2011 11th IEEE-RAS International Conference on Humanoid Robots
Bled, Slovenia, October 26-28, 2011

# A Robotic Model of the Ecological Self

Justin W. Hart
Department of Computer Science
Yale University
New Haven, CT 06520-8285
justin.hart@yale.edu

Brian Scassellati
Department of Computer Science
Yale University
New Haven, CT 06520-8285
scaz@cs.yale.edu

***Abstract*—This paper discusses an integrated model of a robot's sensory and perceptual capabilities based on one of the earliest forms of self-knowledge that humans develop, knowledge of the Ecological Self. The Ecological Self is a cohesive model of the body and senses learned through the experience of using them together. This unified model allows kinematic and sensory data to be combined, producing an intersensory perception grounded in both inputs. Taking inspiration from this Ecological Self, but building on modern engineering practices, this model allows a robot to learn the kinematics of its end-effector by witnessing its motion in its visual field. This property of adaptation through self-observation also allows the model to adapt to changes in the robot's kinematic structure, as in the case of tool use. A final refinement is performed over the combined visual-kinematic model and is demonstrated to improve not only the accuracy of the kinematic model, but also the robot's stereo vision calibration. This refinement is inspired by the hypothetical process by which infants learn about their selves. The system is demonstrated to require fewer than 200 motion samples to fully train, to predict end-effector position within 2.29mm (SD=0.10) and 2.93 pixels (SD=3.83), to learn the lengths of the linkages in the robot's arm to within 1.1mm, and to adapt to tool use after only 52 samples.**

## I. Introduction

Typically, when a robot is designed, it is engineered with all of its capabilities built directly into the system. The thinking about the robot's senses and structure is done by engineers when the machine is designed, and coded into models such as kinematic and vision models. These are then used in black-box subprograms, vision and inverse kinematic routines that are not dealt with directly by the robot's cognitive model. People, on the other hand, learn about their physical and sensory capabilities from first-hand experience. To the infant attempting to grasp objects, the kinematic and sensory capabilities of their body are something to be learned.

Developmental psychologists have suggested that seemingly random infant behaviors - kicking their legs and putting their fists into their mouths - help tune models of their *Ecological Selves* [1]. The Ecological Self is a cohesive model of the body and the senses, learned by using and witnessing them with respect to each other. Its unified nature allows sensory information to be combined and interpreted as a whole. Tactile information, for instance, combined with kinesthetic information, can inform where a touched object is in space.

Infants are born with an early form of their Ecological Selves intact. It has been demonstrated that infants exhibit the *rooting reflex*, orientation of the mouth toward tactile stimuli when touched on the cheek, with greater frequency when the touch is performed by an experimenter, rather than a self-touch, performed by the infant [2]. Infants also, when placing their fists into their mouths, open their mouths in preparation, demonstrating knowledge of their kinematic structure [3]. The pairing between the tactile and kinesthetic senses allows the infant to disambiguate between the self-stimulus of their own touch and the external stimulus of another person touching them, leading to the differential rooting response. The fact that one's own mouth cannot be witnessed in their visual field leads to the conclusion that, in the fist-sucking behavior, interaction between their tactile and kinesthetic senses allows infants to know the distance between the fist and the mouth.

Imitating this human process, the robot discussed in this paper learns a model of its Ecological Self. Starting with a calibrated model of its vision system, the robot learns its arm kinematics and the arm's relationship to its visual field. It learns this model by witnessing the motion of its end-effector in its visual field. The product of this learning process, however, is grounded firmly in modern engineering techniques, outputting the Denavit-Hartenberg parameters [4] of the robot's end-effector, a commonly-used kinematic modeling convention. This process calibrates the robot's kinematic model to its vision model, producing a unified model that allows kinematic and visual information to be meaningfully combined. The presented model is demonstrated to predict end-effector position in the visual field. The robot further refines this combined kinematic-visual model by minimizing the distance between the predicted and observed positions of its end-effector in its visual field. The fact that the robot learns these properties based on first-hand observations of its body, through its sensors, allows the model to adapt to changes in the robot's configuration on-line. This endows the robot with the ability to adapt its self-model for tool use.

This work is related to work in body schemas, for which Hoffmann, Marques, Arieta, Sumioka, and Pfeifer [5] provide a good overview. Several papers in the body schema literature have been devoted to the subject of learning robot kinematics. Hersch, Sausser, and Billard [6] present a robot which learns the parameters of a model describing its kinematic chain. Cantin-Martinez, Lopes, and Montesano [7] present a similar model to Hersch *et al.* [6], improving on the number of samples required for training by several orders of magnitude through the use of better optimization techniques and active

learning. Sturm, Plagemann, and Burgard present a technique utilizing a Bayesian representation of kinematic chains [8].

