Grasp Recognition and Manipulative Motion Characterization from Human Hand Motion Sequences

Sing Bing Kang and Katsushi Ikeuchi

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We are developing a system capable of observing a human performing a task and understanding the task well enough to replicate it. This approach is called Assembly Plan from Observation [6]. In order to replicate the observed task, we have to analyze the entire sequence. This can be done by first segmenting the task sequence into its constituent pre-grasp, grasp, and manipulation phases [8]. This paper describes the different analyses that can be done subsequent to the temporal segmentation. These include human grasp recognition, extraction of object motion, and the spatiofrequency (spectrogram) analysis of the manipulation phase.

1 Introduction

Current conventional methods to robot programming include teach-by-guiding (e.g., [2]) and textual programming (e.g., [5]), which have their deficiencies of inapplicability in hazardous environments and requirement of skills, respectively. Automatic programming methods (e.g., [13]) seek to reduce the task programming burden, but face the problem of combinatorial complexities in path planning and grasp synthesis. We are currently developing a system which has the capability of observing a human performance of a task, understanding it, and subsequently performing that task. This approach, called Assembly Plan from Observation [8], relegates most of programming effort to merely demonstrating the task. (This approach is similar to Kuniyoshi *et al.*'s [11]; their system emulates pick-and-place operations from visual observation.) In order to replicate the task, however, the system has to be able to understand the actions performed in the task. We are interested in analyzing tasks which involve grasping, termed *grasping tasks*.

This paper describes our work on analyzing a grasping task sequence, which is done after temporally segmenting it. The temporal task segmentation is important as it serves as a preprocessing step to identify the frames associated with the phases. This information would then be used to focus on the relevant frames in order to characterize the phases in the task. For example, when the grasp phase has been temporally located, the grasp can then be identified using the location of the object and the hand configuration data [7]. In addition, by analyzing the motion of the object within the manipulation phase, the type of motion can be extracted and determined.

2 Temporal segmentation of task sequences

The first step in analyzing a given task sequence is to break it up into meaningful segments. A grasping task comprises three identifiable phases: pre-grasp, grasp, and manipulation phases. It has been demonstrated [8] that by using the following features, the task sequence can be broken into its constituent phases:

1. *Fingertip polygon area*
   The fingertip polygon is defined as the polygon formed with the fingertips as its vertices.
2. *Speed of the hand*
3. *Volume sweep rate*
   This is defined to be the product of the first two measures, and has been found to be more effective in localizing the breakpoints of the task sequence.

3 Task analysis

Subsequent to the localization of the task breakpoints (which are transitions from one phase to another), we can perform certain useful analyses on the individual phases. They include human grasp identification, object motion extraction, and repetitive motion analysis.

3.1 Task analysis system

Our system comprises the following hardware and software:

- CyberGlove [4] and Polhemus [1] devices. The CyberGlove measures 18 joint angles while the Polhemus 3Space Isotrak sensing device provides the position and orientation of the hand relative to the Isotrak source.
- Ogis light-stripe rangefinder to provide range images
- CCD camera to provide intensity images to aid object localization, and for visualization.
- Vantage geometric modeler [3]
- Knowledge Craft [9]. The grasp hierarchy is represented by frames created using Knowledge Craft, which is a frame-based toolkit with procedural attachment and inheritance.

1050-4729/94 $03.00 © 1994 IEEE

- Graphics and interface software. This software is written mostly in C. A significant portion of it is adapted from the VirtualHand v1.0 software supplied by Virtual Technologies).

3.2 Methodology

A grasping task comprises pre-grasp, grasp, and manipulation phases. A task has at least one of these phases (a collection of which is term a *subtask*). We define a sequence of tasks which has N subtasks as an *N-task*.

Analyzing subtasks is equivalent to analyzing separate 1-tasks, since each 1-task contains a subtask; there is no loss in generality in illustrating the analysis using 1-tasks. Each subtask can be characterized in terms of the grasp used and object motion during the manipulation phase.

