# Chapter 2

# Related Work

## 2.1 Introduction

The problem of robotic object unloading from platforms, has received a high attention in the latest years, and a variety of system addressing it have been devised. Such systems can be broadly classified into two categories: Systems incorporating no vision at all, and systems incorporating vision.

The majority of the systems employed in industrial applications so far, do not contain any vision modules [1], [5], [93], [136], [51], [46]. They usually employ pre-programmed gantry robots for bulk depalletizing tasks. Since they are fast and accurate, such systems may be very effective in strictly controlled environments, but they fail in adverse environments e.g. in a distribution center, where the pallet's position is not well-defined, and the objects to be grasped may be arbitrarily jumbled due to human intervention.

Systems incorporating vision, usually acquire an image (or a set of images) of the object configuration. Given these images, image processing techniques are employed the target of which is the extraction of object grasping information. When acquired, object grasping information is forwarded to the robot which grasps and unloads the objects. The process intervening between image acquisition and object grasping is the core of vision- based object unloading systems, and will be hereinafter referred to as the *object recovery* process.

A solution to the object recovery problem is not an easy task, and the reason for this is that it comprises two sub-problems, addressing each of which is since years known to be strenuous: the image segmentation problem, and the object recognition problem. Solution to the former results into the isolation of the region in the image corresponding to a unique object. Solution to the latter results into the identification of the kind of the particular object, and is strongly related with a way in which the machine internally represents the target objects. A rough categorization of existing systems can be performed with regard to the technique they employ to acquire the data, the *principle* they use to perform segmentation of the input image, and finally the way in which they represent or *model* the target objects.

The purpose of this chapter is dual: Firstly, to give an overview of related work about object recovery from images. Since the literature dealing with this issue is vast, we will focus here on presenting systems for object recovery in images of multiple *three-dimensional* objects, for *robotic grasping* purposes. In particular, we will describe systems able to effectively deal with *box-like* objects, since this is the kind of objects related to our application. This is the subject of sections 2.2, and 2.3. More specifically, section 2.2 details on the issue of object recovery. Section 2.3 presents the principles governing a multitude of existing systems, and illustrates their advantages and drawbacks. The second purpose of this chapter, is to introduce our system and its operation, and discuss its advantages with respect to existing systems. This is the subject of section 2.4.

## 2.2 Multiple object recovery from images

Object recovery from images is admittedly a subproblem of the more general problem of machine perception [7]. Given a scene consisting of one or more objects, its target is the localization and identification of the objects that are sufficiently visible to the sensory system. As discussed in the preceding paragraph, key constituting parts of such systems, are the kind of images used for the properties, the image segmentation principle adopted, and the adopted object representations. In this section, we attempt a further categorization of each of these parts: Data acquisition techniques are further subdivided into passive and active. Object representation techniques can be divided in appearance based, and those based on geometric information. Finally, image segmentation techniques can be divided into techniques in which the recognition task is *decoupled* from the segmentation task, techniques according to which segmentation and recognition are *interwoven* and the *Hough transform.*

### 2.2.1 Sensory data acquisition

Target of a sensory system is to capture the scene information to the highest possible extend, in form of an *image.* Since our final target is recovery of three- dimensional objects, we will mainly discuss about techniques dealing with acquisition of three dimensional or *depth* information. There is a variety of ways to obtain such information, a detailed discussion on which can be found in [15], [17], [114], [33], [80], [53] p. 468. Roughly, two trends exist for acquiring three dimensional information: the *passive* and the *active techniques*

According to the passive techniques, no energy is projected in the environment by the sensor. These techniques employ intensity cameras to acquire grey-level or color images of the scene. A variety of approaches have been so far devised for acquiring three dimensional information from intensity images, for example shape from shading, shape from texture, shape from contour, shape from binocular stereo, shape from photometric stereo, shape from motion, shape from shadows (see [17] for a detailed discussion). The advantages of these techniques are the low cost of the intensity sensors, as well as that they are readily available on the market. However, a major problem is that the robustness of these techniques is questionable, since they are heavily sensitive on the lighting conditions occurring during data acquisition. Besides, they have often considered inadequate for reconstructing the dense and accurate $3D$ data required for many tasks [8]. Finally, the computational costs for shape reconstruction does not usually allow for depth data acquisition in real time.

Active approaches emit energy in the environment for acquiring data in form of a *range image.* A range image is a large connection of distance measurements from a known coordinate system which will be hereinafter referred to as the *sensor coordinate system*, to surface points on objects in the scene. Range images are known by many other names, depending on the context: range maps, depth maps, depth images, or 2.5D images [15]. Active data acquisition approaches can be divided into two categories: laser radar based, and structured light techniques.

The former, transmit a laser beam toward the scene, and part of the light is reflected back to the sensor. The sensor acquires depth information by processing the reflected beam. A rotating mirror deflects the light beam in such a way that the entire scene can be scanned in a raster fashion. This strongly resembles the operation principle of radar devices, and is the reason why these sensors are frequently referred to as *laser radars* (ladars). Laser radar techniques can be further subdivided into three categories according to the way in which the reflected beam is processed: Time of flight laser radars acquire depth information by measuring the transit time of the beam. Frequency and amplitude modulation radars utilize alternative ways of processing. The reader is referred to [15] for more details in the issue. According to the structured light techniques, the scene is illuminated with a pattern of light or a laser beam. A camera registers the illuminated scene points. The pixel coordinates of the illuminated points are subsequently converted into three dimensional coordinates, using appropriate triangulation formulas. The structured light techniques are subdivided into techniques using lines, grids, and coded binary patterns. More details about the operation and properties of structured light techniques can be found at [15], [7], [80].

Active techniques for acquisition of three dimensional scene information are considered to be more robust than passive techniques. The main reason for this is that active techniques in contrast to passive techniques do not require additional processing for delivering three dimensional information. As far as the active techniques are concerned, the structured light methods exhibit drawbacks with respect to the laser radars. The structured light techniques depend on the lighting conditions: They require that the illuminated pattern is the only illumination source for the scene. This is the reason why such image acquisition techniques can not be used outdoors. In addition, calculation of the triangulation parameters needed for determining the three dimensional coordinates of an image pixel is usually a laborious process. This is not the case for laser radars. Besides, laser radars do not suffer from occlusion problems that can sometimes reduce the effectiveness of the structured light sensors: According to the structured light techniques, an object must be visible to *both* the illuminating source and the camera. For the laser radars, it is sufficient for the object points to be visible to just the laser source since no additional camera is required. The main disadvantage of the laser radars is their high cost, which is mainly due to the complexity of their hardware.

### 2.2.2 Object representation

Automatic object recognition via a computer, requires the existence of object representations internally to the machine, which usually suggest a mechanism in which object recognition can be performed. Existing techniques for object representation can be very roughly classified into *appearance-based*, *geometric model-based* techniques, or combination of both. Geometric model-based object representation is based on geometric features, whereas appearance based representation uses a large set of images for training, but does not require any insight on the geometric structure of the objects.

Appearance based techniques, compute a vector of global object attributes or *features.* There are several ways in which these vectors can be generated. They can be defined via spatial frequency descriptions such as the discrete cosine transform, Fourier descriptors, wavelets or eigen-images, or even on features which are invariant to a certain group of rigid or non-rigid transformations [141], [144]. The process of feature vector generation, is known as the *training* or *learning* process. According to the recognition process of appearance based techniques, given a segmented part of an image corresponding to one object, it is analyzed in the same way as in the training phase, and the feature vector is calculated. Then the distance of the feature vector to feature vectors of objects calculated in the training phase is computed. The internal representation having the smallest distance from the input feature vector is assumed to be the one corresponding to the examined object. The advantages of appearance based techniques is their computational efficiency. However the recognition process is heavily depended on the result of the segmentation process. The latter is likely to be inaccurate in the event of clutter in the image. In such a case, the robustness of these techniques significantly drops [7]. Since our application involves images with cluttered objects we decided to avoid using appearance based technique in our recovery process, but rather employ geometric model based techniques, which are known to be more robust against clutter.