The work of Hersch *et al.* [6] and Cantin-Martinez *et al.* [7], as well the work presented in this paper, are *kinematic calibration* techniques. Hollerbach and Wampler [9] provide a good overview, and would classify all three methods as *open-loop methods*, because of the use of an external metrology system (the robot's vision system). The work presented in this paper differs from the approaches presented by Hersch *et al.* [6] and Cantin-Martinez *et al.* [7] in that it utilizes a different representation of the kinematic chain, achieves better spatial resolution by an order of magnitude, and requires fewer samples. The presented system also is initialized using a *screw-axis measurement method* and focuses on the task of predicting the robot's end-effector position in its visual field, which is not attempted in either Hersch *et al.* [6] or Cantin-Martinez *et al.* [7]. In the sense that this work concentrates on the intermodal problem of combining kinematics and vision, it is similar to the work of Yoshikawa, Tsuji, Hosada, and Asada [10] and Stoytchev [11]. It differs in that it utilizes a parameterized model of the robot's kinematics, concentrating on mutually calibrating this model and that of the robot's vision system. While other work on this intermodal perception problem focuses on the biological plausibility of the approach, the presented method is intended to build on classical engineering techniques, such that it may be easily integrated into existing systems [10], [11].

## II. A Model of the Ecological Self

This work focuses on integrating two systems, kinematics and vision. It should be noted that these are not the only senses for which the notion of the Ecological Self is applicable. Yoshikawa *et al.* [12], for instance, focus on the combination of visual and tactile stimuli. Ideally, as many senses as possible would be integrated, yielding a single, cohesive model model of the self and the environment.

### A. Kinematics

The *Denavit-Hartenberg parameters* [4] are a kinematic modeling convention which represents the rotational axes of revolute joints as lines in space. The axes for adjacent joints are described with respect to a line running normal to both. This allows a minimal description of each joint, consisting of four parameters. These parameters describe a transformation between two coordinate frames. The rotational axes are the $z_i$ axes of these frames, and the common normals are the $x_i$'s. The parameters, illustrated in Figure 1, then, are:

- $\theta_i$ The joint angle, or equivalently, the angle between $x_i$ and $x_{i+1}$ about $z_i$.
- $r_i$ The joint radius, measured along $x_{i+1}$.
- $\alpha_i$ The angle between $z_i$ and $z_{i+1}$.
- $D_i$ The distance between $x_i$ and $x_{i+1}$ along $z_i$.

The transformation produced by a single joint is represented as a matrix $M_n$. The position of the end-effector is determined by multiplying these matrices together, where $M_0$ is the transformation to the inertial frame and $E$ is the position of the end-effector in this frame:

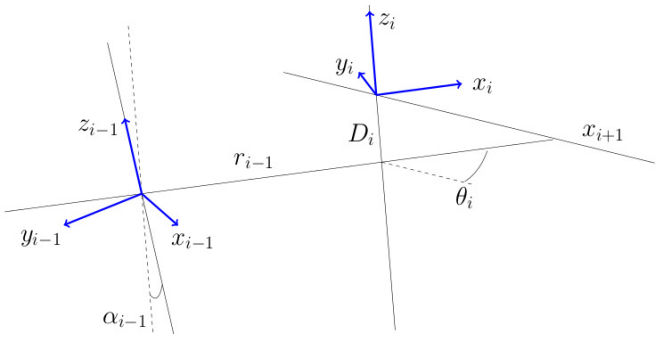



Fig. 1. Engineering diagram showing the Denavit-Hartenberg parameters.

$$E = M_0 \dots M_n [0, 0, 0, 1]^T \tag{1}$$

The present model adds two parameters to this standard set. Since the zero point of the robot's encoder is unlikely to match the corresponding $\theta_i$, the offset between the two is represented as $\hat{\theta}_i$. It is also assumed that gear reduction is not known *a priori*, therefore it is represented as $G_i$. The joint angle passed to the robot's motor, $\breve{\theta}_i$, then, is computed as in Equation 2.

$$\breve{\theta}_i = \hat{\theta}_i + G_i \theta_i \tag{2}$$

### B. Stereo Vision Parameterization

Cameras are parameterized under the *pinhole camera model* [13]. Pixels in an image are modeled as rays of light running through the aperture of a pinhole camera. The *intrinsic parameters*, describing the camera itself, are stated as a $3 \times 3$ matrix, as in Equation 3. The parameters $\alpha$ and $\beta$ describe focal length, their ratio accounting for non-square pixels. The *principal point*, the point at which a ray perpendicular to the camera's image plane would run through the camera center, is at $(u_0, v_0)$. The final parameter, $\gamma$, is a skew factor accounting for non-rectangular pixels. Images are undistorted using two terms of radial distortion, following the model found in Zhang [14]. Distortion is otherwise not addressed by this model.

$$K = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

$$p = K[R| - RC]P \tag{4}$$

The *extrinsic parameters* $R$ and $C$ describe the camera's position and orientation, where $R$ is the rotation of scene points about the camera, and $C$ is the camera's position. Scene points, then, are projected as in Equation 4, where $P$ is the homogeneous representation of a point in 3D, and $p$ is its projection. Given these parameters for a stereo pair of cameras, it is possible to reconstruct the position of a point in 3D, $P$, by determining the intersection of the light rays corresponding to its image in the two cameras, $p$ and $p'$. Prime denotes variables pertaining to the second camera in the stereo pair.

