A 3D grasping system based on multimodal visual and tactile processing

Beata J. Grzyb, Eris Chinellato, Antonio Morales and Angel P. del Pobil
Robotic Intelligence Lab, Department of Computer Engineering and Science, Universitat Jaume I, Castellon, Spain

Abstract
Purpose – The purpose of this paper is to present a novel multimodal approach to the problem of planning and performing a reliable grasping action on unmodeled objects.
Design/methodology/approach – The robotic system is composed of three main components. The first is a conceptual manipulation framework based on grasping primitives. The second component is a visual processing module that uses stereo images and biologically inspired algorithms to accurately estimate pose, size, and shape of an unmodeled target object. A grasp action is planned and executed by the third component of the system, a reactive controller that uses tactile feedback to compensate possible inaccuracies and thus complete the grasp even in difficult or unexpected conditions.
Findings – Theoretical analysis and experimental results have shown that the proposed approach to grasping based on the concurrent use of complementary sensory modalities, is very promising and suitable even for changing, dynamic environments.
Research limitations/implications – Additional setups with more complicate shapes are being investigated, and each module is being improved both in hardware and software.
Originality/value – This paper introduces a novel, robust, and flexible grasping system based on multimodal integration.

Keywords Control technology, Robotics, Control applications

1. Introduction

Traditional research on robot grasp planning, analysis, and control assumes that the layout of the workspace is known in advance, and that models of the objects to manipulate and of the robot hand are readily available. In these conditions the problem of grasping becomes an analytical planning problem, and many theoretical and computational solutions have been proposed for the different stages of a reach and grasp action.

In service robotics applications the above assumptions normally do not hold, and real world scenarios are usually unstructured and prohibitively costly to model. In these cases, theoretical analytical solutions are not directly applicable, and more flexible and versatile approaches have to be pursued. In dealing with unstructured environments, the main sources of uncertainty for grasping actions come from the attempt to manipulate unmodeled objects, which pose and physical characteristics can be variable and not known in advance. The use of sensors allows to acquire information about the environment and hence reduce uncertainty during action execution.

Within the field of grasp planning and execution, the use of sensors focuses on three main stages: first, on the object model acquisition, allowing a most traditional grasp planning after a model is built; second, on the approaching phase; and third, on the control loop of the grasp execution phase, with the purpose of obtaining a stable grasp. For what concerns the two first stages, vision is the most widely used modality. Many different strategies have been developed to estimate shape and pose of target objects from visual input (Jang et al., 2005; Wang et al., 2005). Successful approaches, with certain limitations, are available for grasp planning on 2D planar objects (Davidson and Blake, 1998; Morales et al., 2006), but no completely satisfactory solutions have been provided for the full 3D case. Visual feedback is often employed also when approaching the target objects, by using various techniques of visual servoing and active vision.

However, when the effector gets in touch with the object, vision leaves its leading role, and other sensory modalities, mostly contact based sensors, take over the control of the action. Pressure and/or force sensors are mainly employed as feedback for the grasp control execution loop (Platt et al., 2002), and in object exploration strategies (Teichmann and...
Mishra, 2000), and these tasks are often bundled together. From the available research works, an interesting conclusion can be drawn: robust and stable grasps can be obtained even though a detailed model of the object is not available. This is the main assumption on which the research presented here relies.

This paper introduces a grasping system which is able to deal with various objects, even novel ones, thanks to the purposeful use of sensory cues from different modalities. Objects which shape and location are unknown in advance can be robustly grasped through the subsequent use of visual and tactile feedback. Instead of exact shape models of objects, only a “common sense” knowledge of some object classes is assumed. Such approach is ideal for manipulation tasks in service robotics, where a robot can be asked to grasp many different kinds of objects in various conditions. The expected working environment can be correspondingly unstructured, but possibly not cluttered, as in principle the robot should be able to deal with a large variety of situations, but not extremely complex or potentially dangerous ones, such as those faced by rescue-robots. Similar conditions can be found in flexible manufacturing environments, where a possibly large but finite number of objects with different shape, size, and properties have to be handled in controlled but variable conditions. Our proposal constitutes a step toward a manipulation system able to operate autonomously in such conditions, and which main requirements are flexibility and robustness, while maintaining a good performance in movement precision and speed.

The design of the proposed system consists of three main parts. The first component is a manipulation framework that represents grasps in terms of hand preshapes, approaching directions, and control policies, rather than as sets of target contact points, as in most robot grasping applications. The second component is a vision system that makes use of algorithms based on a computational model of primate visual mechanisms. This stage allows to estimate pose, location, and size of a target object through the integration of binocular (stereoptic) and monocular (perspective) visual cues (Chinellato and del Pobil, 2008). The third component is a robust grasp execution controller that uses tactile feedback to execute the planned grasp, and corrects the grasp program if divergences are found between the expected and the actual situation of the object. A metric that measures the stability of a grasp and a decision procedure that produces a sequence of suitable corrections is implemented. The final outcome is a system that shows very good precision in approaching and preshaping in normal conditions, but that can also react to unexpected situations in an adaptive manner.