In the context of the geometric model-based techniques, the objects are represented in terms of their geometric properties. Numerous model-based object representation techniques are reported in the computer vision literature. The reader is referred to [16], [123], [24] for excellent reviews on the issue. A big amount of work towards the representation of free form objects has been performed (e.g. [17], [151], [82]). The advantages of such techniques is their generality. However for objects with simple shape, application of such strategies for representation introduces unnecessary computational costs. For this reason, we will deal with popular approaches adopted for representing relatively simple shapes. Such techniques can be broadly classified into two categories: Techniques where the representation is an aggregation of lower level geometric features, and techniques where the representation is parametric.

The former techniques [31], [65], [160], [7], employ special data structures, in which simple geometric entities such as a vertex, an edge or even a low parametric surface can be stored. The most fundamental structure used for this purpose is the *attribute graph*, the nodes of which represent each of the low level features, and the edges represent adjacency or any other relationship between the features. The process of constructing such models is usually interactive: The user constructs the attribute graph by establishing correspondences between the features contained in the graph and the values they have in the target objects. The recognition task is realized as follows: Given a segmented part of an image which correspond to a unique object, geometric features are extracted. The extracted geometric features are of the same kind with the features used for creating the models. Each of the image features, could belong to more than one model. Moreover, if it is given that a specific image feature corresponds to a specific model, there might be various ways in which the model could be placed in space so that the image feature exactly matches with its corresponding model feature. For this reason the image features are *grouped* into feature sets, in such a way so that each feature set comprises exactly the number and the kind of features, which uniquely determine the identity and pose of the model in space. Given a feature set, the pose and identity of the object are determined by an *alignment* process. The computed model pose and identity comprise an *hypothesis* about the real values are then *verified*, by examining if it coincides with all image features. This framework for object recognition is widely known as the *hypothesis generation and verification framework*, and has been successfully applied with small variations in a plethora of working systems [126], [31], [65], [160], [86]. The main advantage of the hypothesis generation and verification techniques is that they are much more robust in the presence of occlusion than their appearance based techniques [7]. There is a trade-off however between robustness and computational efficiency in such techniques: If robustness is required, the number of features that should be examined in the verification phase is high [66], hence the computational costs are increased. Inversely, when fast responce is required, the robustness of the system considerably reduces.

According to the parametric model representations, the model is described by a function having a set of parameters, which completely define its shape and its position in space. Popular three dimensional parametric models are the B-splines (NURBS) [91], generalized cylinders [19], Geons [18], spherical harmonic surfaces [140], Fourier surfaces [150], symmetry seeking models [155], blob models [111], Hyperqadrics [69], implicit (fourth degree) polynomials [89], as well as superquadrics [121], [148], [99], [77]. The object recognition task in the case of geometric parameter models is equivalent to a parameter estimation task: Given a segmented part of an image corresponding to a unique object, the parameters of the models can be computed using standard estimation techniques like likelihood maximization. Once estimated, the parameter values describe both the identity and the pose of the object. The parameter estimation process is in general both robust and computationally efficient, when the number of parameters of the employed models is low. However, models with a low number of parameters have limited expressive ability. If we want to describe a bigger variety of shapes, then more complicated models are needed, which not only reduces the robustness of the parameter estimation, but its computational efficiency as well. However, as we will see shortly, the big advantage of using parametric geometric models is that image segmentation can be *interwoven* with parameter estimation, in such a way so that the robustness of both processes increases.

### 2.2.3 Image segmentation

Image segmentation techniques are strongly related to the models used for recognizing the objects in the image, and the reason for this is that target of these techniques is to isolate image areas corresponding to objects described by the models. Since, as shown in the previous section, geometric model-based object representations can provide more robust recognition results in the event of images with clutter, and since our application should be able to effectively deal with clutter, we will focus here on image segmentation techniques usually applied when geometric models for object representation are utilized. Such approaches are roughly divided into the following three categories [7], [77]. Techniques according to which the model is not directly involved in the segmentation process, techniques according to which the model is involved in the segmentation, and the Hough transform.

The former techniques *decouple* object recognition from image segmentation [126], [31], [65], [160], [168], [90], [86]. These segmentation techniques, usually employ some low level, local criteria, usually concerning pixel homogeneity, in order to extract low level image *features.* Hence, target of image segmentation in this case is feature extraction rather than complete object isolation. Such approaches are adopted by systems for which the models are constellations of low level features. For such systems, object recognition follows the feature extraction and is performed using the hypothesis generation and verification strategy. Usually, object recognition from low level features can be a fast process. However, as discussed in the previous paragraph its robustness is questionable, when computational efficiency is required.

Techniques in which object recognition and image segmentation are *interwoven*, are more robust [99], [77], [173], [119], [142], [34], [131], [30] [106], [113]. These are used in conjunction with geometric parametric entities for object modeling. The idea behind these techniques, is that the models can be directly recovered from the data without the meditation of low level features, so that problems of mismatch between features and models are avoided. They pose the recovery problem as a parameter *optimization* problem and perform the segmentation task by maximizing the *posterior probability* of the parameters of all models given the image data. This is the reason why these techniques are usually referred to as *probabilistic approaches* for image segmentation. The robustness of these techniques stems from the fact that segmentation is based on taking optimal statistical decisions [15]: by estimating the model parameters via maximization of the posterior, the probability of error in the segmentation or rather the Bayes *risk* of not correctly segmenting the image is minimized. Iterative global optimization approaches are employed for the maximization (e.g. *simulated annealing* [61], *graduated nonconvexity* [20], *deterministic annealing* [60], *mean field annealing*, [59], and *Hopfield neural network* construction [73]), which are in general computationally expensive. Fortunately, the computational efficiency can be considerably increased when we are given a good *initial value* of the sought model parameters. This initialization is considered to be a *hypothesis* on the parameter values, which is then *refined* by maximizing the posterior using *local* and thus efficient optimization approaches (e.g. gradient descent, Levenberg- Marquardt [64]). The framework we roughly described will be hereinafter referred to as the *hypothesis generation and verification* framework for image segmentation, and an in depth discussion about the way in which it is used for box-like object recovery, will be performed in chapter 5.

The *Hough Transform* (HT) [75], [96] is as well a technique according to which both geometric parametric models are employed, and segmentation and parameter estimation are interwoven. The key idea underlying the Hough transform, is to map a the difficult segmentation problem into a simple *peak detection* problem in the model parameter space. Maxima in the parameter space correspond to possible model instances. The major advantage of the method is that when the number of model parameters is small, it combines robustness and computationally efficiency without requiring some initialization of the sought model parameters, as is the case for the probabilistic segmentation approaches. When the number of parameter increases, as is the case for almost all three dimensional models, the HT shows problems like high memory consumption, computational inefficiency, and sensitivity to errors in localization of image points. In some cases however, it is possible to recover three dimensional models by *decomposing* the Hough transform in lower dimensional subproblems [116], [117]. Chapter 7 shows how we applied a Hough transform decomposition to recover multiple cardboard boxes from range images.

