Experimental prediction of the performance of grasp tasks from visual features

Antonio Morales, Eris Chinellato
Robotic Intelligence Lab.
Universitat Jaume I
Castellon, E-12071
Spain
{morales, eris}@icc.uji.es

Andrew H. Fagg
Laboratory for Perceptual Robotics
University of Massachusetts
Amherst, Massachusetts 01003
USA
fagg@cs.umass.edu

Angel P. del Pobil
Robotic Intelligence Lab.
Universitat Jaume I
Castellon, E-12071
Spain
pobil@ieee.org

***Abstract*— This paper deals with visually guided grasping of unmodeled objects for robots which exhibit an adaptive behavior based on their previous experiences. Nine features are proposed to characterize three-finger grasps. They are computed from the object image and the kinematics of the hand. Real experiments on a humanoid robot with a Barrett hand are carried out to provide experimental data. This data is employed by a classification strategy, based on the k-nearest neighbour estimation rule, to predict the reliability of a grasp configuration in terms of five different performance classes. Prediction results suggest the methodology is adequate.**

I. Introduction

For a service robot to be truly autonomous in every-day human environments, it must be capable of performing a set of basic fundamental tasks in a robust and adaptive way, so that more complex behaviors can be built on top of them. This paper addresses one of such fundamental tasks; namely, vision-based grasping of unmodeled objects.

Our emphasis has been on exploiting the use of on-line visual sensing, in contrast to previous approaches which typically assume that a parameterized model of the entire object is known before grasp planning begins. When such a model exists, there is a well-founded corpus of analytical methods for grasp analysis and synthesis that are computationally expensive but could be applied off-line [1]. Even if the model is not previously available, such analytical methods could be applied after an exhaustive reconstruction phase from the visual input. However, this is not only too costly in terms of sensing and computation time to be feasible for real-world applications, but also in disagreement with neurophysiological findings [11].

Our challenge is, then, to develop a strategy suitable to compute and execute a reliable grasp for unmodeled objects presented in an unstructured manner by using only visual sensing as input and keeping computation to a minimum in order to achieve real-time performance. Moreover, the robot should learn and adapt its behavior by using its previous experience. There are two key issues in this formulation: the first one is how to define the intrinsic features that can be computed from an object image and that would ideally be necessary and sufficient to completely characterize a particular grip. The second concerns a procedure to relate those features to the nature of a feasible grasp in order to predict the possible outcome of its execution.

In this paper we propose such a set of features that characterize a three-finger grasp (Sec. II). We devise a strategy within this feature space so as to predict the success of a particular grasp configuration (Sec. IV). These predictions are based on real data gathered from experiments conducted on a humanoid robot hand (Sec. III-A). Our results (Sec. V) suggest that our methodology is good enough to predict the reliability of a grasp within a reasonable error margin. An important additional contribution is the fact that we use both features that are dependent on the object image and also on the particular kinematic configuration of the hand, whereas previous work described in the literature ignores the hand geometry by considering only the object geometry.

II. Vision-Based Grasp Features

In the particular case of planar grasp determination, i.e. for objects resulting from the extrusion of a planar shape, we showed in [8], [9], how vision information can be used to select a set of feasible grips that meet certain stability criteria, including the particular kinematics of the three-fingered Barrett hand [4]. This approach typically yields a large set of triplets of contact points, out of the infinite geometric possibilities. However, only one grip can be finally executed, and this choice can be mediated by further considerations such as the particular robot intentions, the task to be performed, additional reachability constraints, etc. An adequate characterization of the grips is called for in order to be able to predict their reliability and adjust practical aspects of the manipulation activity (e.g.: arm accelerations torques, etc) accordingly during the execution of the grasp and subsequent movements. In this section we first describe some basic descriptors (see [8], [9] for more details) and then introduce the nine features used to characterize a grasp configuration. Fig. 1 shows a schematic representation of the kinematics of the Barrett hand.