The three parts of the system are presented in Sections 2-4, respectively, whilst the system behavior in experimental tests is described in Section 5.

2. Manipulation framework

The proposed manipulation framework consists of a set of different a priori defined manipulation primitives that establish the control strategy and the sensory feedback to use in the execution. A manipulation primitive determines several key aspects of the grasp execution. First, it defines the hand preshape, i.e. the posture of the hand while approaching the target. Second, it describes the control strategy to be used for executing the action. This also includes which sensory information is used and how it is interpreted. It also determines the metrics that evaluate the degree of accomplishment of the action.

The set of manipulation primitives to be developed depends on the particular features of the robot hand, and on the different tasks (pulling objects, opening/closing doors, etc.) to be performed by this hand. In any case a detailed study of the hand constraints, tasks and objects is necessary.

In the case of the Barrett Hand, used in this research, Miller et al. (2003) presented a study on the possible hand preshapes that can be obtained with it. The taxonomy of grasp primitives, we derived from such set of preshapes is the following (Figure 1):

- **Cylindrical grasp.** All fingers close around a cylindrical object. The thumb finger opposes the other two (Figure 1(a)).
- **Spherical grasp.** All fingers close around a ball-shaped object (Figure 1(b)).
- **Pinch grasp.** The grasp is characterized by the opposition of the two mobile fingers. The thumb does not participate. This primitive is appropriate for grasping small objects (Figure 1(c)).
- **Hook grasp.** In this grasp the hand opposes the gravity. All fingers form a hook that would enclose a cylindrical shaped object. The palm might exert force opposing the fingers (Figure 1(d)).

In this work, we fully implement only the cylindrical preshape primitive. Note that, due to the lack of sensors on the Barrett Hand palm, it is impossible to check if the contact with the object has been made, therefore the grasp is planned between fingertips, and the palm is not expected to touch the object as it would do in human grasping.

In practical terms, a grasp is an instantiation of a manipulation primitive described by the following features (Figure 2):

- **Manipulation primitive.** A qualitative description of the grasp action to be performed, e.g. power grasp, pushing a button (with finger index), pulling, and others.
- **Distance and approaching direction.** Once the hand is positioned in the vicinity of the object it reaches toward it following the estimated direction for the necessary distance. The approaching line is the path followed by the robot hand when it approaches the object.
- **Hand orientation.** The hand can rotate around the approaching direction. The rotation angle is a relevant parameter to define the action initial configuration. The hand is correctly oriented for grasping before the execution of the last reaching phase, along the approaching line.
- **Object size.** The estimated size and proportion of the object affects the practical execution of the action, and is included as an input parameter to the execution controllers.
- **Force limits.** Depending on the estimated weight and compliance of the object, maximum and minimum grasping force limits can be established.

3. Visual analysis

Our experimental setup consists of a stereo camera mounted on the robot wrist. This configuration allows our fixed pair of cameras to simulate vergence movements of the eyes. The procedure we adopt is to center the lowest point of the object...
in one of the images first, and rotate the camera, in order to center again the same point on the other image without changing the actual distance.

The features of the object and its pose with respect to the effector are extracted through a visual analysis which is based on a computational model of distance and orientation estimation inspired by human visual mechanisms, and based on the integration of perspective and stereoscopic cues (Chinellato and del Pobil, 2008; Chinellato et al., 2008).

Applying such model to our robotic setup, we were able to reproduce with a reasonable level of approximation effects described in neuropsychological data. At the same time, we could provide our robotic grasping system with a very reliable and robust visual estimation of slant, distance, and size of target objects. The system is able to perform such estimation without using models, only exploiting the assumption, supported by neuroscience studies, that what looks like a trapezoid is most likely a slanted rectangular shape having parallel and equal edges. The same assumption of supposed parallelism holds for cylindrical shapes. The third class of objects that the system is able to grasp are spheres of different size, that only requires distance and size estimation, being the pose rotation invariant.

The process of pose estimation begins with the arm moving until the lowest point of the target object blob is placed at the center of the image of one camera, in order to minimize distortions due to the cameras’ optic. Given the left and right object images (Figure 3), the contour extraction and salient point detection are performed. Object faces are not segmented separately, so the number of detected corners ranges from 4 to 6 depending on object pose for a box-like shape, and is usually bigger for cylindrical and spherical shapes.

The coordinates of the defining points are transformed in angles with respect to the center of the image, using the camera focal lens and image size in pixels as parameters. These angles, for both left and right images, are thus used to perform the actual slant estimation. We make use of eight different estimators, both stereoptic and perspective.

