Available online at www.sciencedirect.com
SCIENCE DIRECT
Robotics and Autonomous Systems xxx (2005) xxx–xxx
Robotics and Autonomous Systems
www.elsevier.com/locate/robot

Vision for robotic object manipulation in domestic settings

Danica Kragic*, Mårten Björkman, Henrik I. Christensen, Jan-Olof Eklundh

Computer Vision and Active Perception Lab, Centre for Autonomous Systems, Royal Institute of Technology, Stockholm, Sweden

Received 5 March 2005; accepted 21 March 2005

Abstract

In this paper, we present a vision system for robotic object manipulation tasks in natural, domestic environments. Given complex fetch-and-carry robot tasks, the issues related to the whole *detect-approach-grasp* loop are considered. Our vision system integrates a number of algorithms using monocular and binocular cues to achieve robustness in realistic settings. The cues are considered and used in connection to both foveal and peripheral vision to provide depth information, segmentation of the object(s) of interest, object recognition, tracking and pose estimation. One important property of the system is that the step from object recognition to pose estimation is completely automatic combining both appearance and geometric models. Experimental evaluation is performed in a realistic indoor environment with occlusions, clutter, changing lighting and background conditions.

Keywords: Cognitive systems; Object recognition; Service robots; Object manipulation

1. Introduction

One of the key components of a robotic system that operates in a dynamic, unstructured environment is robust perception. Our current research considers the problem of mobile manipulation in domestic settings where, in order for the robot to be able to detect and manipulate objects in the environment, robust visual feedback is of key importance. Humans use visual feedback extensively to *plan* and *execute* actions. However, planning and execution is not a well-defined one-way stream: how we plan and execute actions depends on what we already know about the environment we operate in, what we are about to do, and what we think our actions will result in. Complex coordination between the eye and the hand is used during execution of everyday activities such as pointing, grasping, reaching or catching. Each of these activities or actions requires attention to different attributes in the environment—while pointing requires only an approximate location of the object in the visual field, a reaching or grasping movement requires more exact information about the object's pose.

In robotics, the use of visual feedback for motion coordination of a robotic arm or platform motion is termed *visual servoing*, Hutchinson et al. [1]. In general, visual information is important at different lev-

* Corresponding author. Tel.: +46 87906729; fax: +46 87230302.
E-mail addresses: danik@nada.kth.se (D. Kragic); celle@nada.kth.se (M. Björkman); hic@nada.kth.se (H.I. Christensen); joe@nada.kth.se (J.-O. Eklundh).

0921-8890/$ – see front matter
doi:10.1016/j.robot.2005.03.011

els of complexity: from scene segmentation to object's pose estimation. Hence, given a complex fetch-and-carry type of task, issues related to the whole *detect-approach-grasp* loop have to be considered. Most visual servoing systems, however, deal only with the *approach* step and disregard issues such as *detecting* the object of interest in the scene or retrieving its three dimensional (3D) structure in order to perform grasping. A so called *teach-by-showing* approach is typically used where the desired camera placement with respect to the object is well defined and known before hand.

Our goal is the development of an architecture that integrates different modules where each module encapsulates a number of visual algorithms responsible for a particular task such as recognition or tracking. Our system is heavily based on the *active vision* paradigm, Ballard [2] where, instead of passively observing the world, viewing conditions are actively changed so that the best results are obtained given a task at hand.

In our previous work, Björkman and Kragic [3] we have presented a system that consists of two pairs of stereo cameras: a peripheral camera set and a foveal one. Recognition and pose estimation are performed using either one of these, depending on the size and distance to the object of interest. From segmentation based on binocular disparities, objects of interest are found using the peripheral camera set, which then triggers the system to perform a saccade, moving the object into the center of foveal cameras achieving thus a combination of a large field of view and high image resolution. Compared to one of the recent systems, Kim et al. [4], our system uses both hard (detailed models) and soft modeling (approximate shape) for object segmentation. In addition, choice of binocular or monocular cues is used depending on the task. In this paper, we formalize the use of the existing system with respect to Fig. 1—how to utilize the system with respect to different types of robotic manipulation tasks.

This paper is organized as follows. In Section 2, a problem definition is given. In Section 3, a short overview of the current system is given and in Section 4 hypotheses generation is presented. In Section 5 we deal with the problem of manipulating known objects and in Section 6 with the problem of manipulating unknown objects. Some issues related to object grasping are given in Section 7. Experimental evaluation is presented in Section 8 and final conclusion given in Section 9.

2. Problem definition

In general, vision based techniques employed in visual servoing and object manipulation depend on:

- Camera placement: Most visual servoing systems today use *eye-in-hand* cameras and deal mainly with the *approach* object step in a *teach-by-showing* manner, Malis et al. [5]. In our approach, we consider a combination of a stand-alone stereo and an eye-in-hand camera systems, Kragic and Christensen [6].
- Number of cameras: In order to extract metric information, e.g. sizes and distances, about objects observed by the robot, we will show how we can benefit from binocular information. The reason for using multiple cameras in our system is the fact that it simplifies the problem of segmenting the image data into different regions representing objects in a 3D scene. This is often referred to as *figure-ground segmentation*. In cluttered environments and complex backgrounds, figure-ground segmentation is particularly important and difficult to perform and commonly the reason for experiments being performed in rather sparse, simplified environments. In our work, multiple cameras are used for scene segmentation while a single camera is used for visual servoing, object tracking and recognition.
- Camera type: Here we consider systems using zooming cameras or combinations of foveal and peripheral ones. With respect to these, very little work has been reported in visual servoing community, Benhimane and Malis [7]. In this paper, we demonstrate how a combination of foveal and peripheral cameras can be used for scene segmentation, object recognition and pose estimation.

In our current system, the robot may be given tasks such as "Robot, bring me the raisins" or "Robot, pick up this". Depending on the prior information, i.e. task or context information, different solution strategies may be chosen. The first task of the above is well defined since it assumes that the robot already has the internal representation of the object, e.g. the *identity* of the object is known. An example of such a task is shown in Fig. 2: after being given a spoken command, the robot locates the object, approaches it, estimates its pose and

"Pick Up ..."		WHERE (location)	
		known	unknown
WHAT (identity)	known	"This Cup"	"The Cup"
	unknown	"*This* Object"	"Something"

Fig. 1. Robotic manipulation scenarios.

finally performs grasping. More details related to this approach are given in Section 5. For the second task, the spoken command is commonly followed by a pointing gesture—here, the robot does not know the *identity* of the object, but it knows its approximate *location*. The approach considered in this work is presented in Section 6. Fig. 1 shows different scenarios with respect to prior knowledge of object *identity* and *location*, with the above examples shaded. A different set of underlying visual strategies is required for each of these scenarios. We have considered these two scenarios since they are the most representative examples for robotic fetch-and-carry tasks.

