Learning Grasping Points with Shape Context

Jeannette Bohg, Danica Kragic

Computer Vision and Active Perception Lab
Centre for Autonomous Systems
School of Computer Science and Communication
Royal Institute of Technology, 10044 Stockholm, Sweden
{bohg, danik}@csc.kth.se

Abstract

This paper presents work on vision based robotic grasping. The proposed method adopts a learning framework where prototypical grasping points are learnt from several examples and then used on novel objects. For representation purposes, we apply the concept of shape context and for learning we use a supervised learning approach in which the classifier is trained with labelled synthetic images. We evaluate and compare the performance of linear and non-linear classifiers. Our results show that a combination of a descriptor based on shape context with a non-linear classification algorithm leads to a stable detection of grasping points for a variety of objects.

Key words: Grasping, Shape Context, Affordances, SVM

1. Introduction

Robotic grasping of unknown objects remains an open problem in the robotic community. Although humans master this skill easily, no suitable representations of the whole process have yet been proposed in the neuroscientific literature, making it difficult to develop robotic systems that can mimic human grasping behaviour. However, there is some valuable insight. Goodale [1] proposes that the human visual system is characterised by a division into the dorsal and ventral pathways. While the dorsal stream is mainly responsible for the spatial vision targeted towards extracting action relevant visual features, the ventral stream is engaged in the task of object identification. This dissociation also suggests two different grasp choice mechanisms dependent on whether a known or unknown object is to be manipulated. Support for this can be found in behavioural studies by Borghi [2], Creem and Proffitt [3]. The authors claim that in the case of novel objects, our actions are purely guided by affordances as introduced by Gibson [4]. In the case of known objects, semantic information (e.g., through grasp experience) is needed to grasp them appropriately according to their function. However as argued in [1, 5, 6] this division of labour is not absolute. In case of objects that are similar to previously encountered ones,

Preprint submitted to Robotics and Autonomous Systems *November 2, 2009*

the ventral system helps the dorsal stream in the action selection process by providing information about prehensile parts along with their afforded actions.

In this paper, we review different approaches towards solving the object grasping problem in the robotic community and propose a vision based system that models several important steps in object grasping. We start by proposing three ways for approaching the problem, namely grasping of:

- *Known Objects*: These approaches consider grasping of a priori known objects. The goal is then to estimate object's pose and retrieve a suitable grasp, e.g., from an experience database, [7, 8, 9].
- *Unknown Objects*: Approaches that fall into this category commonly represent the shape of an *unknown* object and apply rules or heuristics to reduce the number of potential grasps [10, 11, 12, 13, 14].
- *Familiar Objects*: These approaches try reusing preexisting grasp experience from similar objects. Objects can be *familiar* in different ways, e.g, in terms of shape, colour or texture. A common assumption is that new objects similar to the old ones can be grasped in a similar way [15, 16, 17].

A general observation considering the related work is that there is a trade-off between the quality of an inferred grasp and the applicability of the method in a real world scenario. The more precise, accurate and detailed an object model, the more suitable it is for doing grasp planning. Then criteria such as, e.g., form or force closure can be taken into account to plan a stable grasp. However, when facing noise and outliers common in real world data, more assumptions regarding object geometry or generated grasps have to be introduced. Figure 1 outlines a rough taxonomy for the related work with respect to object representation on which we will elaborate in Section 2. Systems that either rely exclusively on 2D or on 3D data have some disadvantages in terms of introduced assumptions or strong dependency on the quality of the sensor data. We develop a representation that integrates data from both modalities as a way to overcome these issues.

The representation has to be rich enough to allow for the inference of the most important grasp parameters. In our case that is

- the grasping point on the object with which the *tool centre point* (TCP) should be aligned [1],
- the *approach vector* [7] which describes the 3D angle that the robot hand approaches the grasping point with and
- the wrist orientation of the robotic hand.

In our approach, a grasping point is detected based on the global shape of an object in a single image. Research in the area of neuropsychology emphasises the influence of global shape when humans choose a grasp [18, 19, 20].

[1] In this paper the TCP is at the centre of palm of the robotic gripper.

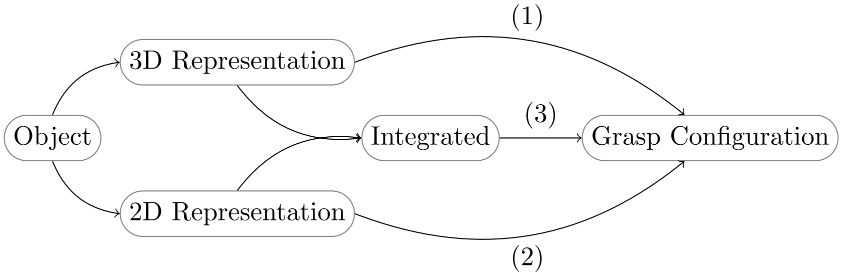


Figure 1: Grasp inference systems with respect to employed representation. (1) Approaches that use 3D data for representing the object are usually strongly dependent on good quality of the sensory data. See Section 2.2 and 2.3.1.(2) Approaches that use 2D information only for inferring a grasp usually have to make strong assumptions about applicable actions or the 3D shape of the object. See Section 2.3.2. (3) Systems that integrate 2D and 3D data are able to remove some assumptions made when using 2D only and are less dependent on the quality of the 3D data.

Matching between stereo views is then used to infer the approach vector and wrist orientation for the robot hand. We further demonstrate how a supervised learning methodology can be used for grasping of *familiar* objects.

The contributions of our approach are:

i) We apply the concept of shape context [21] to the task of robotic grasping which to the best of our knowledge has not yet been applied for that purpose. The approach is different from the one taken in [15, 17] where only local appearance is used instead of global shape.

ii) We infer grasp configurations for arbitrarily shaped objects from a stereo image pair. These are the main difference to the work presented in [16, 22] where either only planar objects are considered or three views from an object have to be obtained by moving the camera.

iii) We analyse how stable our algorithm is in realistic scenarios including background clutter without trained scenario examples as in [15].

iv) We apply a supervised learning algorithm trained using synthetic labelled images from the database provided by [15]. We compare the classification performance when using a linear classifier (logistic regression) and a non-linear classifier (*Support Vector Machines* (SVMs)).

The remainder of this paper is organised as follows: In the next section, we present related work. In Section 3, the method of applying shape context to grasping is introduced. We also describe and comment on the database that we used for training and give some background knowledge on the two different classification methods. The section concludes with a presentation on how a whole grasp configuration can be derived. In Section 4 we evaluate our method both on simulated and real data. The last section concludes the paper and gives an outlook on future work.

2. Related Work

There is a significant body of work dealing with grasp selection. We use the division proposed in the previous section to review the related work.

2.1. Grasping Known Objects

The main problem in the area of grasp planning is the huge search space from which a *good* grasp has to be retrieved. Its size is due to the large number of hand configurations that can be applied to a given object. In the theory of contact-level grasping [23, 24] a good grasp is defined from the perspective of forces, friction and wrenches. Based on this different criteria are defined to rate grasp configurations, e.g., force closure, dexterity, equilibrium, stability and dynamic behaviour.

Several approaches in the area of grasp planning exists that apply these criteria to find a good grasp for an object with a given 3D model. Some of them approximate the object's shape with a number of primitives such as spheres, cones, cylinders and boxes [25] or superquadrics (SQ) [26]. These shape primitives are then used to limit the number of candidate grasps and thus prune the search tree for finding the most stable grasp. Ciorcarlie et al. [27] exploited results from neuroscience that showed that human hand control takes place in a much lower dimension than the actual number of its degrees of freedom. This finding was applied to directly reduce the configuration space of a robotic hand to find pre-grasp postures. From these so called eigengrasps the system searches for stable grasps. Borst et al. [28] reduce the number of candidate grasps by randomly generating a number of them dependent on the object surface and filter them with a simple heuristic. The authors show that this approach works well if the goal is not to find an optimal grasp but instead a fairly good grasp that works well for " everyday tasks". Quite a different approach is taken by Li and Pollard [29]. Although, the method is independent of the ideas of contact-level grasping it still relies on the availability of a 3D object model. The authors treat the problem of finding a suitable grasp as a shape matching problem between the hand and the object. The approach starts off with a database of human grasp examples. From this database a suitable grasp is retrieved when queried with a new object. Shape features of this object are matched against the shape of the inside of the available hand postures.

All these approaches are developed and evaluated in simulation. However, Ekvall and Kragic [7] and Morales et al. [8] combine real and simulated data for the purpose of grasping *known* objects, i.e. their 3D model is available. In a monocular image a known object is recognised and its pose within the scene is estimated. Given that information, an appropriate grasp configuration can be selected from a grasp experience database. This database was acquired offline through simulations of grasps on 3D models of a set of these known objects. While Ekvall and Kragic [7] still apply the selected grasp in simulation, Morales et al. [8] ported this approach to the robotic platform described in Asfour et al. [30]. Glover et al. [9] consider known deformable objects. For representing them probabilistic models of their 2D shape are learnt. The objects can then be detected in monocular images of cluttered scenes even when they are partially occluded. The visible object parts serve as a basis for planning a stable grasp under consideration of the global object shape. However, all these approaches are dependent on an a priori known dense or detailed object model either in 2D or in 3D.

2.2. Grasping Unknown Objects

If the goal is to grasp an *unknown* object these approaches are not applicable since in practise it is very difficult to infer its geometry fully and accurately from measurements taken from sensor devices such as cameras and laser range finders. There are various ways to deal with this sparse, incomplete and noisy data. Hübner and Kragic [10], Dunes et al. [11] for example approximate an object with shape primitives that provide cues for potential grasps. Hübner and Kragic [10] decompose a point cloud derived from a stereo camera into a constellation of boxes. The simple geometry of a box reduces the number of potential grasps significantly. Dunes et al. [11] approximate the rough object shape with a quadric whose minor axis is used to infer the wrist orientation, the object centroid serves as the approach target and the rough object size helps to determine the hand pre-shape. The quadric is estimated from multi-view measurements of the rough object shape in monocular images. Opposed to the above mentioned techniques Bone et al. [14] made no prior assumption about the rough shape of the object. They applied shape carving for the purpose of grasping with a parallel-jaw gripper. After obtaining a model of the object, they search for a pair of reasonably flat and parallel surfaces that are best suited for this kind of manipulator. Richtsfeld and Vincze [12] use a point cloud of an object that is obtained from a stereo camera at a fixed viewpoint. They are searching for a suitable grasp with a simple gripper based on the shift of the top plane of an object into its centre of mass. Kraft et al. [13] also use a stereo camera to extract an object model. Instead of a raw point cloud, they are processing it further to obtain a sparser model consisting of local multi-modal contour descriptors. Four elementary grasping actions are associated to specific constellations of these features. With the help of heuristics the huge number of resulting grasp hypotheses is reduced.