A series of experiments featuring 1-tasks were conducted as follows:

1. *Take the range image of the scene before the subtask.*
2. *Perform the 1-task (which comprises only a set of pre-grasp, grasp, and manipulation phases) while its intensity image sequence and the CyberGlove and Polhemus readings are being recorded.*
3. *Take the range image of the scene after the 1-task has been performed.*

The general approach in analyzing the task is dictated by the imperfect data. The most significant problem faced is the inaccuracies in the Polhemus device due to nearby ferromagnetic material. In addition, the exact moment of grasping cannot be pinpointed due to the discrete sampling of the hand location and configuration. As a result, extra preprocessing is required, specifically adjusting the orientation of the hand at the grasp frame.

The processes of segmenting the task and determining the grasp and manipulative (i.e., object) motions are done using a three-pass approach. The first pass establishes the motion breakpoints while the second pass involves adjusting the pose of the hand and subsequently determining the grasp employed in the 1-task. Finally, the effect of the reorientation of the hand is propagated throughout the 1-task sequence and the object motion is then extracted using the approach delineated in the following subsection. The details of the three-pass approach are as follow:

Pass 1:

1. *Estimate pose of object from the before-task range image.*
   The initial gross position (but not the orientation) of the object of interest is determined by subtracting the 3D elevation map of the scene after the task from that before the task. The 3DTM[1] program [14] is then used to localize the object. Two refinements were made: (a) use three orthogonal initial poses and pick the final estimated pose with the least RMS fit error; and (b) use coarse-to-fine stepsizes.
2. *Calculate the motion profiles (speed, fingertip polygon area, and volume sweep rate).*
3. *Determine the motion breakpoints from the motion profiles as described earlier.*

Pass 2:

1. *From known motion breakpoints (determined in Pass 1), calculate the object motion associated with the manipulation phase (which is bordered by the grasp and ungrasp transitions).*
2. *At the grasp frame, determine the grasp employed.*
   Due to the errors in the Polhemus and CyberGlove readings, the oriented hand may intersect the object. The hand is reoriented (subject to the fixed position of the Polhemus sensor) until: no interpenetration between the hand and object occurs; and the weighted sum of distances between the hand contact points and the object is minimized.

   The determination of the "optimal" hand pose is done with direct search with rotational increments of 1.15^0 and limited to a maximum of 60^0 rotation about discretely sampled axes (80 directions sampled on a once-tesselated icosahedron).

   The object is stored as a collection of oriented surface points (position and normal information) whose spacing is typically between 4.0-7.5 mm. This spacing of the object depends on the object size - it is increased for a larger object size. The nearest distance of each hand contact point to the object is then estimated using this oriented point representation.

   Once the "optimal" pose of the hand and the object-contact information is found, the grasp is then recognized using the classification scheme described in [7].

Pass 3:

1. *Propagate adjustment in both distal and proximal motions throughout the task due to hand reorientation in Pass 2.*
2. *The gross after-the-task pose of the object is determined by successively applying the distal transformations in the frames composing the manipulation phase to the original object position (i.e., prior to the task). This after-the-task pose is refined using the 3DTM program.*

3.3 Grasp Recognition

The grasp is represented by the *contact web* [7]. It is a 3D spatial representation of effective contact points between the

[1] Short for 3D template matching. It is a least-squares distance error minimization technique using a Lorentzian error distribution.

segments of the hand and the grasped object. Each effective contact point, when in contact with the object, has positional and orientation information. A taxonomy based on this contact web has also been proposed. This taxonomy, in conjunction with a mapping which groups fingers into functionally equivalent ones, enables a given grasp to be identified at the grasp frame [7].