2.1. Experimental platform

The experimental platform is a Nomadic Technologies XR4000, equipped with a Puma 560 arm for manipulation (see Fig. 3). The robot has sonar sensors, a SICK laser scanner, a wrist mounted force/torque sensor (JR3), and a color CCD camera mounted on the Barrett Hand gripper. The palm of the Barrett hand is covered by a VersaPad touch sensor and, on each finger, there are three Android sensors. On the robot's shoulder, there is a binocular stereo-head. This system, known as Yorick, has four mechanical degrees of freedom; neck pan and tilt, and pan for each camera in relation to the neck. The head is equipped with a pair of

Fig. 2. Detect-approach-grasp example.

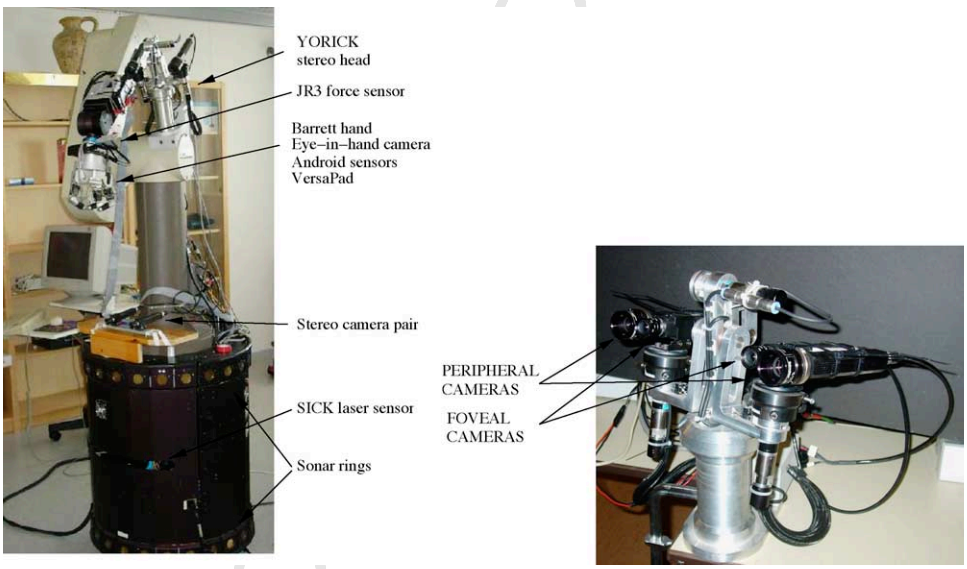

Fig. 3. (Left) Experimental platform Nomadic Technologies XR4000, and (Right) Yorick stereo-head.

Sony XC999 cameras, with focal length of 6 mm. Additional pair of Sony XC999 cameras with focal length of 12 mm is placed directly on the robot base.

For some of the experimental results that will be presented further on, a stand-alone binocular stereo-head system shown in Fig. 3 was used. Here, the head is equipped with two pairs of Sony XC999 cameras, with focal lengths 28 and 6 mm, respectively. The motivation for this combination of cameras will be explained related to the examples.

3. The system

Fig. 4 shows a schematic overview of the basic building blocks of the system. These blocks do not necessarily correspond to the actual software components, but are shown in order to illustrate the flow of information through the system. For example, the visual front end consists of several components, some of which are running in parallel and others hierarchically. For example, color and stereo information are extracted in parallel, while epipolar geometry has to be computed prior to disparities. On the other hand, action generation, such as initiating 2D or 3D tracking, is distributed and performed across multiple components.

The most important building blocks can be summarized as follows:

- The Visual Front-End is responsible for the extraction of visual information needed for figure-ground segmentation and other higher level processes.
- Hypotheses Generation produces a number of hypotheses about the objects in the scene that may be relevant to the task at hand. The computations are moved from being distributed across the whole image to particular regions of activation.
- Recognition is performed on selected regions, using either corner features or color histograms, to determine the relevancy of observed objects.
- Action Generation triggers actions, such as visual tracking and pose estimation, depending on the outcome of the recognition and current task specification.

Due to the complexity of the software system, it was partitioned into a number of smaller modules that communicate through a framework built on a interprocess communication standard called CORBA (Common Object Request Broker Architecture), Vinoski [8]. The current version of the system consists of about ten such modules, each running at a different frame rate. The lowest level frame grabbing module works at a frequency of 25 Hz, while the recognition module is activated only upon request. In order to consume processing power, modules are shut down temporarily when not been accessed by any other module within a time frame of 10 s.

With limited resources in terms of memory storage and computational power, biological and robotic sys-

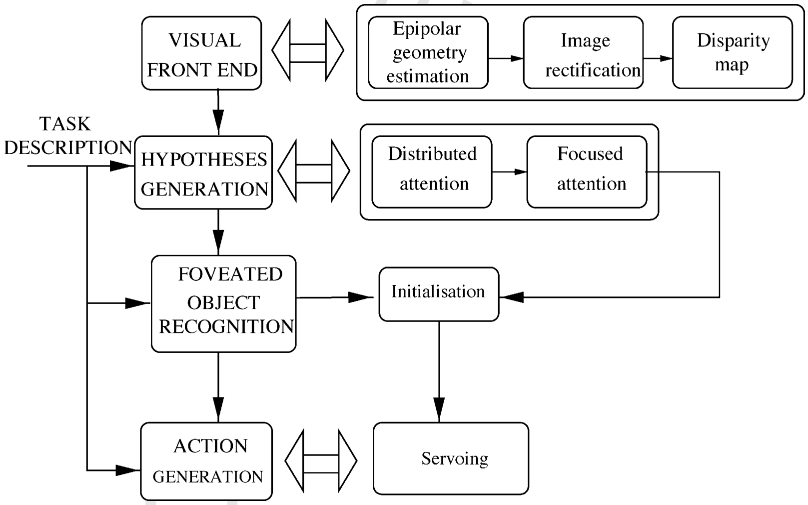

Fig. 4. Basic building blocks of the system.

tems need to find an acceptable balance between the width of the visual field and its resolution. Otherwise, the amount of visual data will be too large for the system to efficiently handle. Unfortunately, this balance depends on the tasks the systems have to perform. An animal that has to stay alert in order to detect an approaching predator, would prefer a wide field of view. The opposite is true if the same animal acts as a predator itself. Similarly, a robotic system benefits from a wide field of view, in order not to collide with obstacles while navigating through a cluttered environment. A manipulation task on the other hand, requires a high resolution in order grasp and manipulate objects. That is, to find objects in the scene a wide field of view is preferable, but recognizing and manipulating the same objects require a high resolution.