## 2.3 Existing systems

We are now going to attempt a presentation of systems dealing with recovery of multiple objects for robotic grasping operations. The presentation here will be done according to the sensor type used for the data acquisition. We will first start with systems using intensity sensors for data acquisition and then describe the ones using range imagery. For each system, we focus on the description of its object recovery part, based on the categorizations presented in the preceding paragraphs. In addition, when possible we comment on its computational efficiency accuracy and robustness, and discuss its advantages and problems.

In [112], mainly intensity imagery is used for recovering neatly placed bags (sacks). An intensity camera is placed above the pallet. The camera recognizes logos on the surfaces of the sacks. Thus the coordinates of the center of gravity of the exposed surface of the object in the image are calculated. The depth of the object is acquired by a proximity sensor which is placed above the pallet as well. A gripper based on vacuum principles is used for grasping. No robustness, computational efficiency or accuracy measurements are presented.

The system in [88], attempts recovery of neatly placed bags as well. In general, the same principle is used for object localization as in [112]. The main difference is that the boundaries of the objects instead of logos in their exposed surfaces are employed for the localization. In addition, it is assumed that the number of objects in the pallet is known. This assumption, although results to an increased robustness of the recovery algorithm, limits the application range of the system to a considerable extend. Edges and logos on the exposed surfaces of neatly placed boxes and bags are as well used for localizing neatly placed bags and boxes in [165]. A grasping accuracy of $1mm$ in $x$ and $y$ directions and $2cm$ in the depth direction $z$ are mentioned. In [49], a robotic system is demonstrated which is able to depalletize flat, neatly placed cardboard boxes. The system employs a unique intensity camera for localizing the objects. The height of the workpieces in this system is calculated according to the size that the objects have inside the intensity image. This implies that the dimensions of the objects in the platform are known in advance. We have to note at this point that in all three [88], [165], and [49], no statements about the robustness of the systems are reported.

[76] attempts automatic unloading of jumbled sheet metal parts from a platform. The sheets in the pile are rectangle objects with the same dimensions. Object grasping is performed using a magnetic gripper installed at the hand of the robot. The authors assume that due to the fact that the height of the objects are very small, they all reside at the same distance from the camera. If so, only images from one CCD camera above the pallet are enough for object localization. According to [76], three images from the configuration are obtained, each one with different lighting conditions, in order to eliminate unwanted illumination effects. Then, edge detection is performed in each image, and the edge pictures are merged in one image. Further, image filtering is performed in order to remove noise and problematic image features like holes on the surfaces of the objects. Subsequently, a distance transform is employed [147], [39], and the transformed picture is binarized. The binarized image is then used for blob analysis in order to find candidate positions for grasping. Output of this procedure is a number of regions in the image, each corresponding to an object that can be grasped by the robotic hand. The object corresponding to the region with the largest area is selected for grasping. The authors report that the execution time of the recovery framework was 2.5

seconds in a Pentium 166MHz PC. No accuracy or robustness measurements are presented. In the last section of their paper, they mention that more work towards the insensitivity of their system to the lighting conditions is required. In addition they point out the problems their system encounters when the height of the objects is not negligible, and when tilted sheets are attempted to be grabbed. Finally, they comment on the need of a system which will be able to successfully deal with objects of arbitrary dimensions.

In [126], a camera stereo pair is employed for recovery of jumbled alternator covers. A region based segmentation method is applied to each image of the stereo pair to extract *features* on which the recovery is based. Such features are the large bearing alternator hole, referred to as seed feature, and the set of small screw holes at the perimeter of the bearing hole, referred to as supporting features. Extracted features are inputs of a stereo correspondence module which matches estimates from both views and determines their pose. This module first looks for correct correspondences between available seed features from the left and right images using the epipolar and geometric constraints. Then for each accepted pair of seed features, the module optimizes the correspondence by associating supporting features. Model alignment with the data in the images determines the position in the model. For each matching, appropriate motion commands are generated for robotic grasping. The accuracy of the recognition is mentioned to be $7mm$ in translation and 10 deg in orientation. The processing time is 1.5 min per image pair on a SUN Sparc 1000 server. In terms of robustness, the system demonstrates a true positive rate of 91.3 % and a false negative rate of 8.7 %. In no case the algorithm recognized some arbitrary object as a target object, which implies a 0 % false positive rate. The authors did not experiment with actual grasping of objects. Nevertheless their intention is to use a parallel jaw gripper for object grasping.

In [152], a pair of intensity cameras is used for recognition of jumbled objects, using *segment-based* stereo vision. This method can deal with almost any type of object, such as boxes of arbitrary dimensions, polyhedra, bodies of revolution such as cylinders, and free form objects. However, the system is efficient from the computational view, only when dealing with planar objects. The recovery approach used in the system is based on boundary information. Firstly, edge detection on both images acquired by the stereo camera pair is performed. The edge images contain object boundaries, but gaps in the boundaries exist, due to noise and illumination. Subsequently, an *edge-linking* operation is performed, the target of which is to close the gaps in the boundaries and form closed contours. Then, the closed contours in the image are segmented into straight, convex or concave segments. The image points between these segments and the segments themselves are the features upon which the reconstruction of three dimensional information, as well as the recognition is based: These features are used to construct a *boundary representation* of the closed contour to which they correspond, which resembles the attribute graphs discussed in the previous section. Three dimensional information is obtained by finding corresponding boundary representations in the two images, and intersecting them in space. Matching the three dimensional boundary representations with the boundary representations of the models acquired by means of a training phase, results into the generation of hypotheses about the identity and pose of the objects in the image. These hypotheses are then refined using an Iterative Closest Point (ICP)- type algorithm, which is guaranteed to converge to the real pose of the object, since the pose obtained during the hypothesis generation is close to the true object pose. Measurements on the computational efficiency of the system in the event that planar objects are considered are very encouraging: In a double processor UltraSPARC-2, 400Mhz, the processing time for $3D$ reconstruction, hypotheses generation, and hypotheses refinement using ICP was 1.9, 0.1, and 0.6 seconds respectively. However, when the objects are not rigid the computational efficiency drops significantly. Regarding reconstruction accuracy, the authors give data concerning the minimum and maximum standard deviation of the position of a polyhedral object in the $Z$ axis of the world coordinate system, obtained after processing a set of measurements which is 0.52mm and 0.21mm respectively, and the standard deviation of measurements concerning rotation along the $X$ and $Y$ axes, which was found to be 0.51, and 0.61 degrees respectively. For non planar objects the accuracy considerably degrades. No statements about the robustness of the system are reported in the paper.

The system [71], attempts localization of boxes and bags, which are placed in distinct layers, but are allowed to have an arbitrary orientation within a layer. A pair of intensity cameras are used for acquiring depth information. The authors question the robustness of feature based scene reconstruction, and for this reason perform depth acquisition in a different way: They project light patterns into the scene using a special light projector. Given the acquired pair of images, they obtain depth information via *disparity* calculation. The advantage of this technique is that depth reconstruction can be achieved even when no texture is present on the surfaces of the objects. The output depth image of the scene is firstly roughly segmented: Object candidates are extracted with a template matching technique. Given the extracted candidates, the exact position of the objects is extracted via a high resolution grey-scale image, obtained by one of the two cameras used for the acquisition of the range image, using again template matching. The authors report accuracy of 10.7 and 16.1mm in the $x$ and $y$ directions and 32.8mm in the depth direction, and a recovery success rate of 99.8%. The execution time of the recongition system is reported to be 5 seconds. However, no hardware platform on which the computational efficiency measurements have been performed is mentioned. In addition, the authors show no images of bags, so it is not clear if experiments with such objects or only with rigid boxes have been conducted.