### C. A Combined Model

Combining these two models is a matter of assuring that their origins, orientations, and scales match. As will be seen in Section III, this is accomplished by learning the Denavit-Hartenberg description of the robot's kinematics by witnessing its end-effector motion in its visual field, assuring that the measurements of this motion are stated in the basis of the vision system. Given an accurate calibration between these two models, the position of the end-effector in the visual field can be determined by combining the forward-kinematic model with Equation 4, for both cameras. This is done by substituting the projected point in 3D, $P$, with the forward-kinematic model, as in Equation 5.

$$p_{end-effector} = K[R|-RC]M_0 \dots M_n[0,0,0,1]^T \quad (5)$$

## III. Estimating the Model Parameters

The initial approach taken in this work to learning this model is to learn the robot's kinematics under its stereo vision system. Observe that the stereo vision and kinematic models may be unified by representing them in the same basis, calibrating them to each other. By measuring the position of the robot's end-effector in its stereo vision system, the kinematic model is computed from points already represented in the correct basis. Learning the robot's kinematics in this way assures that the scale, orientation, and origin of the model of the kinematic chain match those of the stereo vision system.

### A. Kinematic Learning

The stereo vision system is first calibrated using standard techniques. An initial calibration is performed via Zhang's method [14], then refined through *bundle adjustment* [13].

To learn the robot's kinematics, end-effector motion is first tracked through space. The robot's cameras track this motion using techniques described in Section IV-B. Stereo vision is used to reconstruct the end-effector's position in 3D. The kinematic-chain parameters are inferred from samples of this motion. It is assumed that the number of joints in the system is known *a priori*, but that the values of the parameters describing them, the Denavit-Hartenberg parameters, $G_i$, and $\hat{\theta}_i$, are not.

The process of kinematic learning then proceeds in two steps. The first is a derivation of *Circle Point Analysis* (CPA) developed for this work. CPA provides a good initial estimate of the robot's kinematic chain. Initializing the kinematic model with an accurate estimate of its parameters prevents the second step, which performs a nonlinear refinement of this estimate, from stopping at local minima.

*1) Circle Point Analysis:* Circle Point Analysis refers to techniques proceeding from the observation revolute joint motion that traces a circle in space. The rotational axis is treated as a line perpendicular to the plane in which this circle lies, running through its center. The present derivation of CPA assumes that a single, home position for the kinematic chain is first determined. As three points are required to uniquely identify a circle, a circle is sampled for each joint by moving it into at least two poses away from this position. These poses are used to reconstruct the circle uniquely identifying each joint with respect to the rest of the chain. A summary can be found in Algorithm 1. In this implementation, circle fitting is performed using an in-house implementation of the technique from the NIST Algorithm Testing System [15]. Non-linear optimization is performed using FindMinimum from Wolfram Mathematica, Version 7.0.1.0, using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method for the initial 2D estimates, and the Levenberg-Marquardt method for the final 3D refinement.

**Algorithm 1** Circle Point Analysis

1: Determine an initial, home position for the kinematic chain
2: **for** $i = 1$ to $n$ where $n$ is the number of joints in the chain **do**
3: Move kinematic chain to home position
4: Move joint $i$ through at least 2 additional positions along its arc of motion
5: Fit a circle to the set of 3 or more sampled points for this joint.
6: **end for**
7: **return** The set of measured circles

Determining a home position for the kinematic chain, then finding the set of rotational axes with respect to this position, gives us a set of lines in space. Finding the Denavit-Hartenberg parameters, then, is a matter of determining the relationships between these axes. For clarity, the following variables are added to the parameterization:

- $p_i$ The endpoint of the motion of joint $i$.
- $C_i$ The center of the reconstructed circle.
- $w_i$ The distance from $C_i$ to $p_i$.
- $e_i$ The distance from $C_i$ to $p_{i-1}$.

