Multimodal Voxelization and Kinematically Constrained Gaussian Mixture Models for Full Hand Pose Estimation: An Integrated Systems Approach

Shinko Y. Cheng   Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
University of California, San Diego
9500 Gilman Drive, La Jolla CA 92037
{sycheng,mtrivedi}@ucsd.edu

Abstract

We propose a new approach to the hand pose estimation problem using only volume information. We describe a thermal and color image-based approach to generate silhouettes of the hand with which voxel images are produced using shape-from-silhouette. We assume a 16-component, 27 degree-of-freedom kinematically constrained Gaussian mixture model, and fit it over the voxel images. We constrain the otherwise freely-arranged components of this model by a system of equations describing joint characteristics parameterized by component centroids and orientations. We use the EM algorithm and steepest descent to estimate the parameters of the model such that they yield the maximum likelihood and kinematically correct pose estimates. We demonstrate the effectiveness of the proposed system on synthesized as well as captured voxel images of the hand, and show that given appropriate initial conditions, the iterative model parameter estimation procedure effectively converges to and tracks the "voxelized" articulated hand.

1. Introduction

Voxel-based articulated body pose estimation is the task of fitting a kinematic model of joints and segments to observed volumetric reconstructions of the articulated body, such as a hand or human body. The proposed method assumes the use of only the voxel data acquired using shape-from-silhouette [5, 4, 2, 12, 13, 17, 16]. The primary challenge in articulated body pose estimation lies in extracting joint parameters that describe the pose and location of the articulated body from this unlabeled 3D voxel data. Accurate body pose estimation have several foreseeable applications, including a markerless motion capture system for human-computer interfaces, analyzing gait in physiotherapy, 3D animation, ergonomics and robot control.

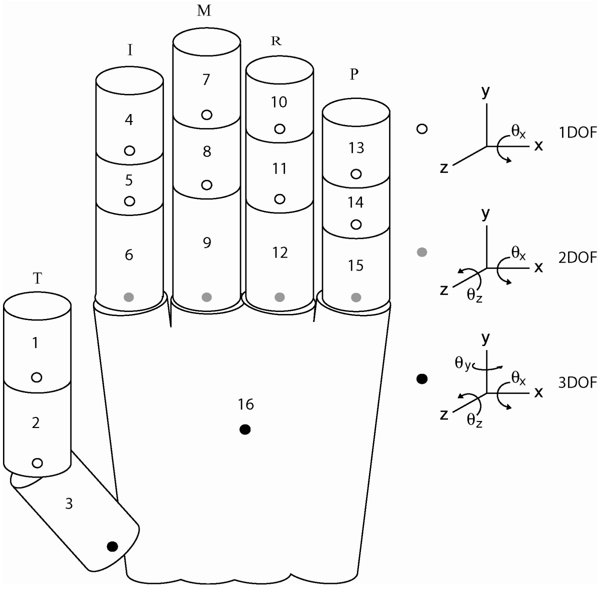


Figure 1. Hand model illustrating locations of degrees-of-freedom.

We propose a technique that assumes a 16 class trivariate Gaussian mixture model to describe the spatial distribution of the voxels. Each class describes a segment that is kinematically constrained according to a pre-specified skeletal model with 27 degrees-of-freedom and 16 segments, shown in fig. 1.

This model lends itself easily to the estimation of two sets of attributes of the hand. Assuming that the segment and joint structure of the articulated hand consists of 5 fingers, each with 3 segments, and another palm segment, then the model parameters describe 1) the dimensions of each segment in height, width and depth, which is assumed fixed in this paper, and 2) the location and orientation of each segment or equivalently, the joint angles and hand position. The proposed algorithm estimates the second set of attributes: the location and orientation of each of the segments, i.e. pose parameters.

The pose parameters are learned using the EM algorithm modified such that the M-step is followed by an independent optimization step. This extra step ensures a kinematically valid estimate at the end of each iteration. It has the effect of constraining the optimization of the M-Step by projecting EM iterates onto the feasible pose space. One can view it as EM where the solution manifold is reduced in size.

The Expectation-Constrained Maximization (ECM) algorithmic and overall system flow is shown in fig. 2. Four images are captured and segmented, resulting in silhouettes of the subject's hand. Knowing the position of each camera in the real-world from calibration, volume reconstruction is found using shape-from-silhouette. The resulting voxel reconstruction is used to estimate the pose of the hand using the proposed algorithm.

To demonstrate the effectiveness of the proposed algorithm and system, we evaluated it on ideal synthesized and actual voxel images of the hand. The synthesized data is constructed from cylinders of voxels. Experiments using synthesized data also conveniently provides access to ground-truth for the pose estimates.

