Estimating 3D Hand Position and Orientation Using Stereo

Afshin Sepehri Yaser Yacoob Larry S. Davis
Institute for Advanced Computer Studies
University of Maryland, College Park
{afshin, yaser, lsd}@umiacs.umd.edu

Abstract

We present an approach for estimating the 3D absolute position and orientation of the hand using a planar model of the back of the hand. We use stereo cameras to first build a sparse disparity map of the estimated area of the back of the hand. Then, the best fitting plane to the disparity points is computed using robust estimation. Finally, the 3D hand plane is calculated based on the disparity plane and the six parameters of position and orientation of the back of the hand are estimated . Experimental results demonstrate the accuracy of this technique.

1. Introduction

In virtual reality applications, the need arises to facilitate manipulating virtual objects through hand motions. Accurate determination of the position and orientation of the hand is an essential step. Having the absolute position and orientation instead of a relative one (such as can be recovered from a tracking algorithm) improves interactions of the hand with objects in a virtual environment as well as the use of the hands as a means of communication. It also helps in successive processing steps such as tracking of the fingers.

The analysis of the hand pose is a challenging objective due to the combination of a complex structure with many degrees of freedom and very low texture in its appearance. Researchers have studied the problem of estimating hand and finger poses from different perspectives. One objective has been to recognize hand gestures such as sign language [1, 6]. Others tracked general hand and finger movements [7, 18]. Different hand and fingers models have been used [13, 14]. In most of the reported research a single image is used. Delamarre and Faugeras [5] used a sequence of stereo images to estimate the pose of the hand. They proposed a 3D articulated model of the hand and tracked the forces that would attract the model. Others used a sequence of images as a dual approach to stereo [17, 12].

We present a method for estimating the absolute position and orientation of the hand using a planar model of the back of the hand. We use a stereo system consisting of two small digital cameras with a baseline of $65mm$, a typical baseline of the human eyes. We assume that the cameras are looking at the back of the hand while the fingers and the arm may or may not be visible. The scene has an arbitrary background. We first detect the area of the back of the hand in the image-pairs, then build a *sparse disparity map* using a fast method and fit a plane to the disparity points using *M-estimation* [10]. Finally, the *hand plane* in 3D is computed based on the *disparity plane* and the six parameters of the position and orientation of the hand are estimated.

2. Approach

Stereo vision is a common and an effective way for 3D information recovery. The low-textured hand as well as imaging conditions such as noise levels of the sensors, different brightness, white balance, and other parameters that differ between the two cameras can make the point-matching process and finding a dense disparity map challenging. To overcome the ambiguity in pixel to pixel matching, we propose to employ a model that approximates the shape of the back of the hand as a 3D plane. Theoretically, this plane can be determined using three conjugate point pairs, however for two reasons such an ideal plane is hard to obtain: first, practically it is very challenging to find reliable point pairs with sufficient accuracy and second, since the back of the hand is not ideally a plane, finding a plane fitting just a few points may not approximate the rest of the points well. In fact, we seek a plane which passes through the majority of the points and approximates the remaining points with an acceptable error. Our proposed algorithm finds a considerable number of conjugate point pairs (with high probability of being correct) and then estimates an optimal planar model of the disparity points. This disparity map may include some wrong conjugate pairs due to the problems mentioned earlier, but they are eliminated using a robust estimation of the plane-fitting method, called *M-estimation*. This method computes an optimal disparity plane having the disparity values of a considerable fraction

of the points in the region of the back of the hand.

Finally, the X, Y and Z coordinates of the position and the orientation angles yaw, $pitch$, and $roll$ are calculated for a coordinate frame attached to the back of the hand. The 3D plane used in these transformations is calculated from the computed disparity plane.

3. Estimating the Area of Back of the Hand

As a first step, We segment the hand region by applying a skin-detection algorithm to each image [16]. Since the region of interest is the back of the hand, we need to detect its area as accurately as possible to locate good points to fit with a plane. We also need to remove the fingers and the arm from the segmented hand area. We rely on the fact that the area of the back of the hand is usually the widest part of the hand with the exception of some of the upper areas of the arm. Also, due to the presence of the fingers, the number of curvature maxima in the neighborhood of the back of the hand is more than the arm areas. These facts allow us to model the area of the back of the hand as union of a set of intersecting circles which is a simple, yet accurate enough model.

The following summarizes the estimation process:

1. Segment the area of the hand. This includes detecting the hand via a skin detection algorithm, removing the background area, and filtering the rest of the image using a floodfill algorithm to fill the holes in the area (e.g. hand hairs) and achieve a uniform area.

