Grasping Real Objects Using Virtual Images *

Rahul Singh Richard M. Voyles David Littau Nikolaos P. Papanikolopoulos

Artificial Intelligence, Robotics, and Vision Laboratory
Department of Computer Science and Engineering
University of Minnesota, Minneapolis, MN 55455

Abstract

This paper describes a fundamentally new approach to vision-based control in robotics. In particular, the problem of grasping different objects using a robot with a singe camera mounted on the end-effector (an eye-in-hand system) is considered. In the proposed approach the recognition of the object, and thereby the identification of its grasp position, the subsequent translational and rotational pose alignment of the manipulator with the object, and its movement in depth are controlled by using image morphing. We use a model-based framework where the image of each object at a graspable pose is stored in a database. Given an unknown object in the workspace, its identity is established by morphing its contours to the shapes in the database and using a quantification of the morph as a dissimilarity measure. From the morph a sequence of virtual (synthesized) images are obtained, which describe the progressive transformation of the input (both in terms of its shape and pose) to the template. These images are used as sub goals to guide the eye-in-hand robotic system to attain the desired orientation and height from which the grasp can be executed.

1 Introduction

Sensing and recognition may be considered to be two primary prerequisites for meaningful interaction of an agent, either cognitive or cybernetic, with the real world. A typical example of such an interaction involving a cybernetic agent is robotic grasping. Of the many different sensor modalities in use, the choice of vision as a sensor has many significant advantages; vision is intuitive, it is a passive non-contact sensor modality, and cameras support a high information bandwidth and can provide a large field of view. Moreover, many well developed theoretical and implementational models are available for controllability of vision. It has also been observed [9, 12], that the incorporation of vision in a grasping methodology often leads to increased robustness.

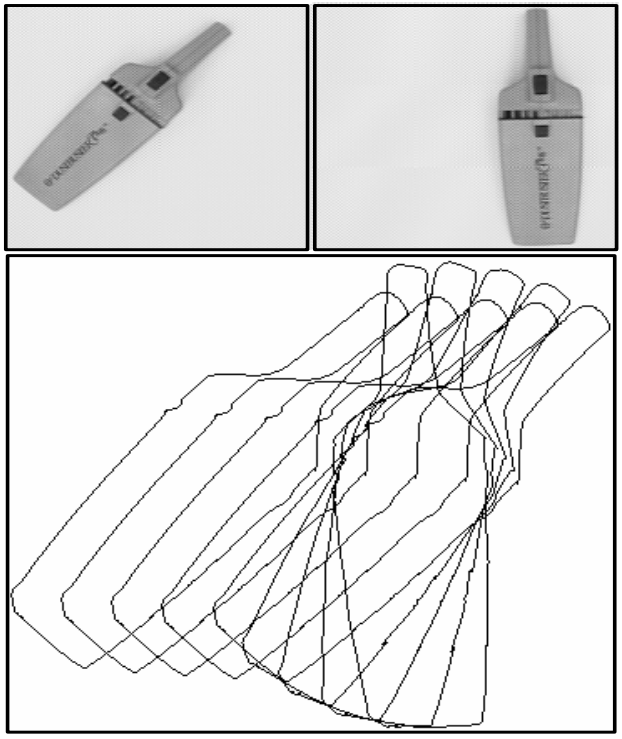

Figure 1: **Starting state (top left), goal state (top right), and the trajectory (bottom), as represented by successive synthetic images, in the morph plane.**

In a realistic setting, the grasp location as well as the alignment position from which a grasp can occur, vary with each object. Given an object in the workspace, grasping it involves solving the following sub-problems:

*This research was partially supported by the National Science Foundation through Grants #IRI-9410003 and #IRI-9502245 and the Defense Advanced Research Projects Agency, Electronics Technology Office, *Distributed Robotics Program*, ARPA order No. G155, program code 8H20, issued by DARPA/CMD under contract # MDA972-98-C-0008.

