# Hand Pose Estimation Using Expectation-Constrained-Maximization From Voxel Data

Shinko Y. Cheng and Mohan M. Trivedi
{sycheng, mtrivedi}@ucsd.edu
Computer Vision and Robotics Research (CVRR) Laboratory
University of California, San Diego
La Jolla CA 92037-0434
cvrr.ucsd.edu

## Abstract

*This paper describes voxel-based hand posture estimation using the Expectation Constrained-Maximization framework to estimate the parameters to a 16 segment 24 degrees-of-freedom hand model. This method uses only voxels to estimate finger segment locations and orientations. The hand model consists of a mixture of kinematically constrained 3D gaussian distributions. Results on simulated and real data show that given a good initial conditions, this proposed iterative model parameter estimation process converges to the "voxelized" hand.*

## 1. Introduction

Voxel-based hand posture estimation is the task of fitting an articulated kinematic model to observed voxel reconstructions of the hand. The proposed method assumes volumetric data of the hand is collected in voxel form. The challenge lies in extracting from this unlabeled 3D voxel data the 27 parameters that describe the posture and location of the hand. Accurate hand posture estimation have several foreseeable applications including human-computer interface design, measuring range of motion in physiotherapy, or as a markerless motion capture system in 3D animation and robot training.

Given only the volumetric representation of the hand, we are most concerned with the estimation of two sets of hand attributes: 1) dimensions of each segment in height, width and depth, including the palm, thumb and finger segments, and 2) the location and orientation of each segment. The second problem is the one we address in this paper.

The proposed technique assumes a 16 class Gaussian mixture model to describe the spatial distribution of the voxels. Each class describes a finger segment or palm that is kinematically constrained according to a 24 degree-of-freedom (DOF) hand model. The model parameters are estimated using a modified Expectation-Maximization algorithm. The modification is instead of a global optimization of parameters in the M-step, a constrained optimization takes place to project EM iterates onto the feasible posture space [8], hence its name Expectation-Constrained-Maximization (ECM). The principles of maximum likelihood estimation still hold. The solution manifold is merely reduced.

The algorithm is evaluated on simulated and real voxelized hand data. Simulated hand data are constructed from cylinders of voxels. A real-time shape-from-silhouette system [4, 13] generates the actual hand data. Experiments using simulated hand data provides the benefit of making performance measurements simple by having both the estimated from actual segment location and joint angles available.

### 1.1. Previous Work

Hand posture estimation has been fairly well studied in the recent past. We describe three of such relevant efforts below.

This work borrows from Hunter, et al. [8] on their proposed object/observer modeling and ECM estimation framework for estimating posture parameters in a 15 component 31 DOF articulated model of the human body. Instead of operating on single 2D silhouette images as their work emphasizes on processing video, our work uses several silhouette images to generate the visual hull and using that as the observed process in our posture estimation. This requires a modification in the observed process model, making the model actually simpler. Simulated data is introduced to provide a measure of accuracy of the pose estimation technique by being able to compare actual and estimated joint angles and finger segment positions.

Our work is very similar in goal to the work by Ueda, et al. [15] who also proposes the use of shape-from-silhouette generated voxel reconstructions of the hand as well as using the hand voxel data alone to estimate the hand posture parameters. The two efforts differ in the estimator, whereas in their work, a skeletal hand model of the subject's hand is used to generate torques to fit the model to voxel data, we used a statistical model and a system of constraining equations to ensure a kinematically valid posture estimate.

The remaining sections are organized as follows. Section 2 explains the process of real-time shape-from-silhouette voxel reconstruction of the hand. Sections 3 and 4 describes the kinematically constrained gaussian mixture model used to describe the hand and the ECM algorithm to estimate the model parameters. Sections 5 and 6 describes the experimental results and general conclusions of the hand posture estimation method.

## 2. SFS Voxel Reconstruction

Images of the hand from various known viewpoints are captured simultaneously. Each image is then segmented using a statistical background subtraction with shadow suppression [6]. The result of the segmentation are silhouettes of the hand with which the voxel reconstruction of the hand is found using HybridVolume [13]. The captured, segmented images and a "voxelized" hand are shown in figure 1.