The parameters are found as follows:

- $z_i$ Runs perpendicular to the plane in which the circle lies. $z_0 = (0,0,1)$. For convenience, $z_{n+1}$ is parallel to $z_n$.
- $x_i$ The normal between $z_{i-1}$ and $z_i$. In the case of parallel axes, choose $x_i$ to run through the centers of the two circles.
- $\alpha_i, \theta_i$ Found as the dot product over the unit vectors between the relevant axes, $z_i \cdot z_{i+1}$ and $x_i \cdot x_{i+1}$, respectively.
- $r_i, D_i$ The relationship between $C_i$ and $p_i$ is described by Equation 6, allowing these parameters to be determined by the Singular Value Decomposition (SVD) described by Equation 7. Note that the matrix in Equation 7 is rank 3, making the SVD underconstrained. The distance, $w_i$, can be found from $C_i$ and $p_i$, allowing the solution vector to be normalized, yielding the correct solution.

$$C_i + e_i z_i + r_i x_{i+1} + d_{i+1} z_{i+1} = p_i \quad (6)$$
$$C_i - p_i = F_i$$

$$\begin{bmatrix} F_{i_x} & z_{i_x} & x_{i+1_x} & z_{i+1_x} \\ F_{i_y} & z_{i_y} & x_{i+1_y} & z_{i+1_y} \\ F_{i_z} & z_{i_z} & x_{i+1_z} & z_{i+1_z} \end{bmatrix} \begin{bmatrix} w_i \\ e_i \\ r_i \\ d_{i+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (7)$$

Given that each $\theta_i$ for each sample must be estimated (allowing the parameters $\hat{\theta}_i$ and $G_i$ to be estimated), it can be seen that the number of unknown variables defining the system (including those defining the inertial frame) is equal to the minimum number of measurements sampled by this method.

*2) Offset and Gear Reduction:* Estimates of the parameters $\hat{\theta}_i$ and $G_i$ are determined by minimizing the squared difference between the $\theta_i$ estimates yielded by CPA and the product of Equation 2 for joint angles passed to the robot during point sampling. The presented implementation uses the BFGS method, as implemented in FindMinimum in Wolfram Mathematica 7.0.1.0.

*3) Nonlinear refinement:* To refine the model yielded by CPA, an additional set of randomly distributed poses is sampled. The squared distance between the set of predicted end-effector positions yielded by the forward-kinematic model and measured end-effector positions is minimized over this combined dataset. Minimization is performed over the set of Denavit-Hartenberg parameters, $\hat{\theta}_i$'s, and $G_i$'s. Optimizations in the presented results use LevMar [16], an implementation of the Levenberg-Marquardt algorithm in C++.

### B. Kinematic/Visual Refinement

During testing it was determined that an improvement in system performance is obtained by simultaneously refining the kinematic and visual parameters. Using the dataset from Section III-A3, the squared distances between the projections of the predicted end-effector position and the two-dimensional end-effector positions from each camera are minimized. LevMar [16] is used in the presented implementation. Section IV-C shows that this can be used as a global refinement of both the kinematic and visual parameters. A similar global optimization for refinement of kinematic and vision calibration is performed in work by Pradeep, Konolige, and Berger [17].

## IV. Results

The upper-torso humanoid robot, Nico, was used to validate this system, Figure 2. The robot includes a stereo vision system, with two $640\times480$ resolution cameras. The kinematics of four degrees of freedom (DOFs) in the robot's right arm, two at the upper shoulder, and two at the elbow, are modeled.

The two basic measures of this system's performance are the distance from the predicted end-effector position to the measured end-effector position in 2D (presented in pixels) and 3D (presented in millimeters). The standard deviation in reconstructions of a target of known dimensions is reported as an estimate of stereo vision performance. Finally, a measurement of how closely the robot's internal model matches external measurements is presented. While it is possible to compare the 3D results to the primary results of Hersch *et al.* [6] and Cantin-Martinez *et al.* [7], the authors attempted neither the 2D task presented, nor to perform stereo vision refinements in those experiments. This work is unique in its approach, which allows the system to accomplish the 2D and 3D tasks simultaneously.

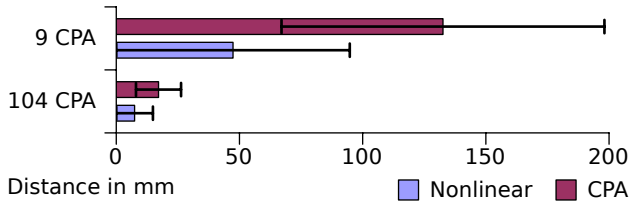



Fig. 3. Results for kinematic learning test using Vicon motion tracker, expressed as distance from predicted end-effector position to measured end-effector position. The test is performed over 100 random samples. "CPA" shows results after circle point analysis with no nonlinear refinement, whereas "Nonlinear" shows performance after refinement.

The test itself proceeds in four steps. First, it is verified that the kinematic learning algorithm is able to learn the robot's kinematic structure. A second test characterizes the performance of kinematic learning under the robot's stereo vision system. The third tests the integrated, full-model learning method from Section III-B. The final test demonstrates the adaptation to tool use.