2.3. Grasping Familiar Objects

A promising direction in the area of grasp planning is to re-use experience to grasp *familiar* objects. Many of the objects surrounding us can be grouped together into categories of common characteristics. There are different possibilities what these commonalities can be. In the computer vision community for example, objects within one category usually share characteristic visual properties. These can be, e.g., a common texture [31] or shape [32, 21], the occurrence of specific local features [33, 34] or their specific spatial constellation [35, 36]. These categories are usually referred to as *basic level categories* and emerged from the area of cognitive psychology [37].

In robotics however, and specifically in the area of manipulation, the goal is to enable an embodied, cognitive agent to interact with these objects. In this case, objects in one category should share common affordances [17]. More specifically, this means that they should also be graspable in a similar way. The difficulty then is to find a representation that can encode this common affordance and is grounded in the embodiment and cognitive capabilities of the agent.

Our method, and all of the following presented approaches, try to learn from experience how different objects can be grasped given different representations. This is different from the aforementioned systems in which unknown objects are grasped. There the difficulty lies in finding appropriate rules and heuristics. In the following, we will present related work that tackle the grasping of familiar objects and specifically focus on the applied representations.

2.3.1. Based on 3D Data

First of all, there are approaches that rely on 3D data only. El-Khoury and Sahbani [38] for example segment a given point cloud into parts and approximate each part by an SQ. An artificial neural net ANN is used to classify whether or not the grasp is prehensile. The ANN has been trained beforehand on labelled SQs. If one of the object parts is classified as prehensile, an n-fingered force-closure grasp is applied on this object part. Pelossof et al. [39] instead directly use a single SQ to find a suitable grasp configuration for a Barrett hand consisting of the approach vector, wrist orientation and finger spread. An SVM is trained on data consisting of feature vectors containing the parameters of the SQ and of the grasp configuration. They were labelled with a scalar estimating the grasp quality. When feeding the SVM only with the shape parameters of the SQ, their algorithm searches efficiently through the grasp configuration space for parameters that maximise the grasp quality. Curtis and Xiao [40] build upon a database of 3D objects annotated with the best grasps that can be applied to them. To infer a good grasp for a new object, very basic shape features, e.g., the aspect ratio of the object's bounding box, are extracted to classify it as similar to an object in the database. The assumption made in this approach is that similarly shaped objects can be grasped in a similar way.

2.3.2. Based on 2D Data

All of the following approaches were performed in simulation where the central assumption is that accurate and detailed 3D models are available. As mentioned previously, this assumption may not always be valid particularly with real world data gathered from sensors. like laser range finders or stereo cameras. However, there are experience based approaches that avoid this difficulty by relying mainly on 2D data. Saxena et al. [15] proposed a system that infers a point at where to grasp an object directly as a function of its image. They apply machine learning to train a grasping point model on labelled synthetic images of a number of different objects. The classification is based on a feature vector containing local appearance cues regarding colour, texture and edges of an image patch in several scales and of its 24 neighbouring patches in the lowest scale. The system was used successfully to pick up objects from a dishwasher after it has been specifically trained for this scenario. However, if more complex goals are considered that require subsequent actions, e.g., pouring something from one container into another, semantic knowledge about the object and about suitable grasps regarding their functionality becomes necessary [2, 3, 41]. Then, to only represent graspable points without the conception of *objectness* [13, 42] is not sufficient.

Another example of a system involving 2D data and grasp experience is presented by [17]. Here, an object is represented by a composition of prehensile parts. These so called *affordance cues* are obtained by observing the interaction of a person with a specific object. Grasp hypotheses for new stimuli are inferred by matching features of that object against a codebook of learnt *affordance cues* that are stored along with relative object position and scale. However, how exactly to grasp these detected prehensile parts is not yet solved since hand orientation and finger configuration are not inferred from the affordance cues. More successful in terms of the inference of full grasp configurations are Morales et al. [16] who use visual feedback to even predict fingertip positions. The authors also take the hand kinematics into consideration when selecting a number of planar grasp hypotheses directly from 2D object contours. To predict which of these grasps is the most stable one, a KNN-approach is applied in connection with a grasp experience database. However, the approach is restricted to planar objects.

2.3.3. Integrating 2D and 3D Data

In Figure 1, we divided the related the work in the area of grasp inference systems into three different kinds dependent on the employed modality of object representation. As already mentioned above, we believe that systems in which both 2D and 3D data are integrated are most promising in terms of dealing with sensor noise and removing assumptions about object shape or applicable grasps.

There are approaches in the community that have taken this path. In [43], two depth sensors are applied to obtain a point cloud of a tabletop scene with several objects. The authors extend their previous work to infer initial 2D grasping point hypothesis. Then, the shape of the point cloud within a sphere centred around a hypothesis is analysed with respect to hand kinematics. This enhances the prediction of a stable grasp and also allows for the inference of grasp parameters like approach vector and finger spread. In their earlier work [15], only downward or outward grasp were possible with the manipulators in a fixed pinch grasp configuration. Speth et al. [22] showed that their earlier 2D based approach [16] is also applicable when considering 3D objects. The camera is used to explore the object to retrieve crucial information like height, 3D position and pose. However, all this additional information is not applied in the inference and final selection of a suitable grasp configuration. In this paper, we are also proposing an approach that falls into the 3rd path of Figure 1. We see the result of the 2D based grasp inference as a way to search in a 3D object representation for a full grasp configuration. Here, we will focus on the development of the 2D method and demonstrate its applicability for searching in a minimal 3D object representation.

3. Using Shape Context for Grasping

A detailed flow chart of the whole system and associated hardware is given in Figure 2. First, scene segmentation is performed based on a stereo input resulting in several object hypotheses. Shape context is then computed on each of the object hypotheses and 2D grasping points are extracted. The models of grasping points are computed beforehand through offline training on an image database. The points in the left and in the right image are associated to each other to infer a 3D grasping point via triangulation. In parallel with the grasping point detection, the segments are analysed in terms of rough object pose. By integrating the 3D grasping point with this pose, a full grasp configuration can be determined and then executed. In the following sections, the individual steps of the system are explained in more detail.

3.1. Scene Segmentation

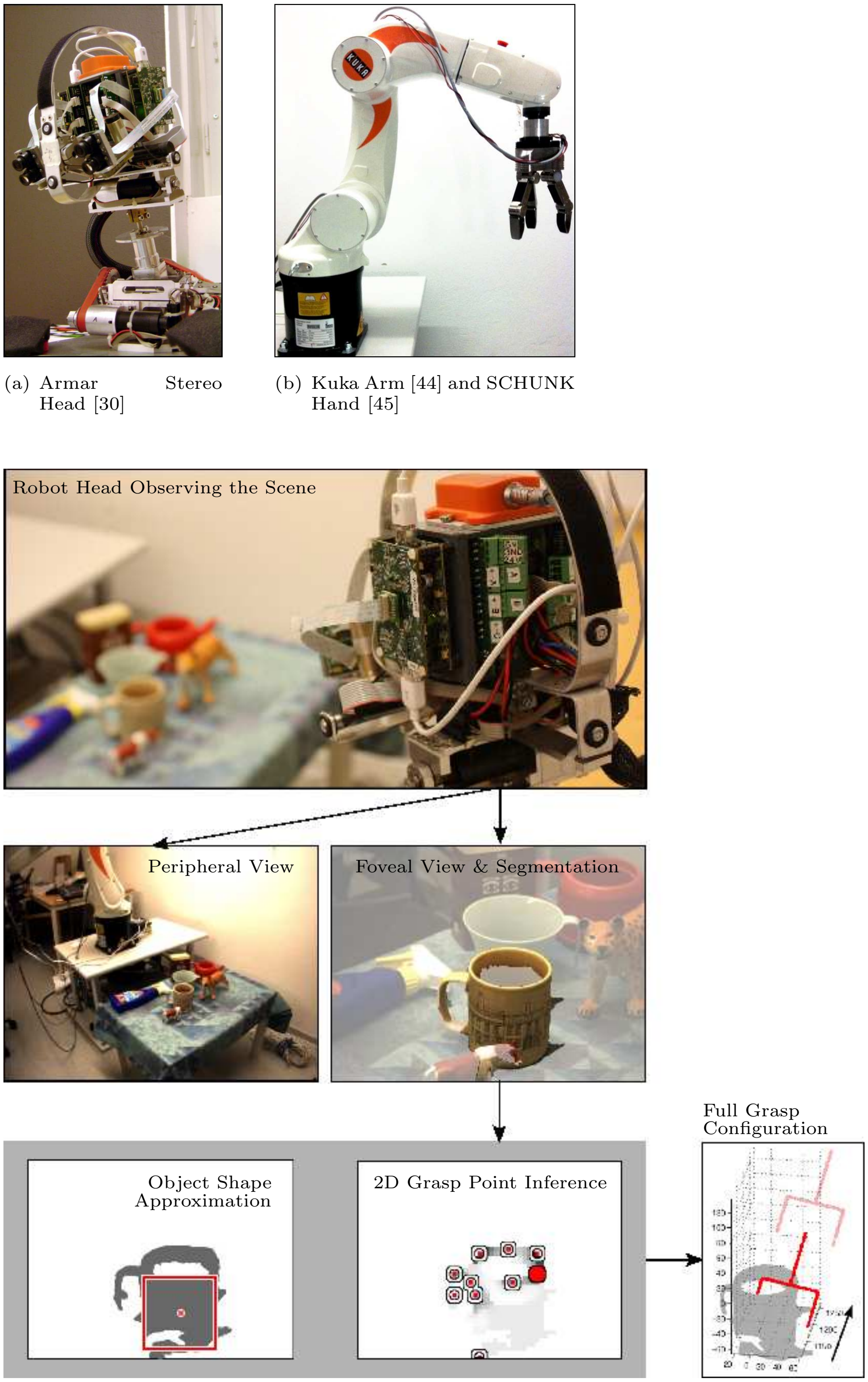
The system starts by performing the *figure-ground segmentation*. Although the problem is still unsolved for general scenes, we have demonstrated in our previous work how simple assumptions about the environment help in segmentation of table-top scenes, [46, 47, 48]. The segmentation is based on integration of stereo cues using foveal and peripheral cameras. In the below, we shortly refer to the different steps of the segmentation process.