3.4 Determining object motion during the manipulation phase

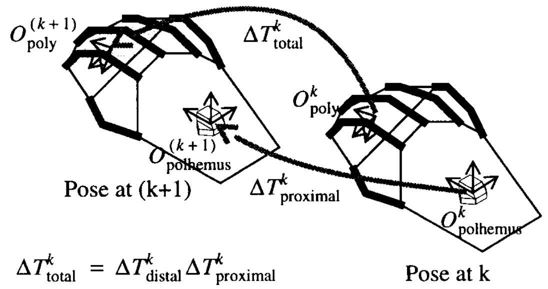


Fig. 1 **Total and proximal motions from frame k to k+1 during the manipulation phase**

It may be useful to determine the *proximal motion* (which corresponds to the motion of the arm and wrist) and *distal motion* (which corresponds to motions of the fingers, otherwise referred to as "precision handling" [12]). The *total motion*, which is the overall effect of both the proximal and distal motions, directly yields the object motion. Meanwhile, the proximal and distal motions yield information on which component of the hand/arm motion is contributing to the object motion.

We can determine the object motion transformations (i.e., the total motion) in the manipulation phase once we have identified the task motion breakpoints. Suppose the *k*th frame has been identified as the grasp frame and the *(k+j)*th frame the ungrasp frame in the task sequence of *N* frames. The object changes in pose at frames between *k* and *(k+j)* (i.e., during the manipulation phase) can be be determined (Fig. 2) from (1):

$$\Delta T_{k,k+j} = T^{k+j}_{hand}(T^k_{hand})^{-1} \qquad (1)$$

where T^k_{hand} is the total transform associated with the motion of both the fingers and hand at the *k*th frame.

Based on (1), we can then calculate the object pose transformations at each frame within the manipulation phase as shown in Fig. 3. The pose of the object at the end of the manipulation phase is most likely not very accurate, due to measurement inaccuracies. This pose is refined using the 3DTM program [14]. Note that directly using 3DTM on the final pose is not generally feasible, since object localization is a local process, and the object may exhibit geometric symmetry.

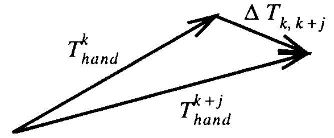


Fig. 2 **Determining the differential motion between two frames in the manipulation phase**

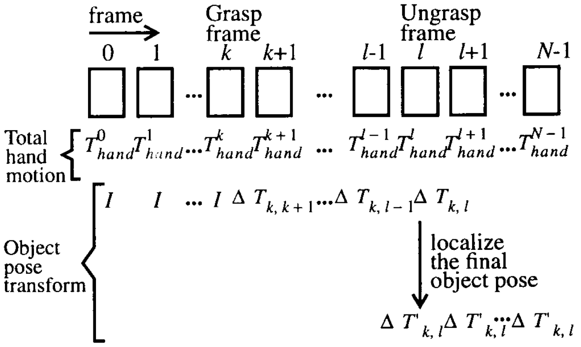

frame
Grasp frame
Ungrasp frame
0 1 k k+1 l-1 l l+1 N-1
Total hand motion
$T^0_{hand} T^1_{hand} \dots T^k_{hand} T^{k+1}_{hand} \dots T^{l-1}_{hand} T^l_{hand} T^{l+1}_{hand} \dots T^{N-1}_{hand}$
$I \quad I \quad \dots I \; \Delta T_{k,k+1} \dots \Delta T_{k,l-1} \Delta T_{k,l}$
Object pose transform
localize the final object pose
$\Delta T'_{k,l} \Delta T'_{k,l} \dots \Delta T'_{k,l}$

Fig. 3 **Determining the pose of the object throughout the task sequence of N frames**

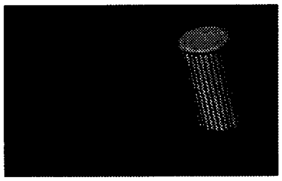


Fig. 4 **Initial pose of the cylinder (1-task #1)**

3.5 Results of applying the 3-pass algorithm

We have applied the 3-pass algorithm on two real 1-tasks to determine the motion breakpoints, identify the grasp employed, and recover the object motion. The first 1-task involves picking up a cylinder from one location and placing it on a different location. The results of the first pass are shown in Fig. 4 and Fig. 5. The pose of the object prior to the performance of the 1-task has been estimated from the range image. As shown in Fig. 5, the motion breakpoints (grasp and ungrasp points) as well as the pregrasp, manipulation, and depart phases are all located. (The duration of each frame is about 0.5 sec; the fingertip polygon area is in cm^2 while the speed is in cm/frame.)

