

The Robutler: a Vision-Controlled Hand-Arm System for Manipulating Bottles and Glasses

Ulrich Hillenbrand, Bernhard Brunner, Christoph Borst, and Gerd Hirzinger

Institute of Robotics and Mechatronics

German Aerospace Center (DLR), 82234 Wessling, Germany

Ulrich.Hillenbrand@dlr.de

Abstract— We present an experimental service robot, a hand-arm system with anthropomorphic features as well as some capability for autonomous behavior. The system integrates light-weight construction principles of the arm and an articulated dexterous four-finger hand with real-time scene analysis by stereo vision, compliant torque control, and an intuitive man-machine interface. It thus comprises many of the key features required to realize the vision of a robotic servant that acts and interacts in a human environment. The task we consider here is manipulation of bottles and glasses as required for preparing and serving drinks. The purpose of this article is to provide a comprehensive view of the whole system, revealing the synergies between its components that lead to real-world performance within the addressed problem domain.

I. INTRODUCTION

In the field of service robotics, systems are needed that can handle everyday objects in unconstrained environments, in much the same way they are handled by humans. Interaction with changing and, hence, only partially known environments requires intelligent sensor-data processing and advanced control architectures. Addressing this grand challenge for robotics, we present a hand-arm system with anthropomorphic features as well as some capability for autonomous behavior. The system integrates light-weight construction principles of the arm and an articulated dexterous four-finger hand with real-time scene analysis by stereo vision, compliant torque control, and an intuitive man-machine interface.

The world of the *Robutler* consists of various bottles and glasses that are freely arranged on a table. Its task is to prepare and offer drinks. Initially, an interpretation of the scene is computed within a few seconds. Object identities and poses are estimated by matching empirically learned models against 3D-point data that have been obtained from stereo processing. Different types of bottles and glasses are located and distinguished under varying conditions of lighting. Once the scene is understood, the *Robutler* autonomously executes a sequence of actions. For instance, it may unscrew the cap from a bottle, grasp the bottle, and pour a drink into a glass; see Figure 1.

A hierarchical command interface enables the user to instruct the *Robutler* at various levels of abstraction. In particular, the interface provides graphical representations of the objects as detected in the scene, displayed in a virtual-reality environment. It thus allows the user to issue high-level commands about actions to take w.r.t. the objects in an intuitive

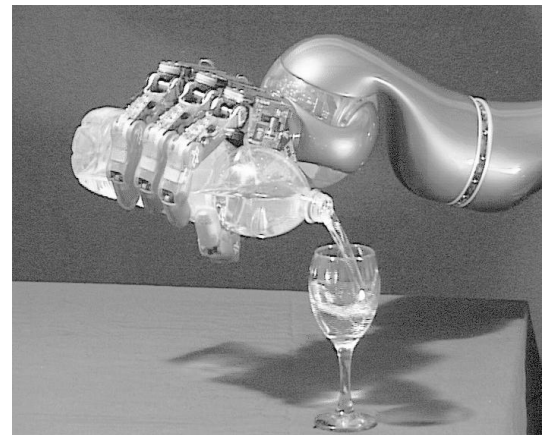


Fig. 1. The DLR Light-Weight Arm with Four-Finger Hand, pouring water into a glass.

manner. Alternatively, the *Robutler* can be instructed via a speech interface.

We consider the present experimental system a significant step towards designing service robots usable in human environments.

II. PREVIOUS WORK

The core idea of the German national research projects DIROKOL and MORPHA [1] was to provide advanced mechatronic systems, particularly service robots, with the capability to communicate, interact, and collaborate with human users in a natural and intuitive manner. This capability should enable a robot to cooperate with and assist the human user in a variety of tasks, under the user's guidance and control. The robot assistant in the home should work together with the user to perform simple housework. In addition to fetch-and-carry duties, this includes tasks such as setting the table or performing basic cleaning.

So far, robot systems have only been able to deal with the high complexity and variability of everyday surroundings to a limited extent. This complexity and variability poses great demands on the robot's intelligence and autonomy, demands that have not been fully satisfied by current technology. The capability to interact with a human user offers the robot system the possibility of making use of human guidance and support to expand its effective competence.