3.1.1. Zero-Disparity

The advantage of using an active stereo head lies in its capability to fixate on interesting parts of the scene. A system that implements an attentional mechanism has been presented by Rasolzadeh et al. [49]. Once the system is in fixation, zero-disparities are employed as a cue for figure-ground segmentation through different segmentation techniques, e.g., watershedding , Björkman and Eklundh [48]. The assumption made is that continuity in reconstructed depth results from an object. However, Figure 3 shows that such a simple assumption results in bad segmentation of the object from the plane on which it is placed.

3.1.2. Planar Surfaces

The environment in which service robots perform their tasks are dominated by planar surfaces. In order to overcome the above segmentation problem, we use an assumption of the dominant plane. In our examples, this plane represents the table top objects are placed on. For that purpose, we fit a planar surface to the disparity image. The probability for each pixel in the disparity image to belong to that plane or not depends on its distance to the most likely plane. In that way, objects standing out of a plane are well segmented. Problems can arise with non-textured objects when the disparity image has large hollow regions. When the table plane assumption is violated through, e.g., clutter, the segmentation of the object is more difficult. Examples are shown in Figure 3.

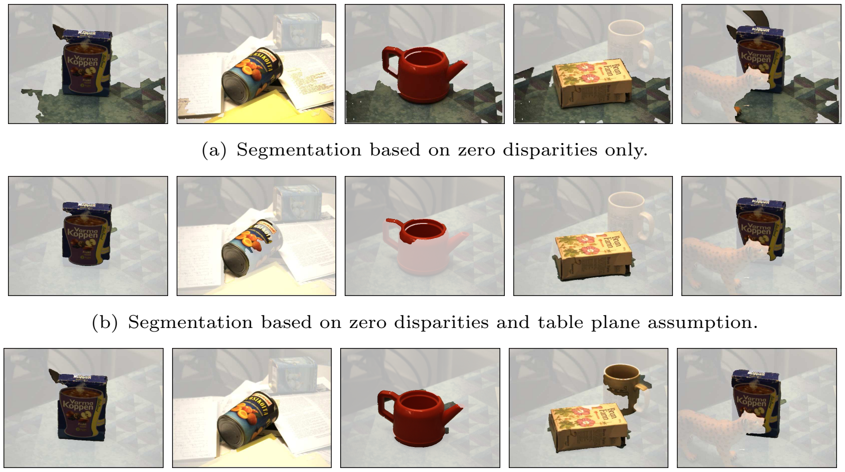

Robot Head Observing the Scene
Peripheral View
Foveal View & Segmentation
Object Shape Approximation
2D Grasp Point Inference
Full Grasp Configuration

(a) Armar Stereo Head [30]

(b) Kuka Arm [44] and SCHUNK Hand [45]

(c) Flow Chart

Figure 2: Components of the Stereo Vision based Grasp Inference System



(a) Segmentation based on zero disparities only.

(b) Segmentation based on zero disparities and table plane assumption.

(c) Segmentation based on zero disparities, table plane assumption and known hue.

Figure 3: Segmentation results for: 1st column) One textured object. 2nd column) Cluttered table scene. 3rd column) Non-textured object. 4th column) Two similarly coloured objects. 5th column) Occlusion.

3.1.3. Uniform Texture and Colour

An additional assumption can be made on the constancy of object appearance properties, assuming either uniformly coloured or textured objects. Introducing this cue in conjunction with the table plane assumption, the quality of the figure-ground segmentation increases. The probability that a specific hue indicates a foreground object depends on the foreground probability (including the table plane assumption) of pixels in which it occurs. This holds equivalently for the background probability of the hue. The colour cue contributes to the overall estimate with the likelihood ratio between foreground and background probability of the hue. The examples are shown in Figure 3. Judging from the examples and our previous work, we can obtain reasonable hypotheses of objects in the scene. In Section 4 and Section 4.3 we analyse the performance of the grasp point detection for varying quality of segmentation.

3.2. Representing Relative Shape

Once the individual object hypotheses are made, we continue with the detection of grasping points. In Section 3.5 we show how to further infer the approach vector and wrist orientation. Grasping an object depends to a large extent on its global shape. Our approach encodes the global property of an object with a local, image based representation. Consider for example elongated objects such as pens. A natural grasp is in its middle, roughly at the centre of mass. The point in the middle divides the object in two relatively similar shapes. Hence, the shape *relative* to this point is approximately symmetric. In contrast to that, the shape *relative* to a point at one of the ends of the object is highly asymmetric. Associating a point on the object with its *relative shape* and

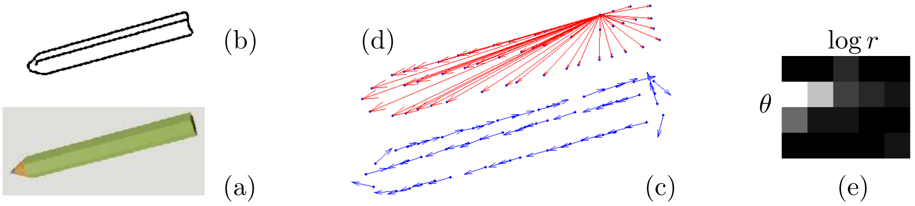


Figure 4: Example of deriving the shape context descriptor for the image of a pencil. (a) Input image of the pencil. (b) Contour of the pencil derived with the Canny operator. (c) Sampled points of the contour with gradients. (d) All vectors from one point to all other sample points. (e) Histogram with four angle and five log-radius bins comprising the vectors depicted in (d).

the natural grasp is the central idea of our work. For this purpose we use the concept of shape context commonly used for shape matching, object recognition and human body tracking, [50, 51, 52]. In the following, we briefly summarise the main ideas of shape context. For a more elaborate description, we refer to [21].

The basis for the computation of shape context is an edge image of the object. N sample points are taken uniformly from the contours, considering both inner and outer contours. For each point we compute the vectors that lead to all other sample points. These vectors relate the global shape of the object to the considered reference point. We create a compact descriptor comprising this information for each point by a two dimensional histogram with angle and radius bins. In [21] it is proposed to use a log-polar coordinate system in order to emphasise the influence of nearby samples. An example for the entire process is shown in Figure 4.

A big advantage of shape context is that it is invariant to different transformations. Invariance to translation is intrinsic since both the angle and the radius values are determined relative to points on the object. To achieve scale invariance, [21] proposed to normalise all radial distances by the median distance between all N^2 point pairs in the shape. Also rotation invariance can be easily achieved by measuring the angles relative to the gradient of the sample points. In the following, we will describe how to apply the *relative shape* representation to form a feature vector that can later be classified as either graspable or not.

3.2.1. Feature Vector

In the segmented image, we compute the contour of the object by applying the Canny edge detector. This raw output is then filtered to remove spurious edge segments that are either too short or have a very high curvature. The result serves as the input for computing shape context as described above. We start by subdividing the image into rectangular patches of 10 × 10 pixels. A descriptor for each patch serves as the basis to decide whether it represents a grasping point or not. This descriptor is simply composed of the accumulated histograms of all sample points on the object's contour that lie in that patch.

Typically only few sample points will be in a 10×10 pixel wide window. Furthermore, comparatively small shape details that are less relevant for making a grasp decision will be represented in the edge image. We therefore calculate the accumulated histograms in three different scales centred at the current patch. The edge image of the lowest scale then contains only major edges of the object. The three histograms are concatenated to form the final feature descriptor of dimension 120.

3.3. Using Feature Vector for Classification

The detection of grasping points applies a supervised classification approach utilizing the feature vector described in the previous section. We examine two different classification methods: a linear one (logistic regression) and a non-linear one (SVMs), [53]. We describe these briefly below.

Logistic Regression. Let g_i denote the binary variable for the ith image patch in the image. It can either carry the value 1 or 0 for being a grasping point or not. The posterior probability for the former case will be denoted as $P(g_i = 1|D_i)$ where D_i is the feature descriptor of the ith image patch. For logistic regression, this probability is modelled as the sigmoid of a linear function of the feature descriptor:

$$P(g_i = 1|D_i) = \frac{1}{1 + e^{-wD_i}} \tag{1}$$

where w is the weight vector of the linear model. These weights are estimated by maximum likelihood:

$$w = \arg\max_{w'} \prod_i P(g_i = 1|D_i, w') \tag{2}$$

where here g_i and D_i are the labels and feature descriptors of our training data, respectively.

Support Vector Machines. SVMs produces arbitrary decision functions in feature space by a linear separation in a space of higher dimension compared to the feature space. The mapping of the input data into that space is accomplished by a non-linear kernel function K. In order to obtain the model for the decision function when applying SVMs, we solve the following optimisation problem:

$$\max \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j g_i g_j K(D_i, D_j) \tag{3}$$

subject to $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i g_i = 0$ with the solution $w = \sum_i^{N_s} \alpha_i g_i D_i$. As a kernel we have chosen a *Radial Basis Function* (RBF):

$$K(D_i, D_j) = e^{-\gamma ||D_i - D_j||^2}, \gamma > 0 \text{ and } \gamma = \frac{1}{2\sigma^2} \tag{4}$$

The two parameters C and σ are determined by a grid search over parameter space. In our implementation, we are using the package libsvm [54].

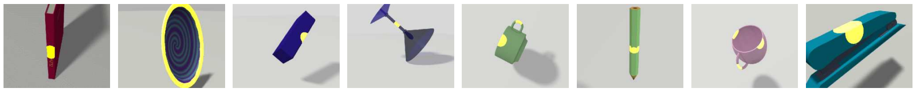


Figure 5: One example picture for each of the eight object classes used for training along with their grasp labels (in yellow). Depicted are a book, a cereal bowl, a white board eraser, a martini glass, a cup, a pencil, a mug and a stapler. The database is adopted from Saxena et al. [15].

Training Database. For training the different classifiers we will use the database developed by Saxena et al. [15] containing ca. 12000 synthetic images of eight object classes depicted along with their grasp labels in Figure 5. One drawback of the database is that the choice of grasping points is not always consistent with the object category. As an example, a cup is labelled at two places on its rim but all the points on the rim are equally well suited for grasping. The eraser is quite a symmetric object. Neither the local appearance nor the relative shape of its grasping point are discriminative descriptors. This will be further discussed in Section 4 where our method is compared to that of [15].