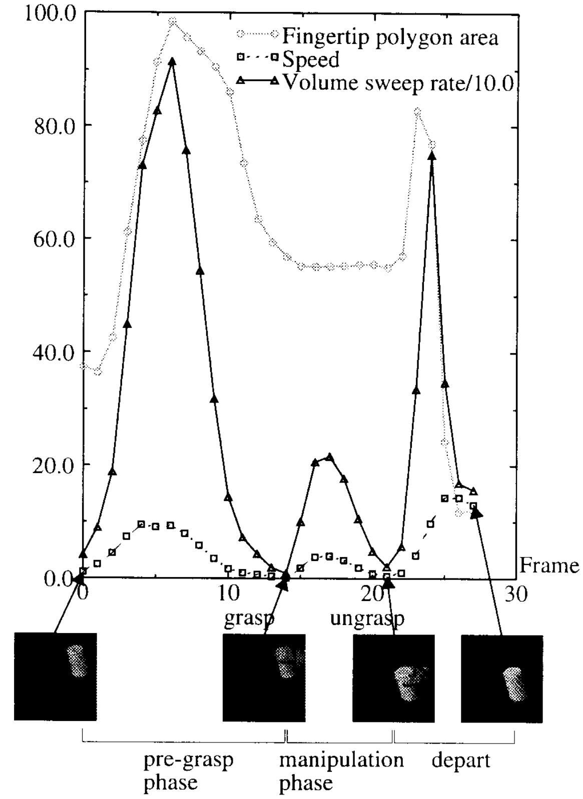
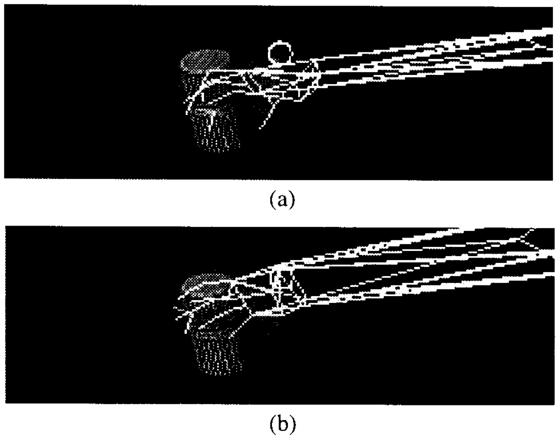


Fig. 5 **Motion profiles and the identified motion breakpoints (1-task #1)**

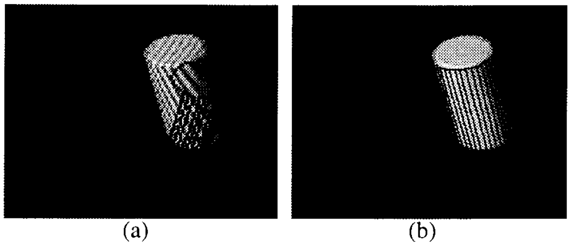

(a)

(b)

Fig. 6 **Reorienting the grasp in Pass 2: (a) initial pose of the hand relative to the object; (b) final pose of hand relative to cylinder**

Once the hand was reoriented (Fig. 6), the grasp was then correctly identified as a type 2 'coal-hammer' cylindrical grasp[2] using the grasp classification scheme described in [7]. By propagating the extracted object motion during the manipulation phase, the object pose was then estimated (Fig. 7(a)). The pose is subsequently refined (Fig. 7(b)).

The second 1-task considered is picking up a stick and inserting it through a hole in a castle-shaped object. The two objects involved in this 1-task and the superimposed model of the stick are shown in Fig. 8. Fig. 9 depicts the extracted motion breakpoints and phases of this task.