2. Find the largest interior circle (LIC) of the segmented area using the distance transform. This circle is likely to be located on the back of the hand. However to avoid circles in the area of the arm, we find the center of gravity of the curvature maxima of the hand contour and consider only those circles that contain this point. Since the fingers create more curvature maxima than the smooth edges of the straight arm, this tend to place the center point on the back of the hand.

3. Find other large interior circles with a radius larger than a given threshold (e.g. 0.8 of the radius of the LIC). The fingers inherently will not belong to a circle with such a radius even if a few of them are joined; To avoid including circles on the arm, we discard circles that do not intersect the LIC.

4. Compute the union of the area of all the obtained circles and consider it as the estimated area of the back of the hand. We do not expect this area to cover the back of the hand perfectly. Also, the largest interior circles in the two images may not exactly correspond to the same actual hand region. Nevertheless, they will have a high percentage of overlap. Figure 1 (middle row) shows back-hand areas for the sample image set of the top row.

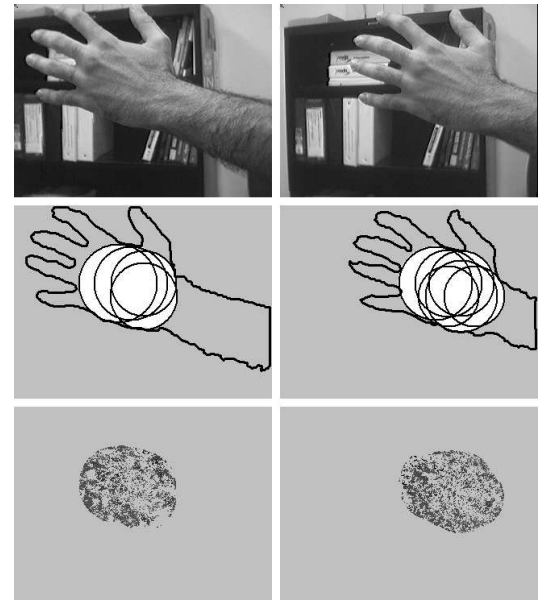

Figure 1. A Sample Image Pair. Top: Input images. Middle:Hand regions and largest interior circles. Bottom: Sparse disparity maps.

4. Computing Sparse Disparity Map

After estimating the area of the back of the hand, we find the sparse disparity map using a correlation-based method where the conjugate point pairs are detected. We identify and remove wrongly matched corresponding points by relying on the *uniqueness* of the matching and *left to right consistency* of the corresponding points. Uniqueness means each point in the left image should match to one and only one point in the right image and vice versa. Consistency of the corresponding points means that if a point p_r in the right image is the best match to point p_l in the left image, point p_l should also be the best match for p_r. An important issue here is that to have a reliable plane fit, we seek disparity values of a considerable number of the points as participants in the estimation.

An important issue is that we need a uniformly distributed sparse disparity map such that we give equal opportunity to both high-textured and low-textured areas to participate in planar fitting of next section. Instead of using a traditional method for calculating disparity for all the points in the region, we develop a new technique that is both fast and reliable. The required steps follow:

1. Select a pixel with random position (x, y) from the region of interest in left image where its disparity has not been estimated yet.

2. Search for the corresponding pixel in the right image using a correlation-based search algorithm.

3. If a unique conjugate pixel at position (x_2, y_2) was found and if right disparity map at position (x_2, y_2) is not estimated yet, search for a correspondence in the left image for pixel at (x_2, y_2).

4. If a unique conjugate pixel at position (x_3, y_3) was found and $x_3 = x_1$ and $y_3 = y_1$ then left to right consistency is satisfied and we can assign disparity value $d = x2 - x1$ to both positions (x_1, y_1) and (x_2, y_2) in left and right image disparity map respectively. If our input images are not rectified (i.e. $y_1 \neq y_3$), save $y2 - y1$ as disparity value in y direction as well.

5. Repeat steps 1 through 5 until a desired fraction of the pixels (e.g. 20%) have their disparities calculated or a certain fraction of the total number of points in the region (e.g. 70%) are tried.

The images of the bottom row in figure 1 show the final estimated sparse disparity maps for the sample image pair. Statistics for sample sequences will be given in section 7.