- Identification of the grasp points for the object.
- Trajectory generation for the manipulator from its current position to the position from which the grasp can occur. This step includes the following issues:
    1. Translational and rotational positioning of the end-effector with respect to the object.
    2. Movement in depth to execute the grasp.
- Calibration between the image plane and the world.

In this paper we consider the problem of vision-directed grasping of objects based on a conceptually new framework involving image morphing. The proposed approach allows us to treat issues related to recognition, trajectory generation and on-line calibration in a unified manner. We begin with a operational division, similar to [6], of the positioning and grasping problem:

1. *Off line learning*: The eye-in-hand system is moved to the alignment position with respect to an object. The alignment is such that the gripper is directly above the grasp position of the object. The view of the object is captured and stored in an image database.
2. *Visual servoing*: After the camera and/or the target has moved, the camera motion is controlled so that the visual error between the current view of the object and its reference view (obtained in the previous step) is reduced to zero.

The first step is used to create a database of images where each image represents the desired view of the object and specifies the pose which the manipulator needs to attain. The alignment and grasp positions can be arbitrarily defined and are neither limited by assumptions on object shape, nor depend on hand-crafted rules for their definition.

During the second step, a given unknown input is compared at run time with the images acquired and stored in the database. The comparison is based on using a quantification of the non-rigid deformations of the input object contours to the models stored in the database by using image morphing. Along with the recognition result the morph provides a sequence of *virtual images* describing the progressive transformation of the input (both in terms of its shape and pose) to the template (see Figure 1). These images are used as sub-goals to guide an eye-in-hand robotic system to attain a desired pose and grasp arbitrary objects in its workspace.

We begin this paper with an overview of the prior research in the area of vision-based grasping systems (Section 2). This is followed by the formulation of the proposed method in Section 3. Experimental results are presented in Section 4 and the work is summarized and some possible directions of future work are outlined in Section 5.

2 Issues and Prior Research

Depending on whether the control error function is computed in the Cartesian space or in the image plane, most vision-based positioning and grasping systems can be classified as *position-based* or *image-based* respectively [5, 6]. The basic idea of image-based robotic visual servoing lies in defining a visual error between the current and the desired positions of the manipulator in image coordinates [11], such that zero error implies that the desired end-effector position is reached, regardless of the camera position [3]. In contrast position-based methods allow the direct specification of the desired relative trajectories in the end-effector Cartesian coordinate frame.

The problem of robotic visual servoing around a static object has been a subject of active research. A model reference adaptive control scheme was proposed in [10]. A pre-computed Jacobian was used in conjunction with a an adaptive control law to control the motion of a robot-camera system to attain a certain pose with respect to a static object in [1]. In [4], a neural-network based approach was proposed to learn the inverse perspective transformation based on feature points. It may be noted that the above mentioned approaches were all tested in simulations. A real robot (Direct Drive Arm II) was used in [7], where the *controlled active vision* framework was used for positioning the robot. An extension of this approach for grasping static and moving objects can be found in [9]. Based on geometric-cues obtained from the environment, a calibration free approach for manipulating a randomly placed part is formulated in [2]. In this work calibration is performed as a part of the estimation, planning and tracking process. In [3] results from projective geometry are used to design a calibration free approach. A pure position based approach is presented in [12]. A hybrid approach is proposed in [6], where the control error function is computed partly in the 2D image plane and partly in the 3D Cartesian space.

3 The Proposed Method

3.1 Recognition and Contour Morphing

Encoding shapes and comparing them using deformations is a possible approach to the recognition problem. We propose a strategy for shape recognition using a framework based on image morphing. Each shape is represented by its contour. Conceptually, recognition is achieved by quantifying and using as a dissimilarity measure, the morph of an unknown input shape to models stored in a database. The morph is defined in terms of the stretching and bending of the contours by using a physics-based model and describes the minimum amount of deformation required to transform one shape into another. Using a quantification of the morph as a dissimilarity measure emphasizes the intuitive fact that two shapes that are similar do not have to go through an extensive metamorphosis in order for one to assume the form of the other.