Accordingly, a lot of research activity has been invested in designing human-machine interfaces [2], [3]. Another focus of research has been on navigation capabilities, mainly but not exclusively in buildings [4], [5]. The present study is complementary to most previous work in that fine manipulation of objects from the human living environment is emphasized. The aim is to command a robot, in an intuitive and abstract manner, to autonomously complete a complex task like serving a drink. Such a task involves perception of a scene and execution of a sequence of delicate actions.

III. THE ARM

Service robots need to operate in the human environment or even cooperate with humans. Therefore, some basic requirements have to be met by the manipulators, making them quite distinct from industrial hardware. Service robotic applications require light-weight arms for safety reasons and human-friendly interaction, as well as to support mobility. Interaction with not fully known environments demands compliant arms and fingers, controlled by information extracted from many different sensors. To integrate multisensory components in arms and hands, sophisticated mechatronic concepts and a flexible control architecture are needed.

Over the last years, our focus on service robotics was driven by strong considerations, how to push robotic technologies for real-world applications. The design philosophy of DLR's light-weight robots [6] has been to realize a type of manipulator similar to the kinematic redundancy and dynamics of the human arm, i.e., with 7 degrees of freedom (DOFs), a load-to-weight ratio of better than 1:2, and a high dynamic performance; see Figure 2. All joints have position and torque sensors. Like in all modern approaches to robot control, commanding joint torques has been considered as essential, allowing programmable impedance, stiffness, and damping.

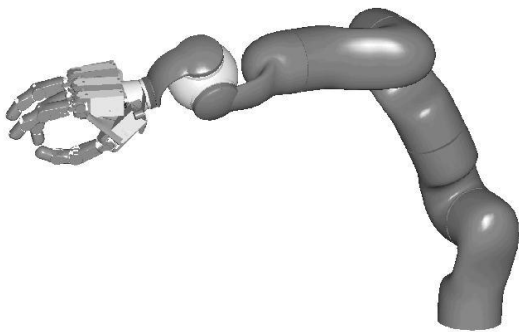


Fig. 2. The 3rd generation of the DLR Light-Weight Arm.

IV. THE HAND

For humans, hands and arms are universal actuators to interact with their environment. Grasping, carrying, and manipulating a wide variety of objects, using tools, catching and throwing things are basic abilities that we need them for. As

robots start to serve and collaborate with man, it is natural to equip robots with artificial hands.

The objects we want to manipulate, bottles and glasses, are challenging: glass can break and drinks can be spilled. A sensitive and compliant manipulator is therefore required.

The DLR hand features 4 fingers with anthropomorphic kinematics: 4 joints per finger with a collective 3 DOFs (distal joints coupled). An additional DOF resides in the palm for adaptation to power or precision grasps; see Figure 3. Joint control is accomplished with position, speed, and torque sensors for each finger DOF. High-level control can be driven by signals from a 6-DOF force-torque sensor in each finger tip.

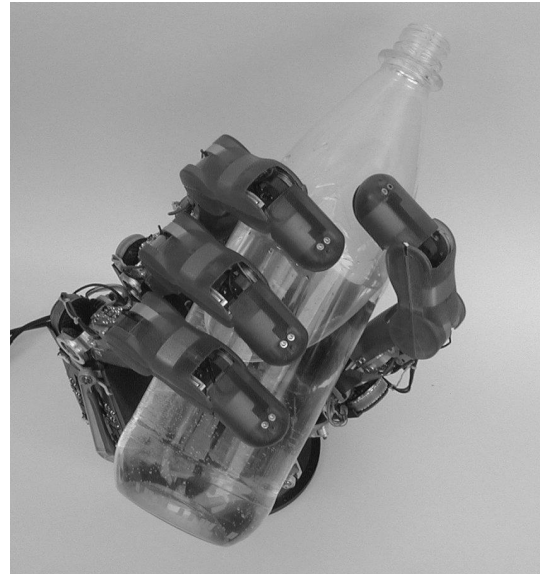


Fig. 3. DLR Hand II - power grasp.

In the case of precision grasps, usual finger design only allows for point contacts with the object at the finger tips. The fingers on the DLR hand, however, can bend backwards to create a much more robust line contact at the distal finger links (pinch grasp); see Figure 8.

V. THE VISUAL SYSTEM

The task of vision for a service robot is to deliver an interpretation of the scene that contains both geometric and semantic information. In other words, the *Robutler* has to know *where* objects are placed and *what* can be done with them. Clearly, the semantics of an object is not generally conceivable from vision alone. This kind of knowledge can presently be only attached to objects in a database. It is thus natural for a service robot to work with an a-priori known set of objects.