There are several robust ways of segmenting images to generate silhouettes of hands. It is therefore natural that silhouettes are commonly used for pose estimation. Research in this area can be loosely categorized under monocular [1, 10] and multi-perspective approaches. A recent survey describes the many hand pose estimation approaches [7]. Of the latter approaches, many efforts [4, 13, 8] focus on detailed human body pose estimation but disregard the task of estimating hand pose. Of the research pertaining to estimating the finer hand pose [17, 14], a rigid skeletal-surface model of the hand are relied upon to generate torque-like adjustments to fit the model to voxel data. We propose to use a more flexible Gaussian mixture model with dimension parameters that have physical meaning, and more conducive towards hand dimensions estimation as well as hand pose.

Furthermore, a motivating factor for using an ellipsoidal (Gaussian) model, over a more rigid model, is the flexibility to use multiple resolution models, i.e. some with more components than others, for the same articulated body. The use of any particular resolution model depends on the needs of the system, as each resolution model has its uses [3], or depends on the information present in the voxel images. It presents a natural initialization procedure which starts with coarse models and progresses towards finer models when sufficient information arises. It is conceivable that the hand is first modeled as a single segment. Then, as fingers become distinguishable by separation between voxel masses, 5 more segments are used in its place. As a finger bends, the algorithm may notice that 3 segments might fit better than 1. Other work in the past [5, 13] have utilized this idea.

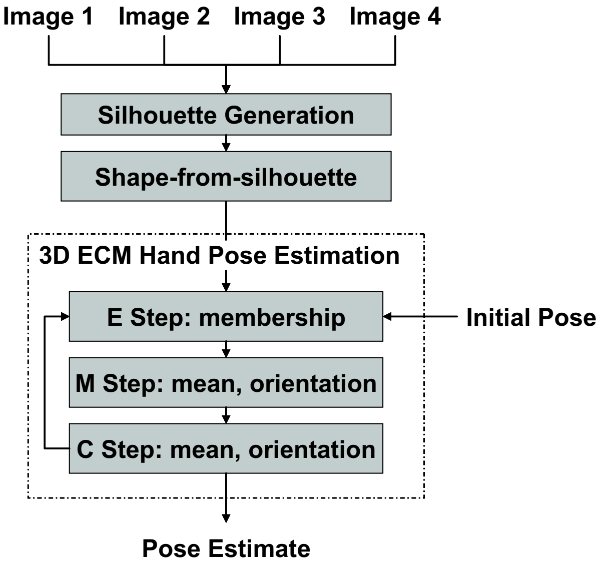


Figure 2. Flow diagram of ECM hand pose estimation algorithm.

Hunter describes the use of ECM framework for estimating pose parameters in a 15 component 31 DOF articulated model of the human body in [10]. It uses monocular 2D silhouettes as the observed process for the Gaussian mixture model, which implies there are two problems to overcome: One, it requires careful resolution of layer ordering ambiguities resulting from self-occlusion. Two, it assumes orthographic projection to simplify scale estimation. In our research, we use multiple silhouette images to generate the 3D visual-hull and using that as the observed process in our pose estimation. This effectively eliminates the two issues. It should also be pointed out that [10] does not address a critical step between relating the covariance matrices of each component to the Euler representation of its orientation in the likelihood function, which we propose to solve using steepest descent. We also introduce a way to measure estimator accuracy by comparing actual and estimated component orientation and positions using synthesized voxel images.

The remaining sections are organized as follows. Section 2 explains the process of real-time shape-from-silhouette using color and thermal infrared imagery. Sections 3 and 4 describes the kinematically constrained Gaussian mixture model used to describe the hand and how the EM algorithm is employed to estimate the model parameters. Sections 5 and 6 describes the experimental results and general conclusions of the hand pose estimation method.

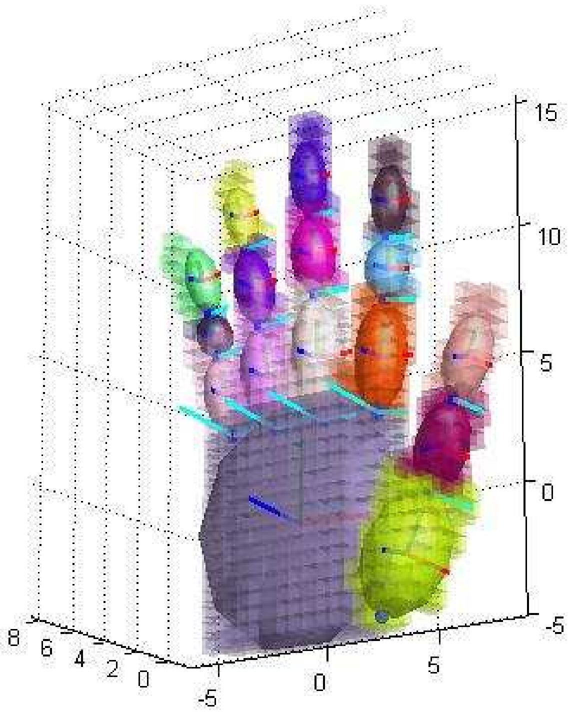


Figure 3. Example result from proposed algorithm on synthesized data.

