans. Robotics Automat.*, 13(4):607–617, August 1997.
- [12] C. V. Stewart. Bias in robust estimation caused by discontinuities and multiple structures. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(8):818–833, August 1997.
- [13] A. Ude. Nonlinear least squares optimisation of unit quaternion functions for pose estimation from corresponding features. In *Proc. 14th Int. Conf. Pattern Recognition*, Brisbane, Australia, August 1998.
- [14] C. R. Wren, A. Azarbayejani, T. Darell, and A. P. Pentland. Pfunder: Real time tracking of the human body. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(7):780–785, July 1997.
- [15] C. R. Wren and A. P. Pentland. Dynamic model of human motion. In *Proc. Third Int. Conf. Automatic Face and Gesture Recognition*, pages 22–27, Nara, Japan, 1998.