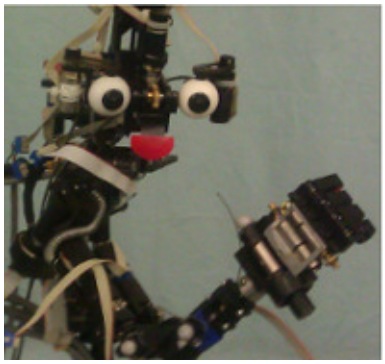
### A. Accuracy of the Kinematic Learning Method

First, the accuracy achieved by the kinematic learning method described in Section III-A is examined in order to establish that the method works as expected and to estimate the number of samples required to adequately train the kinematic model. To test this method in isolation of the stereo vision system, the robot was instrumented with markers for use with the Vicon MX motion tracker, as in Figure 2(a).[1] In this test, the tracker acts as an independent, externally validated 3D position tracker. The robot sampled a dataset consisting of points for Circle Point Analysis, sampled along the arcs of motion of individual joints as described in Section III-A1, and points with the arm in random poses for the purpose of evaluation. The dataset consists of 104 points for CPA and 100 random test samples. Points such that the end-effector is obstructed from the field of view of the Vicon MX cameras were discarded.

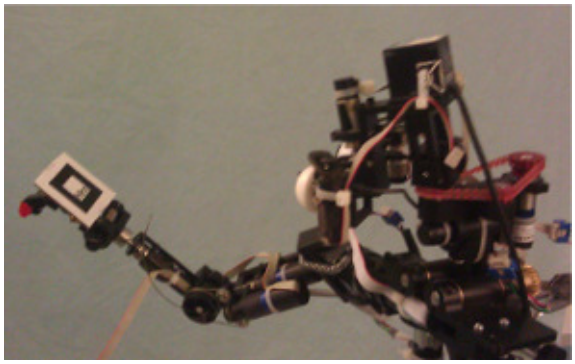
Figure 3 shows two important points regarding the algorithm. First, additional points improve the performance of CPA greatly. The minimum number, 9 points, yields a model accurate to within 132.60mm (SD=65.53), while the model trained on 104 points is accurate to within 17.15mm (SD=9.15). Second, nonlinear refinement over the learned kinematic model significantly improves performance in both cases. In the 9 training sample case, it reduces error to 47.38mm (SD=34.24), and in the 104 sample case to 7.44mm (SD=3.51).

An additional set of 500 random arm poses was sampled. It was added to both the 9 and 104 point training sets during

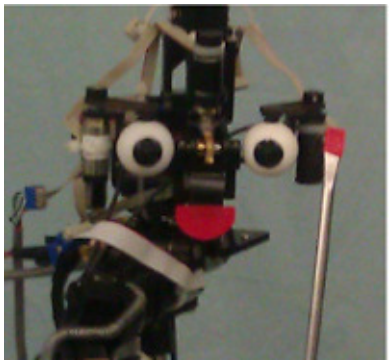
[1]Only the marker on the hand was used to train the system. The tracker requires a minimum of three markers placed on the tracked object.



(a) The robot instrumented with reflective markers for the Vicon MX motion tracker.



(b) A fiducial marker is mounted on a piece of cardboard, then taped to the back of the robot's end effector.



(c) Colored electrical tape is wrapped around the tip of the screwdriver to test tool use performance.

Fig. 2. Images of the humanoid robot, Nico, configured for the three trackers used in this experiment.

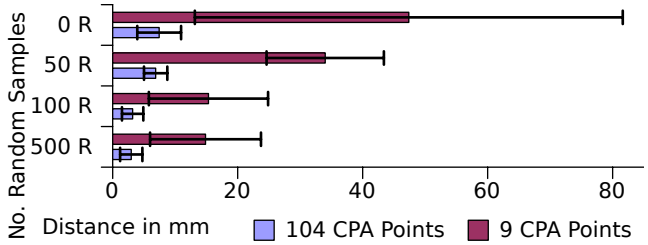



Fig. 4. Results for the kinematic learning test using points used for CPA, plus $0, 50, 100$, and $500$ randomly sampled points for nonlinear refinement.

nonlinear refinement in subsets of $50$, $100$, and $500$. This step evaluates the importance of additional random training samples to the algorithm and establishes an estimate of how many samples are required to fully train the system. Figure 4 shows that the performance of nonlinear refinement is sensitive to the quality of the initial estimate provided by CPA. Having a better initial estimate significantly improves performance. Additionally, performance peaks after approximately $100$ randomly distributed points have been sampled. In the remaining tests, kinematic models are refined using $100$ randomly distributed points, in addition to the points sampled for CPA.

### B. Learning Kinematics Through Stereo Vision

To test the capability of kinematic learning to create a mapping between the kinematic model and the stereo vision system, the robot's kinematic model was trained on a dataset sampled through its eye cameras. To track the end-effector visually, the robot's software is configured to use a fiducial tracker implemented using ARToolKit [18], with the robot instrumented as in Figure 2(b).