No finite number of silhouette images from at one time instant will be able to perfectly reconstruct the 3D shape of the hand. The primary difficulty lie in detecting the space between fingers. According to [9], two factors contribute to the need of an infinite number of silhouette images. They are the fact that the finger are round and the camera positions lie outside the so-called outer visual hull making concave surfaces not reproducible. However, it is possible to capture a volume of space that is necessarily the upper-bound of the actual volume that the images represent.

## 3. Articulated Hand Model

The gaussian mixture model assumes that the observed unlabeled data is produced by any number of $N$ hidden point generators, and that these points follow a gaussian distribution with a particular mean and variance. Voxel reconstructions of the hand clearly do not follow the gaussian distribution assumption; however, the voxels can be thought of as produced from distinct finger segment generators producing point data that follow a uniform distribution confined to the cylindrical shape of the finger segment, also describable with a particular mean and covariance matrix. To partially compensate for the discrepancy between model and phenomenon, we rely on kinematic constraints which confine the components to the proximity of the observed finger segment data. This is in the form of constraints on the optimization in the M-step, which means the estimated result is still a maximum likelihood estimate.

The proposed hand model consists of 16 tri-variate gaussian distributions parameterized by their means and

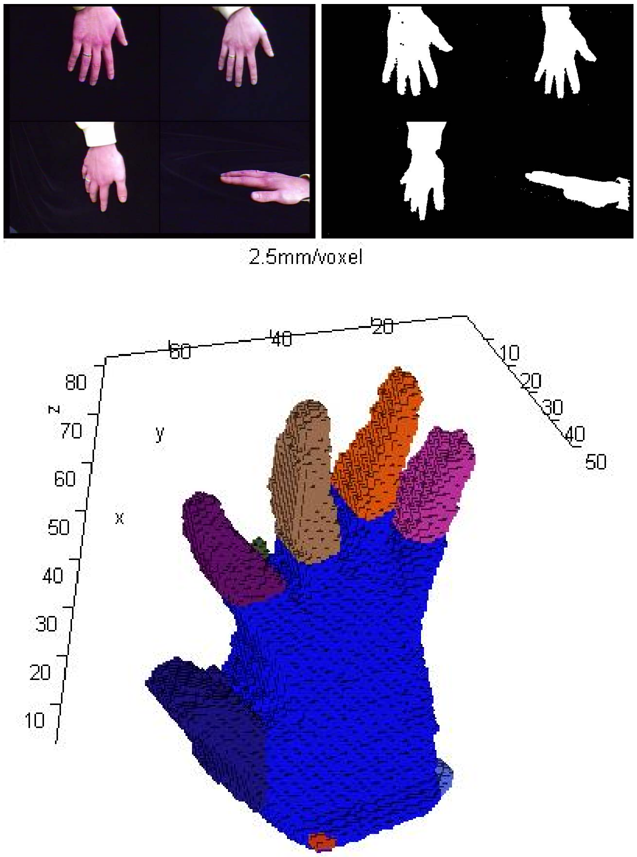

Figure 1: Actual color and segmented images of the hand serving as input to the shape-from-silhouette volume reconstruction system

variances. The probability distribution function is given by

$$P(\mathbf{x_n}) = \sum_{\mathbf{i=1}}^{\mathbf{16}} \alpha_{\mathbf{i}} \phi(\mathbf{x_n}|\mu_{\mathbf{i}}, \mathbf{\Sigma_i}) \tag{1}$$

where $\mathbf{x_n}$ is the 3D coordinates of voxel $\mathbf{n}$, $\alpha_i$ is the *a priori* probability that the $i$-th process produced $x_n$, and finally $\mu_i$ and $\Sigma_i$ are the mean and covariance matrix of gaussian process $i$.