The objects we are dealing with, bottles and glasses, are challenging not only for manipulation, but also from the visual point of view. The reason is that visual data of transparent objects are generally more ambiguous than of opaque, colored or textured objects.

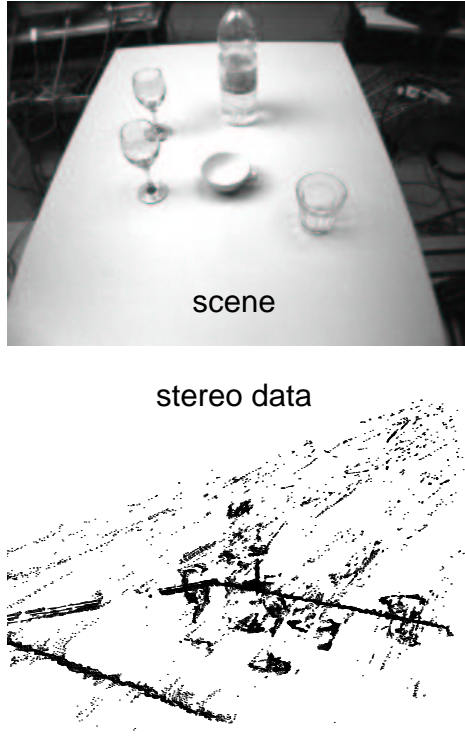


Fig. 4. Example scene and corresponding stereo data. The scene features a bottle, two wine glasses, a water glass, and a cup on a table, as seen in the top image taken by one of the cameras. The bottom image shows the 3D-data points obtained from stereo processing, viewed from the close left table corner. The objects and the table are mainly represented by contour points. Note the large amount of background data that arise as artifacts from stereo processing under such difficult conditions (transparent objects). This data set comprises 21,043 points.

A. Stereo Processing

We use a three-camera system (Digiclops, Point Grey Research Inc.) to perform stereo processing with a horizontal and a vertical stereo pair (baseline 10 cm each). Each image has a resolution of 640×480 pixels. The algorithm employed is a straightforward local correspondence search by minimizing the sum of absolute differences over square patches of the LoG-filtered (Laplacian-of-Gaussian filter) images. We have set a high threshold for valid regions of the LoG-filtered images, such that only regions of high scene contrast are taken into account. As a result, mainly surface creases, sharp bends, and depth discontinuities contribute to further processing. The output from stereo processing thus is a sparse representation of the scene by rather few 3D-data points (around 20,000), outlining the objects on the table; see Figure 4.

B. Object Recognition

The sparse scene representation that we obtain from stereo processing has two major advantages for the processing steps to follow: it keeps the computational load low and it allows for a reliable scene interpretation by simple methods adapted from fuzzy-set and probability theory; see [7] for details. It turns out that objects standing on a weakly textured plane,

like our table, are visually well described by a few 3D-data points restricted to high-contrast regions of the scene.

Let $\mu_i(\omega, d) \in [0, 1]$ be the probability for a data point to arise at position $d \in \mathbb{R}^3$ in data space from an object i with pose parameters ω . These functions can be empirically learned by sampling histograms of the stereo data produced by the objects of interest. In fuzzy-set theory, $\mu_i(\omega, d)$ would be called the ‘membership function for data points of object i with pose ω ’. Traditional fuzzy logic, however, is independent of any probabilistic interpretation of membership values.

The match of n object models $\{\mu_1(\omega, d), \mu_2(\omega, d), \dots, \mu_n(\omega, d)\}$ to N data points $\{d_1, d_2, \dots, d_N\}$, or, in fuzzy-set slang, the degree of membership of the data to the objects, is given by

$$\begin{aligned} M(\omega_1, \omega_2, \dots, \omega_n, d_1, d_2, \dots, d_N) \\ &:= \sum_{i=1}^n M_i(\omega_i, d_1, d_2, \dots, d_N) \\ &:= \sum_{i=1}^n \sum_{j=1}^N \mu_i(\omega_i, d_j), \end{aligned} \quad (1)$$

where $\omega_1, \omega_2, \dots, \omega_n$ are the objects’ pose parameters. A scene interpretation is obtained by optimizing the function $M(\omega_1, \omega_2, \dots, \omega_n, d_1, d_2, \dots, d_N)$ with respect to the pose parameters of the n objects. The pose parameters are constrained to take consistent values, that is, objects cannot intersect each other. Moreover, an object i is included in a scene interpretation, only if its contribution $M_i(\omega_i, d_1, d_2, \dots, d_N)$ to the best match exceeds some object-specific threshold.