2. SFS Voxel Reconstruction

Silhouette images from multiple known viewpoints are the required input for generating visual-hulls of the subjects using real-time shape-from-silhouette (SFS) [16, 15]. We describe a color and thermal camera-based method of generating silhouettes, each having their benefits and drawbacks.

To generate silhouettes using color images, images of the hand from various known viewpoints (using calibrated cameras) are captured simultaneously. Each image is then segmented using a statistical background subtraction with shadow suppression [9]. The result of the segmentation are silhouettes of the hand with which the voxel image of the hand is found. The captured, segmented images and a ”voxelized” hand are shown in fig. 4.

An alternative method of generating voxel images is to not use multiple color cameras, but multiple thermal long-wavelength infrared cameras. Skin is generally within a small variable range of temperatures, between 26° and 28° C degrees [11], and thermal imaging devices have a direct relationship between pixel intensity and temperature Therefore, it is extremely simple to generate hand silhouette images from thermal imaging sources by choosing an upper and lower threshold.

For thermal imaging devices that have automatic-gain-control that cannot be disengaged, Yalcin, et al. describes how gain change can be estimated using an affine model between the current and reference pixel intensities [18]. We use the estimate to adjust the thresholds accordingly, or equivalently reverse the gain change with the gain estimate and apply the same thresholds. We found the slope that relates the current pixel to the background pixel using the relation $I_t = mI_0$. We found that estimating the y-intercept does not improve the gain estimate as much as to increase the sensitivity of noise due to the small but inevitable section of foreground in the current frame. It is important that the pair of points represent the same object in the scene, so a mask around the image is used to disregard the pixels in the center of the image, where the hand is likely to be. The pixels near saturation in both the background and current points are also disregarded. The results of the thermal silhouette generation are shown in fig. 5.

The primary advantage of thermal-based over color-based silhouette generation technique is its invariance to visible illumination, which is why it is widely used for night surveillance. To gain that added robustness, one must pay the cost of multiple thermal cameras.

It can be proven that no finite number of silhouette images from 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 [12], two factors contribute to the need of an infinite number of silhouette images. They are the fact that the fingers are round and the camera positions lie outside the so-called outer-visual-hull making concave surfaces impossible to reproduce. However, it is possible to capture a volume of space that is necessarily the upper-bound of the actual volume that the images represent. And as one can see from fig. 4, with certain camera arrangements, a very clear hand voxel image can be generated from just four silhouette images.

3. Articulated Hand Model

The Gaussian mixture model assumes that the observed unlabeled data is produced by any of $P = 16$ hidden point generators. These points follow a Gaussian distribution parameterized by their means and component orientation parameters derived from the covariance matrix. Voxel images of an articulated body clearly do not exactly follow the Gaussian distribution assumption; however, the voxels can be thought of as following a uniform distribution confined to the cylindrical shape of the finger segment, which is 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 an optimization

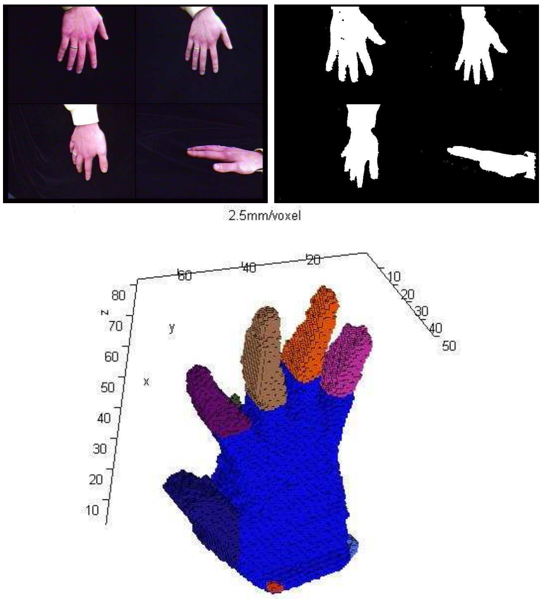
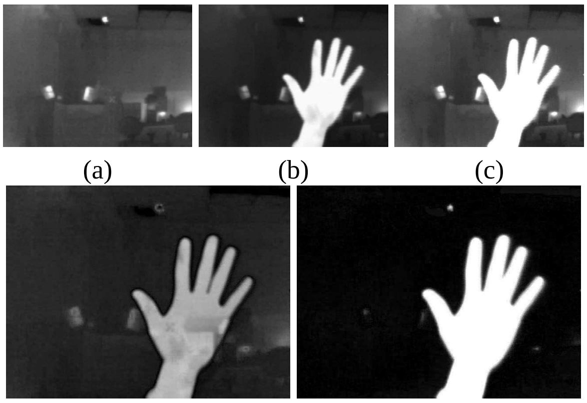


Figure 4. Actual color and segmented images of the hand serving as input to the shape-from-silhouette volume reconstruction algorithm.