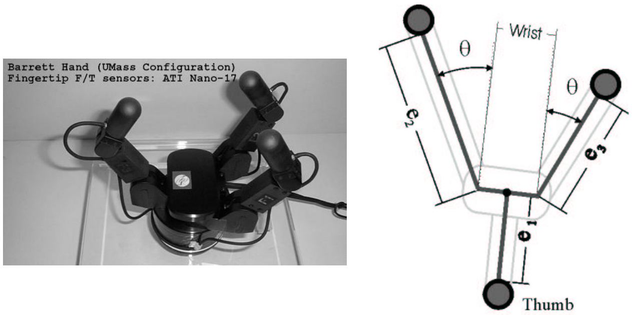
Barrett Hand (UMass Configuration)
Fingertip F/T sensors: ATI Nano-17
θ
Wrist
θ
e2
e3
e1
Thumb

Fig. 1. Barrett hand kinematics. The hand has a thumb and two opposing fingers that spread symmetrically along the axis defined by the thumb.

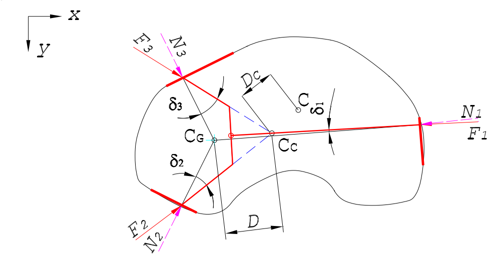

Fig. 2. Geometrical representation of the variables involved in the computation of features 1 (δ_1, δ_2, and δ_3); feature 2 (D, C_C, and C_G); and feature 5 (D_C).

A. Grasp descriptors

- **Grasp regions.** The portions of the object contour where the three fingers are placed. They are modeled as short straight segments and described by the coordinates of their extreme points.
- **Contact points.** The three points where the fingers are supposed to touch the object, each lying on one of the three grasp regions (P_1, P_2, P_3).
- **Force directions.** The real force directions F_1, F_2, F_3 exerted by the fingers of the Barrett hand are usually different from the ideal normal directions N_1, N_2, N_3.
- **Force focus.** The intersection of the directions of the real forces C_C.
- **Finger extensions.** The opening of the fingers (e_i in Fig. 1 and 3).
- **Spread angle.** The spread angle (θ in Fig. 1) of the opposing fingers.

B. Feature definitions

The previous descriptors are somehow low level. Now, we propose nine high-level features computed from the grip descriptors that try to measure different properties of each grip. Note that the inputs for their computation come only from the object contour extracted from the image along with the knowledge about the hand geometry.

All features have been designed to have similar ranges. More precisely, they are defined so as the best grips correspond with the lower values, with a theoretical best value of 0. Also, a preprocessing stage is performed based on physical and numerical considerations. This consists in a normalization dependent on the distributions and ranges of each feature, so that a middle quality grip for a certain feature is expected to have a quality value of 1. More details can be found in [2].

All variables used in the different features are indicated in Fig. 2, 3 and 4.

1. FORCE LINE: This feature [9] considers the deviations δ_i of the real forces F_i from the ideal condition of being perpendicular to the contour at the grasp points. Low deviations indicate low risk of instability: $q_1 = \frac{4}{3}(\delta_1^2 + \delta_2^2 + \delta_3^2)/arctan^2(\mu)$. μ is an estimation of the friction coefficient.

2. REAL FOCUS DEVIATION: This feature measures the distance D between the focus of the ideal forces C_G and the real focus of the grip C_C. The feature is computed as $q_2 = \frac{2D}{\eta\mu}$ where η is the maximum possible finger extension.

3. FINGER EXTENSION: If the fingers contact the object with different extensions, they probably exert a torque out of the horizontal plane of the object. This feature estimates the risk given by the differences in the finger extensions: $q_3 = \frac{1}{\eta 2}((e_1 - e_2)^2 + (e_2 - e_3)^2 + (e_3 - e_1)^2)$.

4. FINGER SPREAD: An equilibrated grip should have its three forces roughly equally separated by 120^o angles [10]. This feature measures the equilibrium of the grasps. $q_4 = (\frac{\pi}{6}/(\frac{\pi}{2} - \theta)) - 1$ for $\theta > \frac{\pi}{3}$ or else 0, where θ is the opening angle of the fingers of the Barrett hand in opposition to the thumb.