In [63], recovery of rectangular and triangular shapes in cluttered configurations is attempted. In this system, range data acquisition is based on depth from defocus, which is a passive technique for obtaining depth information using intensity images. More specifically, the depth from defocus techniques use the direct relationship between the depth of the scene, camera parameters and the degree of blurring in several images to acquire depth information. In the particular implementation, two images from one intensity camera have been utilized, to acquire a depth image, which is registered with the intensity image in a common sensor coordinate system. The recovery process is as follows: Firstly, edge detection in the intensity image is applied and regions are extracted by edge linking. Then, region attributes are calculated from the range image. Perimeter, area and minimum and maximum distances from the regions centroid to the regions borders are extracted. Subsequently, using these attributes the region corresponding to the exposed surface of the object lying on the top of the pile is extracted. Based on the attributes of this region, the identity of the object is determined. Then, appearance based techniques are used for determining the pose of the object. The system is able to recover only the top most object each time although more graspable objects may exist. The reason for this is that appearance based techniques have a limited robustness against occlusion. The authors present experiments in a controlled laboratory environment with non textured objects.

All systems examined up to now employ intensity imagery for object recognition and exhibit advantages like high accuracy computational efficiency and low cost. However all these systems are very sensitive to the lighting conditions occurring at the installation sites. It is noteworthy that all experiments performed with the systems were conducted in environments with strictly controlled lighting and/or with dark background. A description of systems employing range imagery is now presented, which in general do not suffer from such problems.

In [92], a new solid state camera, based on technology with a laser flash and an ultra fast electronic shutter which sits in front of a standard CCD chip is used to acquire range images of cluttered card-board boxes. Despite the fact that the camera acquires a grey-value image co-registered with the range image, the former is not employed in the object recovery process. The recovery process involves a segmentation of the range image into planar regions, using the well known approach [79]. Then the boundaries of the planar regions are extracted. Subsequently, extracted contours are used to find straight line segments and corners. Given these segments, box localization is performed by means of a probabilistic graph matching technique. The features of interest for the matching is the length of the linear segments found in the contours of the range image, which correspond to nodes in the graph, and the graphs extracted from the range image are matched to the graphs corresponding to the model objects. This of course assumes that this approach can only work if the dimensions of the boxes expected to be found in the configuration is known in advance. Given the centers of gravity of the matched objects, and the orientation of the plane inside the matched contour, the robotic hand grasps the object. The authors conduct experiments with a Pentium 200MHz PC, and report an overall image processing time of less than 2 seconds. In addition, they report an accuracy of the grasping position of 1 cm for all three directions $x, y, z$ relative to the ideal grasp position. No robustness measurements are reported. According to our view, the main problem of the approach is that despite the recovery is based on the boundaries of the objects, a region based segmentation approach is used to determine this boundaries, which results to inaccuracy in boundary localization and thus reduces the effectiveness of the overall recovery framework.

In [160], a system for recognition of jumbled postal parcels and box-like objects of unknown dimensions is described. The authors attempt to overcome scene over-segmentation problems and deal with object recognition, object dimension detection and complete scene interpretation at the same time. Range data are acquired with a structured light approach based on single-slit projection. A region based data segmentation approach detects surface regions. These surface regions generate object location and dimension hypotheses, which are verified, refined or rejected by examination of hypotheses generated by other scene surfaces. The dimensions of the objects are calculated via a geometric reasoning process, that is by virtually extending objects dimensions initially hypothesised until they contravene other object hypotheses. The control structure of the system is rather complex. Computational efficiency measurements are not presented by the authors, but is expected to be high. Moreover, generating object location hypotheses from single surfaces is somewhat problematic, since surface information alone, does not satisfy the necessary constraints for accurate formation of an object location hypothesis. This may result in non accurate object detection.

One of the fastest recognition systems for 3D convex objects developed so far, is described in [31]. Structured light is used for data acquisition. Neighboring features are grouped into sets, which are named local feature sets (LFSs). The matching of a model LFS with a scene feature group assumes the existence of the model in the pile and generates a unique value for the pose transform which takes the particular model object to the scene. This model location hypothesis is then rejected or verified by checking if the position of neighbouring scene feature groups match positions of other model LFS. The authors have chosen to use 3D vertices as LFSs, since they provide constraints for calculating the pose transform. The accurate calculation of the vertex position is of extreme importance for the accuracy of recognition. The authors employ a region based range image segmentation technique and the vertex position is determined by intersecting surfaces. In the event that the objects expose three surfaces to the sensor, the calculation of the vertex point is as accurate as desired. The problem is that in many object configurations this is not the case. If only two surfaces of the object are exposed, no reliable method for computing the coordinates of the vertex points is proposed. What is more, in configurations where objects expose only one surface, the calculation of vertex points based on surface information is impossible. In [84], the authors claim that the principle used in [31] can be used to create a system for automatic unloading or rigid boxes. In [86], an experimental system based on [31] is presented. However, this system can only deal with rigid convex objects. In the event that deformations in the objects occur, the vertex detection process delivers inaccurate results and the reliability of the system decreases.

In [65], recognition of various jumbled, uniformly colored square and round shaped objects in piles is described. A prototype sensor which acquires registered range and intensity images is used. The system makes use of color information as an important attribute for distinguishing between surface features. A hash table which hashes over the object color attribute is used in order to decrease the number of hypotheses that a scene group of features generates. In this respect improvements on the speed of the system with regard to the system [31] are achieved. For geometric feature detection (e.g. vertices) as well as for hypothesis verification the system works in the same way as [31]. Concerning the system robustness, the authors present results according to which the system achieves 70 % true positive rate. In 30 % of the cases no objects at all are detected. Since the system uses the feature detection philosophy of [31] it inherits the problems of this system mentioned in a preceding paragraph. This means that if the objects do not expose three surfaces to the source no vertex can be detected. This explains the high percentage of no identification of objects.

In [90], single-slit structured light projection is used for detecting a variety of objects in piles. In the feature extraction phase, both surface regions and edges are extracted and attributes like surface area, centroids, normals, edge lengths etc are calculated. Object recognition is based on graph matching. A graph for the scene object is created. Nodes of the graphs are the detected surfaces, whilst the arcs represent the detected edges. Since the system is model based, graphs for all the models in the model database are created off-line. Bipartite graph matching is used to find distinct models whose features match with every scene feature. For all possible complete matchings, only those that correspond to feasible transformations between scene and model objects are extracted. The authors present good results for single object scenes. In occluded environments however, partially occluded scene features are declared matchable with a larger number of model features, which reduces the system robustness. Nevertheless, when the number of objects in the library is large and each object possess a large number of surfaces with different attributes, this method could be used for implementing a fast recovery approach.

In [9], bar code parallel structured light projection is used as a range data acquisition method of a system that deals with detecting and grasping postal parcels in piles. This system performs object recovery by finding planar visible object surfaces in the input range image as follows: Firstly, edge detection is applied to the range image. Output of the edge detection step are jump edges (see chapter 4 for a definition of various edge types in range images). Secondly, a connected component labeling algorithm is applied, to find connected regions in the range image. For every detected region a verification step is executed, target of which is to determine if the particular region corresponds to a planar surface. This is done by fitting a plane in the region and examining if the fitting error residual has a small value. Note, that since the edge detection algorithm is not able to detect crease edges, one non planar region may contain range points corresponding to two or three visible sides of the same box. The planar regions contained within a non planar region, if any, are acquired by means of a technique based on clustering of the surface normals, resulting in connected components which have about the same surface normals, thus correspond to planar object sides. Given all extracted planar region, the ones which correspond to graspable object sides are found. These are considered to be rectangular planar regions, and are determined via a rectangularity test to the input regions.