On a binocular head, Björkman and Kragic [3] we overcame this problem by using a combination of two pairs of cameras, a peripheral set for attention and a foveated one for recognition and pose estimation. In order to facilitate transfers of object hypotheses from one pair to the other, and replicate the nature of the human visual system, the pairs were placed next to each others. The camera system on the robot is different in that the two pairs are widely separated and placed on an autonomously moving platform, see Fig. 3: a stereo head on a shoulder and another pair on the base. The search pair is located on-top of the robot overlooking the scene and the manipulation pair is at waist height, such that the gripper will not occlude an object while it is being manipulated. In the original version, hypothesis transfers were based on matched corner features and affine geometry. Hence, with the cameras related pairwise, the position of hypotheses seen by the peripheral cameras could be transferred to the images of the foveated stereo set.

This way of transferring positions is no longer feasible in the robot camera configuration. With the cameras separated by as much as a meter, the intersections between visual fields tend to be small and the number of features possible to match is low. Furthermore, a feature seen from two completely different orientations is very difficult to match, even using affine invariant matching. Instead we exploit the fact that we can actively move the platform such that an object of interest, found by the search pair, will become visible by the manipulation pair. For this to be possible we have to approximately know the orientation and position of the cameras in relation to the base. Hypotheses are found by the search pair, the 3D positions are derived using triangulation and finally projected onto the image planes of the manipulation pair. For the 3D position to be accurately estimated, the search pair is calibrated on-line, similarly to the original version of the system, Björkman and Eklundh [9]. The precision in depth ranges from about a decimeter to half a meter depending on the observed distance.

3.1. Stereo system modeling—epipolar geometry

With a binocular set of cameras, differences in position between projections of 3D points onto the left and right image planes (disparities) can be used to perform figure-ground segmentation and retrieve the information about three-dimensional structure of the scene. If the relative orientation and position between cameras is known, it is possible to relate these disparities to actual metric distances. One of the commonly used settings is where the cameras are rectified and their optical axes mutually parallel, Kragic and Christensen [6]. However, one of the problems arising is that the part of the scene contained in the field of view of both cameras simultaneously is quite limited.

Another approach is to estimate the epipolar geometry continuously from image data alone, Björkman [10]. Additional reason for this may be that small disturbances such as vibrations and delays introduce significant noise to the estimation of the 3D structure. In fact, an error of just one pixel leads to depth error of several centimeters on a typical manipulation distance. Therefore, for some of the manipulation tasks, the epipolar geometry is estimated robustly using Harris' corner features, Harris and Stephens [11]. Such corner features are extracted and matched between the camera images using normalized cross-correlation. The vergence angle α, gaze direction t, relative tilt r_x and rotation around the optical axes r_z, are iteratively sought using

$$\begin{pmatrix} \mathrm{d}x \\ \mathrm{d}y \end{pmatrix} = \begin{pmatrix} (1+x^2)\alpha - yr_z \\ xy\alpha + r_y + xr_z \end{pmatrix} + \frac{1}{Z}\begin{pmatrix} 1 - xt \\ -yt \end{pmatrix}, \quad (1)$$

where Z is the unknown depth of a point at image position (x, y). The optimization is performed using a combination of RANSAC [12] for parameter initialization, and M-estimators [13] for improvements.

This optical flow model [14] is often applied to motion analysis, but has rarely been used for stereo. The reason for this is because the model is approximate and only works for relatively small displacements. In our previous work we have, however, experimentally shown that this model is more robust than the essential matrix in the case of binocular stereo heads, Björkman and Eklundh [9], even if the essential matrix leads to a more exact description of the epipolar geometry, Longuet-Higgins [15].

4. Hypotheses generation

The purpose of this component is to derive qualified guesses of *where* the object of interest is located in the current scene. As mentioned earlier, this step is performed using the peripheral cameras while the recognition module uses the foveal ones. This requires a transfer from peripheral to foveal vision, or from distributed to focused attention Palmer [16].

4.1. Distributed attention

Unlike focused attention, distributed attention works on the whole image instead of being concentrated to a particular image region. Using the available visual cues a target region, that might represent an object of interest, is identified. Even if the current system is limited to binocular disparities, it is straightforward to add additional cues, such as in the model of Itti et al. [17]. Here, we have concentrated on disparities because they contain valuable information about object size and shape. This is especially important in a manipulation task, where the color of an object might be irrelevant, whereas the size is not.

The only top-down information needed for hypotheses generation is the expected size of an object of interest and the approximate distance from the camera set. More information about the attention system can be found in Björkman and Eklundh [18]. A binary map is created containing those points that are located within a specified depth range. The third column of Fig. 9 shows two such maps overlaid on-top of the corresponding left peripheral images. Initial hypotheses positions are then generated from the results of a difference of Gaussian filter applied to the binary map. The scale of this filter is set so as to maximize the response of image blobs representing objects of the requested size and distance. The depth range is continuously updated so that hypotheses are obtained for objects at different depths. In our system, the depths typically vary between 1 and 3 m.

4.2. Focused attention

From the generated hypotheses, a target region is selected so that the gaze can be redirected and recognition performed using the foveal cameras. This selection is done automatically from the hypothesis of largest strength. However, before the strongest hypothesis is selected, a small amount of noise equivalent to about 20% of the largest possible strength is added. This is done in order to prevent the system from getting stuck at a local maximum. Due to occlusions, the requested object might otherwise never be visited.

Since hypotheses are described in the peripheral cameras frame and recognition is performed using the foveal ones, the relative transformations have to be known. These are found applying a similarity model to a set of Harris' corner features similar to those used for epipolar geometry estimation in Section 3.1. On the stereo head system shown in Fig. 3, the relative rotations, translations and scales are continuously updated at a rate of about 2 Hz. For the manipulator system, the robot first has to rotate its base while tracking the hypotheses until visual fields overlap. Knowing the transformations, it is possible to translate the hypotheses positions into the foveal camera frames.

Before a saccade is finally executed, fixating the foveal cameras onto the selected hypothesis region, the target position is refined in 3D. During a couple of image frames, a high-resolution disparity map is calculated locally around the target area. A mean shift algorithm, Comaniciu et al. [19], is run iteratively updating the position from the cluster of 3D points around the target position, represented by the disparity map. The maximum size of this cluster is specified using the top-down information mentioned above. The first two images of Fig. 5 show these clusters highlighted in the left peripheral images before and after a saccade. The foveal images after the saccade can be seen to the right.

Fig. 5. The first two images show a target region before and after a saccade (the rectangles show the foveal regions within the left peripheral camera image) and the foveal camera images after executing a saccade are shown in the last two images.