5. REAL FOCUS CENTERING: This feature aims to measure the effect of gravitational and inertial forces endorsing grasps with short distances between the real focus C_C and the center of mass of the object C. The feature definition is $q_5 = \frac{4D_C}{M_L + m_L}$, where M_L and m_L are the sizes of the major and minor inertia axes computed from the shape.

6. FINGER LIMIT: When trying to grip large objects, there is a limit in the extension of the fingers . Due to the way the Barrett Hand grips objects, there is a finger extension value that, if overcome, causes the grip to shift from a fingertip grip to a fingerside grip on the part edge, which is more risky and less stable although still possible (see Fig. 3). Therefore, a threshold on the maximum optimal finger extension ϵ has been set in order to avoid marginal contacts: $q_6 = \epsilon_1 + \epsilon_2 + \epsilon_3$ where $\epsilon_i = (\frac{e_i - \eta}{\lambda})^2$ if $e_i > \eta$, else 0. The threshold λ is an estimation of the positioning error.

7. POINT ARRANGEMENT: Similarly to [5], [10], we assess the likeness of the grasp triangle to an equilateral one to obtain better grip balance. Each angle is compared with

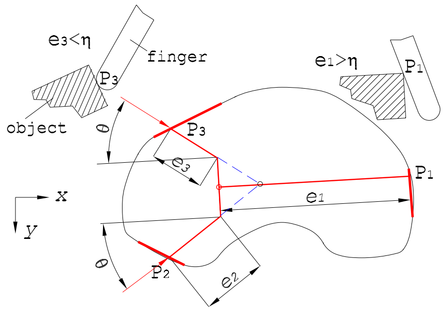

Fig. 3. Geometrical representation of the variables involved in the computation of features 3 (e_1, e_2, and e_3); feature 4 (θ); and feature 6 (η).

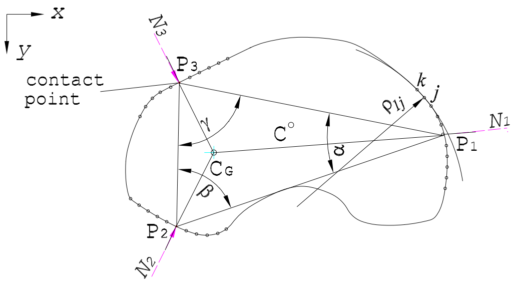

Fig. 4. Geometrical representation of the variables involved in the computation of features 7 (α, β and γ); 8; and 9 (ρ_{ij}).

a 60^o ($\pi/3$ rad) angle typical of an equilateral triangle: $q_7 = \frac{3}{2\pi}(|\alpha - \frac{\pi}{3}| + |\beta - \frac{\pi}{3}| + |\gamma - \frac{\pi}{3}|)$.
8. TRIANGLE SIZE: The larger the area of the grasp triangle, the more stable a grip is [5]. The quality measure is $q_8 = \frac{A}{4A_{S2}}$, where A_{S2} is the area of the grasp triangle, and A is the area of the object.
9. CONTACT CURVATURE: A concave surface is a better place to put a finger for grasping purposes than a convex one [7]. This feature takes into account the curvature of the three grasp zones. All the points closer to the grasp point than the positioning error threshold are considered, and their local curvature values are summed. The sum is weighted by the actual distance of each point from the contact point, in a way that the more we approach the expected point of contact the more the local curvature value becomes influent on the total. The curvature ρ is positive for concavities, negative for convexities and 0 for planar zones. We define the overall grip quality as: $q_9 = 3 * \psi - (\rho_1 + \rho_2 + \rho_3)$, where ψ is the curvature threshold value, that is the best (most concave) possible curvature allowed for a contact point. $\rho_i = \sum_{j=-k}^{k} (1 - \frac{|j|}{k}) * \rho_{ij}$, with ρ_{ij} local curvature of a point that is at distance j along the contour from the point i. In practice the distance in measured in discrete steps. The maximum distance k depends on the positioning error λ.