3.4. Approximating Object Shape

Shape context provides a compact 2D representation of objects in monocular images. However, grasping is an inherently three-dimensional process. Our goal is to apply a pinch grasp on the object using a 3D coordinate of the grasping point and known position and orientation of the wrist of the robot. In order to infer the 6D grasp configuration, we integrate 2D grasping point detection with the 3D reconstruction of the object that results from the segmentation process.

We approach the problem by detecting the dominant plane $\hat{\Pi}_d : D_d = A_d x + B_d y + C_d d$ in the disparity space $d = I_d(x, y)$. The assumption is that the object can be represented as a constellation of planes in 3D, i.e. a box-like object has commonly three sides visible to the camera while for a cylindrical object, the rim or lid generates the most likely plane. We use RANSAC to estimate the dominant plane hypothesis $\hat{\Pi}_d$ and we also determine its centroid M_d by calculating the mean over all the points in the plane $\{(x, y)|e(x, y) > \theta\}$ where

$$e(x, y) = I_d(x, y) - \frac{(D_d - A_d x - B_d y)}{C_d}, \tag{5}$$

the error between estimated and measured disparity of a point, and θ a threshold. By using the standard projective equations between image coordinates (x, y), disparity d, camera coordinates (X, Y, Z), baseline b and the focal length f

$$x = f\frac{X}{Z}, \ y = f\frac{Y}{Z}, \ d = \frac{bf}{Z}, \tag{6}$$

we can transform $\hat{\Pi}_d$ into the camera coordinate frame:

$$\hat{\Pi}_C : -C_d bf = A_d fX + B_d fY - D_d Z. \tag{7}$$

The normal of this plane is then defined as

$$n_C = (A_C, B_C, C_C) = (A_d f, B_d f, -D_d). \tag{8}$$

Equations 6 are also used to convert the plane centroid M_d from disparity space to M_C in the camera coordinate frame.

3.5. Generation of Grasp Hypotheses

In the following, we describe how the dominant plane provides the structural support for inferring a full grasp configuration.

3.5.1. 3D Grasping Point

The outputs of the classifier are candidate grasping points in each of the stereo images. These then need to be matched for estimation of their 3D coordinates. For this purpose we create a set $B_l = \{b_{(i,l)} | i = 1 \cdots m\}$ of m image patches i in the left image representing local maxima of the classifier $P(g_i = 1|D_i)$ and whose adjacent patches in the 8-neighbourhood carry values close to that of the centre patch. We apply stereo matching to obtain the corresponding patches $B_r = \{b_{(i,r)} | i = 1 \cdots m\}$ in the right image. Let $P(b_{(i,l)}|D_{(i,l)})$ and $P(b_{(i,r)}|D_{(i,r)})$ be the probability for each image patch in set B_l and B_r to be a grasping point given the respective feature descriptors $D_{(i,l)}$ or $D_{(i,r)}$. Assuming naïve Bayesian independence between corresponding patches in the left and right image, the probability $P(b_i|D_{(i,l)}, D_{(i,r)})$ for a 3D point b_i to be a grasping point is modelled as

$$P(b_{(i,l)}|D_{(i,l)}, D_{(i,r)}) = P(b_{(i,l)}|D_{(i,l)}) \times P(b_{(i,r)}|D_{(i,r)}). \tag{9}$$

As already mentioned in the previous section, the approach vector and wrist orientation are generated based on the dominant plane. Therefore, the choice of the best grasping point is also influenced by the detected plane. For this purpose, we use the error $e(b_{(i,l)})$ as defined in Equation 5 as a weight w_i in the ranking of the 3D grasping points. The best patch is then

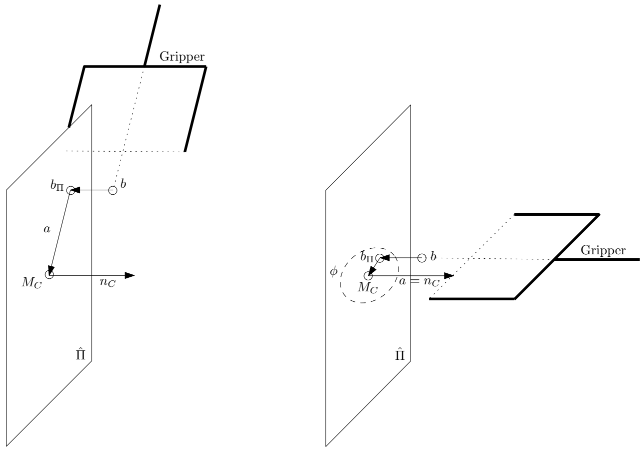
$$b = \arg\max_i w_i \times P(b_{(i,l)}|D_{(i,l)}, D_{(i,r)}). \tag{10}$$

3.5.2. Orientation of the Schunk Hand

Given a 3D grasping point b, the dominant plane $\hat{\Pi}_C$ with its centroid M_C and normal n_C, there are two possibilities for the approach vector a:

(i) $a = v_C$ where $v_C = M_C - b_\Pi$ and b_Π is the projected grasping point on $\hat{\Pi}_C$ along n_C
(ii) $a = n_C$.

This is illustrated in Figure 6. Which of them is chosen depends on the magnitude of v_C. If $|v_C| > \phi$, the wrist of the hand is chosen to be aligned with the normal of the plane. If $|v_C| < \phi$, such that b_Π is very close to the centroid of the plane, we chose n_C as the approach vector, i.e., the hand is approaching the dominant plane perpendicularly. In this case, we choose the wrist orientation to be parallel to the ground plane. In Section 4.4 we present results of object grasping on our hardware.



(a) Approach vector a is equal to $M_C - b_\Pi$. The wrist is aligned with n_C.

(b) Approach Vector a is equal to the normal. The wrist is aligned with the horizon.

Figure 6: Visualisation of the 6 DoF Grasp Configuration with respect to the estimated dominant plane $\hat{\Pi}$

4. Experimental Evaluation

We start by comparing our method to the one presented in [15]. The goal is to show the performance of the methods on synthetic images. This is followed by an in-depth analysis of our method. Finally, we investigate the applicability of our method in real settings.

4.1. Evaluation on Synthetic Images

In this section, we are especially interested in how well the classifiers generalise over global shape or local appearance given synthetic test images. For this purpose we applied four different sets of objects to train the classifiers.

- *Pencils* are grasped at their centre of mass.
- *Mugs & cups* are grasped at handles. They only differ slightly in global shape and local grasping point appearance.
- *Pencils, white board erasers & martini glasses* are all grasped approximately at their centre of mass at two parallel straight edges. However, their global shape and local appearance differ significantly.
- *Pencils & mugs* are grasped differently and also differ significantly in their shape.

We divided each set into a training and test set. On the training sets we trained four different classifiers.

- *Shape context & SVM* (SCSVM). We employed twelve angle and five log radius bins for the shape context histogram. We sample the contour with 300 points. The same parameters were applied by [21] and have proven to perform well for grasping point detection.
- *Local appearance features & logistic regression* (OrigLog) is the classifier by [15].
- *Local appearance features & SVM* (OrigSVM) applies an SVM instead of logistic regression.
- *Shape context, local appearance features & SVM* (SCOrigSVM) integrates shape context features with local appearance cues. The resulting feature vector is used to train an SVM.

4.1.1. Accuracy

Each model was evaluated on the respective test sets. The results are shown as ROC curves in Figure 7 and as accuracy values in Table 1. Accuracy is defined as the sum of true positives and true negatives over the total number of examples. Table 1 presents the maximum accuracy for a varying threshold.

The first general observation is that SVM classification outperforms logistic regression. On average, the classification performance for each set of objects rose about 4.32% when comparing OrigSVM with OrigLog. A second general observation is that classifiers that employ global shape (either integrated or not integrated with appearance cues) have the best classification performance for all training sets. In the following we will discuss the results for each set.

- *Pencils*. The local appearance of a pencil does not vary a lot at different positions along its surface whereas the relative shape does. Therefore, local appearance based features are not discriminative enough. This is confirmed for the models that are only trained on images of pencils. SCSVM performs slightly better than OrigSVM. The classification performance grows when applying an integrated feature vector.
- *Mugs & Cups*. These objects are grasped at their handle which is characterised by a local structure that is rather constant even when the global shape changes. Thus, OrigSVM outperforms slightly the classifier that applies shape context only. However, an integration of both features leads to an even better performance.
- *Pencils, white board erasers & martini glasses*. For this set of objects the position of the grasp is very similar when considering their global shape whereas the local appearance of the grasping points differs greatly. Also here, the models based on shape context performs best. Feature integration degrades the performance.
- *Pencils & mugs*. The performance of the different classifiers for the previous set of objects is a first indication for a weaker generalisation capability of OrigSVM and OrigLog over varying local appearance compared

to SCSVM and SCOrigSVM. This is further confirmed for the last set where not just local appearance but also global shape changes significantly. SCSVM improves the performance of OrigSVM about 6.75% even though the grasping points are very different when related to global object shape. Feature integration increases the performance only moderately.

4.1.2. Repeatability

Our goal is to make a robot grasp arbitrary and novel objects. Thus, we are also interested in if the *best* grasping point hypotheses correspond to points that in reality afford stable grasps. Thus, our second experiment evaluates whether the best hypotheses are located on or close to the labelled points. We constructed a set of 80 images from the synthetic image database with ten randomly selected images of each of the eight object classes (Figure 5). Thus, also novel objects that were not used for training the different classifiers are considered. On every image we run all the aforementioned models and for each one picked out the best ten grasping points b_i. In the database, a label is not a single point, but actually covers a certain area. We evaluated the Euclidean distance d_i of each of the ten grasping points measured from the border of this ground truth label at position p_j and normalised with respect to the length l_j of its major axis. This way, the distance is dependent on the scale of the object in the image. In case there is more than one label in the image, we choose the one with the minimum distance. If a point b_i lies directly on the label, the distance $d_i = 0$. If a point lies outside of the label, the distance d_i gets weighted with a Gaussian function ($\sigma = 1, \mu = 0$) multiplied with $\sqrt{2\pi}$. The number of hits h_m of each model m on the picture set is counted as follows:

$$\begin{aligned} h_m &= \sum_{k=1}^{K} \sum_{i=1}^{N_k} e^{-\frac{d^2_{(i,k)}}{2}} \\ \text{with } d_{(i,k)} &= \min_{j=1}^{M_k} \frac{dist(b_{(i,k)}, p_{(j,k)})}{2l_{(j,k)}} \end{aligned}$$

where K is the number of images in the set, M_k is the number of grasp labels in that picture and N_k is the number of detected grasping points. Grasping points whose distance d_i exceeds a value of $3 * \sigma$ are considered as outliers. Figure 8 shows the number of hits, that is, the amount of good grasps for each model.