Though ARToolKit is able to provide estimates of 3D position, these estimates require knowledge of the robot's camera calibration. The described system, however, is based on stereo vision and performs visual calibration refinement. Modifications to ARToolKit were made to perform 2D tracking in the absence of this calibration, in order to eliminate the possibility of this *a priori* knowledge affecting the results of this test. The final system uses this 2D tracker output from both cameras to reconstruct 3D end-effector position.

To facilitate comparison to other work and simplify result interpretation, the vision system was calibrated to a metric using a chessboard calibration target. Knowing that the sides of the squares on the target are $28$mm, measurements of the 3D reconstruction of this target in several poses are used to compute a conversion to millimeters. The system is accurate to within $1.66$mm.[2]

Because stereo vision can only sample poses witnessed in the robot's visual field, the robot sampled only $60$ CPA points. Having established that $100$ random training samples is sufficient, the dataset contains $100$ training points and $100$ test points for this test. After training, the system predicts end-effector position to within $2.92$mm (SD=$0.10$). It also predicts the position of the end-effector in the visual field to within a mean of $4.21$ pixels (SD=$6.40$) in both cameras.

### C. Full-Model Learning

While testing the system described in Section III-A, the learning method in Section III-B was developed to determine if it would improve performance on the task of predicting end-effector position in the visual field by simultaneous adaptation of kinematic and visual parameters. This technique is applied as a refinement of the model learned in the previous section, over the fully-trained model. In this experiment, the model was also separately refined with pinned camera intrinsic parameters[3] to better illustrate improvements due improved camera calibration.

*1) Impact of Full-Model Learning on the Vision System:* Estimates of the accuracy of stereo reconstruction were computed before and after full model learning, using the method discussed in Section IV-B. The full-model learning procedure improved the quality of stereo reconstructions, improving estimates of the accuracy of stereo reconstructions from within $1.66$mm to within $1.09$mm.

[2]Standard deviation of width of squares in reconstructions of the calibration target.

[3]To pin a parameter is to disallow the optimizer from changing its value.

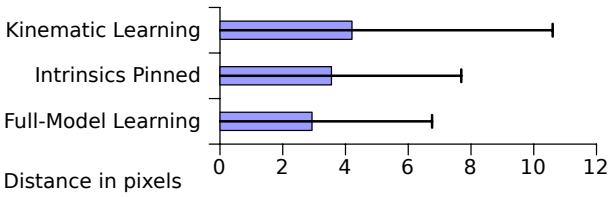



Fig. 5. Comparison of performance in 2D between kinematic and full-model learning. The test is performed over 100 random samples. Results labeled "Kinematic Learning" use CPA and nonlinear refinement. Results labeled "Full Model Learning", and "Intrinsics Pinned" use the technique outlined in Section III-B, to improve on the "Kinematic Learning" results. The "Intrinsics Pinned" case does not attempt to refine the camera intrinsic parameters.

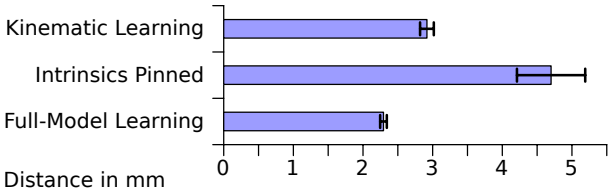



Fig. 6. Comparison of performance in 3D between kinematic learning and full-model learning.

*2) Performance in Predicting End-Effector Position:* Figures 5 and 6 compare the performance of kinematic learning and full-model learning in 2D and 3D, respectively. Figure 5 shows that full-model learning consistently improves 2D performance. This is expected, as the metric being minimized is also the one being measured. When the camera intrinsic parameters are pinned, however, it harms 3D performance. Overfitting 2D performance on an imperfectly calibrated vision system comes at the expense of 3D performance. By training the kinematic and visual models to each other, however, the two converge. The continually improving estimates of the robot's kinematic structure improve its ability to serve as a stereo calibration target,[4] while the refinements in stereo reconstruction inform the accuracy of the kinematic model. Note that the extrinsic parameters can be manipulated in both cases. This is, in part, an improvement in the intrinsic calibration of each camera. The manner in which improvements in the accuracy in one model inform the calibration of the other is reminiscent of Rochat's theory of the development of the Ecological Self [1]. The final model is accurate to within 2.29mm (SD=0.10) and 2.93 pixels (SD=3.83).

*3) Estimates of Linkage Lengths:* The arm of the robot used in this experiment comprises two main segments with paired joints at the intersection of those segments. To verify the estimated model of the robot's kinematics, external measurements of the two main segments were obtained for comparison against the robot's internal estimates. The first segment goes from the robot's shoulder to its elbow and is 130mm long. The second goes from the elbow to the end-effector and is 127mm long. As shown in Figure 7, estimates of linkage lengths are accurate to within 1.1mm (0.85% of the length of the linkage) for both linkages, respectively (First: 130.79 (0.61%), Second: 128.09 (0.85%)), when trained using full-model learning.