III. METHODOLOGY

In this section we describe the system setup, and the protocol followed to gather the experimental data.

A. Experimental Setup: the UMass Torso

Our experiments have been implemented using the UMass Torso. This humanoid robot (Fig. 5) consists of two Whole Arm Manipulators from Barrett Technologies, two Barrett hands with tactile sensors and a BiSight stereo head.

The stereo vision system estimates the two-dimensional location of the target object on the table, and provides a monocular image for surface curvature analysis (see [9] for more details). Once a grip is selected (consisting of contact locations and a hand posture), the hand is preshaped and positioned above the object. It moves down, closes the fingers so that the object is grasped, lifted and transported to a designated location.

B. Experimental protocol

A set of real objects has been built for this experiment. They are planar objects with a constant height made of an homogeneous material. Moreover, the colors of the objects have been selected to simplify the image processing. An important feature is that their shape is *unknown* for the system. The only programmed assumptions about the objects is that they are planar. The rest of the information, in particular the shape and location, is obtained from the images.

Moreover, in order to study the grasping performances in different circumstances several characteristics of the environment are tested. These are the weight of the objects and the friction coefficient. Two qualitative categories for each of both conditions are distinguished: heavy and light objects, and high and low friction. The different weight is obtained with two different object sets similar in appearance, but made of different material. Different contact friction is achieved by using a latex fingertip to envelope the fingers.

In order to perform the experiments, a single object is placed on a table within the robot workspace. Using the stereo-visual information the robot locates the object and computes a set of feasible grasp configurations. One of the configurations is selected, either manually by a human operator, or automatically by the robot, and executed.

If the robot has been able to lift the object safely, a set of stability tests are applied in sequence. These are aimed at measuring the stability of the current grasp. They consist of three consecutive shaking movements of the hand which are executed with an increasing acceleration. After each

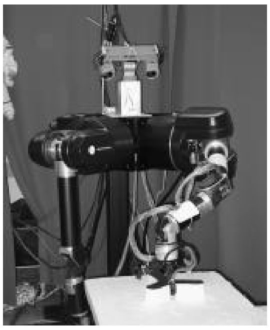

Fig. 5. The UMass Torso. A humanoid robotic system at the Laboratory for Perceptual Robotics in the University of Massachusetts.

movement the tactile sensors are used to check whether the object has been dropped off.

This protocol provides us with a qualitative measure of the success of a grasp. Thus, an experiment may result in five different reliability classes: *E* indicates that the system was not able of lifting the object at all; *D, C, B* indicate that the object was dropped, respectively, during the first, second, or third series of shaking movements; finally *A* means the object did not fall and was returned successfully to its initial position on the table. Hence, we define $\Omega = \{A, B, C, D, E\}$ as the set of reliability classes.

The number of feasible grips that are computed for a single object is usually large, varying from several dozens to more than one hundred. In addition a particular execution of a grasp configuration can be influenced by many unpredictable factors. To avoid this problem, each configuration is executed a sufficiently large number of times, by varying the location and orientation in the presentation of the object. In this way, statistically significant conclusions can be reached.

Nevertheless, this repetition could lead to a non practical number of executions, so for each object only a few configurations are selected to be executed. This selection consist of the most representative configurations of each object. Each configuration is executed 12 times, 4 times for three different orientations of the object.

IV. Prediction strategies

The data collected during the experiments comprises a large amount of information. Several analyses can be carried out over this data, specially those regarding the appropriateness and usefulness of the different features. Here, however, we are more interested in the predictive capabilities that can be inferred from these data and the methods that can make the best use of it.

In theoretical terms a data set is composed of N executed triplets. Each grip $g_i, i = 1 \ldots N$ is described by the nine visual features $q_1, \ldots q_9$ introduced in subsection II-B. The space Q_S is formed by the ranges of the values of the features. Moreover, we have also recorded the performance of the grip and have assigned it to a class $\omega_i \in \Omega$ for each g_i.