The means and variances conform to a kinematic structure illustrated in figure 3. Using an ellipsoid to represent each segment, the kinematic structure of our hand model is fully specified by a the set parameters $S = (\Lambda, E, \Omega)$. $\Lambda = \{\Lambda_i\}_{i=1,2,...,16}$ is the set of diagonalized matrices $\Lambda_i = \text{diag}(\lambda_{i,x}, \lambda_{i,y}, \lambda_{i,z})$, one associated with each of 16 ellipsoids to describe its 3-D shape. $E = \{(j, k, \mathbf{a}_{j,k}, \mathbf{a}_{k,j})_l\}_{l=1,2,...15}$ with $j, k \in \{1, 2, ..., 16\}, j \neq k$ and $\mathbf{a}_{j,k}, \mathbf{a}_{k,j} \in \mathbf{R}^3$ which describes the connectivity structure of the model's 16 segments: $\mathbf{a}_{j,k}$ is a vector from the centroid of the $j$-th segment to the centroid of the $k$-th segment in canonical form. Finally, $\Omega = \{c, \mathbf{\Theta}\}$ is the set of posture parameters where $c \in \mathbf{R}^3$ is the center of the reference frame–in this case the center of the palm–and $\Theta = \{\theta_n\}_{n=1,2,...24}$ is the set of angles describing the relative orientation of a segment relative to another higher in the hierarchy of segments, in this case, to another segment closer to the palm segment. Together, $\Omega$ consists of 24+3 parameters that will be estimated from voxel data.

# 4. Estimation of Posture Parameters

The estimation procedure begins as usual with the estimation of the class associations in the E-step in gaussian mixture model estimation. The usual M-step produces estimates of the prior probabilities of each component, $\alpha_i^*$ and of the statistics, $\mu_i^*$ and $\Sigma_i^*$, belonging to the 3D gaussian observation process given in equation 1. The asterisks denote intermediate estimates. The eigenvalues are clamped to the measured finger segment values to enforce the shape of the estimated segment. After eigenvalue decomposition,

$$\Sigma_i = A_i \begin{bmatrix} \lambda_{i,1} & 0 & 0 \\ 0 & \lambda_{i,2} & 0 \\ 0 & 0 & \lambda_{i,3} \end{bmatrix} A_i^T \tag{2}$$

the new covariance is computed by replacing the eigenvalues consistent with the original ordering

$$\Xi_i = A_i \begin{bmatrix} \zeta_{i,1} & 0 & 0 \\ 0 & \zeta_{i,2} & 0 \\ 0 & 0 & \zeta_{i,3} \end{bmatrix} A_i^T \tag{3}$$

where $\{\zeta_{i,n}\}_{n=1,2,3}$ are measured height,width and depth dimensions of the finger segment, thereby enforcing the shape of the component.

## 4.1. Constrained Maximization

The remainder of the M-step consists of a constrained maximization of the expected likelihood, confining $\mu_i^*$ and $\Sigma_i^*$ of each finger segment to within the manifold of solutions consisting of physically possible postures that a human hand can make. This requires the extraction of a valid rotation matrix from the covariance matrix $\Sigma_i^*$ and using it in the set of constraint equations.

The constraint equations are expressed in terms of model parameters. For each joint, there is the spherical joint constraint used for each joint

$$\begin{aligned} \mu_i - \mathbf{R}_{0,i}\mathbf{a}_{i,j} &= \mu_j - \mathbf{R}_{0,j}\mathbf{a}_{j,i} \\ \mu_i - \mathbf{R}_{0,i}\mathbf{a}_{i,j} &= \mu_j - \mathbf{R}_{i,j}\mathbf{R}_{0,i}\mathbf{a}_{j,i} \end{aligned} \tag{4}$$

where $\mathbf{R}_{o,j}$ is the rotation matrix between the world frame to the $j$-th segment frame while $\mathbf{R}_{i,j}$ is the rotation matrix between the $i$-th to the $j$-th segment frame.

A *revolute* joint with 1 degree-of-freed requires two additional constraints along with to confine the rotation matrix to only 1 degree of freedom. This constraint is given by

$$\mathbf{R}_{0,i}\mathbf{q}_{i,j} = \mathbf{R}_{0,j}\mathbf{q}_{j,i} \tag{5}$$

where $\mathbf{q}_{i,j}$ is the unit vector parallel to the hinge axis in the $i$-th segment frame. Enforcing this constraint ensures that the two axes of rotation align. Since $\|\mathbf{q}_{i,j}\|\|\mathbf{q}_{j,i}\| = 1$, only two of the three constraints are independent; choose any two for this constraint.