[4]Compare to classical photogrammetric techniques, in which the projection of a target of known shape is computed from images of it.

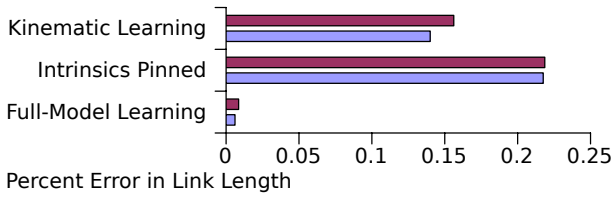



Fig. 7. Estimate of linkage lengths expressed as a percentage of the externally-measured linkage length.

### D. Tool Use

The system is able to adapt to tool use by retraining an already-initialized model using the techniques from Sections III-A3 and III-B. A test of the system's adaptation to tool use was performed in which the robot was instrumented using red electrical tape to be tracked by a color blob detector built into the robot's vision system. For this test, a screwdriver was placed into the robot's end-effector, its tip marked with colored electrical tape, as in Figure 2(c). Initial training was performed with electrical tape wrapped around a finger on the robot's end-effector. Most of the tests in this paper are done using fiducial markers, which are more accurate. Both camera intrinsic and extrinsic parameters must be pinned when training using color segmentation data, as the accuracy of the tracker is not sufficient for camera calibration. Time did not permit re-performance of this test, due to the need to repair a joint in the robot's arm. Color blob detection results are omitted for other tests, due to space restrictions.

A total of 100 random test points and 52 random training points were sampled. The robot's kinematic model was updated using the nonlinear refinement method from Section III-A3 and full-model learning. The system, as trained on the robot's hand, tracked its end-effector to within 4.04mm (SD=0.10), and 3.09 (SD=3.44) pixels. Upon retraining with the screwdriver, the system adapted, tracking the end-effector to within 7.18mm (SD=1.17), 4.69 (SD=7.74) pixels.

## V. Discussion

This work is uniquely situated in the literature on body schemas. Focusing on the concept of the Ecological Self, this model centers on the interaction between the body and the senses. Much as infants learn about their bodies through their senses, so does the robot in this study. This model focuses on integrating kinematics and the visual sense, creating a predictive model of the position of the end-effector in the visual field. In this sense, this work is similar to Yoshikawa *et al.* [10], and other related work. Yoshikawa *et al.* [10] focus, however, on statistical methods and correlations between visual and tactile stimuli and motor states, while this paper focuses on bridging the gap between a parametric model of the visual system and one of the kinematics. A trade-off is made between the ease with which this system can be integrated into existing robotic systems, using off-the-shelf inverse-kinematic solvers and vision techniques, and the biological plausibility of systems such as the one developed in Yoshikawa *et al.* [10].

In that this system learns a parametric model of the robot's kinematics, it is similar to work by Hersch *et al.* [6] and

Cantin-Martinez *et al.* [7]. It requires significantly fewer training samples than Hersch *et al.* [6], who post results as learning curves requiring on the order of $10^6$ training examples, and outperforms both in predicting end-effector position in 3D. Peak performance in the systems presented by Hersch *et al.* [6] and Cantin-Martinez *et al.* [7] determines end-effector position to within about 5cm. This system predicts end-effector position to within $47.4$mm (SD=$34.24$) after only $9$ training examples using the Vicon tracker, achieving peak a precision of $2.29$mm (SD=$0.10$) using stereo vision.

In comparing these results to those presented by Hersch *et al.* [6] and Cantin-Martinez *et al.* [7], it is important to note that all three systems use different tracking methods. Hersch *et al.* [6] use stereo reconstructions based on a color blob tracker. Cantin-Martinez *et al.* [7] use 3D position estimates provided by ARToolKit [18] and report that their vision system is only accurate to within around 5cm. As such, part of the present system's performance can be explained by the accuracy its vision system. Additionally, all three systems sample kinematic configurations differently. Hersch *et al.* [6] use random motion, and Cantin-Martinez *et al.* [7] use active learning to explore the space. The present system is initialized using the structured motion of CPA, followed by random samples. All three systems post results based on random samples. The systems were also evaluated on different robots with differing mechanical complexities. Hersch *et al.* [6] first validated their model on a simulated 24-DOF robot, then demonstrated its ability to adapt to tool use by initializing a HOAP-3 robot with an accurate kinematic model and learning the adapted kinematics of a 5-DOF arm. Cantin-Martinez *et al.* [7] validate their algorithm on simulated 6 and 12-DOF robots, and test on 4-DOFs of the robot Baltazar's arm. In this paper, all tests are performed on 4-DOFs of the robot Nico's arm. Cantin-Martinez *et al.* [7] use only 4 of the 6 DOFs due to accuracy limitations of their vision system. In this paper, only 4 of the 6 DOFs are used due to difficulty obtaining unobstructed images of the fiducial marker in the range of motion of the two distal joints.