For the case of our table-top scenes, the pose parameters ω_i of an object can take values in intervals $[a_1, a_2] \times [b_1, b_2] \times [0, 2\pi)$ for translation on the table and rotation around a table-orthogonal axis. Rotationally symmetric objects have poses only in $[a_1, a_2] \times [b_1, b_2]$. Search and optimization across all translations $(x, y) \in [a_1, a_2] \times [b_1, b_2]$ is efficiently implemented by the fuzzy extension of a generalized Hough transform with a single template $\mu_i(x, y, d)$ [8]. We have quantized the space of translations at intervals of 2 mm in both directions. The achieved translational accuracy of estimated object positions is usually within 2 quanta, i.e., 4 mm, up to distances of 1 m from the cameras.

For different rotational views $\phi \in [0, 2\pi)$ of object i , various templates $\mu_i(x, y, \phi, d)$ for discrete viewing angles $\phi \in \{\phi_1, \phi_2, \dots\}$ have to be employed. The orientation Φ of an object i at table-top position (x, y) can then be estimated by weighted averaging over the template orientations, that is,

$$\begin{aligned} \Phi = \arg \bigg[& e^{i\phi_{k-1}} \sum_{j=1}^N \mu_i(x, y, \phi_{k-1}, d_j) \\ & + e^{i\phi_k} \sum_{j=1}^N \mu_i(x, y, \phi_k, d_j) \\ & + e^{i\phi_{k+1}} \sum_{j=1}^N \mu_i(x, y, \phi_{k+1}, d_j) \bigg]. \end{aligned} \quad (2)$$

Here $\mu_i(x, y, \phi_k, d)$ is assumed to be the best-matching template from the set $\{\mu_i(x, y, \phi_1, d), \mu_i(x, y, \phi_2, d), \dots\}$, and ϕ_{k-1} and ϕ_{k+1} are the two cyclic neighbors of orientation ϕ_k . Provided the best-matching template is indeed the one with closest orientation, the orientation-estimate error will always be less than $\Delta\phi/2$ for an angular increment $\Delta\phi = |\phi_1 - \phi_2| = |\phi_2 - \phi_3| = \dots$ between templates.

An example of a visual scene interpretation is given in Figure 5. The scene features a bottle, two wine glasses, one partially occluding the other, a water glass, and a cup, as seen in Figure 4. The cup is included as an outlier that is not sought, i.e., the system has no model of cups. As shown in the figure, the bottle, wine glasses, and water glass are correctly located and identified, and the system is not confused by the cup. The cup may be marked as an unknown obstacle. Processing time, all the way through from image acquisition to scene interpretation, is for this scene around 5.7 seconds on a Pentium 4 CPU at 2.4 GHz under Linux.

Figure 6 shows the case of an oriented object, a bottle of cleaning liquid with an ellipsoidal cross section. Because both position and orientation have to be estimated, the uncertainty with respect to position is somewhat higher than in the previous example.

We have found that the described method for scene interpretation can cope with challenging objects such as glasses under partial occlusion and over large variations of lighting (daylight, fluorescent light, spotlight).

VI. GRASP PLANNING

A highly desirable functionality for a service robot with a dexterous hand is autonomous grasping. In the following, we want to briefly introduce the problem of grasp planning and point out aspects that have led us to our grasp-planner design.

A. What is Needed for Grasp Planning?

Any approach to grasp planning needs kinematic and geometric information about the hand and the objects to be grasped. In other words, it has to be known how the hand can move and how hand and objects are shaped.

The kinematic and geometric model of an artificial hand is usually available from CAD. When choosing the representation for the object model, we had in mind that the model can be obtained autonomously in real time. Thus, we use a polyhedral model and allow gaps and unconnected faces. Such models can be created with 3D-reconstruction methods based on structured light, laser scanners, or stereo vision.

For the *Robutler*, we currently handle only objects from an a-priori known set. Only the composition of the scene with objects from this set has to be established in real time by the visual system; cf. section V. The semantic knowledge on these objects that is available to the *Robutler* influences the grasps used to manipulate them. For instance, a glass with water should be handled so as to prevent spilling.