(a)

(b)

Fig. 7 **Pose of the cylinder after the task subsequent to Pass 3: (a) pose obtained by successively applying total motion transformations in the manipulation phase; (b) refined pose using the 3DTM program [14].**

Fig. 8 **Initial pose of the stick (1-task #2)**

Fig. 10 shows the pose of the hand relative to the stick at the grasp frame before and after reorientation. The grasp was identified as a precision grasp. However, because the middle segments of the four fingers are within the tolerance range of the object (which is set at 1.0 cm), the grasp is classified as a composite nonvolar grasp [7], specifically a prismatic pinch grasp. The grasp that was actually employed in the task is a five-fingered prismatic precision grasp; it can be seen from this result that the while the general grasp classification is correct, the specific category is sensitive to orientation and position errors.

Fig. 11(a) shows the estimated object pose at the end of the task from extracted total motion. Fig. 11(b) shows the refined final object pose.

[2] A 'coal-hammer' cylindrical grasp is one in which the thumb is highly abducted (i.e., significantly deviated from the plane of the palm). This 'coal-hammer' cylindrical grasp is of type 2 because the thumb touches the object. See [7] for more details.

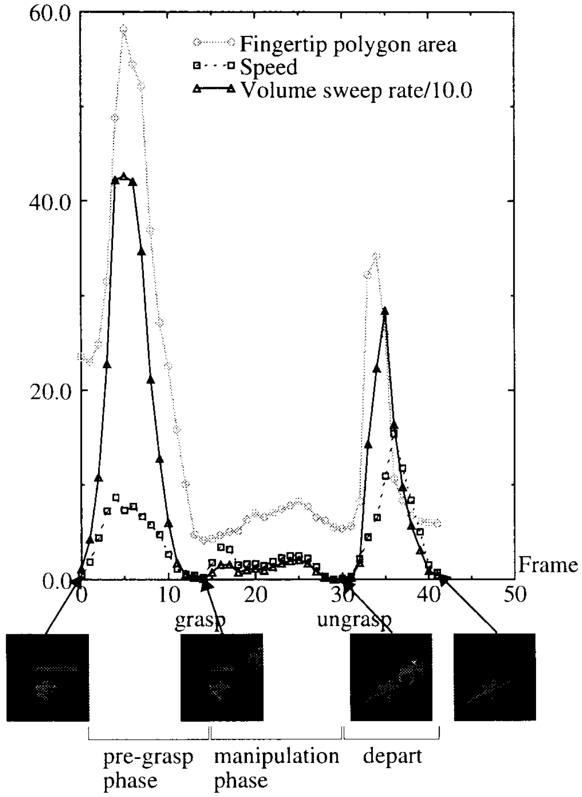
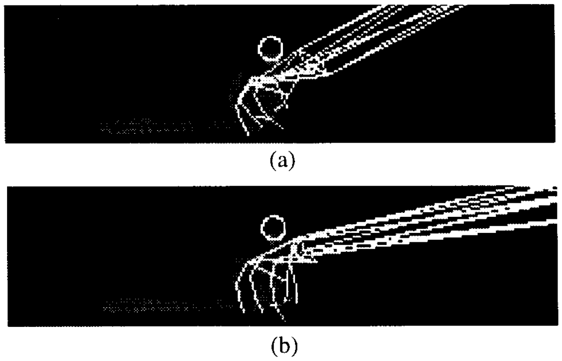


Fig. 9 **Motion profiles and the identified task breakpoints (1-task #2)**



(a)

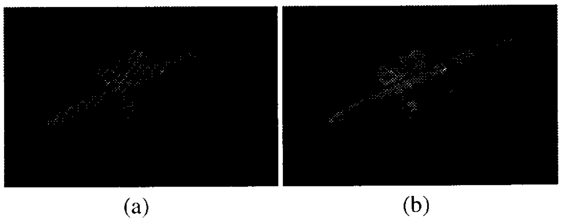
(b)

Fig. 10 **Reorienting the grasp in Pass 2: (a) initial pose of the hand relative to the object; (b) final pose of hand relative to stick**

3.6 Detection and localization of repetitive motion using the spectrogram

Many industrial tasks involve turning screws; it would be useful if the system is able to detect such a repetitive action. The spectrogram is a useful tool for this purpose.