4.3. Active search

For mobile manipulation tasks, it is important that the visual system is able to actively search for the object of interest. The search system includes two necessary components, an attentional system that provides hypotheses to where an object of interest might be located, and a recognition system that verifies whether a requested object has indeed been found, as presented above. Even if the attentional system works on a relatively wide field of view, 60° is still limited if alocation is completely unknown to the robot. In our system, we have extended this range by applying an active search strategy, that scans the environment and records the most probable locations. Five images from such a scan can be seen on the last row of Fig. 6. The crosses indicate hypothesis positions when the robot actively searches for and locates an orange package that is in fact located on the table seen on the first and fourth image.

5. Manipulating known objects

If a robot is to manipulate a known object, some type of representation is typically known in advance. Such a representation may include object textural and/or geometrical properties which are sufficient for theobject to be located and manipulation task to be performed. For realistic settings, a crude information about objects location can sometimes be provided from the task level. e.g. “Bring me red cup from the dinner table”. However, if the location of the object is not provided, it is up to the robot to search the scene. The following sections give examples of how these problems are approached in the current system.

5.1. Detect

If we can assume that the object is in the field of view from the beginning of the task, a monocular recognition

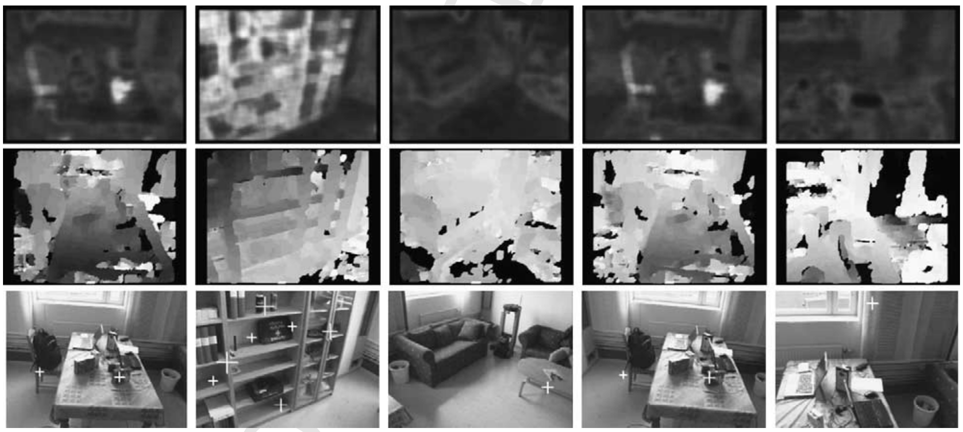

Fig. 6. First row: hue-saliency map with orange package as requested object, second row: peripheral disparity map, and third row: strongest hypotheses marked with crosses.

system can be used to locate the object in the image, Zillich et al. [20].

However, when a crude information about object's current position is not available, detecting a known object is not an easy task since a large number of false positives can be expected. Candidate locations have to be analyzed in sequence which may be computationally too expensive, unless the robot has an attentional system that delivers the most likely candidate locations first, using as much information about the requested object as possible.

A natural approach here is to employ a binocular system that provides metric information as an additional cue. Since the field of view of a typical camera is quite limited, binocular information can only be extracted from those parts of the 3D scene that are covered by both cameras' peripheral field of view. In order to make sure that an object of interest is situated in the center of each camera's field of view, the head is able to actively change gaze direction and vergence angle, i.e. the difference in orientation between the two cameras. In our system, stereo based figure-ground segmentation is intended for mobile robot navigation and robot arm transportation to the vicinity of the object. More detailed information about an object's pose is provided using a monocular model based pose estimation and tracking, Kragic [21].

The visual front-end is responsible for delivering 3D data about the observed scene. Such information is extracted using a three-step process, which includes the above mentioned epipolar geometry estimation, image rectification and calculation of dense disparity maps. The generation of this data is done continuously at a rate of 8 Hz, independently of the task at hand and used by more high-level processes for further interpretation. Further information on this part of the system can be found in Björkman [10]. Since most methods for dense disparity estimation assume the image planes to be parallel, image rectification has to be performed using the estimated epipolar geometry before disparities can be estimated. The current system includes seven different disparity algorithms, from simple area correlation, Konolige [22] to more complicated graph-cut methods, Kolmogorov and Zabih [23]. The benefit of using a more advanced global method, is the fact that they often lead to denser and more accurate results. However, even if density is important, the computational cost of these methods makes them infeasible for our particular application which means that correlation based methods are typically used in practice. Currently, we use two kinds of visual cues for this purpose, 3D size and hue histograms using the procedure described in Section 4.1. These cues were chosen since they are highly object dependent and relatively insensitive to changing lighting conditions, object pose and viewing direction. The images in Fig. 6 show examples where the orange package is requested. The upper images illustrate the saliency maps generated using the hue histograms of this object. From the disparity maps (second row) a number of candidate locations are found, as shown in the last row.

We further use recognition to verify that a requested object has indeed been found. With attention and recognition applied in a loop, the system is able to automatically search the scene for a particular object, until it has been found by the recognition system. Two recognition modules are available for this purpose: (i) a feature based module based on Scale Invariant Feature Transform (SIFT) features Lowe [24], and (ii) an appearance based module using color histograms, Ekvall et al. [25].

Most recognition algorithms expect the considered object to subtend a relatively large proportion of the images. If the object is small, it has to be approached before is can be detected. Possible solution would be using a eye-in-hand camera and only approach the object through the manipulator, keeping the platform itself static. A more efficient solution is a system equipped with wide field as well as foveal cameras, like the stereo-head system used for the example presented here. Hypotheses are found using the wide field cameras, while recognition is done using the foveal ones.

5.2. Approach

Transporting the arm to the vicinity of the object, considering a closed-loop control system, requires registration or computation of spatial relationship between two or more images. Although this problem has been studied extensively in the computer vision society, it has rarely been fully integrated in robotic systems for unknown objects. One reason for this is that high real-time demand makes the problem of tracking more difficult then when processing image sequences off-line. For cases where the object is initially far away from the robot, a simple tracking techniques can be used to keep the object in the field of view while approaching it. For this purpose we have developed and evaluated methods based on correlation and optical flow, Kragic et al. [26] as well as those based on integration of cues such as texture, color and motion, Kragic and Christensen [27]. The latter approach is currently used for tracking.

Performing final approach toward a known object depends also on the number of cameras and their placement. For eye-in-hand configuration we have adopted a *teach-by-showing* approach, where a stored image taken from the reference position is used to move the manipulator so that the current camera view is gradually changed to match the stored reference view. Accomplishing this for general scenes is difficult, but a robust system can be made under the assumption that the objects are piecewise planar. In our system, a wide baseline matching algorithm is employed to establish point correspondences between the current and the reference image, Kragic and Christensen [27]. The point correspondences enable the computation of a homography relating the two views, which is then used for 2 1/2D visual servoing.