A *two-rotational dof* joint, such as the knuckle joints, with 2 degrees-of-freedom, requires additional constraints given by

$$\begin{aligned} 0 &= \mathbf{R}_{0,i}\mathbf{q}_{i,j} \cdot \mathbf{R}_{0,j}\mathbf{q}_{j,i} \\ &= \mathbf{q}_{i,j}^T\mathbf{R}_{0,i}^T\mathbf{R}_{0,j}\mathbf{q}_{j,i} \end{aligned} \tag{6}$$

where here $\mathbf{q}_{i,j}$ and $\mathbf{q}_{j,i}$ are the unit vectors along the two rotational degrees of freedom. One can think of it as concatenating two segments each with one degree of freedom.

The collection of these constraints for each joint completes the set of constraint equations used to project the EM iterate onto the solution manifold of possible postures $\Omega^*$ using Newton-Raphson procedure.

## 4.2. Extracting the Rotation Matrix

¿From each EM iterate, we need to apply a set of rules to remove the ambiguity in extracting the rotation matrix $\mathbf{R}_{o,j}$ from $\Sigma_i$ for the constraint equations. Given the 3 lengths for each dimension of a component, a simple sort reveals the association between eigenvalue and corresponding dimension, but 2 of 3 lengths for every component is assumed to be equal, and neither direct application of SVD nor eigenvalue decomposition typically produce rotation matrices

$\mathbf{R}_{o,j}$ that are members of the special orthogonal group of matrices, SO(3), that is $\mathbf{R}\mathbf{R}^T = I$ and $\det(\mathbf{R}) = +1$, *and* in the right orientation for constrained optimization. The first source of ambiguity is the order of the principal components, which is in the order of strength from eigenvalue decomposition. The second source of ambiguity is in the sign of the principal component direction. By rearranging the eigenvectors and their signs, there are a total of 24 different rotation matrices in SO(3) that can be constructed with only one that is closest to the desired solution. To resolve the ambiguity, we need involve the kinematic model of the hand itself:

- For the palm segment, choose the rotation matrix that aligns the shortest axis to the z-axis (or third column of the rotation matrix) and at the same time minimizes the constraint equations for the palm joints shared with segment 3, 6, 9, 12, and 15. Knowing which principal component belongs to the third column reduces the number of ambiguities to 4.
- For thumb and finger segments, choose the rotation matrix that aligns the longest axis to the y-axis (or 2nd column of the rotation matrix), and minimize the constraint equations for the joints closest to the palm. For finger segment 9, find the rotation matrix that minimizes joint constraint equations between 9 and 16.

The danger of disregarding the ambiguity and choosing the first valid rotation matrix is the danger of delaying convergence or not converging at all. Suppose the model has neared convergence. After the M-step, the correct rotation matrix will have been forgotten by reconstituting it into a covariance matrix. Upon the next M-step, a rotation matrix that is still valid but oriented different than before will require iterating through the Newton-Raphson procedure again, potentially converging at a different and wrong local minimum.

## 5. Results and Discussion

The test currently consists of several sequences of simulated data and several voxel images of real hand data. The simulated data is constructed by placing cylinders into position and filling each of 20 cylinders (4 for the palm) with voxels. All parameters are known in the simulated case. Actual hand data is collected from the voxel reconstruction system using shape-from-silhouette [4, 13].

In both cases, the initial posture estimates are assumed to be in the proximity of the actual posture values. In the real data case, this amounts to limiting the initial hand search to within a small window before tracking commences, providing a starting hand posture is sufficient. Subsequent hand posture estimations are initialized with the previous posture estimate.

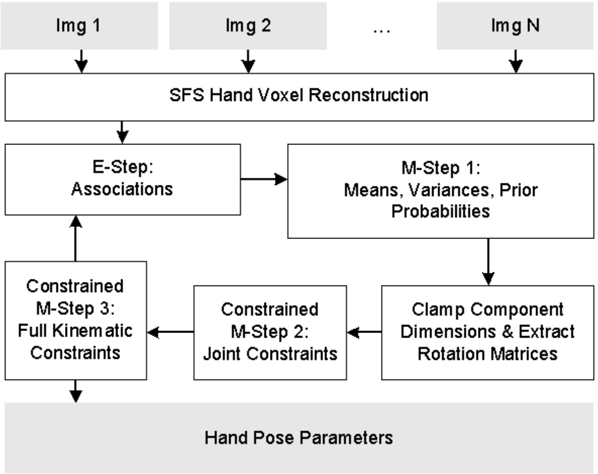
Img 1
Img 2
...
Img N
SFS Hand Voxel Reconstruction
E-Step: Associations
M-Step 1: Means, Variances, Prior Probabilities
Constrained M-Step 3: Full Kinematic Constraints
Constrained M-Step 2: Joint Constraints
Clamp Component Dimensions & Extract Rotation Matrices
Hand Pose Parameters