(a) (b) (c)

(d) (e)

Figure 5. (a) Background, (b) current, (c) gain-compensated, (d) background-subtraction without compensation, and (e) background-subtraction with compensation frames. The hand is better differentiated from the background with than without gain compensation.

step following the M-step.

Let $\mu_i \in \mathrm{I\!R}^3$ and $\Sigma_i \in \mathrm{I\!R}^{3\times3}$ be the mean and covariance matrices of each Gaussian component that represents segment i of the articulated body. The covariance is parameterized by the Euler angles that describes the components orientation with respect to the world coordinate frame, and three values in the diagonal matrix that describes the components' length, width and depth. Namely, the covariance matrix

$$\begin{aligned}\Sigma_i &= \mathbf{R}\Lambda_i\mathbf{R}_i^\top \\ &= \mathbf{R}_{zi}\mathbf{R}_{yi}\mathbf{R}_{xi}\Lambda_i\mathbf{R}_{xi}^\top\mathbf{R}_{yi}^\top\mathbf{R}_{zi}^\top \\ &= e^{\hat{\omega}_z\theta_{zi}}e^{\hat{\omega}_y\theta_{yi}}e^{\hat{\omega}_x\theta_{xi}}\Lambda_i e^{-\hat{\omega}_x\theta_{xi}}e^{-\hat{\omega}_y\theta_{yi}}e^{-\hat{\omega}_z\theta_{zi}} \\ &= \Sigma(\theta_{xi},\theta_{yi},\theta_{zi},\lambda_{xi},\lambda_{yi},\lambda_{zi}) \qquad (1)\end{aligned}$$

This parametrization allows us to easily take the derivative of Σ with respect to θ by simply left (or right) multiplying the exponential by the constant term $\hat{\omega}$ in the exponent.

The proposed model consists of P such trivariate Gaussian distributions parameterized by their means, dimensions, and orientation. The probability distribution function is given by

$$P(\mathbf{x}|\Theta) = \prod_{i=1}^{N} \alpha_i^{y_{n,i}} P(\mathbf{x}_n|\mu_i,\Sigma_i)^{y_{n,i}} \qquad (2)$$

where $\mathbf{x}_n$ is the 3D coordinates of the voxel n, α_i is the *a priori* probability that the i-th process produced $\mathbf{x}_n$, and $y_{n,i} \in [0,1]$ is the hidden random variable describing the membership assignment where $y_{n,i} = 1$ if x_n is a member of component i and 0 otherwise. By taking the log of the this density function, we are able to separate the dependencies of the *a priori* class probabilities α and the parameters Θ to form the log-likelihood function as formulated by Dempster, et al [6],

$$\mathcal{L}(\Theta,\mathbf{x}) = \sum_{i=1}^{N} \hat{\mathbf{Y}}_k^\top \big(\log .\alpha + \log .P(\mathbf{x}_n|\Theta)\big) \qquad (3)$$

where $\log .\alpha = (\log\alpha_1, ..., \log\alpha_N)^\top$ and $\log .P(\mathbf{x}_n|\Theta) = (\log P(\mathbf{x}_n|\mu_1,\Sigma_1), ..., \log P(\mathbf{x}_n|\mu_N,\Sigma_N))^\top$.

The kinematic structure is defined by the set of vectors $\{\mathbf{a}_{ij}\}$ that locally define the location of the joint, and $\{\mathbf{q}_{ij}\}$ that define the axes of rotation at the joints, both as it is written in the i-th frame. Specifically, the vector $\mathbf{a}_{ij} \in \mathrm{I\!R}^3$ points from the centroid of component i towards the joint shared between components i and j, and $\mathbf{q}_{ij} \in \mathrm{I\!R}^3$ where $\|\mathbf{q}_{ij}\| = 1$ points along one of the component's axis of rotation. This is the case for all pairs of components (i,j) that are connected by a joint.

To summarize, the kinematic structure of a P component articulated body is fully specified by the set of parameters $S = (\Lambda, E, \Omega)$. The first element $\Lambda = \{\Lambda_i\} = \{diag(\lambda_{xi}, \lambda_{yi}, \lambda_{zi})\}$ $\forall i \in \{1...P\}$ describes the component dimensions. The second element $E = (i, j, \mathbf{a}_{ij}, \mathbf{a}_{ji}, \mathbf{q}_{ij}, \mathbf{q}_{ji}) \forall i, j \in \{1...P\}$ describes the connectivity structure of the model's P segments. Finally, $\Omega = \{c_0, \Theta\}$ is the set of pose parameters, where $c_0 \in \mathbb{R}^3$ is the reference center of the articulated body, which is the position of the root component (palm) in the hierarchy of components. And $\Theta = \{\theta_i\}$ $\forall i$ is the set of angles describing the relative orientation of the segment relative to another segment higher in the hierarchy of segments. For the root component, the world coordinate frame is used. Together, Ω consists of the number of angular and translational degrees-of-freedom (24+3 for our hand model) which will be estimated from voxel data via the knowledge of E and initial component orientations and mean locations. In this paper, Λ and E are known beforehand, and Ω is estimated.