Let $\mathbf{S}^I = [S_0^I, \ldots, S_n^I]$ and $\mathbf{S}^T = [S_0^T, \ldots, S_n^T]$ be the segmentation points of the input and the target shape contours. Contour metamorphosis of $\mathbf{S}^I$ to $\mathbf{S}^T$ is defined by a sequence of intermediate shapes obtained by a linear interpolation between $\mathbf{S}^I$ and $\mathbf{S}^T$. The intermediate images can be expressed as:

$$\begin{aligned} \mathbf{S}(t) &= u\mathbf{S}^I + t\mathbf{S}^T \\ &= [uS_0^I + tS_0^T, uS_1^I + tS_1^T, \ldots, uS_n^I + tS_n^T] \\ &= [S_0(t), S_1(t), \ldots, S_n(t)] \end{aligned} \quad (1)$$

where $u = 1 - t$. $S_i(t)$ is the ith segmentation point of the intermediate shape, formed at time t. The time parameter t is normalized to the interval $[0, 1]$.

The contour morphing process can be used as the basis of a method to compare the similarity of contours. It may be noted that the morph as described by Eq. (1), requires a correspondence of the segmentation points of the two contours. By defining the optimal correspondence between two contours to be the one that involves minimal deformation, the shape recognition problem can be reduced to solving the following two sub-problems:

- Definition of a quantification of the deformation.
- Algorithmic establishment of the optimal point correspondence between any two shapes, by minimizing the above quantification.

Based on the physics-based model [8], we quantify the deformations using the energy spent in stretching and bending the input contour to the target. The stretching energy is computed for every segment (pair of points) and is defined as:

$$E_s = k_s \frac{(L_T - L_O)^2 - (L_I - L_O)^2}{(1 - c_s)L_{min} + c_s L_{max}} \quad (2)$$

where

$$\begin{aligned} L_{min} &= min(L_O, \ldots, L_I, L_T) \\ L_{max} &= max(L_O, \ldots, L_I, L_T). \end{aligned}$$

In Eq. (2) E_s denotes the stretching energy spent in the current deformation, L_O, L_I, and L_T denote the segment lengths at the beginning, before the current deformation, and after the current deformation, respectively. The term c_s corresponds to the penalty for segments collapsing to points and k_s is the stretching stiffness parameter. The bending energy E_b is computed for point triplets and denotes the cost of angular deformation.

$$E_b = \mid k_b[(\phi_T - \phi_O)^2 - (\phi_I - \phi_O)^2] \mid \quad (3)$$

where k_b indicates bending stiffness, ϕ_O represents the original angle, and ϕ_I and ϕ_T denote respectively, the angles before and after the current deformation. By constraining the deformations at the segmentation points, the following optimal substructure property may be observed: The optimum cost of the point correspondence (S_i^I, S_j^T) equals the optimum cost of the previous point correspondence (S_{i-1}^I, S_j^T) or (S_{i-1}^I, S_{j-1}^T) or (S_i^I, S_{j-1}^T) and the cost of establishing the correspondence (S_i^I, S_j^T). Based on the above, an efficient $(O(mn))$ dynamic programming scheme can be constructed for obtaining the optimal correspondence. Two issues need to be pointed out here. First, the optimal correspondence as described above is dependent on the starting correspondence. This dependence can be removed by minimizing the energy required for deformation over all starting correspondences. Second, the point correspondence is invariant to Euclidean transformations and can be used to recover the translation and the rotation between the input and the target. Using this information, the pose error between successive virtual images and the target can be reduced.

3.2 Alignment and Grasping using Virtual Images

Let $\{W\}$ denote the world frame, $\{R\}$ the robot end-effector frame, and $\{C\}$ the camera tool frame. Let also the morph plane be denoted by $\{I\}$ and the

object frame by $\{O\}$. The relations between these coordinate frames can be described by the following homogeneous transformations.