5. Planar modelling

We consider back of the hand as a plane in 3D with equation

$$Z = B_0 + B_1 X + B_2 Y = B_0 + B_1(\frac{x}{f}Z) + B_2(\frac{y}{f}Z) \quad (1)$$

where $P(X, Y, Z)$ is a point on the plane and $p(x, y, f)$ is the image of P on image plane. Since Z is inversely proportional to the disparity value d (i.e. $Z = \frac{\alpha}{d}$)

$$d = \frac{\alpha}{B_0} + (-\frac{B_1\alpha}{fB_0})x + (-\frac{B_2\alpha}{fB_0})y = b_0 + b_1 x + b_2 y \quad (2)$$

which means that points (x, y, d) we have found should also lie on a plane. To cope with outliers, we employ robust estimation to find the parameters of the planar model. M-estimation is a robust method of estimating the regression plane which works well in the presence of outliers. Considering the plane model

$$d_i = b_0 + b_1 x_i + b_2 y_i + e_i = \mathbf{x_i}^T \mathbf{b} + e_i$$

with $\mathbf{x_i} = (1, x_i, y_i)^T$ and $\mathbf{b} = (b_0, b_1, b_2)^T$, the general M-estimator which corresponds to the *maximum-likelihood estimator* [10], minimizes the objective function

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(d_i - \mathbf{x_i}^T \mathbf{b}) \quad (3)$$

where n is the number of points and ρ is the *influence function* [8].

Let $\psi = \rho'$ be the derivative of ρ. To minimize (3), we need to solve the system of three equations

$$\sum_{i=1}^{n} \psi(d_i - \mathbf{x_i}^T \mathbf{b})\mathbf{x_i}^T = \mathbf{0} \quad (4)$$

Defining the weight coefficients $w_i = \psi(e_i)/e_i$, the estimating equations may be rewritten as

$$\sum_{i=1}^{n} w_i(d_i - \mathbf{x_i}^T \mathbf{b})\mathbf{x_i}^T = \mathbf{0} \quad (5)$$

The solution $\mathbf{b}$ to (5) can be found using an iterative algorithm (called *iteratively reweighted least-squares, IRIS*) as follows [8]:

1. Select initial estimates $\mathbf{b}^{(0)}$ such as the least-square estimates.

2. At each iteration t, calculate residuals $e_i^{(t-1)}$ and associated weights $w_i^{(t-1)}$ from the previous iteration.

3. Solve for the new weighted-least-squares estimates

$$\mathbf{b}^{(t)} = [\mathbf{X}^T \mathbf{W}^{(t-1)} \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W}^{(t-1)} \mathbf{d} \quad (6)$$

where $\mathbf{X}$ is the matrix of points with $\mathbf{x}_i^T$ as the ith row and $\mathbf{W}^{(t-1)} = diag\{w_i^{(t-1)}\}$ is the weight matrix.

For fitting a 3D plane to our disparity data, we choose *Geman-McClure* function for ρ [9]

$$\rho(x, \sigma) = \frac{x^2}{\sigma + x^2} \quad (7)$$

Since this function has a differentiable ψ-function, it provides a more gradual transition between inliers and outliers than some other influence functions [4].

To achieve a fast convergence as well as to avoid local minima, we initialize weights $w_i^{(0)}$ with values proportional to the *confidence* of each point in disparity calculation process. This confidence can be defined as reciprocal of the sum of differences of the pixel values in the correlation windows.

6. Estimating Hand Position and Orientation

Finding the disparity plane, we can map it to the hand plane. Then by locating a coordinate frame on the hand, the position and orientation of the hand can be calculated.

We find the hand plane in 3D using the calibration information and the disparity plane. We can use (2) to find (B_0, B_1, B_2) when we have rectified images. A simple way of mapping for unrectified images is to find three points lying on this plane in 3D and then fit a plane to these three points. To find points in 3D we identify corresponding points from the disparity plane and use a simple triangulation process with the camera calibration information [15].