This paper is inspired by Rochat's [1] narrative of an infant learning about their body and senses by witnessing them mutually through each other. At peak performance the system identifies the position of the end-effector in the visual field to within $2.93$ pixels (SD=$3.83$). Creating a pairing between the system's vision and kinematic models that is strong enough to do this is an important goal of this paper, as sensorimotor integration of this nature is one of the properties of the Ecological Self [1]. Results demonstrate that this system is able not only to learn a highly-accurate model of the robot's kinematics, but also that it is able to improve on the system's already accurate vision calibration. The model is based on modern engineering techniques and can be easily integrated into existing robotic systems.

## VI. Conclusion

This paper presents a robotic model of the Ecological Self. It focuses on the idea that this is an intersensory model, capable of cohesively pairing the robot's stereo vision and kinematic systems. The system outperforms the most similar existing systems in the literature, both in terms of the precision of the learned model and the number of training examples required to train the system. It is based on modern engineering techniques, and could be easily incorporated into the software of many existing robots. This work represents a step towards our group's goal of building robots that incorporate their physical selves as first-class components which can be modeled, reasoned about, and modified as needed, into their cognitive models.

*Acknowledgements*

This material is based upon work supported in part by the National Science Foundation under grants SES-0835767 and IIS-0968538. The authors also acknowledge the generous support of Microsoft and the Sloan Foundation.

## References

[1] P. Rochat, *The Infant's World*. Cambridge, Massachusetts and London, England: Harvard University Press, 2001.

[2] P. Rochat and S. J. Hespos, "Differential rooting response by neonates: Evidence for an early sense of self," *Early Development and Parenting*, vol. 6, no. 3–4, pp. 105–112, 1997.

[3] P. Rochat, E. M. Blass, and L. B. Hoffmeyer, "Oropharyngeal control of hand-mouth coordination in newborn infants," *Developmental Psychology*, vol. 24, no. 4, pp. 459–463, 1988.

[4] J. Denavit and R. S. Hartenberg, "A kinematic notation for lower-pair mechanisms based on matrices," *Trans. of the ASME J. of Applied Mechanics*, vol. 23, pp. 215–221, 1955.

[5] M. Hoffmann, H. Marques, A. H. Arieta, H. Sumioka, M. Lungarella, and R. Pfeifer, "Body schema in robotics: a review," *IEEE Trans. Auton. Mental Develop.*, vol. 2, no. 4, pp. 304–324, December 2010.

[6] M. Hersch, E. Sauser, and A. Billard, "Online learning of the body schema," *Intl. J. of Humanoid Robot.*, vol. 5, no. 2, pp. 161–181, 2008.

[7] R. Cantin-Martinez, M. Lopes, and L. Montesano, "Body schema acquisition through active learning," Alaska, USA, 2010.

[8] J. Sturm, C. Plagemann, and W. Burgard, "Body schema learning for robotic manipulators from visual self-perception," *J. of Physiology-Paris*, vol. 103, no. 3-5, pp. 220–231, Sept. 2009.

[9] J. Hollerbach and C. Wampler, "The calibration index and taxonomy for robot kinematic calibration methods," *Intl. J. of Robotics Research*, vol. 14, no. 573–591, 1996.

[10] Y. Yoshikawa, Y. Tsuji, K. Hosada, and M. Asada, "Is it my body? - body extraction from uninterpreted sensory data based on the invariance of multiple sensory attributes," in *Proc. of 2004 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2004, pp. 2325–2330.

[11] A. Stoytchev, "Toward video-guided robot behaviors," in *Proc. of 7th Intl. Conf. on Epigenetic Robotics*, 2007, pp. 165–172.

[12] Y. Yoshikawa, K. Hosoda, and M. Asada, "Unique association between self-occlusion and double-touching towards binding vision and touch," *Neurocomput.*, vol. 70, pp. 2234–2244, August 2007.

[13] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[14] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 1330–1334, 2000.

[15] C. M. Shakarji, "Least-squares fitting algorithms of the nist algorithm testing system," *J. of Research of the National Institute of Standards and Technology*, vol. 103, no. 6, pp. 633–641, 1998.

[16] M. Lourakis, "Levmar: Levenberg-marquardt nonlinear least squares algorithms in C/C++," [web page] `http://www.ics.forth.gr/~lourakis/levmar/`, Jul. 2004, [Accessed on 31 Jan. 2005.].

[17] V. Pradeep, K. Konolige, and E. Berger, "Calibrating a multi-arm multi-sensor robot: A bundle adjustment approach," in *Intl. Symp. on Experimental Robotics (ISER)*, New Delhi, India, 12 2010.

[18] I. Poupyrev, H. Kato, and M. Billinghurst, *Artoolkit user manual, version 2.33*, 2000.