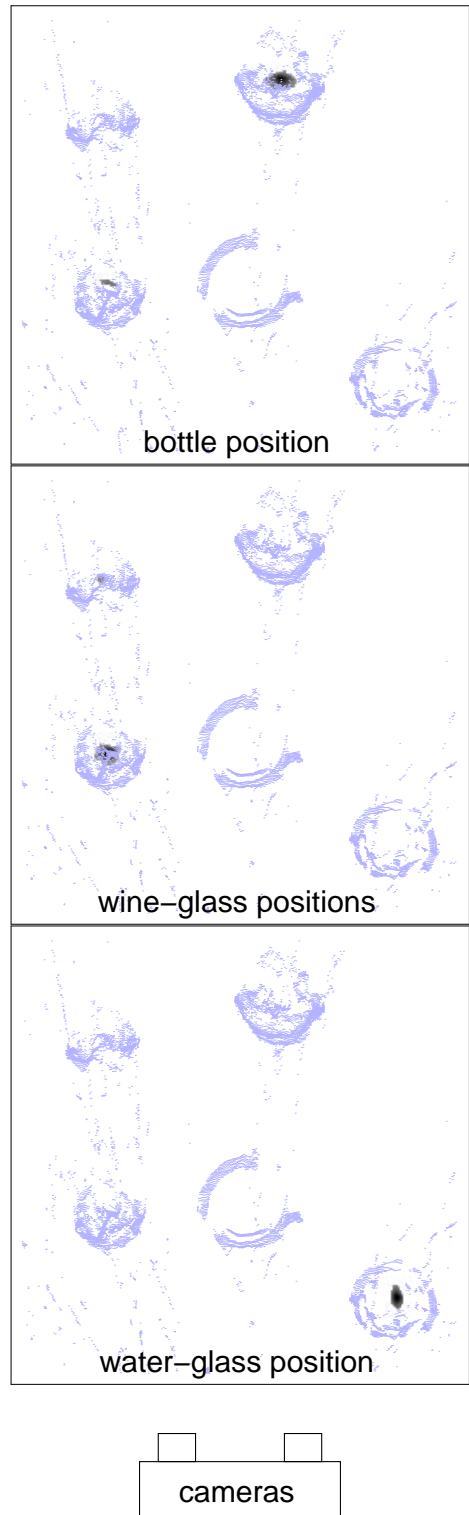


Fig. 5. The three panels show the distribution of evidence across the table (top view) for positions of bottles, wine glasses, and water glasses for the scene shown in Figure 4. A darker shade means higher evidence, white means zero evidence. Superimposed are the 3D-data points. The system did not look for cups. The evidence $M_i(x, y, d_1, d_2, \dots, d_N)$ for object $i \in \{\text{bottle, wine glass, water glass}\}$ is calculated according to equation (1) for all discrete positions (x, y) on the table. The spurious evidence for a bottle near the position of one of the wine glasses (see top panel) is suppressed by the larger evidence for a wine glass at that position (see middle panel). In fact, all sought objects are correctly located and identified, while the cup is ignored.