We define the hand plane as a transformed plane found after two rotations and one translation applied to the camera X-Y plane. Specifically, we rotate the X-Y plane with equation $Z = 0$ first about X axis and then about Y axis to transform it to $Z = B_1 X + B_2 Y$. These two rotations are called *yaw* and *pitch* respectively. Then, we translate the

plane along Z axis by constant value B_0 which makes the plane equation $Z = B_0 + B_1X + B_2Y$. Coefficient values B_1, B_2 which were already found through the plane fitting process, are used to determine the two rotation angles ψ and θ corresponding to *yaw* and *pitch* respectively as follows:

$$\psi = tan^{-1}(\frac{B_2}{\sqrt{1+B_1^2}})$$

$$\theta = tan^{-1}(-B_1)$$

Using the two rotation angles ψ and θ, and the translation vector $(0, 0, B_0)^T$, we compute the transformation matrix P, which transforms the X-Y plane to the hand plane

$$P = \begin{pmatrix} cos(\theta) & sin(\theta)sin(\psi) & sin(\theta)cos(\psi) & 0 \\ 0 & cos(\psi) & -sin(\psi) & 0 \\ -sin(\theta) & cos(\theta)sin(\psi) & cos(\theta)cos(\psi) & B_0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (8)$$

This matrix will be used in later stages of the process.

The next step is to assign a coordinate frame to the hand where the X-Y plane of this frame resides on the model plane. This coordinate frame provides the 6 parameters required to determine the position and orientation of the hand in 3D. To determine the position of the hand, we need to locate the origin of the hand frame relative to a fixed point on the hand. A good point is the center of the back of the hand which can be approximated by the center of mass of the estimated area of the back of the hand built as union of the set of circles as explained in section 3. To make a better approximation, we first find c_1 and c_2 corresponding to the center of the mass in the left and right image respectively, and then find c_2', the corresponding point of c_2 in the left image, using the calculated disparity plane of the right image and then find the midpoint of c_1 and c_2' and consider it as o_1, the image of the origin of the hand coordinate frame in the left image plane. We can then easily find o_2, the image of the origin in the right image using the disparity plane of the left image. Finally the position of the origin $O = (O_X, O_Y, O_Z)^T$ in 3D is calculated through a sample triangulation process as done earlier to find the 3D hand plane.

The rotation of the hand about the Z axis of the hand frame, the *roll*, can be computed using the orientation of the 2D silhouette points of the hand in the X-Y plane. Ignoring some infrequent cases where the arm is hidden and all fingers but thumb are bent, *roll* can be computed as the angle of the axis of the least moment of inertia [11] and is calculated as

$$\phi = \frac{1}{2}tan^{-1}(\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}})$$

where

$$\mu_{p,q} = \sum_{(x,y)} \sum_{\in R} (x - \overline{x})^p (y - \overline{y})^q$$

and

$$\overline{x} = \frac{1}{n} \sum_{(x,y)} \sum_{\in R} x$$

$$\overline{y} = \frac{1}{n} \sum_{(x,y)} \sum_{\in R} y$$

when R includes the n points of the silhouette. This will give us the direction of the arm or the rough direction of the fingers in case the arm is missing as a good approximation of true hand *roll*. In the exceptional case mentioned above, to obtain an absolute orientation, the motion between consecutive frames can be used to update the *roll* angle.

7. Experiments and Results

To measure the accuracy of the proposed technique, we compare it with a hand model computed using a set of markers on the back of the hand, finding their positions on the images manually. We compute the coordinates of those points in 3D and fit a plane to them. Figure 2 shows a sample image with markers. As depicted in the figure, positions of the markers are selected such that they cover the area of the back of the hand uniformly. This provides us a better comparison as the region-based method picks points uniformly from all over the region.

Figure 3 shows the position coordinates O_X, O_Y and O_Z and orientation angles yaw, $pitch$ and $roll$ denoted as ψ, θ, and ϕ of this *marker-based plane* as well as the *region-based plane* estimated through disparity analysis. A sequence of 30 frames were used for this experiment. The statistical results are shown in table of figure 4.

Although, the marker-based plane passes through a set of reliable points, this plane may not be the optimal plane as the shape of the back of the hand is not exactly a plane. For this reason we do not call the marker-based plane a ground truth plane as we believe the plane estimated through disparity analysis is a better approximation and gives us more reliable position and orientation parameters of the hand.