The graspable objects are grasped by the robot, using a parallel jaw gripper. Additional proximity sensors placed on gripper, verify if the object is actually grasped or not. Using this approach, and a computer specialized for image processing tasks (SYDAMA-2) the authors managed to perform both processing and grasping on about three seconds per parcel. Regarding robustness, the percentage of correct identifications is mentioned to be 80 %, and of false identifications 20 %. In the event that no objects are detected, the robot is commanded to disturb the pallet and therefore target objects risk being damaged. However, no grasping accuracy measurmenets are mentioned. In our opinion, the grasping accuracy is not expected to be high, since it depends on the accurate recovery of the dimensions of the boxes. The accuracy of the latter, depends on the accuracy in which the boundaries of the boxes are recovered. This is done using region information (connected components and clustering), which when used for this purpose its is known to deliver results of low accuracy. We will deal with the drawbacks of region based approaches when used to recover boundary information in chapter 5.

In [158] a system is described whose goal is grasping of piled, rigid, piecewise smooth objects. Range imagery is used for this purpose. Two range sensors using structured light generated by commercial stripe projectors are employed. Two range images of the configuration are acquired, each from a different viewpoint, so that enough depth information related to the entire area of the exposed surfaces of the objects is obtained. The recovery task involves segmentation of both images into *planar* patches. The recover-and-select framework is employed for this purpose [98], [99], a detailed description of which is given in section 5.1.3. Then, neighboring patches of each image are grouped into object hypotheses. Subsequently, object information in each image is merged with the corresponding information in the other, and global object hypotheses are created. This information is passed to the grasping process: The objects are ordered according to the height of their highest constituent patch, *antipodal* planar patches are determined for the highest object, and the pair of patches which is the nearest to the object's center of gravity is chosen. Objects are grasped from the antipodal patches using a parallel jaw gripper. The authors experiment with only one configuration of non-heavily occluded objects. As far as computational efficiency is concerned, the authors report a duration of 80 seconds in total for both data acquisition and object recovery per image, in a Sun SPARC 2 workstation. No experiments regarding robustness or accuracy are presented. The authors point out that increased efficiency is expected when the target objects are planar, for example boxes. In our point of view, the problems of this system can be summarized as follows: Firstly, the recover-and-select framework as is, does not allow for real-time object recovery. Secondly, as we will see in section 5.1.3, the recover-and-select framework has considerable problems in correctly classifying range points in the area of object boundaries, which implies that the system's grasping accuracy is expected to be low.

From the description of existing systems conducted in the preceding paragraphs, we can draw the conclusion that none of the works attempting multiple object recovery from images for grasping purposes, combines all characteristics pointed out in section 1.2. Our approach for addressing this problem meets these requirements. The way in which this is achieved is described in the subsequent section.

## 2.4 Our system

In this section we present our approach for automatic unloading of box-like objects in piles. Section 2.4.1 presents the hardware components of our robotic system. Subsequently, section 2.4.2 discusses the way in which images of the configuration are acquired. Finally, in section 2.4.3, we present the ways in which object recovery is performed within the context of our application.

### 2.4.1 Hardware

From the hardware point of view we employ a six degree of freedom KUKA 15/2 industrial robot, with very high repeatability (accuracy) of 0.1mm. Besides, we employ a vacuum gripper for grasping the objects. For data acquisition we use a SICK LMS200 laser scanner, which is a laser radar based on the measurement of time-of-flight technique. A pulsed infrared laser beam is emitted and reflected from the object surface. The time between the transmission and the reception of the laser beam is used to measure the distance between the scanner and the object. The device, shown in fig. 2.1, incorporates an internal coordinate system, along the plane $\mathbf{X_s}$-$\mathbf{Y_s}$ of which all measurements are performed. As the laser beam rotates, a set of two dimensional coplanar points $(x, y)$ are acquired and placed in an one dimensional array, which will be hereinafter referred to as *scan line*. Each of these points expresses the distance measured in $mm$ between the origin of the sensor coordinate system and the particular point of an object in front of the sensor which reflects the laser beam. The sensor is connected to a computer via a high-speed $RS422$ interface card with transfer rate up to 500 Kbaud. The time required for generation and transfer of one scan line from the sensor to the computer is about 53ms. The scanner can measure ranges up to 8m with ±15mm system error and 5mm standard deviation, with a cost of about 3KEuro. The reader is referred to [167] for more details on the sensor characteristics.

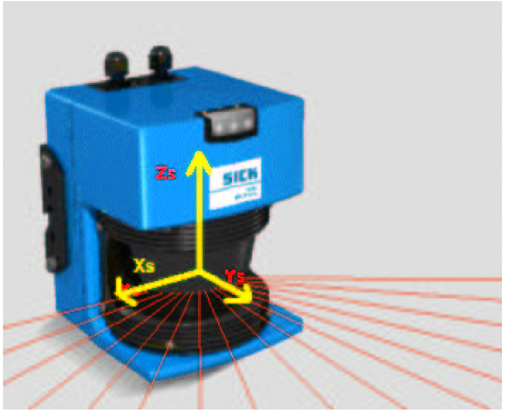


Figure 2.1: The SICK LMS 200 laser sensor

In order to take full advantage of the flexibility for viewpoint selection made possible by a six-degree-freedom robot, the sensor is mounted on the hand of the robot. Hence, within the constraints imposed by manipulator kinematics, the sensor can be oriented in any direction deemed desirable by the robot for the task at hand, and scan either in a translational, or a rotational mode for data collection. In addition, a square vacuum-gripper, is as mounted on the hand of the robot, which grasps the box-like target objects from one of their sides. This configuration is illustrated in figure 2.2. The operation of our system can be summarized as follows: Firstly a range image of the object configuration is acquired. Then the object recovery stage, determines which objects can be grasped. Finally the coordinates of the grasping positions are send to the robot which grasps and removes the objects. The cycle comprising data acquisition, object recovery and object grasping continues until the pallet is empty. In the following paragraphs, each of the three constituting components of our system is discussed in detail.

### 2.4.2 Data acquisition

A range image of the object configuration is acquired by linearly moving the hand of the robot above the pallet with a constant velocity. During this movement, the scanning plane of the sensor is perpendicular to the movement direction, and the sensor faces the upper side of the pallet. This is illustrated in fig. 2.3. More specifically, fig. 2.3 (a), shows the starting position of the data acquisition phase. The origin of the sensor coordinate system is placed about 2m above the pallet. The $\mathbf{Y_s}$ axis of the sensor coordinate system, which will be hereinafter referred to as *depth* axis points towards the ground. The intersection of the depth axis with the rectangle pallet at the starting position, is the mid-point of the small side of the pallet. The particular position is regarded to be the origin of the sensor coordinate system in our application. The end position of the data acquisition phase is shown in fig. 2.3 (b). During the movement from the start to the end position, a set of scan lines is acquired, which yield the range image. Each scan line corresponds to a row of the range image. In the context of our application, a range image comprises 241 scan lines. The number of points within each scan line, or the number of columns of an image is 369. The distance that the robotic hand traverses between two subsequent scans is 7mm. The total