4. Estimation of Hand Pose Parameters

Given a voxel image of an articulated object, the estimation procedure begins with the usual E-Step in Gaussian mixture model estimation, producing membership weights y_n. The initial values of the means μ_i and covariances $\Sigma(\theta_{xi}, \theta_{yi}, \theta_{zi})$ are set approximately where the components are expected. The implications of this step means that the articulated body must start at a reference pose, such as the so-called T-pose for the human body. For the hand,the starting pose is one where the fingers are spread out. Automatic initialization is a problem left for future work.

Given the membership weights, we proceed to the M-Step to solve for a priori probabilities α and means and orientation of each component. For brevity, we describe only the estimation of the orientation parameters, which is the only difference from the usual M-Step for Gaussian mixture model estimation [6]. Note that by taking the derivative of equation 3 with respect to $\theta_i = (\theta_{xi}, \theta_{yi}, \theta_{zi})^\top$, we are left with only one of the components in the sum. Because the derivative is nonlinear with respect to θ_i, we use steepest descent to find to solution. We have the initial angles $\theta_i^{[0]}$ from the initial position assumption above to start the iterative solver.

$$\theta_i^{[n]} = \theta_i^{[n-1]} - 0.0005 \nabla_{\theta_i} \mathcal{L}(\Theta, \mathbf{x}) \quad (4)$$

$$\nabla_{\theta_i} \mathcal{L}(\Theta, \mathbf{x}) = \begin{pmatrix} \bar{y}_{k,i}^\top e^{\hat{\omega}_z \theta_{zi}} e^{\hat{\omega}_y \theta_{yi}} \hat{\omega}_x e^{\hat{\omega}_x \theta_{xi}} \Lambda_i R_i^\top \bar{y}_{k,i} \\ \bar{y}_{k,i}^\top e^{\hat{\omega}_z \theta_{zi}} \hat{\omega}_y e^{\hat{\omega}_y \theta_{yi}} e^{\hat{\omega}_x \theta_{xi}} \Lambda_i R_i^\top \bar{y}_{k,i} \\ \bar{y}_{k,i}^\top \hat{\omega}_z e^{\hat{\omega}_z \theta_{zi}} e^{\hat{\omega}_y \theta_{yi}} e^{\hat{\omega}_x \theta_{xi}} \Lambda_i R_i^\top \bar{y}_{k,i} \end{pmatrix} \quad (5)$$

$$\bar{y}_{k,i} = y_k - \mu_i \quad (6)$$

The algorithm terminates when the norm of the gradient $\|\nabla_{\theta_i} \mathcal{L}(\Theta, \mathbf{x})\| < 0.25$ or 50 iterations have been performed, whichever occurs first.

The EM algorithm does not impose any particular arrangements of the Gaussian mixture components. Here we employ a set of (finger) joint constraints $C(\Theta) = 0$ to confine the pose parameters onto the feasible manifold of kinematically valid poses [10]. The components are forced to connect at the joint locations and oriented such that the relative rotations between the components are non-zero only for the predetermined degrees-of-freedom, whether they be 1, 2 or 3. They are explained next.

4.1. Kinematic Constraints

Three systems of equations describe the feasible pose manifold in the projection step after the M-Step.

$$\mathcal{C}_b(\Theta) = \mu_i + \mathbf{R}_{0i} \mathbf{a}_{ij} - (\mu_j + \mathbf{R}_{0j} \mathbf{a}_{ji}) \quad (7)$$

$$\mathcal{C}_h(\Theta) = \mathbf{R}_{0i} \mathbf{q}_{ij} \cdot \mathbf{R}_{0j} \mathbf{q}_{ji} \quad (8)$$

$$\mathcal{C}_r(\Theta) = [\mathbf{R}_{0i} \mathbf{q}_{ij} - \mathbf{R}_{0j} \mathbf{q}_{ji}]_{2\ rows} \quad (9)$$

Given a 3 DOF (ball) joint between components i and j, the constraint equation is given by equ. 7. If $\mathcal{C}_b(\Theta) = \mathbf{0}_{3\times1}$, then the two components are connected at the joint. The rotation matrices are parameterized by Euler angles as in equ. 1. For each subsequent type of joints, $\mathcal{C}_b(\Theta)$ is used in addition to the constraint equations described below.

Given a 2 DOF (hardy-spicer) joint between components i and j, the constraint equation *in addition to the* $\mathcal{C}_b(\Theta)$ is given by equ. 8. Each of the two axes of rotation is described by $\mathbf{q}_{ij}$ and $\mathbf{q}_{ji}$. For a joint that rotates along the x and y axes in the two components reference orientation, $\mathbf{q}_{ij} = (1, 0, 0)^\top$ and $\mathbf{q}_{ji} = (0, 1, 0)^\top$. If $\mathcal{C}_h(\Theta) = 0$, then the two components' axes of rotation are at right angles with one another, resulting in rotation of either component relative to the other along only those two axes.