The spectrogram of a signal is a space/frequency representation which comprises a series of small-support, Fourier transforms of the signal, each centered around a different point of the signal [10]. For a 1D signal, this space/frequency representation is 2D. It reveals the frequency content of the signal within the vicinity of each different point. By determining the frequency content locally at each point, we can localize the signal having a particular maximum instantaneous frequency.



(a)

(b)

Fig. 11 **Pose of the stick after the task subsequent to Pass 3: (a) pose obtained by successively applying total motion transformations in the manipulation phase; (b) refined pose using the 3DTM program [14].**

It is obvious from the nature of the spectrogram of its utility in detecting and localizing repetitive motion within the manipulation phase. The motion that is most frequently associated with repetitive motion is the screwing motion. We detect the repetitive motion by analyzing the spectrogram of the fingertip polygon area profile throughout the task, with the following conditions:

1. *Ignore low frequencies.*

   The dc component as well as the first non-zero frequency component are ignored.

2. *Consider only parts of the spectrogram that are associated with manipulation phases.*

3. *Find the highest frequency peak at each point within a manipulation phase.*

   This peak has to be higher than a calculated minimum magnitude determined to the average magnitude of the entire spectrum at that point. In addition, this peak has to be associated with a frequency higher than a determined minimum frequency. This minimum frequency is calculated based on the assumption that there has to be at least three spatial peaks within the manipulation phase for the establishment of a repetitive action.

   If the duration of the manipulation phase (within which the point lies) in the task is M frames, then the minimum frequency is

$$F_{min} = \frac{P_{min}}{M}$$

where P_{min} is the minimum number of peaks within a manipulation phase (3 in our case).

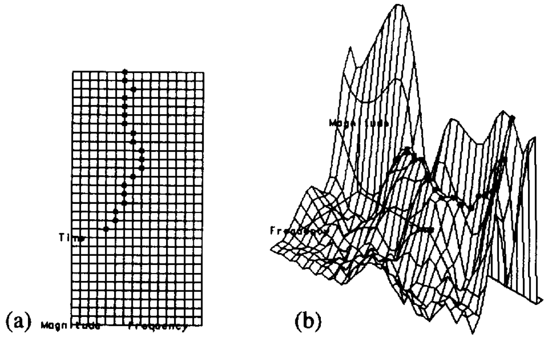


Fig. 12 **Spectrogram of a 1-task involving screwing actions: (a) top view (b) oblique view. The marks on the spectrogram indicate significant frequency peaks**

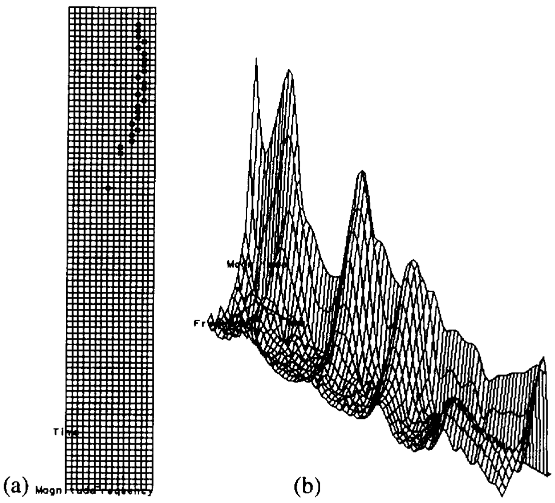


Fig. 13 **Spectrogram of a 4-task involving a manipulation phase with screwing actions: (a) top view (b) oblique view. The marks on the spectrogram indicate significant frequency peaks**

Fig. 12 and Fig. 13 show the spectrogram of two different tasks which involve screw-turning actions. The width of the spectrogram window is 19 frames; the duration of each frame is about 0.5 sec. and the maximum frequency detected in the spectrogram is about 1 Hz. As can be seen, the detected peaks cluttered along a line do indicate the existence of such repetitive movements. In spectrograms of other tasks which do not involve repetitive actions, no prominent peaks were detected, as to be expected.