- ${}^{W}_{R}T_i$: Describes the pose of the end-effector in the world frame at time instant i. This transformation is known at any time instant.
- ${}^{R}_{C}T$: The constant transformation between the robot end-effector and the camera (unknown).
- ${}^{C}_{O}T_i$: Denotes the object in the camera frame and is unknown.
- ${}^{C}_{I}T$: The image projection transformation, which is assumed to be known. In the equations below, we use ${}^{I}_{C}T$ which is the inverse of the trasnformation ${}^{C}_{I}T$
- ${}^{I}_{O}T$: The transformation describing the object in the image plane and is assumed to be known.
- ${}^{W}_{O}T$: The transformation between the object frame and the world frame which is unknown.

A graphical depiction of these transforms and their interrelationship is shown in Figure 2. The series of transformations may be arranged in two loops, denoted as A and B respectively. These loops are coupled mathematically by the transform ${}^{C}_{O}T_i$. The object frame can be related to the world frame and the image frame based on the following relations:

$$ {}^{W}_{O}T = {}^{W}_{R}T\ {}^{R}_{C}T\ {}^{C}_{O}T \quad (4) $$

$$ {}^{I}_{O}T = {}^{I}_{C}T\ {}^{C}_{O}T \quad (5) $$

Denoting the initial and the goal transformation between the object and the image frames by ${}^{I}_{O}T_0$ and ${}^{I}_{O}T_{goal}$ respectively, we have from loop B of Figure 2

$$ {}^{I}_{O}T_0 = {}^{I}_{C}T\ {}^{C}_{O}T_0 \quad (6) $$

Let the transformation from the initial state to the goal state in the morph plane be denoted as ${}^{I}_{M}T_{goal}$. Letting the intermediate transformation between the camera and the object to be ${}^{C}_{O}T_i$, we obtain the following relationships

$$ {}^{I}_{O}T_{goal} = {}^{I}_{M}T^{-1}_{goal}\ {}^{I}_{O}T_0 = {}^{I}_{C}T\ {}^{C}_{O}T_i \quad (7) $$

Replacing from Eq. (5), the value of ${}^{I}_{O}T_0$ in Eq. (7) and solving for ${}^{C}_{O}T_i$, we get

$$ {}^{C}_{O}T_i = {}^{I}_{C}T^{-1}\ {}^{I}_{M}T^{-1}_{goal}\ {}^{I}_{C}T\ {}^{C}_{O}T_0 \quad (8) $$

But from Eq. (4), we have for ${}^{C}_{O}T_0$

$$ {}^{C}_{O}T_0 = {}^{R}_{C}T^{-1}\ {}^{W}_{R}T^{-1}_{0}\ {}^{W}_{O}T_0 \quad (9) $$

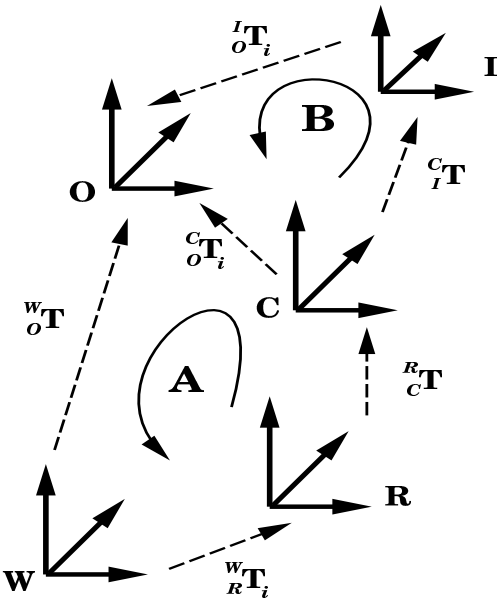

Figure 2: **Various coordinate frames and their relationships.**

Replacing this value in the previous equation, we get

$$ {}^{C}_{O}T_i = {}^{I}_{C}T^{-1}\ {}^{I}_{M}T^{-1}_{goal}\ {}^{I}_{C}T\ {}^{R}_{C}T^{-1}\ {}^{W}_{R}T^{-1}_{0}\ {}^{W}_{O}T_0 \quad (10) $$