Given a 1 DOF (revolute) joint between components i and j, the constraint equation in addition to $\mathcal{C}_b(\Theta)$ is given by equ. 9. Since $\|\mathbf{q}_{ij}\| = 1$, only two equations are independent, so retain any two rows. Now, $\mathbf{q}_{ij}$ and $\mathbf{q}_{ji}$ point in the same direction in the two components reference orientation. If $\mathcal{C}_r(\Theta) = \mathbf{0}_{2\times1}$, then the two components' axes of rotation are directionally aligned with one another.

The collection of constraint equations $\mathcal{C}(\Theta) = 0$ consists of one of at least the first type and depending on the type of joint, either the second or third type of constraint equations for every joint. Newton's method is used to find the solution given the pose parameters following the M-Step.

The iterative optimization of $\mathcal{C}(\Theta) = 0$ terminates when the norm of the change $\|[\nabla_{\theta_i} \mathcal{C}(\Theta)]^{-1} \mathcal{C}(\Theta)\| < 1$ or after 10 iterations. Convergence is attained usually in less than 5 iterations.

The ECM algorithm itself, which embodies the two iterative processes above in 2 of the 3 steps, is terminated when the value of the change of the likelihood function decreases below 1 or after 20 iterations.

5. Results and Discussions

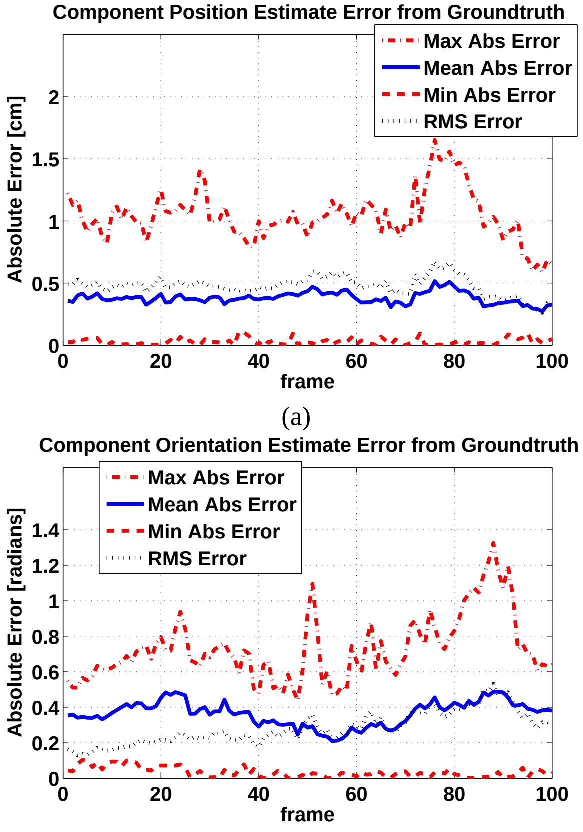
The proposed algorithm was run on several sequences of synthesized and actual hand voxel images. The synthesized data is constructed by placing 20 voxel-filled cylinders into position (4 for the palm), with each voxel 6mm to the side. All parameters are known in the synthesized case. Actual hand data is collected from the thermal images of the hand and shape-from-silhouette. Each voxel in the actual hand voxel image is 2.5mm to the side.

In both cases, the initial pose estimate was manually selected. Subsequent hand pose estimates are used to initialize the algorithm for the next voxel image.

In the synthetic case, several error measures were used to describe the pose estimate accuracy on three pose sequences where each sequence is more complex than the next. Specifically, in each sequence, finger joints underwent bending from -9 to 60 degrees. The first sequence bends the fingers all at once. The second sequence adds rotation to the entire hand over the same range. Finally, the third sequence delays the start of the bending of the fingers for each finger, producing a wave pattern. The accuracy of each run are summarized in table 5 and the error progression over 200 voxel image for the wave sequence is shown in fig. 7.

There are several axes marks displayed in the hand pose estimate in fig. 3. There are 3 axes marks for x, y, and z directions at the centroid of each finger segment, or component, to depict its orientation with respect to the world frame; they represent $\mathbf{R}_i$ for each component i. There are 2 axis marks that depict the axis of rotation at the joints of which the position and orientation are defined by $\mathbf{q}_{ij}$. A kinematically valid result is when these axes of rotation are aligned for joints between finger segments and at right angles at the knuckles except for the thumb knuckle. The thumb knuckle is assumed to have 3 DOF. The voxels are shown as transparent cubes overlapping with the hand model. The progression of hand pose estimation for the wave sequence is shown in fig. 6.