KNN classification rule

A prediction function has the form $F(g) = \bar{\omega}$ where $g \in Q_S$ and $\bar{\omega} \in \Omega$. There exists a wide bibliography on the building of such functions based on the Bayesian decision theory[3]. In this paper we have chosen the approach of the nonparametric techniques in particular the *voting k-nearest neighbor (KNN) rule* [6], [3] for modeling this function. The nonparametric techniques do not assume any density distribution of the features and the classes. To predict the class of a *query* point g_q, the KNN rule counts the K-nearest neighbors and choses the class that most often appears, the most voted.

In our implementation we have introduced some modifications to the basic schema. First we use the euclidean metric for measuring the distance between the points in the Q_S. We weighted the contribution of each of the KNN points according to its distance to the query point. This gives more importance to the closer points. The kernel function used is $K(d) = \frac{1}{1+(d/T)}$, where T is an adjustable parameter, and d is the distance.

We define $knn(g_q) = \{(g_i, \omega_i), i = 1 \ldots k, g_i \in Q_S, \omega_i in \Omega\}$ as the *k* closest points to g_q and d_i its corresponding distances from g_q. The probability corresponding to a class $\bar{\omega}$ are computed using this expression:

$$P(\bar{\omega}, g_q) = \sum_{\substack{g_i \in KNN(g_q) \\ \omega_i = \bar{\omega}}} \frac{K(d_i)}{\sum_{g_j \in KNN(g_q)} K(d_j)}$$

Function P is also an expression of the posterior probability [6]. Our predictor would be defined as $F(g_q) = \omega \in \Omega, MAX\{P(\omega, g_q)\}$.

Error and risk functions

Performance of classification methods is measured in terms of successful or wrong classifications. Our classes have an important particularity, their qualitative order (i.e.: class A means a higher stability for a grip than any other class). Having this in mind, we try not to penalize in the same amount when the failure is qualitatively smaller (i.e.: predicting B when the outcome is C), than larger (i.e.: predicting A when the outcome is D). For this we build the error function $E(\bar{\omega}, \omega)$,being $\bar{\omega}, \omega \in \Omega$, where $\bar{\omega}$ is the predicted outcome and ω the real one. This is easily implemented with a table (see practical cases in Table 2).

A step further is the definition of the *risk function*: $R(\bar{\omega}, g_q) = \sum_{\omega \in \Omega} P(\bar{\omega}) E(\bar{\omega}, \omega)$, where $\bar{\omega} \in \Omega$. The class $\omega \in \Omega$ selected for the prediction is the one that minimizes the risk, $F(g_q) = \omega \in \Omega, MIN\{R(\omega, g_q)\}$. Using the risk function makes it possible to introduce in the prediction step the qualitative ordering of the problem classes.

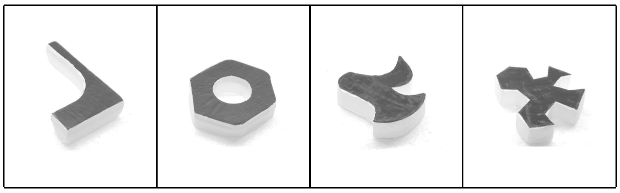

Fig. 6. The four objects used in the experiments

TABLE I

SAMPLE DATA SETS

	E	D	C	B	A	Total
Light	102	84	33	27	18	262
Low	38.6%	31.8%	12.5%	10.2%	6.8%	(22)
Light	51	97	56	38	118	360
High	14.2%	26.9%	15.6%	10.6%	32.8%	(34)
Heavy	95	92	29	2	2	220
High	43.1%	41.8%	13.2%	0.9%	0.9%	(23)

Sample distributions among classes for the different data sets. The figures in brackets in the "Total" column indicates the number of different configurations really tested.