Figure 2: Flow diagram of ECM hand pose estimation algorithm.

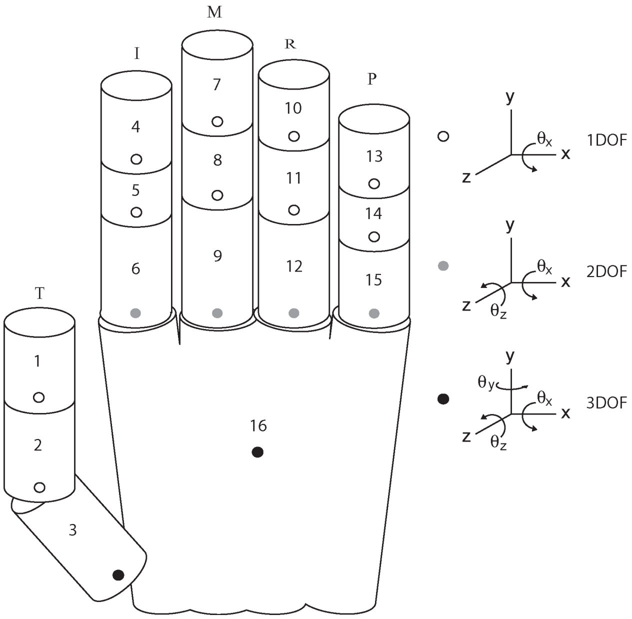

Figure 3: Hand model illustrating locations of degrees-of-freedom.

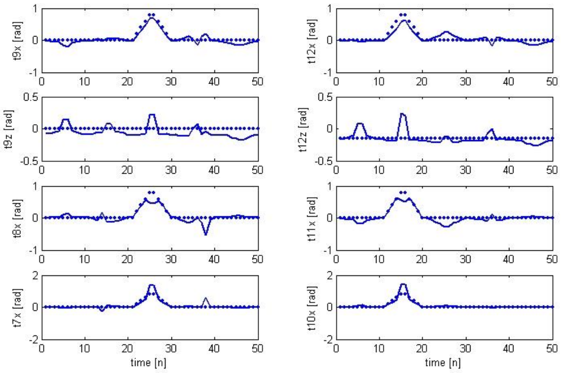

Figure 5: Plot of the angles of middle and ring fingers.

Figure 4 show that estimates from voxel data follow closely to the actual hand voxels, in the form of a point cloud, as a series of 50 voxel images of the hand bending and flexing each finger individually. The entire sequence of 16 frames took 323.4 seconds or 6.46 seconds/frame for EM posture parameter estimation on a Pentium 4HT 2.6GHz PC in Matlab. The voxel reconstruction takes a little more than 150ms per voxel reconstruction. The progression of joint angles can be seen in figure 5; the dotted line indicates actual and solid line indicates estimated joint angles.

In the progression of joint angle estimates, the affects of other fingers moving also affects the estimate of other fingers despite not moving at all. This is because of the method by which all the voxels are considered in the estimate of the joint angles.

Finally, figure 6 illustrates applying ECM hand posture estimation on real voxel reconstructions of the hand. These figures show very promising results. Even with the presence of some wrist mass, or the lack of presence in some palm mass, the model is able to remain neatly centered over the hand voxels.

## 6. Summary and Conclusions

This paper extends the work of Hunter, et al. in the manner of applying constrained optimization in the M-step of the EM algorithm to estimate posture parameters of an articulated model of the hand from unlabeled data. The unlabeled data in our experiments consist of simulated and some real voxel data of hands.