Regarding the source of errors, in the progression of joint angle estimates, the pose of any one finger affects the estimate of other fingers, even if the one finger had not moved at all. This is both a benefit and the cause of some of the errors one can see in the results. Another source of error is the fact that the thumb knuckle is not really a 3 DOF joint. Between the middle and base segment of the thumb is actually a restricted 2 DOF joint and the thumb knuckle where the base segment joins the palm segment is actually a oriented 2 DOF joint. More experiments will be performed with modified hand models. Ideally, the amount of reorientation of the thumb knuckle can be automatically estimated.



(a)

(b)

Figure 7. Maximum, mean, minimum and RMS component (a) translational and (b) angular error over ”wave” sequence.

Finally, one last source of error worthy of mention is the reliance on the close proximity of the previous estimate to the current for successful convergence of the estimate. If the current and previous finger positions are different by half a finger width or more, or when the subject's fingers are too close together, loss of track or finger overlap errors may occur. An example of a finger-overlap error can be seen in fig. 8. This can probably be resolved by employing constrained optimization in the M-Step in the form of lagrange multipliers and the Kuhn-tucker conditions to limit the range of angles at the joints of the fingers, or using assuming a prior for the parameters. This is reasonable to do since fingers can be safely expected to have only a certain range of motion.

Finally, fig. 9 illustrates applying the proposed method 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 of palm mass, the

Table 1. Error results of fold, fold-and-turn, and wave test sequences averaged over all component parameters.

	Fold-Only	Fold-and-Turn	Wave
(min,max) abs error [deg]	(0.021, 61.1)	(0.129, 91.7)	(.0284, 75.9)
mean abs error ± std [deg]	29.9 ± 15.6	31.7 ± 16.0	20.9 ± 11.9
rms error [deg]	21.2	21.3	16.2
(min,max) abs error [cm]	(2.59e-5, 1.1924)	(1.6795e-4, 1.572)	(2.886e-4, 1.6524)
mean abs error ± std [cm]	0.3319 ± 0.2665	0.3427 ± 0.2850	0.38085 ± 0.2939
rms error [cm]	0.4250	0.4441	0.47755

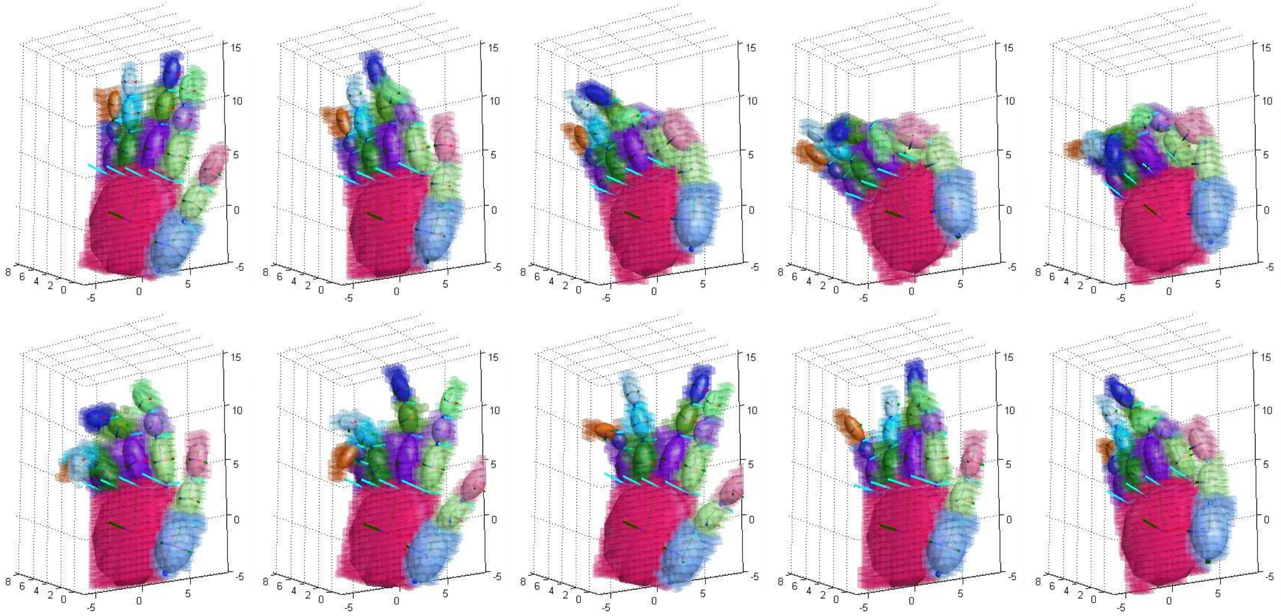


Figure 6. Results of pose estimation over synthesized ”wave” sequence.

model is able to remain neatly centered over the hand voxels.