In cases where the CAD model of the object is available, a full 6D pose estimate is obtained. After the object has been localized in the image, its pose is automatically initiated using SIFT features from the foveal camera image, fitting a plane to the data. Thus, it is assumed that there is a dominating plane that can be mapped to the model. The process is further improved searching for straight edges around this plane. The complete flow from hypotheses generation to pose estimation and tracking is performed fully automatic.

6. Manipulating unknown objects

For general setting, manipulation of unknown objects has rarely been pursued. The primary reason is likely to be that the shape of an object has to be determined in order to successfully grasp it. Another reason is that, even if the location is given by a pointing gesture, the size also has to be known and the object segmented from its background.

6.1. Detect

Numerous methods exist for segmentation of objects in cluttered scenes. However, from monocular cues only this is very difficult, unless the object has a color or texture distinct from its surrounding. Unfortunately, these cues are sensitive to lighting as well as pose variations. Thus, for the system to be robust, one has to rely on information such as binocular disparities or optical flow. A binocular setting is recommended, since the motion that needs to be induced should preferably be parallel to the image plane, complicating the process of approaching the object.

In our current system, binocular disparities are used for segmentation with the foveal camera set. We use this set since the focal lengths have to be relatively large in order to get the accuracy required for grasping. When the resolution in depth increases, so does the range of possible disparities. If only a fraction of these disparities are tested, e.g. the range in which the object is located, a large number of outliers can be expected, such as in the lower-left image of Fig. 7. We apply a Mean-Shift algorithm, Comaniciu et al. [19] to prune the data, using the fact that the points representing the object are located in a relatively small part of 3D space and the center of these points is approximately known. After applying a sequence of morphological operation a mask is found as shown in the lower-right image.

6.2. Approach

Approaching an unknown object can be done either using the stereo-head or with an eye-in-hand camera. Without knowing the identity of the object the latter case is hardly feasible. It would be possible to take a sequence of images, while approaching the object, and from these estimate a disparity map, but this map would hardly be as accurate as using the disparities available from the foveal camera set.

If the stereo-head is used instead, it is essential that the robot gripper itself can be located in disparity space. Using the mask derived in Section 6.1, the elongation and orientation of the object can be determine and the fingers of the gripper be placed on either side of the object. In general we will not be able, from one stereo view only, to retrieve the full 3D shape of the object. In particular, if the extension in depth is significant, it will

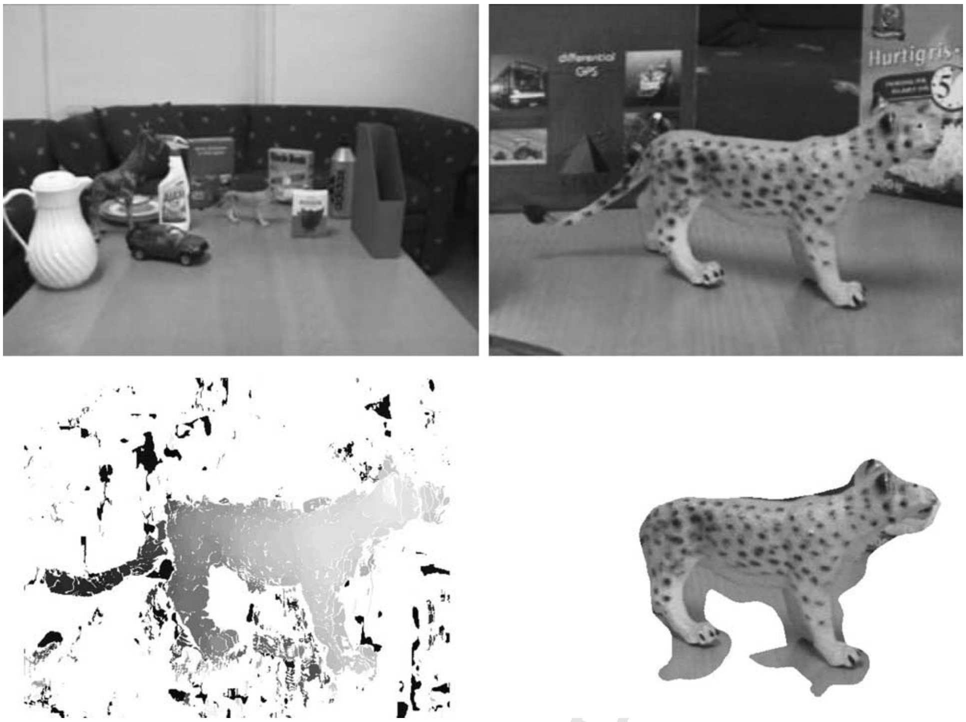

Fig. 7. Left peripheral (upper left) and foveal (upper right) camera images and disparities (lower left) and segmentation (lower right) automatically obtained from the peripheral stereo pair.

be difficult to guarantee that the full closing grasp can be performed. This problem can be solved by moving the stereo-head to another location. This is a topic we intend to investigate further in the future.

7. Grasping

For active grasping, visual sensing will in general not suffice. One of the problems closely related to eye-in-hand configurations is the fact that when the *approach* step is finished, the object is very close to the camera, commonly covering the whole field of view. To retrieve features necessary for grasp planning is impossible. One solution to this problem is to use a wide field eye-in-hand camera, together with a stand-alone mono- or stereo vision system. Our previous work has integrated visual information with tactile and force-torque sensing for object grasping, Kragic and Christensen [28]. We have, however, realized that there is a need for a system that is able to monitor the grasping process and track the pose of the object during execution. We have shown that in this way, even if the robot moves the object, grasping can successfully be performed without the need to reinitiate the whole process. This can be done even for unknown objects where the Mean-Shift strategy suggested in Section 6.1 is applied on consecutive images.

8. Experimental evaluation

As mentioned in Section 3, our system is built on a number of independently running and communicating modules. Since most methods used within these modules have been analyzed elsewhere, we will concentrate on the integrated system as a whole, rather than analyzing each individual method in isolation. The system should be considered as an integrated unit and its performance measured based on the behavior of the complete system. The failure of one particular module does not necessarily mean that the whole system fails. For example, figure-ground segmentation might well fail to separate two nearby objects located on a similar distance, but the system might still be able to initiate pose estimation after recognition.

The following properties of the system have been evaluated, as will be described in more detail in the sections below:

- combined figure-ground segmentation based on binocular disparities and monocular pose estimation,
- combined monocular Cooccurence Color Histograms (CCH) Chang and Krumm [29] based object recognition and monocular pose estimation,
- robustness of figure-ground segmentation,
- robustness toward occlusions using SIFT features,
- robustness of pose initialization toward rotations.