TABLE II

ERROR TABLES

Criterion 2a

	E	D	C	B	A
E	0.0	0.5	1.0	1.0	1.0
D	1.0	0.0	0.5	1.0	1.0
C	1.0	1.0	0.0	0.5	1.0
B	1.0	1.0	1.0	0.0	0.5
A	1.0	1.0	1.0	1.0	0.0

Criterion 2b

	E	D	C	B	A
E	0.0	0.0	1.0	1.0	1.0
D	0.0	0.0	1.0	1.0	1.0
C	1.0	1.0	0.0	0.0	0.0
B	1.0	1.0	0.0	0.0	0.0
A	1.0	1.0	0.0	0.0	0.0

Criterion 2c

	E	D	C	B	A
E	0.0	0.00	0.25	0.50	0.75
D	0.25	0.00	0.00	0.25	0.50
C	0.50	0.25	0.00	0.00	0.25
B	0.75	0.50	0.25	0.00	0.00
A	1.00	0.75	0.50	0.25	0.00

The rows indicate the predicted outcome, and the columns the real outcome. An error 1.0 indicates a failure, and 0.0 a successful prediction.

V. RESULTS AND DISCUSSION

A series of experiments where done following the experimental protocol described in section III-B. Three different combinations of physical properties were tested: light objects and low friction (light/low), heavy objects and high friction (heavy/high); and light objects and high friction (light/high). A set of four different objects were used (fig, 6). Table I shows the number of different grips executed for each case, and the percentages of grips that resulted in each class of Ω. Note that the total number of grips results from the repetition of a smaller number of configurations.

Two basic questions need to be answered about the prediction capabilities of the rule described in section IV: first, is it able to generalize across different objects, and second, did we have enough data to properly construct a function ? To answer these questions we have developed a cross-validation method named *leave-one-grasp-out validation* similar to the well known *leave-one-out validation* and *n-fold cross-validation* [3]. This consist of the following steps: 1) given the whole data set, remove all the points of a particular grasp configuration and use this subset as validation set; 2) use the remaining samples for predicting the outcomes of the validation set and compute the mean error; 3) repeat steps 1) and 2) for all configurations. The validation error will be the mean error of the iterations of step 2). The reason for removing all the points of a configuration from the data set is that all the points of a particular configuration are very close in the Q_S and the KNN rule would be affected by this points instead of using points of unrelated configurations, farther in Q_S.

Moreover, we are also interested in the sensitivity of the error with respect to the size of the data set. We can analyze it by modifying the second step. Instead of using the whole remaining data set, we chose randomly a set of given size. This introduces a random factor, and to reduce the effect of this randomness we repeat this step a sufficiently large number of times.

We have defined three error tables (see table II). The first one *2a* is quite strict. It considers as failure any wrong prediction. The only exception is that it considers half a failure a prediction one class lower that the real output. This is a kind of conservative rule. The second table, *2b*, is a way of reducing the classes to two super-classes: the first one composed of class A, B and C (reliable grasps), and the second, D and E (unreliable). Finally, the third table *2c* tries to penalize errors depending on the qualitative distance between the predicted outcome and the real one.

The different parmeters of the knn prediction rule, *K* and *T* for the kernel function, has been chosen using leave-one-out validation with the full datasets minimizing the errors.

Figure 7 shows the evolution of the prediction error for the *light-high* data set using the three error schemes. The first and most obvious observation that can be drawn from these figures is that the error is reduced as the size of the available data set increases. Moreover, the evolution of the error rates depending on the table error used seems to be equivalent, but with a different scale.

Finally Table III shows the error rates reached in the size sensitivity experiments with the different data sets.

From a practical point of view, when performing a strongly stochastic action like grasping an unmodeled real object with a robotic hand, an error between two neighbor classes can be considered acceptable, especially in the case of a false negative. Indeed, it means that the reliability of the grasp is only slightly better than the predicted one.