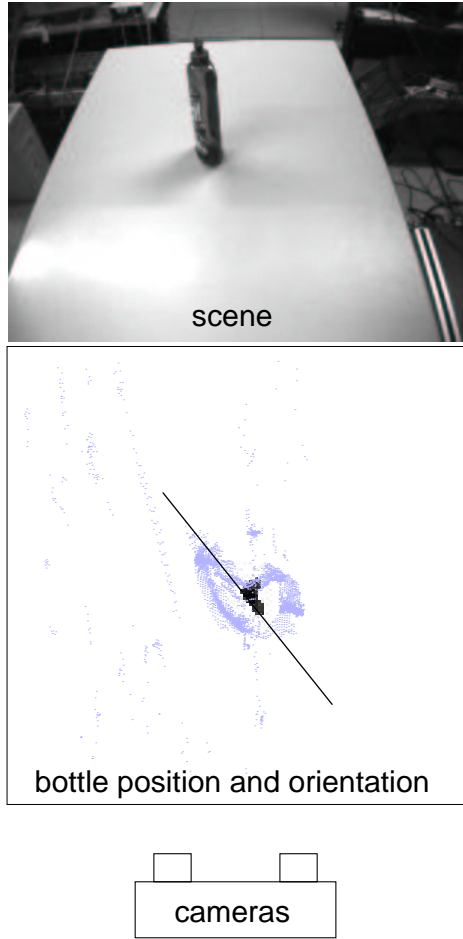


Fig. 6. A scene with a bottle of cleaning liquid with an ellipsoidal cross section, as seen in the top image taken by one of the cameras. The lower panel shows the distribution of evidence across the table for the position of the bottle (cf. Figure 5) and the estimated orientation of its major axis (line). The orientation is estimated according to equation (2) with an angular increment $\Delta\phi = \pi/36$. Note that the data of the bottle do not represent its ellipsoidal cross section, due to self-occlusion and artifacts of stereo processing. Nonetheless, the estimated orientation is well aligned with the bottle. Positional accuracy, however, is lower than for symmetrical objects; cf. Figure 5.

B. The Grasp-Planning Problem

Grasps for dexterous hands are reasonably divided in two main categories [9]: precision grasps for high manipulability where only the fingertips or distal finger links are in contact with the object, and power or enveloping grasps when high forces must be resisted or exerted and where the whole hand can be in contact with the object; see Figure 7.

The planning for these two types of grasps is very different. For precision grasps, one has to search for a few fingertip contacts (3 to 5) that allow for a stable grasp. For most of these contact sets, one can compute more than one valid hand configuration, as we show in [10]. Power grasps, on the other hand, are mainly determined by the geometry of the hand and the object. So one first tries to find a suitable hand configuration to “wrap” the fingers around the object and then calculates the resulting contacts.



Fig. 7. Power (left) and precision (right) grasp configuration of the DLR Four-Finger Hand.



Fig. 8. Planned pinch grasp.

C. The Grasp Planner for the DLR Hand

As stated above, it is possible to start precision grasp planning from finding contacts and then calculate a valid hand configuration to realize a grasp. Moreover, we have observed that about 20% of all grasps computed from four randomly chosen contact points on a set of geometrical and real world objects (cube, sphere, cylinder, coffee mug, martini glass, etc.) result in force closure, which is the most common quality criterion for precision grasps [11], [12]. Therefore, we have decided to implement a random-sampling-based grasp planner. The algorithm can be summarized as follows, for details see [13].

- 1) Choose four contact points randomly on the object.
- 2) Calculate a kinematically valid hand configuration using an optimization approach [10]; see Figure 8 for a pinch-grasp example. Return to step 1) if there is none.
- 3) Perform a collision test between hand and object to see if geometric constraints are satisfied. Return to step 1) in case of a collision.
- 4) Compute a quality measure for the grasp.
- 5) Store the grasp in a list sorted by grasp quality and return to step 1), until there are enough grasps in the list or a given time limit is exceeded.

The most interesting step is the quality computation, step 4). We use a measure introduced by Ferrari and Canny [14] which physically means: The quality of a grasp is as good as the minimal wrench that breaks the grasp if all fingers can press

with unit forces. This measure is physically more justified than the widely-used variant where the sum of all finger forces is limited to unit force. We have developed an incremental convex-hull construction algorithm to calculate the Ferrari-Canny measure. With this technique, we can compute and evaluate about 100 valid and force-closure grasp candidates in 20 to 60 seconds, depending on the complexity of the object, on a Pentium III CPU at 700 MHz under Linux. We then choose the best of these candidates as the final grasp.

VII. THE TASK

We emulate a constrained but realistic domestic scenario. Various bottles and glasses are freely arranged on a table. It is the task of the *Robutler* to prepare and offer drinks. The user may order, via a graphical or a speech interface, a particular drink (water, wine, etc.). The *Robutler* will choose the appropriate bottle, unscrew its cap (if the bottle has a cap), grasp it, and pour into the appropriate glass; see Figure 1. It will then pick up the glass and offer it to the user. Alternatively, the user may command the *Robutler* to pour from a specified bottle into a specified glass and so on.

It is important to note that unscrewing the cap from a bottle is a true fine-manipulation task that requires both visual precision and compliant torque control of the three fingers involved; see Figure 9. Compliance makes the fingers adaptable to residual errors incurred from the visual scene interpretation.