In parallel, with the extraction of the silhouette salient points, a chain code representation of the object contour is fed to a probabilistic neural network which classifies the object into one of three classes: boxes, cylinders, and spheres. This classification permits to establish which pose estimators can be applied on the object. For example, perspective slant estimation can be performed on both visible faces of a parallelepiped, on one face for a cylinder, and cannot be performed for spheres. Stereoscopic slant estimation can be performed for all shapes but slightly differently for each of them. The classification also allows to fill in possible missing salient points or to remove some extra ones according to the shape class (Chinellato et al., 2008).

To estimate the distance of the object we make use of a biologically-inspired estimator based on vergence data: the distance of the object is inversely proportional to the angle between the object and the eyes (Chinellato et al., 2008).

4. Grasp planning and execution

Visual analysis produces an estimation of the pose, distance and size of the object, as well as an identification of its shape type. These parameters are used to produce a grasp primitive instantiation.

In the first stage of the execution, the robot arm reaches toward the object and moves down until the fingertips are at level with the
estimated object centroid. The hand is set in the preshape configuration and orientation. The second stage is the hand closing, and the tactile sensors are used to determine the fingertips contact with the object. The finger movement stops once the contact with the object has been made or when it reaches a previously defined extension threshold in case of missing the object. This stage finishes successfully when all fingertips have contacted the object, as in the case of Figure 4(a).

During the third, reactive stage, the grasp is assessed to verify whether it is stable enough to lift the object. A grasp stability criterion is used to measure this aspect. The criterion used is based on finger extension rule (Chinellato et al., 2005) which compares the extensions of the fingers assuming the fingers with similar extensions act more uniformly on the object. If the criterion does not reach a minimum quality threshold, the grasp is considered not stable. This can be due to any problems with pose estimation or to a change in the workplace occurred after the beginning of the movement. In this situation, correction movements are performed successively until a stable grasp is produced. An example of unstable grip that requires a rotation correction movement is shown in Figure 4(b).

5. Implementation and results

The system is implemented on a robotic setup consisting of a Mitsubishi PA-10 7 d.o.f. arm mounted on an ActiveMedia PowerBot mobile robot. The manipulator is endowed with a three-fingered Barrett Hand and a JR3 force/torque and acceleration sensor mounted at the wrist, between the hand and the end-effector. The hand has been improved by adding on the fingertips arrays of pressure sensors.

Regarding visual analysis, we executed more than 400 experiments, with different values of slant and distance (Chinellato et al., 2008). The global average of orientation estimation errors of all executed experiments is less than three degrees for boxes and nearly six for cylinders. The results show that merging of stereo and perspective cues provides the system with reliable and robust estimation capabilities across working conditions.

5. Implementation and results

For what concerns grasp execution and the work of the tactile system, we tested our robotic system in two different conditions, i.e. without or with small environmental changes. The first condition, corresponding to a normal working situation, usually ends with a successful grasping action without performing any correction movement. In fact, in almost all cases the input provided by the visual system is good enough to allow the execution of the grasping action without the need for correction of hand position or orientation.

While testing the system performance in the second condition, we were introducing on purpose some changes in the object position and/or orientation, to check if the system was able to deal with unexpected and suddenly changing situations. The changes were made after the visual analysis had been completed so that the real pose of the object was much different from the estimated one, like in the example of Figure 4(b). In these situations the robot may not be able to grasp the object without the support of the tactile feedback. Using the information about the finger extension and the hand contact with the surface of the object, the orientation and position of the hand are corrected in a closed-loop manner (Chinellato et al., 2008). If the difference between the real and estimated object pose is big, more than one correction movement might be required. As a limit, the correction of the grasping action cannot be performed when the displacement of the object is so large that, after its reaching movement, the hand does not have any contact with the object surface. We are now working on an additional visual feedback to deal with this situation.

6. Discussion and conclusion

This paper has presented the design of a complete grasp system that includes the initial visual processing of the object and scene, the grasp planning, and the final execution of a grasp. This is done by considering a theoretical framework to describe manipulation actions as controller templates, and using visual and tactile feedback to face the uncertainty of the task.

In normal conditions, the visual pose estimation is precise and robust enough to allow reliable visual analysis and grasp execution on simple but previously unmodeled 3D shapes. Tactile feedback provide the additional robustness required when operating in unstructured environments and unknown scenes. The proposed approach is ambitious, and some of its
modules are only partially implemented. This is the case of the grasp planning that can only deal with basic shapes such as boxes, cylinders and spheres. For this reason, the system is being endowed with better vision and manipulation hardware and software, that should allow to deal with more variable and complex shapes.

Overall, the presented grasping system has demonstrated the feasibility of the approach, and its suitability even for changing, dynamic environments. Further development of each module and a more exhaustive testing and validation in different scenarios are the next steps in this research.

References


Corresponding author

Beata J. Grzyb can be contacted at: grzyb@uji.es