Let us now consider the relationship described in loop A of Figure 2 and expressed by Eq. (9). Let ${}^{R}T_{i_d}$ be the desired motion of the robot at the intermediate stage i. We then have

$$ {}^{W}_{R}T_i = {}^{W}_{R}T_0\ {}^{R}T_{i_d} \quad (11) $$

Replacing in Eq. (9), the value of ${}^{W}_{R}T^{-1}_{i}$, from Eq. (11), we obtain the following relationship for ${}^{C}_{O}T_i$

$$ {}^{C}_{O}T_i = {}^{R}_{C}T^{-1}\ {}^{R}T^{-1}_{i_d}\ {}^{W}_{R}T^{-1}_{0}\ {}^{W}_{O}T_0 \quad (12) $$

From the above equation, we obtain the following expression for ${}^{R}T^{-1}_{i_d}$:

$$ {}^{R}T^{-1}_{i_d} = {}^{R}_{C}T\ {}^{C}_{O}T_i\ {}^{W}_{O}T^{-1}\ {}^{W}_{R}T_0 \quad (13) $$

Using the value of ${}^{C}_{O}T_i$ from Eq. (10), replacing it in the above equation and simplifying, we finally have

$$ {}^{R}T_{i_d} = {}^{R}_{C}T\ {}^{I}_{C}T^{-1}\ {}^{I}_{M}T_{goal}\ {}^{I}_{C}T\ {}^{R}_{C}T^{-1} \quad (14) $$

If the calibration transform ${}^{R}_{C}T$ was known, Eq. (14) could be used to align the eye-in-hand system at the desired position with the object. However, we have assumed no knowledge of this calibration matrix. Instead, we estimate it on-line using the intermediate images that result from the morph sequence. We start