We proposed an observation process which consists of a tri-variate mixture of gaussian densities specified by means and covariances that are constrained to the kinematic structure of the hand. To confine the mixture statistics, we employed the constrained object/observe model and a set of constraint equations at the joints. Experiments on the two kinds of data demonstrates the methods effectiveness in estimating the parameters of the articulated hand model.

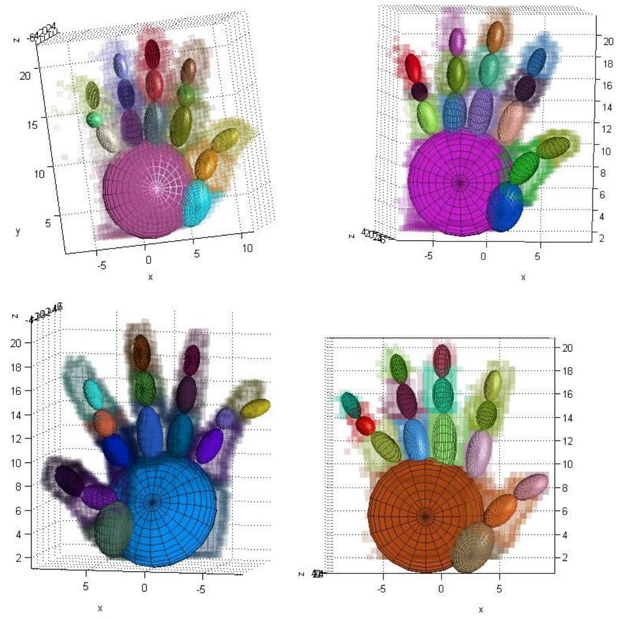

Figure 6: Results of ECM hand posture estimation on real voxel data.

Our long term goal is to not only robustly estimate the posture parameters $\Omega$ in the hand model, but $\Lambda$ and $E$ as well with as few assumptions as possible. For this, we intend to incorporate motion cues to determine the locations of joints, therefore the delimiters of finger segments which yield the lengths and direction of long axis. Also, the ECM framework appears to be rich enough to allow such extensions in the form of ranges of valid principal components values.

## Acknowledgements

This research was support in part by the UC Discovery Program and the Technical Support Working Group of the US Department of Defense. The authors also acknowledge the support and assistance of their colleagues from the CVRR Laboratory.

## References

[1] C. Bregler. Tracking people with twists and exponential maps. In *IEEE Proceedings Computer Vision and Pattern Recognition*, 1998.

[2] S. Y. Cheng and M. M. Trivedi. Occupant posture modeling using voxel data:issues and framework. In *IEEE Proceedings on Intelligent Vehicles Symposium*, 2004.