Apart from the model trained on cups and mugs, the SVM trained only on shape context always performs best. The performance drop for the second object set can be explained in the same way as in the previous chapter: handles have a very distinctive local appearance and are therefore easily detected with features that capture this. In general, this result indicates that classifiers based on shape context detect grasping points with a better repeatability. This is particularly important for the inference of 3D grasping points in which two 2D grasping points in the left and right image of a stereo camera have to be matched.

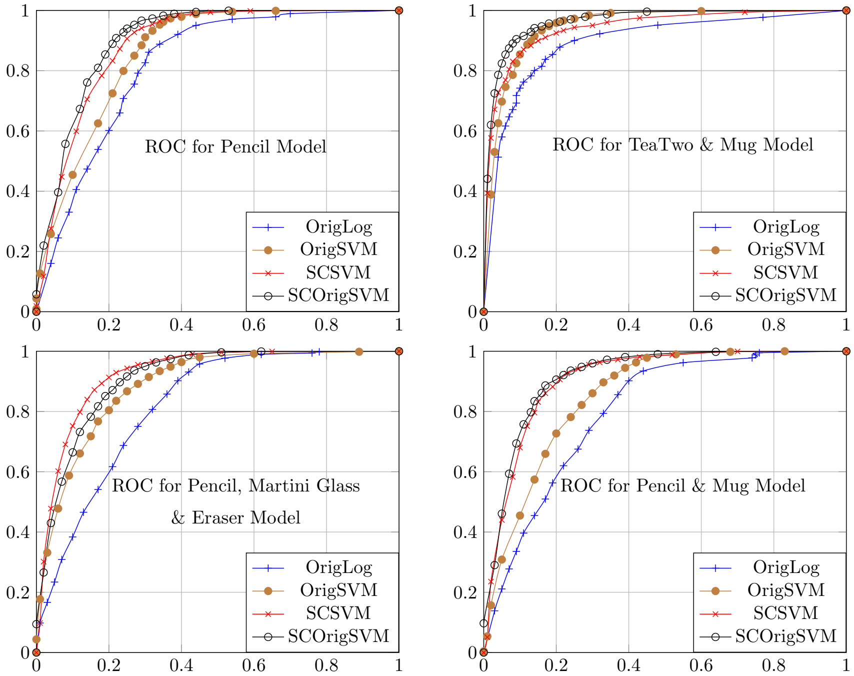


Figure 7: ROC curves for models trained on different objects.

4.1.3. Summary of Results

We draw several conclusion regarding experimental results on synthetic images. First, independent of which feature representation is chosen, SVM outperforms logistic regression. Secondly, our simple and compact feature descriptor that encodes relative object shape improves the detection of grasping points both in accuracy and repeatability in most cases. In case of very distinct local features, both representations are comparable. Integration of the two representations leads only to moderate improvements or even decreases the classification performance.

Table 1: Accuracy of the models trained on different objects.

	SCOrigSVM	SCSVM	OrigSVM	OrigLog
Pencil	**84.45%**	82.55%	80.16%	77.07%
Cup & Mug	**90.71%**	88.01%	88.67%	83.85%
Pencil, Martini & Eraser	84.38%	**85.65%**	80.79%	74.92%
Pencil & Mug	**85.71%**	84.64%	77.80%	74.32%

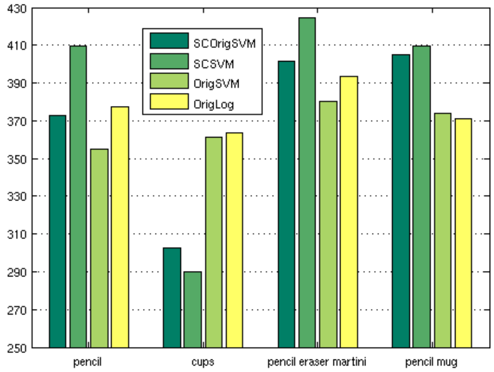


Figure 8: Evaluation of the best ten grasping points of each model on a picture test set containing in total 80 pictures of *familiar* and novel objects (see Figure 5).

4.2. Opening the Black Box

In the previous section, we presented evidence that the shape context based models detect grasping points more accurately than the models trained on local appearance features. As argued in Section 3, we see relative shape as a better cue for graspability than local appearance. In this section, we would like to confirm this intuition by analysing what the different grasping point models encode. We conduct this analysis by applying the Trepan Algorithm by Craven and Shavlik [55] to the learnt classifiers. This algorithm builds a decision tree that approximates a concept represented by a given *black box* classifier. Although originally proposed for neural networks, Martens et al. [56] showed that it is also applicable for SVMs.

We use the same sets of objects as mentioned in the previous section. The extracted trees are binary with leafs that are classifying feature vectors as either graspable or non-graspable. The decisions at the non-leaf nodes are made based on either one or more components of the feature vector. We consider each *positive* leaf node as encoding a prototypical visual feature that indicates graspability. As previously mentioned, the extracted trees are only approximations of the actual models learned. Thus, the feature vectors that end up at a specific leaf of the tree will be of three different kinds:

- *Ground truth.* Features that are graspable according to the ground truth labels in the database.
- *False positives by model.* Features that are not graspable according to the labels but are so according to the classifier.
- *False positives by tree.* Features that are neither labelled in the database nor classified by the model to be graspable, but are considered to be so by the tree.

We will analyse these samples separately and also rate the trees by stating their *fidelity* and *accuracy*. *Fidelity* is a measure of how well the extracted trees approximate the considered models. It states the amount of features vectors whose classification is compliant with the classification of the approximated model. *Accuracy* measures the classification rate for either the tree or the model when run on a test set.

The analysis of these samples is conducted using PCA. The resulting eigenvectors form an orthonormal basis with the first eigenvector representing the direction of the highest variance, the second one the direction with the second largest variance, etc. In the following sections we visualise only those eigenvectors whose *energy* is above a certain threshold and at maximum ten of these. The *energy* e_i of an eigenvector i is defined as

$$e_i = \frac{\sum_{j=1}^{i} \lambda_j}{\sum_{j=1}^{k} \lambda_j} \tag{11}$$

where λ_j is the eigenvalue of eigenvector j with k eigenvectors in total. As a threshold we use $\theta = 0.9$.

The remainder of this section is structured as follows. In Section 4.2.1, we visualise the prototypical features for the local appearance method by applying PCA to the samples at positive nodes. In Section 4.2.1 we do the same for the relative shape based representation.

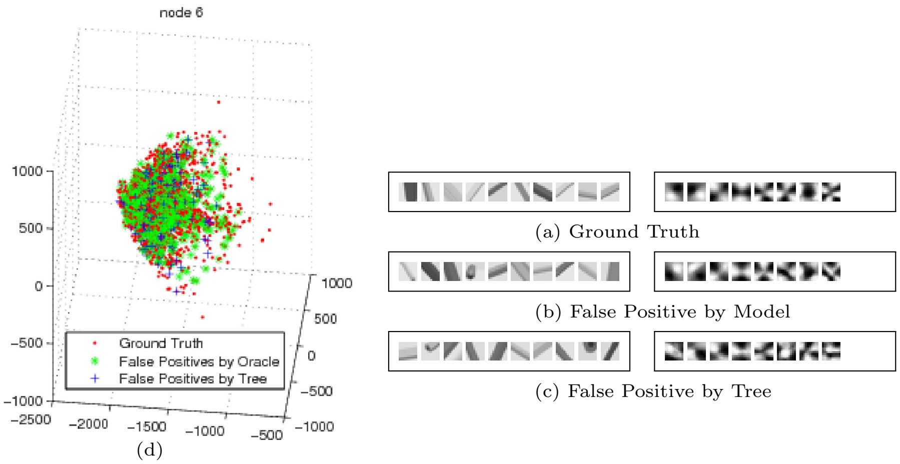
4.2.1. Local Appearance Features

Saxena et al. [15] applied a filter bank to 10 × 10 pixel patches in three spatial scales. The filter bank contains edge, texture (Law's masks) and colour filters. In this section, we depict samples of these 10 × 10 pixel patches in the largest scale. They are taken from every positive node of each tree trained for a specific object set. All feature vectors that end up at one of these positive nodes are used as an input to PCA.

The first set we present consists of images of a pencil (see Figure 5) labelled in its centre of mass. The built tree is rather shallow: it has only four leaf nodes of which one is positive. The decisions on the non-leaf nodes are made based on the output of the texture filters only. Neither colour nor edge information are considered. This means that this part of the feature vector is not necessary to achieve a classification performance of 75.41% (see Table 2). Ten random samples from the positive node are shown in Figure 9(a)-(c) subdivided dependent on whether they are graspable according to the ground truth labels from the database or only according to the model and tree, respectively.

In order to visualise to which visual cues this grasping point models actually respond, we run PCA on the set of feature vectors that ended up at that node. The resulting principal components selected according to Equation 11 are also depicted in Figure 9 (a)-(c). Encoded are close-ups of the body of the pencil and perspective distortions.

However, the majority of the pencil complies with these components. Because of that, the samples from the set of false positives are very similar to the ground truth samples. The appearance of the centre of mass is not that different from the rest of the pencil. This is further clarified by Figure 9 where the false



(a) Ground Truth

(b) False Positive by Model

(c) False Positive by Tree

(d)

Figure 9: Pencils:(a)-(c) Ten samples and PCA components of the positive node of the decision tree. (d) Feature vectors projected into the three dimensional space spanned by the three eigenvectors of the sample set of true grasping points with the highest variance.

Table 2: Accuracy of the models trained on different objects given the local appearance representation.

	Pencil	Cups	Elongated	Pencil & Mug
Fidelity	86.78%	83.97%	87.55%	89.29%
Accuracy Tree	75.41%	82.61%	72.12%	73.30%
Accuracy Model	77.07%	83.85%	74.92%	74.32%

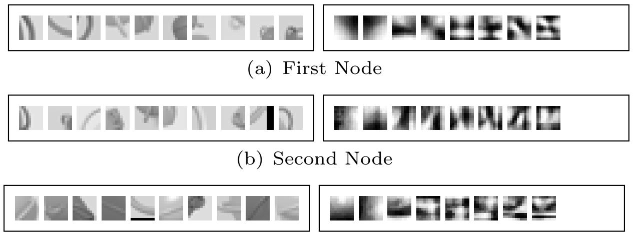
positives by the model and tree are projected into the space spanned by the first three principal components from the ground truth: they are strongly overlapping. We will show later that given our relative shape based representation these three first principal components are already enough to define a space in which graspable points can be better separated from non-graspable points.