Figure 2. A sample image with markers

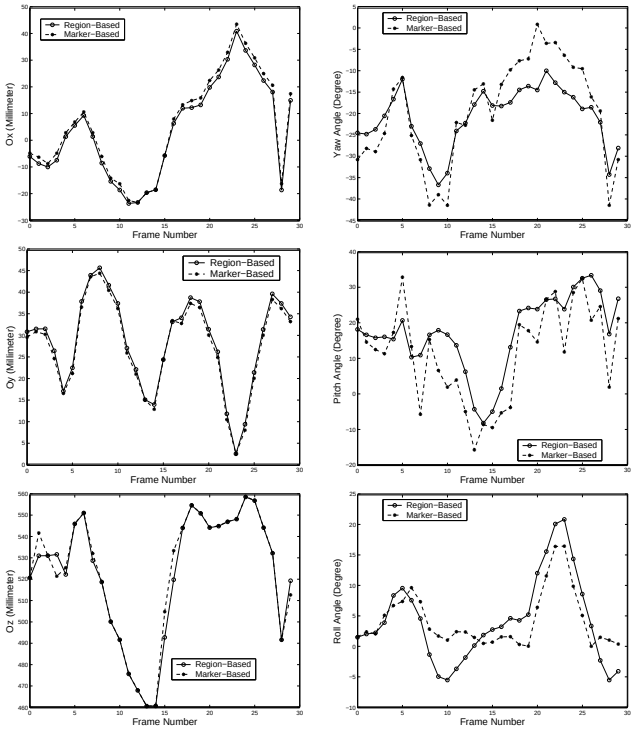

Figure 3. Experimental results. Left: Marker and region-based position values. Right: Marker and region-based orientation values.

Another useful parameter that assesses the accuracy of our algorithm captures the distribution of the disparity errors which measures how far the disparity points are from the fitted disparity plane. In other words how many of the points are outliers. This is an important issue because the M-estimation algorithm breaks down if the percentage of outliers is too high and then it diverges from the optimal plane drastically. Figure 5 shows the distribution of the errors measured by averaging the corresponding distributions over a 30 frame sequence. It is a normal distribution with mean 0.0050 and standard deviation 0.0237 which gives us a 35% rate of outliers if we define inlier-outlier threshold

	mean absolute difference	standard deviation of the absolute difference
O_Z	$1.8135mm$	$0.9215mm$
O_Y	$1.0514mm$	$0.4740mm$
O_Z	$2.0792mm$	$4.0983mm$
ψ	$5.1570°$	$3.2986°$
θ	$6.9515°$	$5.3280°$
ϕ	$3.3571°$	$1.9242°$

Figure 4. Statistical Results

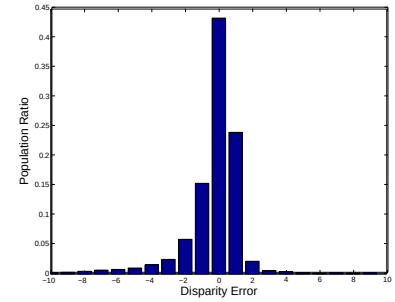

Figure 5. Distribution of the error of disparity values with respect to disparity plane. (Averaged over 30 frames)

1.5 and 9% if threshold is 2.5 levels of disparity. Therefore, the M-estimation algorithm is convergent.

Figures 6 to 9 show sample frames selected from four different image sequences showing a hand in motion. Left image of the image-pairs along with the corresponding models built based on the estimated position and orientation of the hand are depicted in the figures.

The frames in figure 6 show a hand with stretched fingers and visible arm whereas frames in figure 7 do not have visible arm.

In Figure 8 we have fingers moving freely and we can still track the back of the hand. Figure 9 show the result for a low-textured hand.

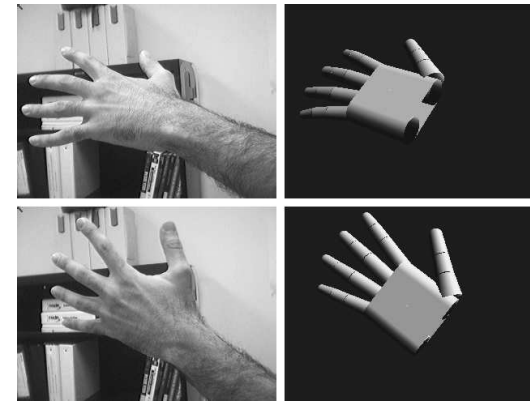

Figure 6. Experimental results: Sample input frames with visible arm

8. Summary and Conclusions

A method for finding the 3D absolute position and orientation of the hand using a planar model of the back of the hand was presented. We computed a sparse disparity map, fitted a plane to it using a M-estimation method, and calculated the plane of the hand in 3D. Also, we estimated the