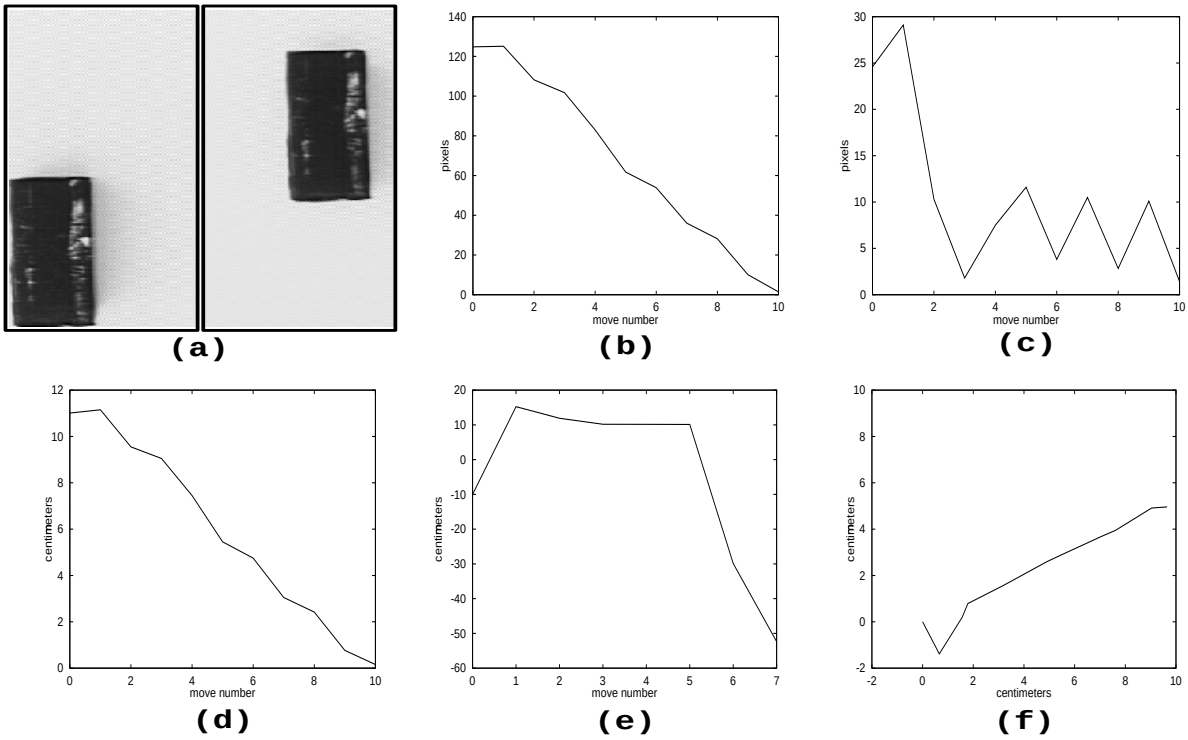

Figure 3: **Results of translational pose alignment and grasping of a prism: (a) view at start (left) and desired view (right), (b) error in the morph plane from the final position, (c) error in the morph plane with respect to each virtual image used as a sub-goal, (d) error in the world coordinates, (e) movement in depth, (f) trajectory during the alignment.**

with an arbitrary estimation of ${}^{R}_{C}T$ and use each intermediate virtual image as a sub-goal. The deviation between the actual and the desired trajectory is then computed by morphing the actual view of the object to the current virtual image being used as a sub-goal. The error transformation thus obtained is used to update (with possible averaging), the estimation of ${}^{R}_{C}T$.

The motion of the manipulator in depth to execute a grasp is initiated after the positional alignment is complete. The amount of movement is controlled by the apparent size of the object in the image plane. It may be noted that at their graspable height, most objects occupy the entire image and render contour extraction impossible. We circumvent this difficulty as follows: during the off-line learning, the view of the object is captured at a known height above it, from where its contours are visible. The image of this object is then annotated with the height information and stored in the database. Starting from an arbitrary height (such that the object contours are visible), the manipulator is first aligned with the object in terms of translation and rotation and then it is moved in depth to bring it to the height at which the image had been stored. From this point, the distance to the grasp height, annotated to the image, is used to execute the grasp. During the recognition and pose alignment phase, the size of the images are normalized. This normalization step is omitted during the movement in depth.

4 Experimental Results

The method has been used for positioning a PUMA 560 manipulator with respect to planar objects as well as positioning and grasping 3D objects using their planar projections. For each object, the alignment and grasp position (if applicable), was stored as described in Section 1. During the on-line visual servoing phase, the unknown object in the work space was identified by morphing it to the images in the database. The images generated by the morph sequence were successively used to guide the manipulator. After each move, the image of the object obtained from the current position was morphed to the virtual image serving as the current sub-goal. The pose difference between the two was used to update the estimated calibration matrix and move the manipulator to a new position. When the pose difference between the actual image and the

virtual images became less than a threshold (empirically defined to be 0.10 radians for rotation and within five pixels in **X** and **Y** for translation), the next virtual image from the morph sequence was used. The process terminated when all the images generated in the morph had been utilized. The movement in depth was controlled such that the size of the object (as defined by the number of pixels inside the contour) was within 3% of the template. From this point, the remaining movement in depth was done using the depth information collected in the learning stage.

Results for pose alignment and grasping of a rectangular prism are presented in Figure 3. Figure 4 contains the graphs for translational and rotational positioning with respect to the vacuum cleaner shown in Figure 1.

5 Conclusions and Future Work

We have presented the theory underlying the control of vision-based robotic tasks involving hand-eye coordination using virtual images created by image morphing. The basic idea of the approach is to map the virtual changes in the morph plane to the real motion of the manipulator. The proposed approach allows the specification of different grasp positions for different objects and does not involve manual feature selection and correspondence. Furthermore smooth trajectories can be obtained by generating arbitrary number of intermediate images. The idea explored in this paper may provide a fundamentally new approach to many problems in vision-based robotics. Foremost amongst its possible extensions are 3-D grasping in a full 6-DOF formulation and its application in a programming by human demonstration framework.