4 Summary

We have described several of the possible analyses on the task sequence subsequent to the identification of the task breakpoints. The analyses include human grasp recognition and extraction of object motion. The spectrogram is seen to be an effective tool in checking the existence of repetitive motion within the manipulation phase.

Acknowledgments

We are grateful to Mark Wheeler for the use of his 3DTM program. The spectrogram code was adapted from John Krumm's code; many thanks to John for the helpful discussions on the spectrogram.

This research was supported in part by the Avionics Laboratory, Wright Research and Development Center, Aeronautical Systems Division (AFSC), U.S. Air Force, Wright-Patterson AFB, Ohio 45433-6543 under Contract F33615-90-C-1465, ARPA Order No. 7597, and in part by NSF under Contract CDA-9121797.

References

[1] *3Space Isotrak User's Manual*, Polhemus, Inc. Jan. 1992.

[2] H. Asada, and Y. Asari, "The direct teaching of tool manipulation skills via the impedance identification of human motions," *Proc. IEEE Int'l Conf. on Robotics and Automation*, 1988, pp. 1269-1274.

[3] P. Balakumar, J.C. Robert, R. Hoffman, K. Ikeuchi, and T. Kanade, *VANTAGE: A Frame-based Geometric Modeling System - Programmer/User's Manual*, Carnegie Mellon University, Dec. 1988.

[4] Cyber*Glove*^{TM} *System Documentation*, Virtual Technologies, June 1992.

[5] R. Finkel, R. Taylor, R. Bolles, R. Paul, and J. Feldman, *AL: A programming system for automation*, Tech. Rep. AIM-177, Stanford University, Artificial Intelligence Lab., 1974.

[6] K. Ikeuchi, and T. Suehiro, "Towards an Assembly Plan from Observation, Part I: Assembly task recognition using face-contact relations (polyhedral objects)," *Proc. Int'l Conf. on Robotics and Automation*, 1992, pp. 2171-2177.

[7] S.B. Kang, and K. Ikeuchi, "Grasp recognition using the contact web," *Proc. IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, 1992, pp. 194-201.

[8] S.B. Kang, and K. Ikeuchi, *Temporal Segmentation of Tasks from Human Hand Motion*, Tech. Rep. CMU-CS-93-150, Carnegie Mellon University, Apr. 1993.

[9] *Knowledge Craft Manual - Vol. I: CRL Technical Manual*, Carnegie Group, Inc., 1989.

[10] J. Krumm, and S.A. Shafer, "Local spatial frequency analysis of image texture," *Proc. 3rd Int'l Conf. on Computer Vision*, 1990, pp. 354-358.

[11] T. Kuniyoshi, M. Inaba, and H. Inoue, "Teaching by showing: Generating robot programs by visual observation of human performance," *Proc. 20th Int'l Symp. on Industrial Robots*, 1989, pp. 119-126.

[12] J.M.F. Landsmeer, "Power grip and precision handling," *Ann. Rheum. Dis.*, Vol. 21, 1962, pp. 164-170.

[13] T. Lozano-Perez, "Automatic planning of manipulator transfer movements," *IEEE Trans. on Systems, Man and Cybernetics*, SMC-11(10), 1981, pp. 681-689.

[14] M.D. Wheeler, and K. Ikeuchi, *Towards a Vision Algorithm Compiler for recognition of partially occluded 3-D objects*, Tech. Rep. CMU-CS-92-185, Carnegie Mellon University, Nov. 1992.