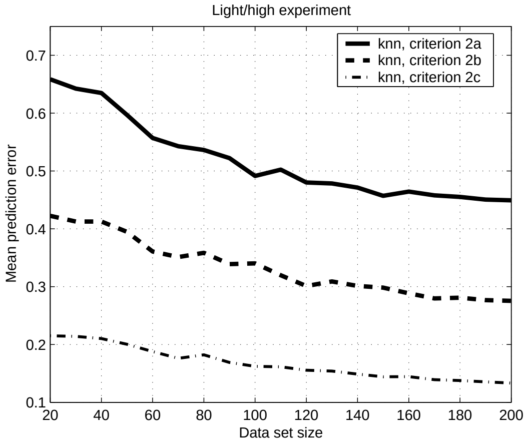

Fig. 7. Size sensitivity validation for the data set of light objects and high friction

This justifies the definition of criteria 2b and 2c. The results summarized in Table 3 suggest that the expected error rates will be around 0.25 (with error criterion 2b) or even close to 0.1 (with criterion 2c). It must be noted that these results have been obtained even though the available data were far from optimal. First they were very unequally distributed across the classes, with some classes poorly represented. More precisely, the low-quality classes D and E strongly prevail on the others. Second, they were very noisy due to uncontrollable errors in sensing, image processing and motor control.

VI. Conclusion

We have presented a contribution to a methodology for computing and executing reliable grasps for unmodeled objects using only visual sensing as input, in such a way that the system can exhibit an incremental adaptive behavior based on its previous experiences. We have proposed a set of intrinsic features that adequately characterize a grip and can be computed by using only the object image and the kinematics of the hand. We have implemented a prediction approach that uses such features to produce as output the reliability class of the grip. Feature space data were obtained from real experiments with a humanoid robot. The obtained prediction results are satisfactory enough to suggest that the methodology is adequate and further progress should be made in this direction.

VII. ACKNOWLEDGMENTS

This work has been funded in part by the Ministerio de Ciencia y Tecnología under project DPI2001-3801, by the Generalitat Valenciana under projects inf01-27, GV01-244, CTIDIA/2002/195, by the Fundació Caixa-Castelló under project P1-1B2001-28. The first and second author have been supported by the Ministerio de Ciencia y Tecnologia under a FPI Program graduate fellowship.

TABLE III

Error rates

	Light/Low	Light/High	Heavy/high
Criterion 2a	0.438	0.449	0.245
Criterion 2b	0.236	0.275	0.115
Criterion 2c	0.110	0.133	0.035

Error rates reached with the three data sets when the size of the data set is 200.

This work is also partly funded by grants CISE/CDA-9703217 and IRI-9704530, DARPA MARS DABT63-99-1-0004, and NASA/RICIS. The authors wish to thank Roderic Grupen, David Wheeler, Robert Platt and Danny Radhakrishnan, who have provided much of the foundation which allowed the robot experiments.

VIII. REFERENCES

[1] A. Bicchi and V. Kumar. Robotic grasping configuration and contact: A review. In *IEEE Intl Conf. on Robotics and Automation*, April 2000.

[2] E. Chinellato. Robust strategies for selecting vision-based planar grasps of unknown objects with a three-finger hand. Master's thesis, School of Artificial Intelligence. Division of Informatics. University of Edinburgh, 2002.

[3] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2nd edition, 2001.

[4] Barrett Technology Inc. http://www.barrett.com/.

[5] B. Mirtich and J. Canny. Easily computable optimun grasps in 2d and 3d. In *IEEE Intl Conf. on Robotics and Automation*, pages 739–747, May 1994.

[6] T. M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[7] D.J. Montana. The condition for contact grasp stability. In *IEEE Intl. Conf. on Robotics and Automation*, Sacramento, California, 1991.

[8] A. Morales, P.J. Sanz, and A.P. del Pobil. Heuristic vision-based computation of three-finger grasps on unknown planar objects. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pages 1693–1698, Lausanne, Switzerland, 2002.

[9] A. Morales, P.J. Sanz, A.P del Pobil, and A. H. Fagg. An experiment in constraining vision-based finger contact selection with gripper geometry. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pages 1711–1716, Lausanne, Switzerland, 2002.

[10] Y.C. Park and G.P. Starr. Grasp synthesis of polygonal objects using a three-fingered robot hand. *International Journal of Robotics Research*, 11(3):163–184, 1992.

[11] G. Rizzolatti and G. Luppino. The cortical motor system. *Neuron*, 31:889–901, 2001.