For recognition, a set of 28 objects was used. Fig. 8 shows a few of them. A database was created consisting of object models based on SIFT features and CCHs. Eight views per object were used for the SIFT models as well as in the case of CCHs. Pose estimation was only considered for the first three box-like objects, automatically starting as one of these objects are recognized. For this purpose, the width, height and thickness of these objects were measured and recorded in the database.

Since the observed matching scores did not significantly differ from those already published in Lowe [24] and Mikolajczyk and Schmid [30] we have chosen not to include any additional quantitative results. A few observations have lead us to believe that recognition would benefit from CCHs and SIFT features being used in conjunction. For example, the blue car is rarely recognized properly using SIFT, since the most salient features are due to specularities. However, the distinct color makes it particularly suitable for CCHs, which on the other hand have a tendency of mixing up the tiger and the giraffe, unlike the recognition module based on SIFT features.

8.1. Binocular segmentation and pose estimation

The first experiments illustrate the typical behavior of the system with binocular disparity based figure-ground segmentation and SIFT based recognition. Results from these experiments can be seen in Fig. 9. The first column shows the left foveal camera images prior to the experiments. It is clear that a requested object would be hard to find, without peripheral vision controlling a change in gaze direction. However, from the disparity maps in the second column the system is able to locate a number of object hypotheses, which can be shown as white blobs overlaid on-top of the left peripheral camera image in the third column of the figure.

The matching scores of the recognition module for these two examples were 66% and 70%, respectively, measured as the fraction of SIFT features being matched to one particular model. Once an object has been recognized, pose estimation is automatically initiated. This is done using SIFT features from the left and right foveal camera images, fitting a plane to the data. Thus, it is assumed that there is a dominating plane that can be mapped to the model. The process is further improved searching for straight edges around this plane. The last two columns show an example of this being done in practice.

Fig. 8. Some of the objects used for experimental evaluation.

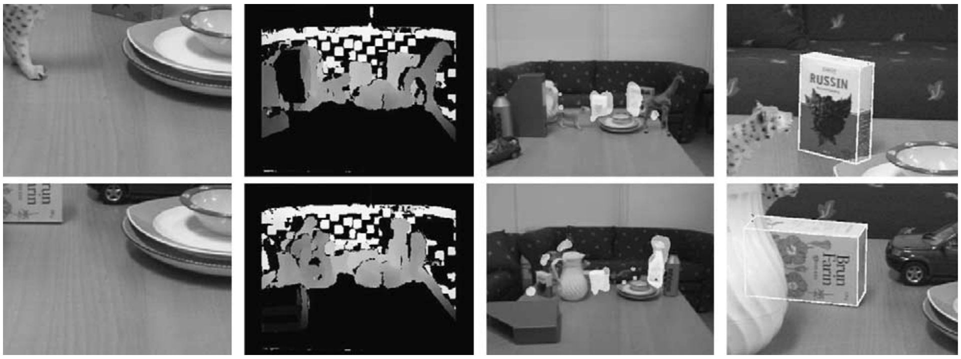

Fig. 9. An example of binocular figure-ground segmentation and pose estimation. The first column shows the foveal images before a saccade has been issued. Disparity maps can be seen in the second column and object hypotheses in third. The last column shows the estimated pose.

8.2. Monocular CCH recognition and pose estimation

Fig. 10 shows two examples of recognition and pose estimation based on monocular CCH. Here, object recognition and rotation estimation serve as the initial values for the model based pose estimation and tracking modules. With the incomplete pose calculated in the recognition (first image from the left) and orientation estimation step, the initial full pose is estimated (second image from the left). After that, a local fitting method matches lines in the image with edges of the projected object model. The images obtained after convergence of the tracking scheme is shown on the right. It is important to note, that even under the incorrect initialization of the two other rotation angles as zero, our approach is able to cope with significant deviations from this assumption. This is strongly visible in the second example where the angle around camera's Z-axis is more than 20°.

8.3. Robustness of disparity based figure-ground segmentation

As mentioned in Section 4, object location hypotheses are found slicing up the disparities into a binary map of pixels located within a given depth range. There are some evident disadvantages associated with such a procedure. First of all, an object might be tilted and extend beyond this range. This can be seen in the upper left image in Fig. 11—but it does not occur in the second image on the same row. However, since a more accurate localization is found through the focused attention process, a saccade is issued to the approximately same location. This is shown in the last two images on the upper row.

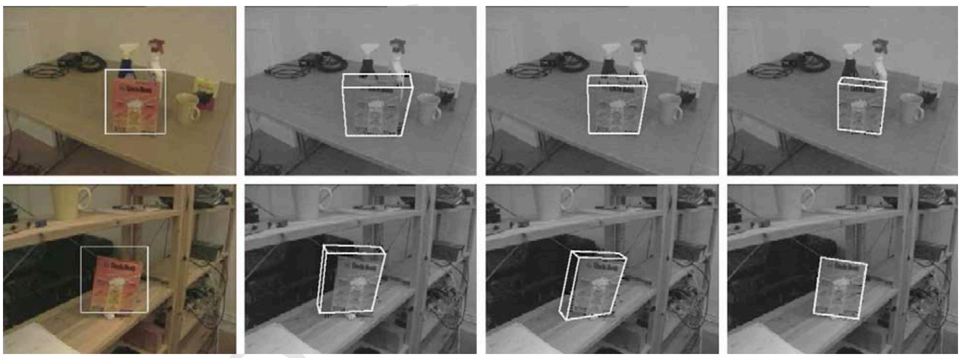

Fig. 10. From object recognition to pose estimation, (from left): (i) the output of the recognition, (ii) initial pose estimation, (iii) after three fitting iterations, (iv) the estimated pose of the object.

Fig. 11. The imperfect segmentation does not effect the final pose estimate of the object. The examples show when: (upper) Only a fraction of the object was segmented, and (lower) Two hypotheses are overlapping.

Another challenge occurs if two nearby objects are placed at almost the same distance, especially if the background lacks sufficient texture. Then the objects might merge into a single hypothesis, which is shown on the second row of Fig. 11. In our experiments this seemed more common when a global disparity method Kolmogorov and Zabih [23] was used and is the reason why we normally use simple area correlation. The global optimization methods tend to fill in the space between the two objects, falsely assuming that rapid changes in disparities are unlikely and thus should be suppressed. In practice, it is preferable if the textureless area between the objects are left unassigned. The right two images on the last row show that pose estimation is still be possible, even when hypotheses are merged. Depending on the density of foveal features, one of the two objects is automatically selected.