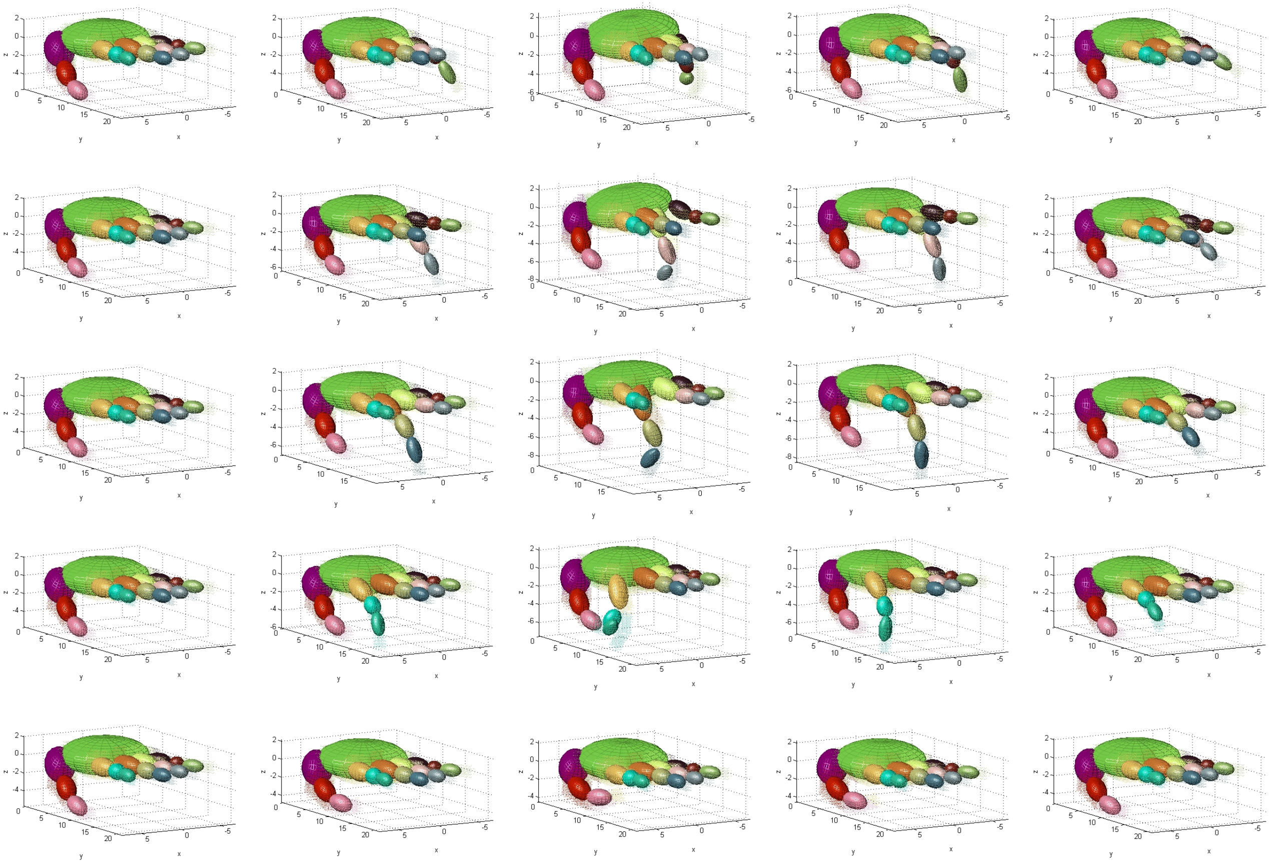

Figure 4: Estimation results using simulated hand data, shown in light dots superimposed with the estimated hand model. The first frame (top left) was initialized with actual means and variances. Subsequent frames are from ECM estimation alone. The motioning begins with the thumb then fingers like a wave.

[3] G. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *IEEE Proceedings Computer Vision and Pattern Recognition*, 2003.

[4] G. K. M. Cheung and T. Kanade. A real-time system for robust 3d voxel reconstruction of human motions. In *IEEE Proceedings Computer Vision and Pattern Recognition*, pages 714–720, 2000.

[5] D. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(82-98), 1999.

[6] T. Horprasert, D. Harwood, and L. S. Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *IEEE Proceedings ICCV Frame-Rate Workshop*, September 1999.

[7] R. Hoshino, D. Arita, S. Yonemoto, and R. Taniguchi. Real-time human motion analysis based on analysis of silhouette contour and color blob. *Springer AMDO*, 2002.

[8] E. A. Hunter, P. H. Kelly, and R. C. Jain. Estimation of articulated motion using kinematically constrained mixture densities. In *IEEE Proceedings Nonrigid and Articulated Motion Workshop*, pages 10–17, 1997.

[9] A. Laurentini. How many 2d silhouettes does it take to reconstruct a 3d object? *CVIU*, 67(1):81–89, July 1997.

[10] I. Mikic and M. M. Trivedi. Vehicle occupant posture analysis using voxel data. In *Ninth World Congress on Intellgient Trasport Systems*, October 2002.

[11] I. Mikic, M. M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisitiona and tracking using voxel data. *IJCV*, 53(3):199–223, 2003.

[12] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268, 2001.

[13] D. E. Small and L. R. Williams. Real-time shape-from-silhouette. Master's thesis, University of Maryland, 2001.

[14] M. M. Trivedi, S. Y. Cheng, E. C. Childers, and S. J. Krotosky. Occupant posture analysis with stereo and thermal infrared video. *to appear in IEEE Transactions on Vehicular Technology*, 2004.

[15] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara. A hand-pose estimation for vision-based human interfaces. *IEEE Transactions on Industrial Electronics*, 50, August 2003.