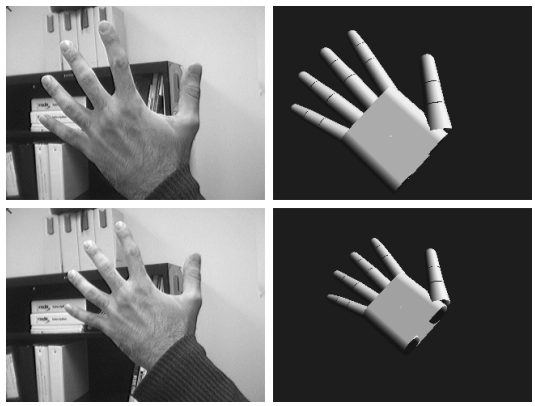

Figure 7. Experimental results: Sample input frames with invisible arm

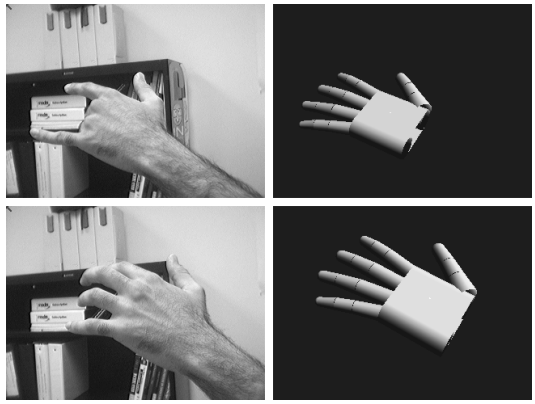

Figure 8. Experimental results: Sample input frames with moving fingers

six parameters of the hand position and orientation. Experimental results and accuracy analysis were also shown. Using the hand model created by this method can make hands a natural means of communication in virtual environments.

References

[1] V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. *IEEE Conference on Computer Vision and Pattern Recognition*, 2(18-20):II–432–9, June 2003.

[2] C. Baillard and A. Zisserman. Automatic reconstruction of piecewise planar models from multiple views. *CVPR*, June 1994.

[3] A. Bartoli, P. Sturm, and R. Horaud. Projective structure and motion from two views of a piecewise planar scene. *ICCV*, July 2001.

[4] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63(1):75–104, Jan 1996.

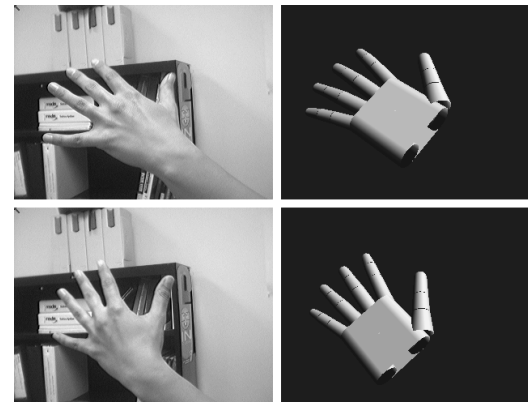

Figure 9. Experimental results: Sample input frames from a low-textured hand

[5] Q. Delamarre and O. Faugeras. Finding pose of hand in video images: a stereo-based approach. *Proceedings of FG'98*, April 1998.

[6] F. R. et al. Hand gesture recognition following the dynamics of a topology preserving network. *5th IEEE International Conference on Automatic Face and Gesture Recognition*, May 2002.

[7] S. L. et al. Using multiple cues for hand tracking and model refinement. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:443–450, 2003.

[8] J. Fox. *Robust Regression: Appendix to An R and S-PLUS Companion to Applied Regression*. SAGE Publications, 2002.

[9] S. Geman and D. E. McClure. Statistical methods for tomographic image reconstruction. *Proc. of the 46-th Session of the ISI, Bulletin of the ISI*, 52:5–21, 1987.

[10] P. J. Huber. *Robust statistics*. John Wiley and Sons, 1981.

[11] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice Hall, 1989.

[12] J. Lin, Y. Wu, and T. Huang. Capturing human hand motion in image sequences. *Workshop on Motion and Video Computing*, December 2002.

[13] J. M. Rehg and T. Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. *3rd ECCV*, volume II, May 1994.

[14] B. Stenger, P. R. S. Mendonça, and R. Cipolla. Model-based hand tracking using an unscented kalman filter. In *BMVC*, volume I, Sep. 2001.

[15] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998.

[16] X. Yin, D. Guo, and M. Xie. Hand image segmentation using color and rce neural network. *IJRAS*, 34:235–250, March 2001.

[17] H. Zhou and T. Huang. A bayesian framework for real-time 3d hand tracking in high cluttered background. *Proc. 10th Intl. Conf. on Human-Computer Interaction*, June 2003.

[18] H. Zhou and T. Huang. Tracking articulated hand motion with eigen dynamics analysis. *IEEE ICCV*, pages 1102–1109, October 2003.