6. Conclusions

We propose a new approach to the articulated hand pose estimation problem using only the volume information. We assume a 16-component, 27 degree-of-freedom kinematically constrained Gaussian mixture model and fitting it over shape-from-silhouette generated voxel images. We constrain the otherwise freely arranged components by a system of joint constraint equations parameterized by component centroids and orientations.

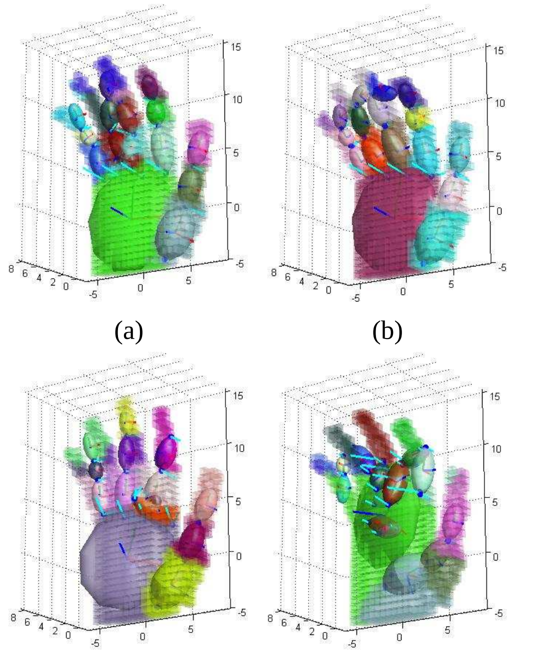
We use the EM algorithm and steepest descent to estimate the parameters of the model such that they yield the maximum likelihood and satisfies the constraint equations producing kinematically correct pose estimates. The algorithm's effectiveness was evaluated on synthesized as well as actual captured voxel data of the hand, and show that given appropriate initial conditions, the iterative model parameter estimation procedure effectively converges to and tracks the ”voxelized” articulated hand with an rms error in of 21 degrees and 4mm deviation.

References

[1] V. Athitsos and S. Sclaroff. An appearance-based framework for 3d hand shape classification and camera viewpoint estimation. In *Proceedings of Face and Gesture Recognition*, 2002.

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

[3] S. Y. Cheng and M. M. Trivedi. Multiperspective thermal ir and video arrays for 3d body tracking and driver activity analysis. In *IEEE International Workshop on Object Track-*



(a) (b)

(c) (d)

Figure 8. Illustration of (a,b) finger-overlap errors from close proximity of finger voxels, and (c,d) resulting complete loss of track.

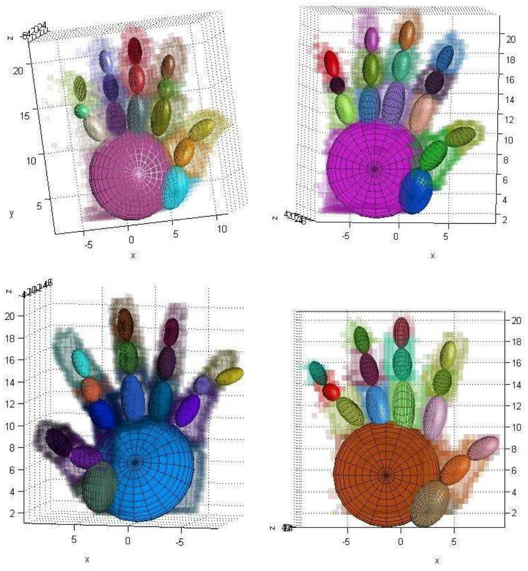


Figure 9. Results of ECM hand posture estimation on real voxel data.

ing and Classification Beyond the Visible Spectrum, pages 132–139, 2005.

[4] 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 of Conference on Computer Vision and Pattern Recognition*, 2003.

[5] 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.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1), 1977.

[7] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. A review on vision-based full dof hand motion estimation. In *IEEE Proceedings of Conference on Computer Vision and Pattern Recognition*, 2005.

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

[9] 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.

[10] E. Hunter. *Visual Estimation of Articulated Motion using Expectation-Constrained Maximization Algorithm*. PhD thesis, University of California, San Diego, 1999.

[11] S. G. Kong, J. H. an dBesma R. Abidi, J. Paik, and M. A. Abidi. Recent advances in visual and infrared face recognition–a review. *Computer Vision and Image Understanding*, 97:103–135, 2005.

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

[13] 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.

[14] K. Ogawara, K. Hashimoto, J. Takamatsu, and K. Ikeuchi. Grasp recognition using a 3d articulated model and infrared images. In *IEEE/RSJ Proceedings of Conference on Intelligent Robots and Systems*, volume 2, pages 27–31, 2003.

[15] M. Potmesil. Generating octree models of 3d objects from their silhouettes in a sequence of images. *Computer Vision, Graphics and Image Processing*, 40(1-20), 1987.

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

[17] 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.

[18] H. Yalcin, R. Collins, and M. Hebert. Background estimation under rapid gain change in thermal imagery. In *IEEE International Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, pages 11–18, 2005.