For the other sets of objects we applied the same procedure. The principal components of the samples at each positive node are shown in Figure 10, 11 and 12. In Table 2, the fidelity of the respective trees in relation to the model and their accuracies are given.

4.2.2. Relative Shape

In this section we evaluate the performance of the shape context in the same manner as the local appearance features were tested in the previous section. The process includes

(i) extracting the contour with the Canny edge detector,
(ii) filtering out spurious edge segments,
(iii) subsampling the contour,

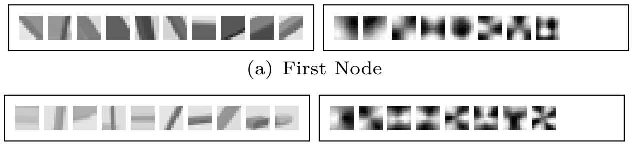


(a) First Node

(b) Second Node

(c) Third Node

Figure 10: Cups: Ten samples and PCA components for each of the positive nodes of the decision tree.



(a) First Node

(b) Second Node

Figure 11: Elongated Objects: Ten samples and PCA components for each of the positive nodes of the decision tree.

(iv) normalising the sampled contour with the median distance between contour points,

(v) rotating the whole contour according to the average tangent directions of all the contour points falling into the patch that is currently considered by the classifier

(vi) and finally plotting the resulting contour on a $20x20$ pixels patch with the grasping point in the centre.

The output of this procedure forms the input for PCA. The sample feature vectors for each node are depicted not as patches but as red squared labels located at the grasping point on the object.

Each of the induced trees in this section is of a slightly worse quality in terms of fidelity when compared with the trees obtained from the logistic regression method (see Table 2). We reason that this is due to the performance of the Trepan algorithm when approximating SVMs. Nevertheless, the purpose of this section is the visualisation of prototypical grasping point features rather than impeccable classification. This performance is therefore acceptable. The results for the induced trees are given in Table 3.

We start by analysing the model trained on the set of pencils. The induced decision tree has one positive node. The samples from this node are depicted in Figure 13 along with the most relevant PCA components to which we will

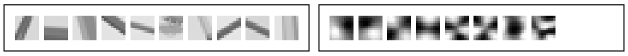
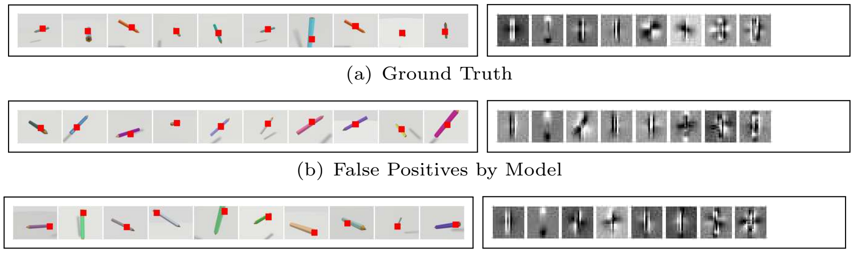


Figure 12: Pencils and Mugs: Ten samples and PCA components for the positive node of the decision tree.

Table 3: Accuracy of the models trained on different objects given the relative shape representation.

	Pencil	Cups	Elongated	Pencil & Mug
Fidelity	78.97%	79.66%	78.79%	80.82%
Accuracy Tree	71.38%	76.89%	73.40%	73.41%
Accuracy Model	82.55%	88.01%	85.56%	84.64%



(a) Ground Truth

(b) False Positives by Model

(c) False Positives by Tree

Figure 13: Pencil: Ten samples and PCA components of the positive node of the decision tree.

refer in the remainder of this paper as *eigencontours*. These components do not encode the local appearance but clearly the symmetric relative shape of the grasping point.

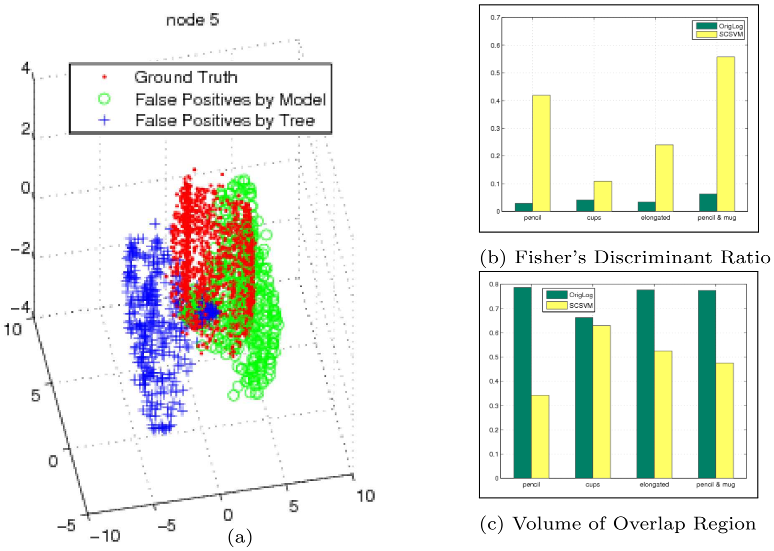
One interesting observation is that the feature vectors projected into the space spanned by the three best principal components of the ground truth samples are quite well separable, even with a linear decision boundary. There is almost no overlap between false positives produced by the tree and the ground truth features and little overlap between false positives produced by the models and the true graspable features. This result is shown in Figure 14.

We applied the same procedure to the models trained on the other sets of objects. The eigencontours for these are shown in Figs. 15-17. For the sets consisting of different objects, each positive node in the decision tree is mainly associated with one of the objects and encodes where they are graspable.

Furthermore, we can observe a better separability compared to the models trained on local appearance. In order to quantify this observation, we analysed the distribution of the samples in the three-dimensional PCA space in terms of linear separability. As measures for that we employed Fisher's discriminant ratio and the volume of the overlap regions. Figure 14 (b) and (c) show a comparative plot of these two measures for all the models considered in this section.

4.2.3. Summary of Results

The evaluation provided a valuable insight into different feature representations. We observed that our compact feature descriptor based on relative shape is more discriminative than the feature descriptor that combines the output of

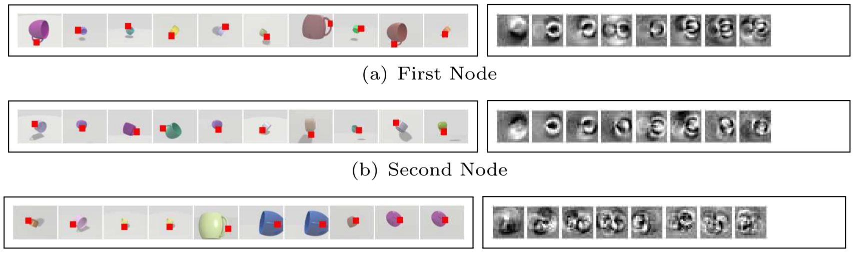


(a)

(b) Fisher's Discriminant Ratio

(c) Volume of Overlap Region

Figure 14: (a)Pencil: Feature vectors projected into the three dimensional space spanned by the three eigenvectors of the sample set of true grasping points with the highest variance. (b) and (c) measure linear separability for models trained on different training sets and with different classification methods. Dark: OrigLog. Bright: SCSVM.

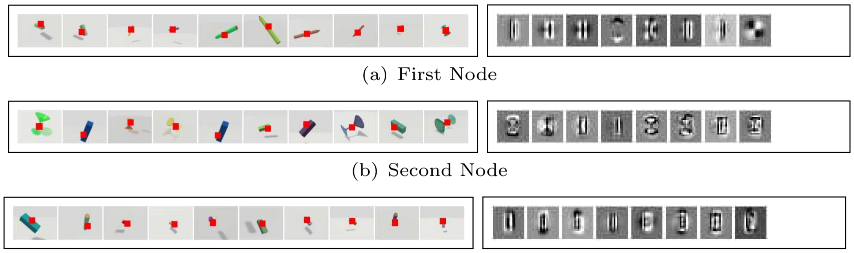


(a) First Node

(b) Second Node

(c) Third Node

Figure 15: Cups: Ten samples and PCA components for each of the positive nodes of the decision tree.

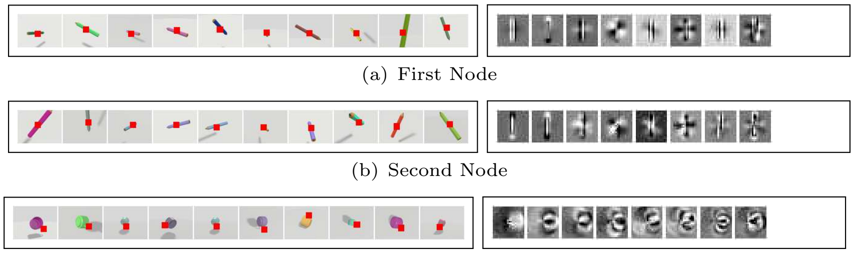


(a) First Node

(b) Second Node

(c) Third Node

Figure 16: Elongated: Ten samples and PCA components for each of the positive nodes of the decision tree.


(a) First Node

(b) Second Node

(c) Third Node

Figure 17: Pencils and Mugs: Ten samples and PCA components of the first positive node of the decision tree.

a filter bank. The dimensionality of our descriptor is almost four times smaller which also has implications for the time needed to train an SVM. The classification performance achieved with an SVM could even be improved by finding a decision boundary in the space spanned by the first three principal components of a set of ground truth prototypical features.

4.3. Evaluation on Real Images

In the previous section, we showed that the performance of the relative shape based classifier is better compared to a method that applies local appearance. In these synthetic images no background clutter was present. However, in a real world scenario this we need to cope with clutter, occlusions, etc. One example is presented in [15]. The authors demonstrated a system for the scenario of emptying a dishwasher. In order to cope with the visual clutter occurring in such a scenario, the grasping point model was trained on hand labelled images of the dishwasher. Although the dishwasher was emptied successfully, for a new scenario the model has to be re-trained to cope with new backgrounds.