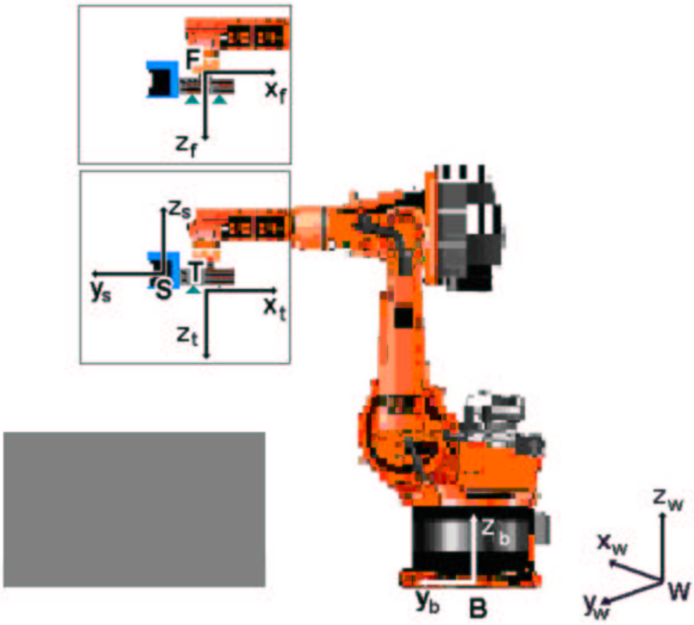


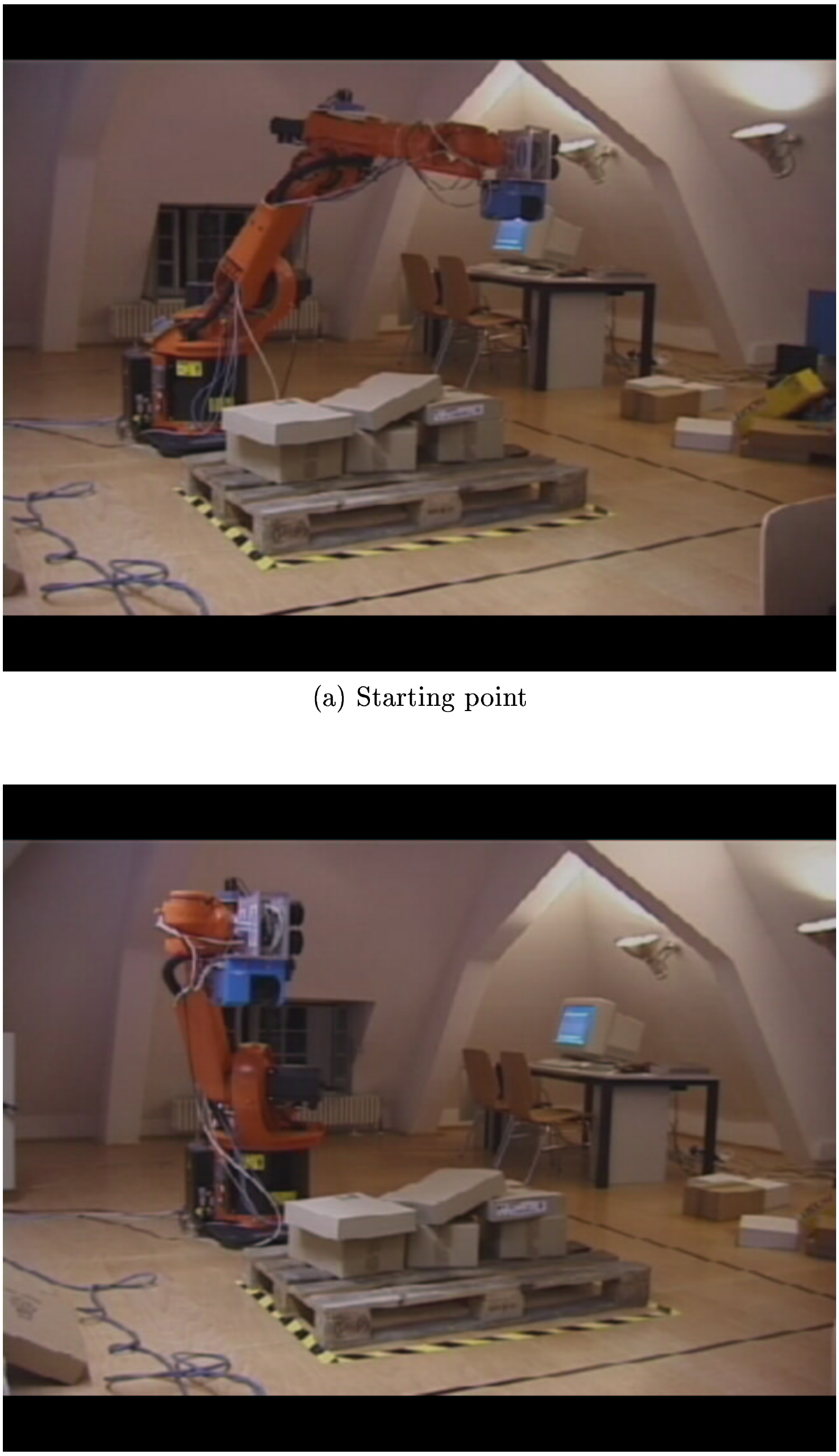
Figure 2.2: Our robotic system

area scanned corresponds to a rectangle pallet with dimensions of about 1.7m length and 1m width. The overall time required for data acquisition is 12 sec. After its acquisition, the range image is forwarded to the object recovery procedures.

Contents of the range image are not two dimensional points as is the case for scan lines but three dimensional. More specifically, each point of the range image expresses the coordinates of the corresponding scan line point in the sensor coordinate system. In detail, assume that at a specific time point the $i$th scan line $S_i$ is acquired. The cell $j$ of the scan line contains a two dimensional point $x_j, y_j$, expressing its position in the current scanning plane, that is, $S_i(j) = (x_j, y_j)$. Assume as well that when $S_i$ is acquired, the distance of the scanning plane from the origin of the sensor coordinate system is $z_i$. This can be easily determined from the current position of the robotic hand. In this respect, the three dimensional coordinates of the point in the sensor coordinate system are $(x_j, y_j, z_i)$, and is stored in the $i$th row $j$th column of the range image. Equivalently, if $I$ the range image then:

$$I(i,j) = (S_i(j), z_i) = (x_j, y_j, z_i) \tag{2.1}$$

In chapter 5 we will need the inverse operation, namely to derive the image coordinates $(i, j)$ of a three dimensional point $(x, y, z)$ measured in the sensor coordinate system. In order to do this, we firstly derive the scan line, or the row in the range image which the point belongs. We employ the $z$ coordinate for this purpose, and additionally take into consideration that the difference in $z$ values between successive scan lines is 7 mm. If so given the $z$ value the row number $i$ is the integer part of the $z/7$. Knowing the scan line to which the point corresponds, we have to find its position within the particular scan line. This is done by examining the $x$ value of the three dimensional point, and considering that firstly the point with $x = 0$, is the mid-point of the scan line, and secondly that equal number of points exist on the left and the right of the mid- point. If so, and given $C = 369$ the number of scan line



(a) Starting point

(b) End point

Figure 2.3: Data acquisition

points, then the position of $(x, y, z)$ inside the scan line, or the column number of the point within the range image will equal to the number originating by subtracting the integer part of $x$ from the position of the mid-point. In conclusion:

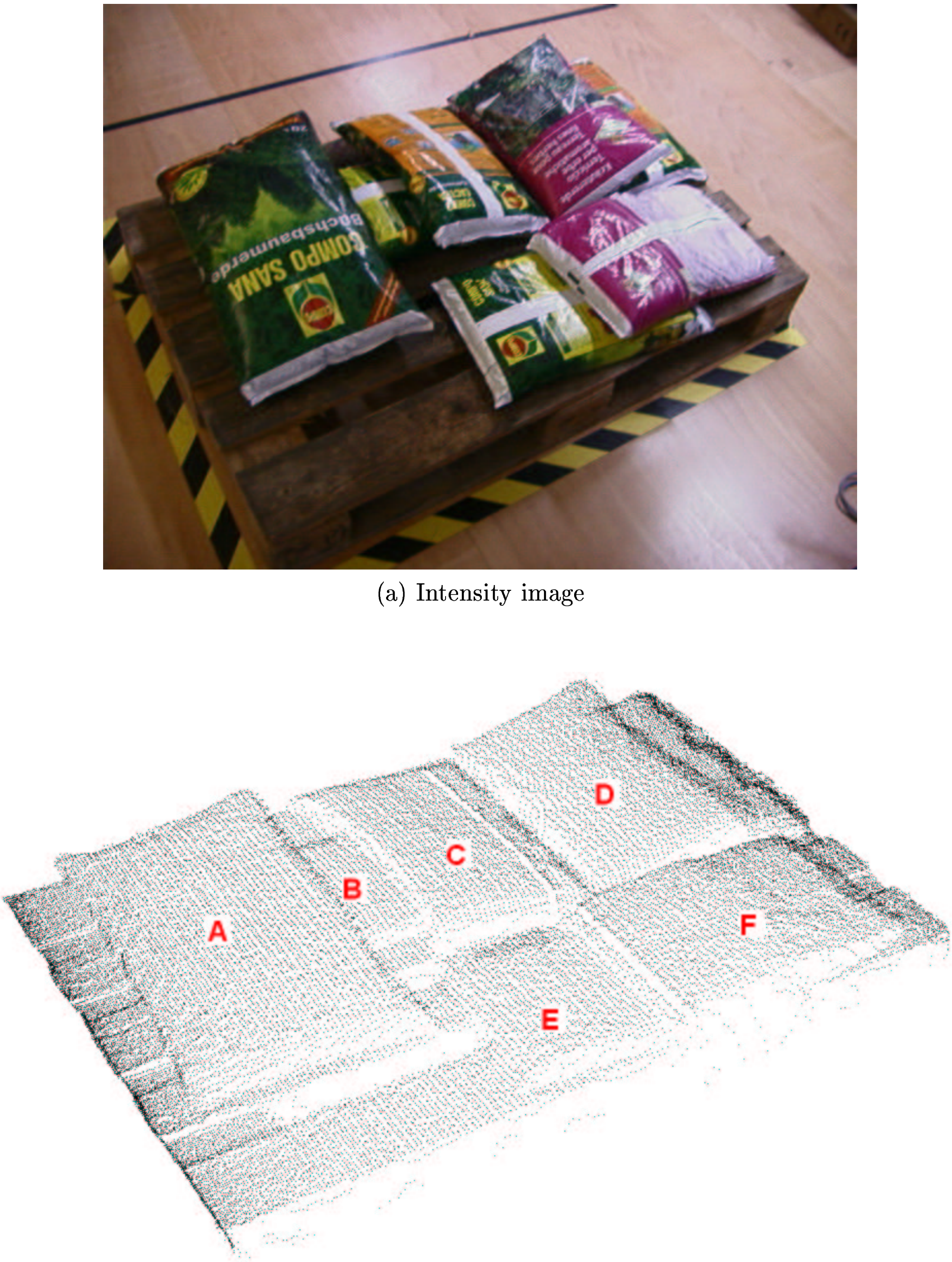
$$\begin{aligned} i &= \left\lfloor \frac{z}{7} \right\rfloor \\ j &= \frac{C+1}{2} - \lfloor x \rfloor \end{aligned} \tag{2.2}$$

### 2.4.3 Object recovery and grasping

Given the acquired range image, the object recovery process determines the objects that will be grasped by the robot. The recovery process may localize more than one objects to be grasped in one range image. In this way, the average time required for object localization is equal to the total time needed for data acquisition and image processing divided by the number of localized objects. However, it is possible that when grasping an object the robotic hand changes the configuration of other correctly localized objects. This can happen for example when the grasped object is partially occluded: Grasping the object will change the position in the pile of the occluding object. This is the reason why we have to ensure that the objects we would like to get grasped are not occluded by other objects. Not occluded objects will be hereinafter referred to as *graspable* objects. Figure 2.4 illustrates. More specifically, fig. 2.4 (a), shows a configuration of box-like objects lying on the top of the rectangle pallet. Fig. 2.4 (b), shows the range image acquired by scanning the configuration. In this figure, the objects $A$, $C$, $D$, $F$ are graspable but the objects $B$ $E$ not. Grasping the object $B$ will result into a change of the poses of $A$, $C$ perhaps of $D$ as well.

In the context of our application which deals with box-like objects, the graspable objects *fully* expose their largest surface to the laser sensor. Hence, our recovery approach recovers objects which fully expose their surface. Its output is the number of graspable objects, as well as the position and the orientation that the robotic gripper should have in order to grasp those objects from the center of gravity of their exposed surface. This information is sent to the robot, which grasps the objects. Object grasping is illustrated in fig. 2.5. Fig. 2.5 (a) shows the robotic hand while approaching one of the recovered objects of fig. 2.3. Fig. 2.5 (b) shows the robotic hand after having grasped the object. Finally, 2.6 shows the robot placing the object in a user defined position. We have to note, that there are cases where grasping of an object which fully exposes its biggest surface to the laser source may disturb the configuration. This problem is related to the sequence in which graspable objects should be removed from the pile. Such cases occur very rare in our application, and we did not extensively deal with addressing them, since scope of the thesis is object recovery and not robotic grasping. The user is referred to [115], [109] for a more formal statement of the problem and strategies towards its solution.

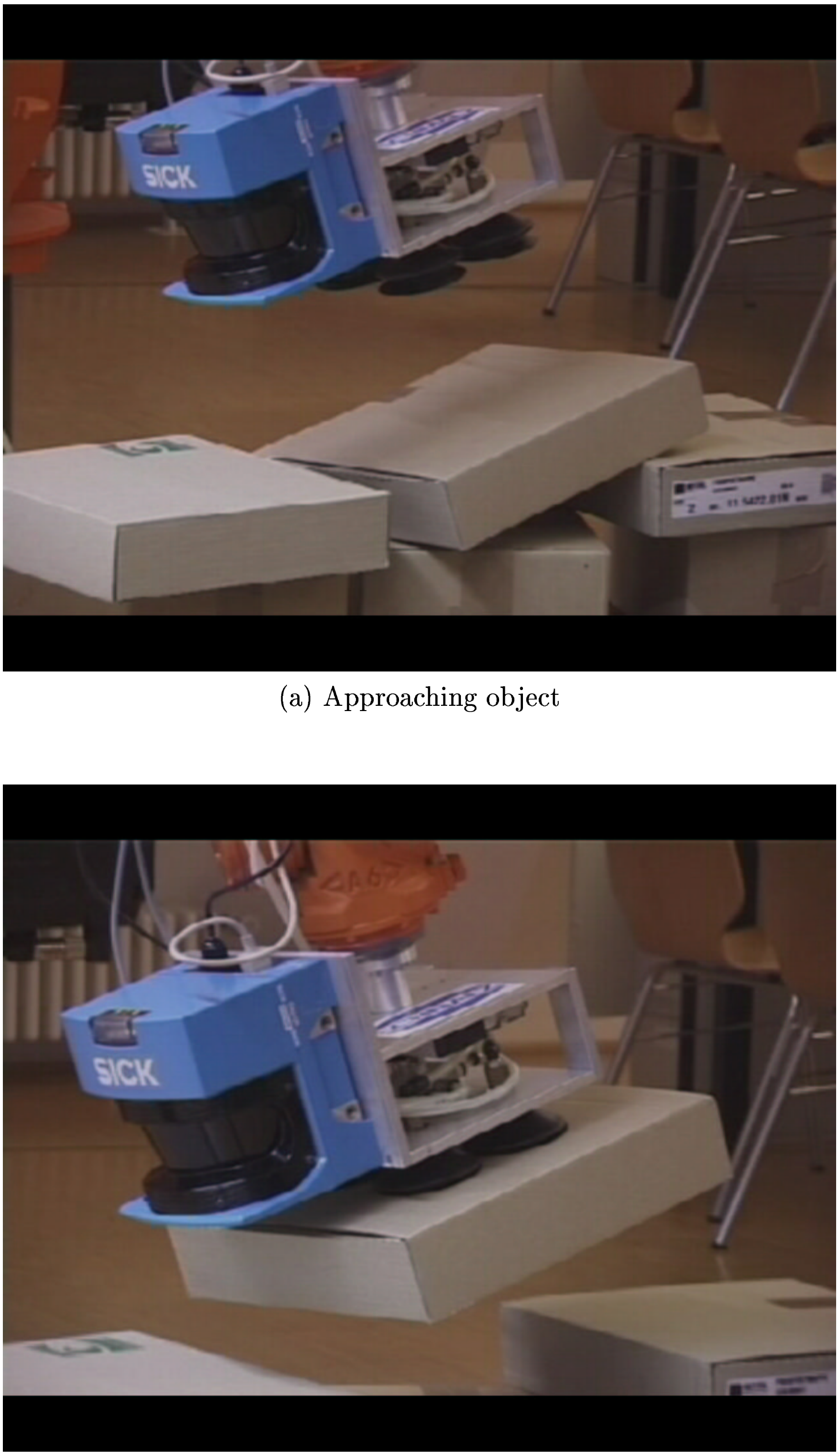
We have developed two approaches for addressing the problem of box-like object recovery. Both approaches utilize both depth information from the input range image, as well as boundary information, obtained via edge detection to the range image. The first approach