8.4. Robustness of SIFT based recognition toward occlusions

In a cluttered environment, a larger fraction of objects are likely to be occluded. These occlusions affect most involved processes, in particular those of recognition and pose estimation. The first two images in Fig. 12 show a scene in which the sugar box is partially occluded behind a bottle. In the first case, the recognition fails because not enough foveal features are available, while successful recognition and pose estimation is

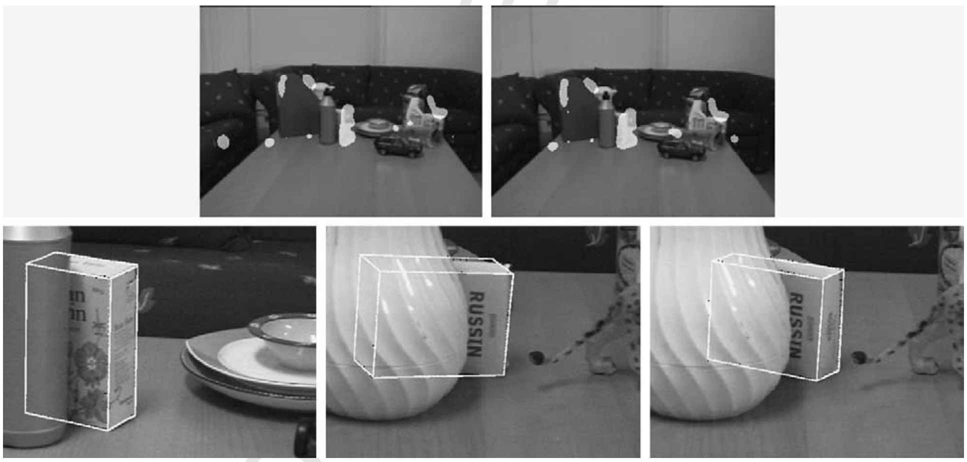

Fig. 12. The system is able to cope with situations where the object of interest is significantly occluded. Too much occlusion can however result in incorrect pose estimation (lower center).

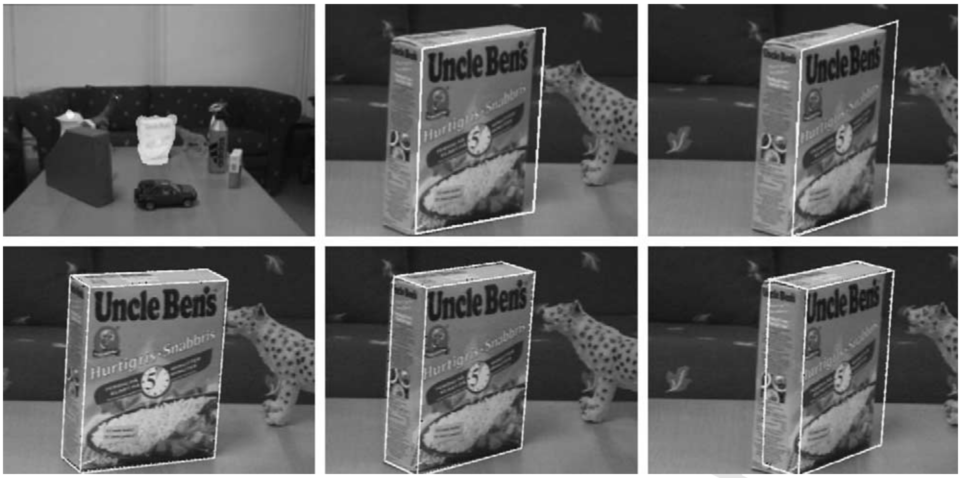

Fig. 13. From object hypotheses (upper left) the orientation of an object is estimated (upper middle/upper right). Pose estimates after three iterations for orientations 20°, 40° and 60° (lower).

possible in the second case as shown in the third image. However, even if recognition is successful, the pose initialization might still fail when not enough edges are clearly visible. This can be seen in the last two images of Fig. 12. As it is apparent from the fourth image that a failure does not necessarily mean that the results are useless, since the location of the object in 3D space is still available.

8.5. Robustness of pose initialization toward rotations

Since, in SIFT based recognition, only one view was available for each object, the sensitivity of the system to rotations was expected to be high. It is already known that for efficient recognition using these features, the relative orientation between query image and object model ought to be less than about 30°. Likely because our model set only consisted of eight objects, our study indicated that slightly larger angles were in fact possible. In the three columns of Fig. 13 an object was rotated about 20°, 40° and 60°, respectively. The rise package was correctly recognized at a score higher than 70%. However, the break-point turned out to be highly object dependent. For example, for an object like the tiger, the breakpoint was as low as 20%. For a more thorough analysis on the SIFT recognition performance we refer to Lowe [24].

As can be seen in the last two images on the upper row of Fig. 13, larger rotations tend to be underestimated when the pose is initialized. However, these errors are still below what is required for the pose estimation to finally converge. The lower row shows the estimated pose after a few initial iterations. Even at an angle of 60° the process will converge, but at a somewhat slower rate. For 40° and below convergence is reached within three frames.

9. Conclusions

In this paper, different visual strategies necessary for robotic hand-eye coordination and object grasping tasks, have been presented. The importance of camera placement and their number have been discussed and their effect on the design and choice of visual algorithms. For realistic, domestic settings we are interested in designing robots that are able to manipulate both known and unknown objects and it is therefore important to develop methods for both cases. We have shown strategies that support both cases.

Reflecting back to Fig. 1, different scenarios can be arranged in a hierarchy depending on prior information. Even if a particular task is given, it is possible to shift between different scenarios and therefore, the underlying strategies used. For example, if the command “Pick Up This Cup” is given, but the system fails to verify the existence of the cup, the execution may still continue as if “Pick up The Cup” was given. A vice-versa example is if the command “Pick Up This Object” was given and the system realizes that the object is, in fact, a known box of raisins. Then, the system automatically changes the task to “Pick Up The Raisins”. In the future, we want to develop a more formal description for the above, in order to design a visual system framework for robotic manipulation in general.

References

[1] S. Hutchinson, G. Hager, P. Corke, A tutorial on visual servo control, IEEE Trans. Robot. Autom. 12 (5) (1996) 651–670.

[2] D.H. Ballard, Animate vision, Artif. Intel. 48 (1) (1991) 57–86.

[3] M. Björkman, D. Kragic, Combination of foveal and peripheral vision for object recognition and pose estimation, Proceedings of the IEEE International Conference on Robotics and Automation, ICRA’04 5, 2004, pp. 5135–5140.

[4] S. Kim, I. Kim, I. Kweon, Robust model-based 3d object recognition by combining feature matching with tracking, Proceedings of the IEEE International Conference on Robotics and Automation, ICRA’03, 2003, pp. 2123–2128.