We argue that we need a way to cope with backgrounds based on more general assumptions. As described earlier in Section 3.1, our method relies on scene segmentation. In this section, we evaluate how the relative shape based representation is affected by different levels of segmentation. For that purpose, we collected images of differently textured and texture-less objects, e.g., boxes, cans, cups, elongated objects, or toys, composed in scenes of different levels of complexity. This ranges from single objects on a table to several objects occluding each other. These scenes were segmented with the three different techniques described in Section 3.1.

Ideally, we would like to achieve two things. First is the repeatability: the grasping points for the same object given different qualities of segmentation have to match. Second is the robustness: the grasping points should be minimally affected by the amount of clutter. Regarding the latter point, a quantitative evaluation can only be performed by applying the inferred grasps in practise. Thus, we demonstrate our system on real hardware in Section 4.4 and present here some representative examples of the grasping point inference methods when applied to different kinds of object situated in scenes of varying complexity.

4.3.1. Examples for Grasping Point Detection

In Figure 18, we show the results of the grasping point classification for a teapot. The left column shows the segmented input of which the first one is always the ground truth segment. The middle column shows the result of the grasping point classification when applying the local appearance based descriptor by [15] and the right one the results of the classification when using the relative shape based descriptor. The red dots label the detected grasping points. They are the local maxima in the resulting probability distribution. Maximally the ten highest valued local maxima are selected.

The figure shows the grasping point classification when the pot is the only object in the scene and when it is partially occluded. Note that the segmentation in the case of local appearance based features is only influencing which patches are considered for the selection of grasping points. In case of the relative shape based descriptor, the segmentation also influences the classification by determining which edge points are included in the shape context representation. Nevertheless, what we can observe is that the detection of grasping points for the representation proposed in this paper is quite robust. For example in Figure 18(b) (last row), even though there is a second handle now in the segmented region, the rim of the teapot is still detected as graspable and the general resulting grasping point distribution looks similar to the cases in which the handle was not yet in the segment. This means, that the object the vision system is currently in fixation on, the one that dominates the scene, produces the strongest responses of the grasping point model even in the presence of other graspable objects.

In Figure 19(a), we applied the models trained on mugs and cups to images of a can and a cup. The descriptor based on local appearance responds very strongly to textured areas whereas the relative shape based descriptor does not get distracted by that since the whole object shape is included in the grasping point inference. Finally in Figure 19(b), we show an example of an object that is not similar to any object that the grasping point models were trained on. In case of the local appearance based descriptor, the grasping point probability is almost uniform and very high valued. In the case of shape context there are some peaks in the distribution. This suggest that the ability of these models to generalise over different shapes is higher than for local appearance based models.

4.3.2. Repeatability of the Detection

One of the goals of the method is the repeatability of grasping point detection. In order to evaluate this, we measured the difference of the detected grasping points in the differently segmented images. For real images, we do not have any ground truth labels available as in the case of synthetic data. Thus, we cannot evaluate the grasp quality as was done in Section 4.1. Instead, we use the detected grasping points in a manually segmented image as a reference to quantify the repeatability of the grasping point detection.

We have a set $B = \{b_i \| i = 1 \dots N\}$ of pictures and three different cues based on which they are segmented: zero disparity, a dominant plane and hue. If we want to measure the difference d_{b_i} between the set of grasping points $G_{b_i} = \{g_{(b_i,j)} \| j = 1 \dots M\}$ and the set of reference points $G_{b_i} = \{g_{(b_i,r)} \| k = 1 \dots R\}$ for a specific kind of segmentation of the image b_i, then

$$d_{b_i} = \frac{1}{K} \sum_{j=1}^{M} e^{\frac{-d_j^2}{2}} \text{ where} \tag{12}$$

$$d_j = \min_{r=1}^{R} dist(g_{(b_i,r)}, g_{(b_i,j)}) \tag{13}$$

where $dist$ is the Euclidean distance and K the length of the image diagonal[2]. The mean and standard deviation of d_{b_i} for all images in the set B that are segmented with a specific cue is then our measure of deviation of the detected from the reference grasping points.

In Figure 20 we show this measure for a representative selection of objects and models. As already mentioned, ideally we would like to see no difference between detected grasping points when facing different qualities of segmentation. In practise, we can observe a flat slope. As expected for both methods, the grasping points detected in the image segmented with zero-disparity cues are the ones that are deviating most from the reference points. Although, the selection of points that are included in our representation is directly influenced by the segmentation, the difference between detected and reference grasping points is not always bigger than for the appearance based method. In fact, sometimes it performs even better. This holds for examples of the models trained on mugs and cups for which both methods show a similar accuracy on synthetic data (Figure 20 (a) and (b)). If the models are applied to novel objects, as can be observed in Figure 20 (c), our descriptors shows a better repeatability. This suggests again a better capability of the models to generalise across different relative shapes. In general, we can say that both methods are comparable in terms of repeatability.

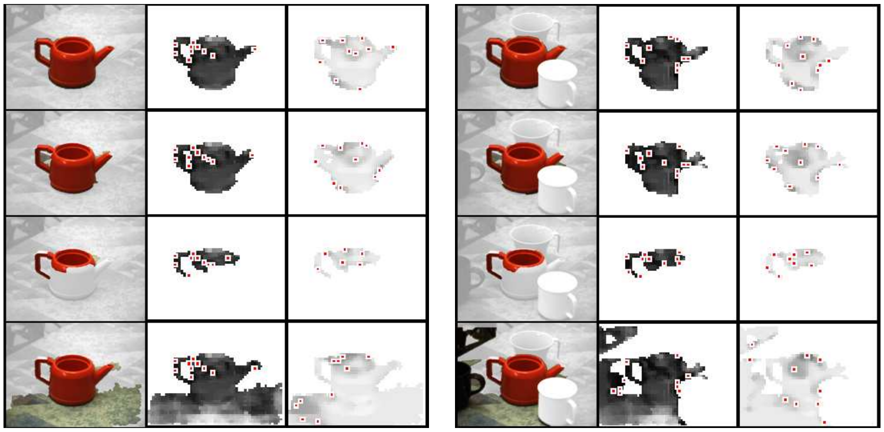
4.3.3. Summary of Results

In this section, we evaluate the performance of our approach on real images. Due to the encoding of global shape, the method is robust against occlusions and strong texture. Although our representation is strongly dependent on the segmentation, we observe that the repeatability of grasping points is comparable to the local appearance based method even when facing imperfect segmentation. The analysis included images of varying qualities of segmentation as well occlusion and clutter.

4.4. Demonstration of Real Grasps

In this section we demonstrate the integration of the 2D grasping point detection with the minimal 3D object representation as described in Section 3.5.

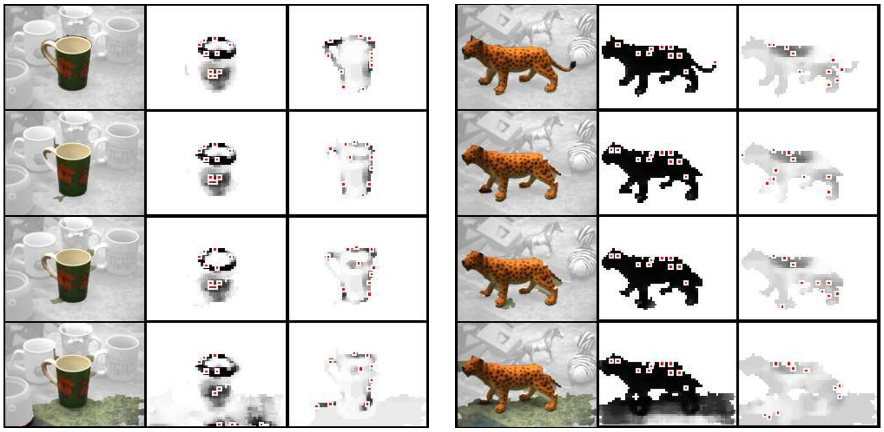
[2] In our case $K = 80$ since we are evaluating 10×10 pixel patches in images of size 640×480 pixels



(a) Single Teapot

(b) Partly Occluded Teapot

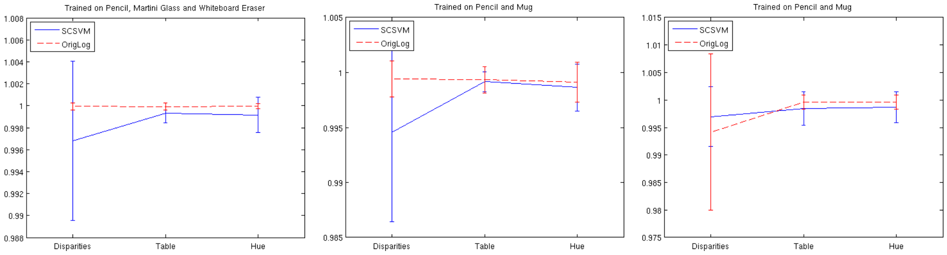
Figure 18: Grasping point model trained on mugs and pencils applied to a textureless teapot. The darker a pixel, the higher is the probability that it is a grasping point.



(a) Grasping point model trained on mugs and cups applied to a textured cup.

(b) Grasping point model trained on pencils, martini glasses and whiteboard eraser applied to a novel object.

Figure 19: The darker a pixel, the higher is the probability that it is a grasping point.



(a) Set of Cans

(b) Set of Textureless Tea Pots

(c) Set of Novel Objects

Figure 20: Comparing the stability of grasp point detection of SCSVM and OrigLog for different sets of objects and different grasping point models when facing imperfect segmentation.

We used the hardware setup as depicted in Figure 2(a) and 2(b): a 6 DoF KUKA robotic arm [44], a three-fingered 7 DoF Schunk Hand [45] and the 7 DoF Karlsruhe Active Head [30]. In Figure 21, video snapshots from the robot grasping three different objects are given along with the segmented input image, inferred grasping point distribution and detected dominant plane [3]. For this demonstration, we rejected grasping points for which the approach vector would result in a collision with the table.

In general, we can observe that the generated grasp hypotheses are reasonable selections from the huge amount of potentially applicable grasps. Failed grasps are due to the fact that there is no closed-loop control implemented either in terms of visual servoing or hand movements as demonstrated in our previous work [57, 58, 59, 60]. Some grasps also fail due to the slippage or collision.