(a) Intensity image

(b) Range image

Figure 2.4: Object configuration



(a) Approaching object

(b) Lifting object

Figure 2.5: Grasping

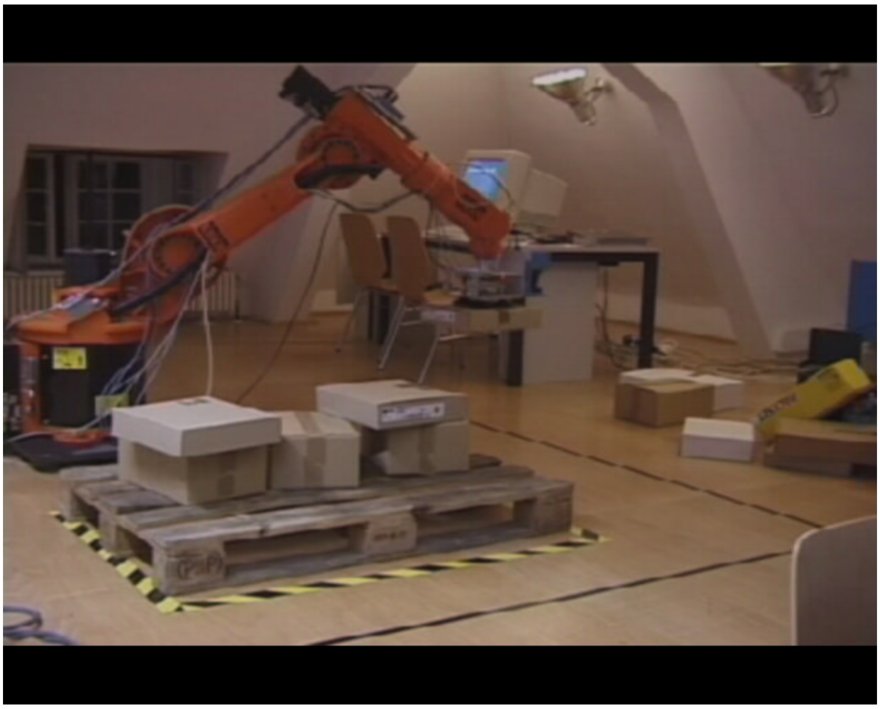


Figure 2.6: Object placement

is based on the hypothesis generation and refinement framework discussed in section 2.2.3 and is more generic, in the sense that is is able to recover both rigid and non-rigid box-like objects, for example bags. The range of target objects with which the second approach is able to deal with is limited: it recovers rigid boxes. Its advantage however, is its computational efficiency. This approach recovers the boxes by *decomposing* Hough transform.

We employ the hypothesis generation and refinement framework for box-like object recovery. As already discussed, application of this framework is connected with usage of parametric geometric entities for objects modeling. We model our objects using superquadric models [121], [148], [99], [77], which resemble or target objects in an remarkable way, which their shape parameters are properly constrained. Our approach extends the popular strategy for multiple superquadric recovery of [99], [77]. The main advantage of this strategy is its robustness: Using this approach we can recover objects like sacks, which when placed on the pallet may locally deform to a considerable extend. In addition, its computational efficiency is high enough so that it allows for implementation of a real time system. Details on the operation of the system are reported in the chapter 5 of this thesis.

In addition, we address the rigid box recovery problem using the Hough transform. The Hough transform [75], [96], is a robust and efficient tool for geometric parametric model recovery, when the number of model parameters is low. For example $2D$ lines (2 parameters), circles (3 parameters), or ellipses (5 parameters) can be both efficiently and robustly recovered from images. Rigid boxes have considerably more parameters. However, the Hough transform can be still used for recovery boxes, if we consider that these objects, or better, the exposed surfaces of these objects can be represented via a set of three dimensional lines.

In short, we employed the Hough transform to detect three dimensional lines in the boundary image. We then used the detected lines to find three dimensional object vertices. We then used the vertices to determine the dimensions of the rectangular surfaces of the boxes. The approach proved to be very robust and efficient, when the boundaries of the objects are linear. The internals of this approach are discussed in chapter 7.

## 2.5 Discussion

In this chapter we presented existing systems dealing with recovery of multiple objects from images for robotic grasping purposes. In particular, we focused on systems having the potential to be used for efficiently recovering multiple box-like objects. Besides, we roughly described our approach for automatic object unloading. Our system is superior to existing approaches because it exhibits the following advantages the combination of which, up to our knowledge, cannot be found in existing works.

Firstly, the data acquisition process is very robust. This is due to the fact that we used a laser sensor for this purpose. In this way, the data acquisition process does not depend on the lighting conditions at the installation sites. What is more, our system can operate even outdoors. Besides, using a laser sensor for data acquisition, we do not any more have to employ computationally expensive (and in many cases of questionable robustness) procedures to generate three dimensional scene information, since it is directly acquired by our sensing devise. It is noteworthy, that our data acquisition mechanism does not require any strenuous sensor calibration process. The relation between the sensor coordinate system and the world coordinate system can be straightforwardly calculated, since the sensor is placed on the hand of the robot.

Secondly, our system is flexible: It is able to deal with configurations where the objects are neatly placed on distinct layers, as well as configurations where the objects are jumbled. Besides, knowledge of the dimensions of the target objects is not required.

Thirdly, our system exhibits accuracy and robustness, due to the fact that segmentation is interwoven with object representation within our recovery processes. In addition, our system is computationally efficient: As discussed in the preceding paragraph, we attack the problem of box-like object recovery using a dual strategy: We use a probabilistic segmentation approach for dealing with both rigid and non-rigid objects, and an approach based on the Hough transform for dealing with configurations of rigid boxes. Computational efficiency in the context of the former approach is achieved via incorporation of the boundary information into the segmentation process, and in the context of the latter approach via decomposition of the recovery problem into subproblems of lower dimensionality. The validity of these statements is will be demonstrated in the chapters that follow.