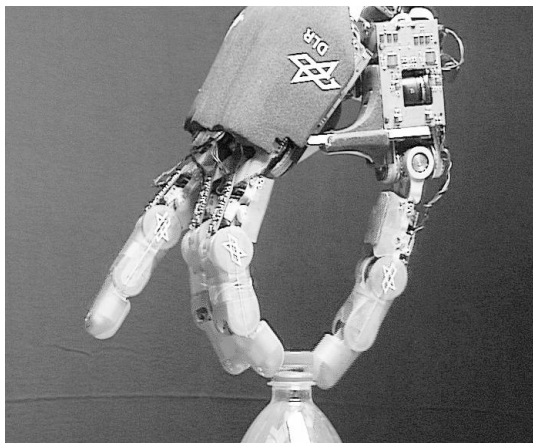


Fig. 9. The DLR Four-Finger Hand, unscrewing the cap from a bottle.

The command interface between the user and the robot is reduced to *what* has to be done. All the information on *how* the task is accomplished is hidden in the autonomous execution layer.

VIII. CONCLUSION

We have presented a hand-arm system with anthropomorphic features, combined with visual perception, dexterous grasping skills, and an intuitive man-machine interface. The *Robutler* thus integrates on the levels of mechatronic design, control strategies, as well as machine intelligence some of the

key components required for the next generation of service robots, that is, robots that will function in the human environment. We have validated robust system performance for a restricted but realistic task domain, preparing and offering drinks. In the future, we expect to expand the skills of the *Robutler* through recognition of more general shape classes, and widen its scope by adding mobility.

REFERENCES

- [1] G. Grunwald, G. Schreiber, A. Albu-Schäffer, and G. Hirzinger, "Programming by touch: The different way of robot human interaction," *IEEE Transactions on Industrial Electronics*, vol. 50, no. 4, 2003.
- [2] F. H. Wullschlegel and R. Brega, "The paradox of service robots: how passers-by can contribute in solving non-deterministic exceptional conditions encountered by service robots," in *Proc. IEEE/RSJ Int. Conference on Intelligent Robots and Systems*, 2002, pp. 1126–1131.
- [3] M. Yoshizaki, A. Nakamura, and Y. Kuno, "Mutual assistance between speech and vision for human-robot interface," in *Proc. IEEE/RSJ Int. Conference on Intelligent Robots and Systems*, 2002, pp. 1308–1313.
- [4] D. Lee, W. Chung, and M. Kim, "A reliable position estimation method of the service robot by map matching," in *Proc. IEEE Int. Conference on Robotics & Automation*, 2003, pp. 2830–2835.
- [5] U. Frese and T. Duckett, "A multigrid approach for accelerating relaxation-based SLAM," in *Proc. IJCAI Workshop Reasoning with Uncertainty in Robotics*, 2003, pp. 39–46.
- [6] G. Hirzinger, A. Albu-Schäffer, M. Hähle, I. Schäfer, and N. Sporer, "On a new generation of torque controlled light-weight robots," in *Proceedings of the IEEE Int. Conference on Robotics and Automation*, Seoul, Korea, May 2001, pp. 1087 – 1093.
- [7] U. Hillenbrand, "On the relation between probabilistic inference and fuzzy sets in visual scene analysis," *Pattern Recognition Letters*, p. submitted, 2003.
- [8] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognit.*, vol. 13, pp. 111–122, 1981.
- [9] M. R. Cutkosky and R. D. Howe, "Human grasp choice and robotic grasp analysis," in *Dextrous Robot Hands*, S. T. Venkataraman and T. Iberall, Eds. Springer Verlag, 1990, ch. 1.
- [10] C. Borst, M. Fischer, and G. Hirzinger, "Calculating hand configurations for precision and pinch grasps," in *Proc. of the 2002 IEEE/RSJ/GI International Conference on Intelligent Robots and Systems*. Lausanne, Switzerland: IEEE, 2002, pp. 1553–1559.
- [11] B. Mishra and N. Silver, "Some Discussion of Static Gripping and Its Stability," *Transactions on Systems, Man and Cybernetics*, vol. 19, no. 4, pp. 783 – 796, 1989.
- [12] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *Proceedings of the IEEE Int. Conference on Robotics and Automation*, San Francisco, California, April 2000, pp. 348 – 353.
- [13] C. Borst, M. Fischer, and G. Hirzinger, "A Fast and Robust Grasp Planner for Arbitrary 3D Objects," in *Proceedings of the IEEE Int. Conference on Robotics and Automation*, Detroit, Michigan, May 1999, pp. 1890–1896.
- [14] C. Ferrari and J. Canny, "Planning Optimal Grasps," in *Proceedings of the IEEE Int. Conference on Robotics and Automation*, Nice, France, May 1992, pp. 2290–2295.