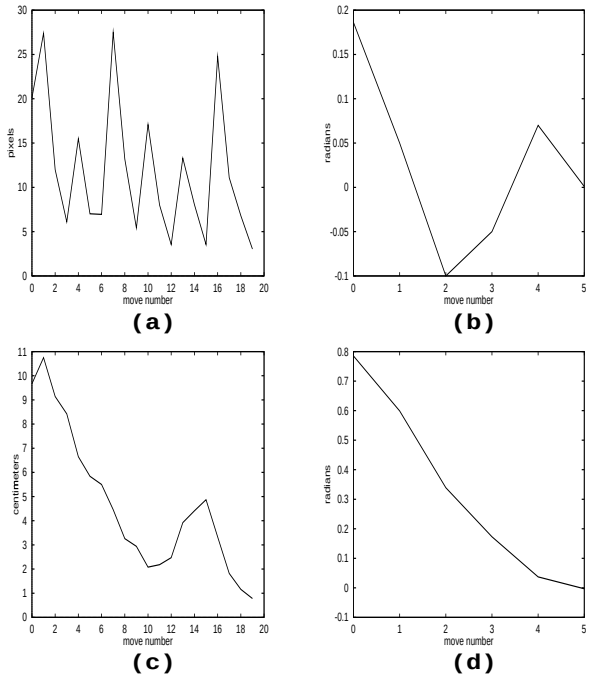

Figure 4: Translational and rotational pose alignment with the vacuum cleaner: (a), (b) translational and rotational error in the morph plane with respect to each virtual sub-goal, (c) translational error in the world coordinates, (d) rotational error in the world coordinates.

References

[1] F. Chaumette, P. Rives, and B. Espiau. "Positioning Of A Robot With Respect To An Object, Tracking It And Estimating Its Velocity By Visual Servoing". In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2248–2253, 1991.

[2] B. K. Ghosh, D. Xiao, N. Xi, and T. J. Tarn. "Calibration-Free Vision-Based Control of a Robotic Manipulator". In *Proceedings of the Thirty-Fourth Annual Allerton Conference on Communication, Control and Computing*, pages 593–602, 1996.

[3] G. D. Hager. "Calibration-Free Visual Control Using Projective Invariance". In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1009–1015, 1995.

[4] K. Hashimoto, T. Ebine, and H. Kimura. "Dynamic Visual Feedback Control For A Hand-Eye Manipulator". In *Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1863–1868, 1992.

[5] S. Hutchinson, G. D. Hager, and P. I. Corke. "A Tutorial on Visual Servo Control". *IEEE Trans. Robotics and Automation.*, 12(5):651–670, 1996.

[6] E. Malis, F. Chaumette, and S. Boudet. "Positioning a Coarse-Calibrated Camera with Respect to an Unknown Object by 2D 1/2 Visual Servoing". In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 2, pages 1352–1359, 1998.

[7] N. Papanikolopoulos and P. Khosla. "Robotic Visual Servoing Around A Static Target: An Example Of Controlled Active Vision". In *Proceedings of the 1992 American Control Conference*, pages 1489–1494, 1992.

[8] T. W. Sederberg and E. Greenwood. "A Physically Based Approach to 2D Shape Blending". *Computer Graphics*, 26(2):25–34, 1992.

[9] C. E. Smith and N. P. Papanikolopoulos. "Vision-Guided Robotic Grasping: Issues and Experiments". In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3203–3208, 1996.

[10] L. Weiss, A. Sanderson, and C. Neuman. "Dynamic Sensor-Based Control Of Robots With Visual Feedback". *IEEE Journal Of Robotics And Automation*, 3(5):404–417, 1987.

[11] L. E. Weiss. "Dynamic Visual Servo Control of Robots: An Adaptive Image-Based Approach". *Ph.D Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University*, 1984.

[12] W. J. Wilson, C. C. W. Hulls, and G. S. Bell. "Relative End-Effector Control Using Cartesian Position Based Visual Servoing". *IEEE Trans. Robotics and Automation.*, 12(5):684–696, 1996.