[5] E. Malis, G. Chesi, R. Cipolla, 2 1/2 d Visual servoing with respect to planar contours having complex and unknown shapes, Int. J. Robot. Res. 22 (10–11) (2003) 841–854.

[6] D. Kragic, H. Christensen, A framework for visual servoing, Proceedings of the International Conference on Computer Vision Systems, ICVS 2003, 2003, pp. 345–354.

[7] S. Benhimane, E. Malis, Vision-based control with respect to planar and non-planar objects using a zooming camera, IEEE IEEE International Conference on Advanced Robotics, vol. 2, 2003, pp. 991–996.

[8] S. Vinoski, CORBA: integrating diverser applications within distributed heterogeneous environments, IEEE Commun. Mag. 14 (2) (1997).

[9] M. Björkman, J.-O. Eklundh, Real-time epipolar geometry estimation of binocular stereo heads, IEEE Trans. Pattern Anal. Mach. Intel. 24 (3) (2002) 425–432.

[10] M. Björkman, Real-time motion and stereo cues for active visual observers, Doctoral dissertation, Computational Vision and Active Perception Laboratory (CVAP), Royal Inst. of Technology, Stockholm, Sweden, 2002.

[11] C. Harris, M. Stephens, A combined corner and edge detector, Proc. Alvey Vision Conference, 1988, pp. 147–151.

[12] M. Fischler, R. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM, vol. 24, 1981, pp. 381–395.

[13] P.J. Huber, Robust Statistics, John Willey and Sons, 1981.

[14] H. Longuet-Higgins, The interpretation of a moving retinal image, Philos. Trans. R. Soc. Lond., B 208 (1980) 385–397.

[15] H. Longuet-Higgins, A computer algorithm for reconstructing a scene from two projections, Nature 293 (1981) 133–135.

[16] S.E. Palmer, Vision Science: Photons to Phenomenology, MIT Press, Cambridge, MA, 1999.

[17] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intel. 20 (11) (1998) 1254–1259.

[18] M. Björkman, J.-O. Eklundh, Attending, foveating and recognizing objects in real world scenes, Proceedings of British Machine Vision Conference, BMVC’04, 2004.

[19] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000, 2000, pp. 142–151.

[20] M. Zillich, D. Roobaert, J.O. Eklundh, A pure learning approach to background-invariant object recognition using pedagogical support vector learning, CVPR-2001, IEEE, Kauai, 2001.

[21] D. Kragic, Visual servoing for manipulation: robustness and integration issues, Ph.D. thesis, Computational Vision and Active Perception Laboratory (CVAP), Royal Institute of Technology, Stockholm, Sweden, 2001.

[22] K. Konolige, Small vision systems: hardware and implementation, International Symposium on Robotics Research, 1997, pp. 203–212.

[23] V. Kolmogorov, R. Zabih, Computing visual correspondence with occlusions using graph cuts, Proceedings of the IEEE International Conference Computer Vision, 2001, pp. 508–515.

[24] D.G. Lowe, Object recognition from local scale-invariant features, Proceedings of the IEEE International Conference on Computer Vision (ICCV 99), 1999, pp. 1150–1157.

[25] S. Ekvall, F. Hoffmann, D. Kragic, Object recognition and pose estimation for robotic manipulation using color cooccurrence histograms, Proceedings of the IEEE/RSJ International Conference Intelligent Robots and Systems, IROS’03, 2003.

[26] D. Kragic, A. Miller, P. Allen, Real-time tracking meets online grasp planning, Proceedings of the IEEE International Conference on Robotics and Automation, ICRA’01 3, 2001, pp. 2460–2465.

[27] D. Kragic, H. Christensen, Weak models and cue integration for real-time tracking, Proceedings of the IEEE International Conference on Robotics and Automation, ICRA’02 3, 2002, pp. 3044–3049.

[28] D. Kragic, H. Christensen, Confluence of parameters in model based tracking, Proceedings of the IEEE International Conference on Robotics and Automation, ICRA’03 3, 2003a, pp. 3485–3490.

[29] P. Chang, J. Krumm, Object recognition with color cooccurrence histograms, Proceedings of the IEEE International Conference Computer Vision and Pattern Recognition, 1999, pp. 498–504.

[30] K. Mikolajczyk, C. Schmid, Indexing based on scale invariant interest points, Proceedings of the IEEE International Conference Computer Vision, ICCV’01, 2001, pp. 525–531.

Danica Kragic is an assistant professor at the Computational Vision and Active Perception Laboratory at the Department of Numerical Analysis and Computer Science at the Royal Institute of Technology, Stockholm, Sweden. She received the MS degree in mechanical engineering from the Technical University of Rijeka, Croatia, in 1995 and PhD degree in computer science from the Royal Institute of Technology, Stockholm, Sweden. Her research interest include computer vision, learning by demonstration, visual servoing, human-machine collaboration and path planning.

Mårtenrten Björkman received in a PhD in computer vision at KTH in Stockholm, Sweden in 2002. Between 1994 and 1997 he was employed by Mentor Graphics. He is currently active as a post-doc within the EC sponsored project MobVis. His primary research interests are stereo vision, cognitive vision systems and image based rendering.

Henrik I. Christensen is a chaired professor of computer science and the director of the Centre for Autonomous Systems at the Swedish Royal Institute of Technology, Stockholm, Sweden. He is also the co-ordinator of the EU network EURON. He does research on systems integration, mapping and sensory fusion. He has published more than 190 contributions on vision, AI and robotics. He serves on the editorial board of IJRR, Autonomous Robots, IJPRAI and AI Magazine.

Jan-Olof Eklundh graduated in mathematics at Stockholm University, 1970. He then joined the newly formed Laboratory for Image Analysis at the National Defence Research Institute, Stockholm, and spent 1977-1979 at the Computer Vision Laboratory, University of Maryland. In 1982 he became associate professor at KTH where he founded the Computer Vision and Active Perception Laboratory, CVAP. In 1996 he initiated the Center for Autonomous Systems, in which CVAP is now a key partner. 1986 he became professor in computer science and in 1995 Dean of the School of Electrical Engineering and Information Technology at KTH. His research interests cover a broad range of topics in computational vision, image processing, and robotics, especially active visual machine perception with relations to human vision, analysis of shape and geometry and motion, behavioral aspects of perception, and perceptually guided autonomous systems. He is and has been on the editorial boards of several journals, including IEEE PAMI, IJCV, CVIU and IVC and chaired ICCV 1990 and ECCV 1994. Professor Eklundh is a member of the Royal Swedish Academy of Science, the Royal Danish Academy of Sciences and Letters and the Royal Swedish Academy of Engineering Science.