5. Conclusions

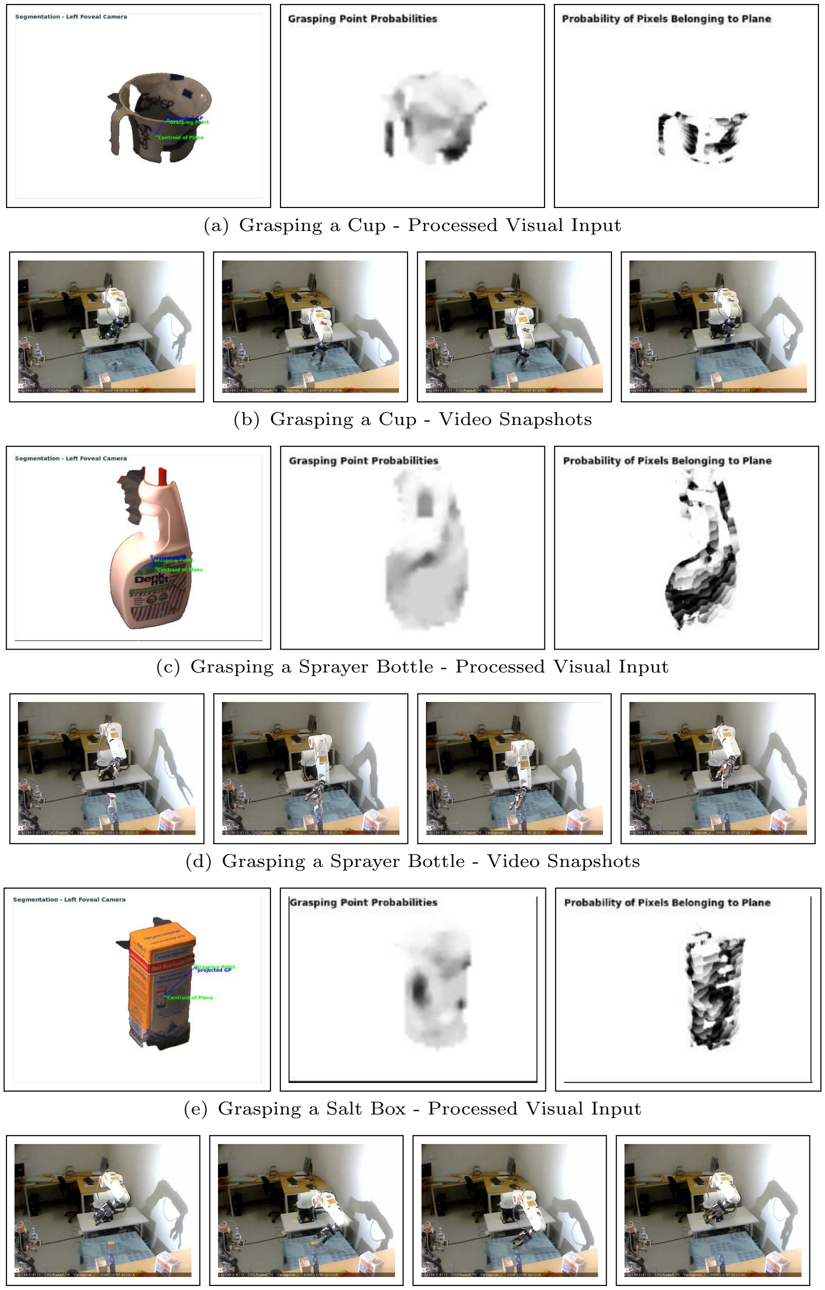
Grasping of unknown objects in natural environments is an important and unsolved problem in the robotic community. In this paper, we have developed a method for detecting a grasping point on an object by analysing it in a monocular image and reconstructing the suitable 3D grasping representation based on a stereo view . Referring to neuropsychological research mentioned in Section 2, we argued that for the purposes of grasping a yet unseen object, its global shape has to be taken into account. Therefore, we applied shape context as a visual feature descriptor that relates the object's global shape to a single point.

The experimental evaluation was performed both in simulation and in the real world. The motivation for the simulated experiments was both to compare our approach with some other state of the art approaches as well as to provide more insight into the complexity of the whole modelling process. We showed that a combination of a relative shape based representation and a non-linear classifier leads to an improved performance of the grasping point classification due to better discriminativity. Evaluation in the real scene has proven the stability of the proposed representation in the presence of clutter. The demonstration on a real robot provides further insight into the difficulty of the object grasping process. We see several aspects to be evaluated in the future work. We will continue to further develop the method but integrate it more on the stereo level for generating the grasping point hypotheses. In addition, we will consider other types of representation that take into account several aspects of 2D-3D information. Our ongoing work presented in [61] demonstrates the use of the method proposed here for grasping unicoloured objects.

Acknowledgement

The authors would like to thank Mårten Björkman for providing the segmentation and Xavi Gratal Martínez for his great help with the hardware system.

[3] A number of videos can be downloaded from `www.csc.kth.se\~bohg`.


Segmentation - Left Foveal Camera
Grasping Point Probabilities
Probability of Pixels Belonging to Plane
Projected GP
Grasping Point
Centroid of Plane

(a) Grasping a Cup - Processed Visual Input

(b) Grasping a Cup - Video Snapshots

(c) Grasping a Sprayer Bottle - Processed Visual Input

(d) Grasping a Sprayer Bottle - Video Snapshots

(e) Grasping a Salt Box - Processed Visual Input

(f) Grasping a Salt Box - Video Snapshot

Figure 21: Generating grasps for different objects: Left: Grasping Point, Projected Grasping Point and Plane Centroid. Middle: Grasping Point Probabilities. Right: Probabilities of Pixels belonging to the Dominant Plane.

This project has been supported by the EU IST-FP7-IP GRASP (2008-2012).

References

[1] M. Goodale, Separate Visual Pathways for Perception and Action, Trends in Neurosciences 15 (1) (1992) 20–25.

[2] A. Borghi, Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking, chap. Object Concepts and Action, Cambridge University Press, 2005.

[3] S. H. Creem, D. R. Proffitt, Grasping Objects by Their Handles: A Necessary Interaction between Cognition and Action, Journal of Experimental Psychology: Human Perception and Performance 27 (1) (2001) 218–228.

[4] J. Gibson, The Ecological Approach to Visual Perception, Lawrence Erlbaum Associates, 1979.

[5] U. Castiello, M. Jeannerod, Measuring Time to Awareness, Neuroreport 2 (12) (1991) 797–800.

[6] M. J. Webster, J. Bachevalier, L. G. Ungerleider, Connections of Inferior Temporal Areas TEO and TE with Parietal and Frontal Cortex in Macaque Monkeys, Cerebral cortex 4 (5) (1994) 470–483.

[7] S. Ekvall, D. Kragic, Learning and Evaluation of the Approach Vector for Automatic Grasp Generation and Planning, in: IEEE International Conference on Robotics and Automation, 4715–4720, 2007.

[8] A. Morales, P. Azad, T. Asfour, D. Kraft, S. Knoop, R. Dillmann, A. Kargov, C. Pylatiuk, S. Schulz, An Anthropomorphic Grasping Approach for an Assistant Humanoid Robot, in: 37th International Symposium on Robotics, 149–152, 2006.

[9] J. Glover, D. Rus, N. Roy, Probabilistic Models of Object Geometry for Grasp Planning, in: IEEE International Conference on Robotics and Automation, Pasadena, CA, USA, 2008.

[10] K. Hübner, D. Kragic, Selection of Robot Pre-Grasps using Box-Based Shape Approximation, in: IEEE Int. Conference on Intelligent Robots and Systems, 1765–1770, 2008.

[11] C. Dunes, E. Marchand, C. Collowet, C. Leroux, Active Rough Shape Estimation of Unknown Objects, in: IEEE International Conference on Robotics and Automation, 3622–3627, 2008.

[12] M. Richtsfeld, M. Vincze, Grasping of Unknown Objects from a Table Top, in: ECCV Workshop on 'Vision in Action: Efficient strategies for cognitive agents in complex environments', Marseille, France, 2008.

[13] D. Kraft, N. Pugeault, E. Baseski, M. Popovic, D. Kragic, S. Kalkan, F. Wörgötter, N. Krueger, Birth of the Object: Detection of Objectness and Extraction of Object Shape through Object Action Complexes, International Journal of Humanoid Robotics 5 (2) (2008) 247–265.

[14] G. M. Bone, A. Lambert, M. Edwards, Automated Modelling and Robotic Grasping of Unknown Three-Dimensional Objects, in: IEEE International Conference on Robotics and Automation, Pasadena, CA, USA, 292–298, 2008.

[15] A. Saxena, J. Driemeyer, J. Kearns, A. Y. Ng, Robotic Grasping of Novel Objects, Neural Information Processing Systems 19 (2007) 1209–1216.

[16] A. Morales, E. Chinellato, A. Fagg, A. del Pobil, Using Experience for Assessing Grasp Reliability, International Journal of Humanoid Robotics 1 (4) (2004) 671–691.

[17] M. Stark, P. Lies, M. Zillich, J. Wyatt, B. Schiele, Functional Object Class Detection Based on Learned Affordance Cues, in: 6th International Conference on Computer Vision Systems, vol. 5008 of *LNAI*, Springer-Verlag, 435–444, 2008.

[18] M. A. Goodale, J. P. Meenan, H. H. Bülthoff, D. A. Nicolle, K. J. Murphy, C. I. Racicot, Separate Neural Pathways for the Visual Analysis of Object Shape in Perception and Prehension, Current Biology 4 (7) (1994) 604–610.

[19] R. H. Cuijpers, J. B. J. Smeets, E. Brenner, On the Relation Between Object Shape and Grasping Kinematics, Journal of Neurophysiology 91 (2004) 2598–2606.

[20] M. Gentilucci, Object Motor Representation and Reaching-Grasping Control, Neuropsychologia 40 (8) (2002) 1139–1153.

[21] S. Belongie, J. Malik, J. Puzicha, Shape Matching and Object Recognition Using Shape Contexts, IEEE Trans. on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.

[22] J. Speth, A. Morales, P. J. Sanz, Vision-Based Grasp Planning of 3D Objects by Extending 2D Contour Based Algorithms, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2240–2245, 2008.

[23] V.-D. Nguyen, Constructing stable grasps, International Journal on Robotics Research 8 (1) (1989) 26–37.

[24] K. Shimoga, Robot Grasp Synthesis Algorithms: A Survey, International Journal of Robotic Research 15 (3) (1996) 230–266.

[25] A. T. Miller, S. Knoop, H. I. Christensen, P. K. Allen, Automatic Grasp Planning Using Shape Primitives, in: IEEE Int. Conf. on Robotics and Automation, 1824–1829, 2003.

[26] C. Goldfeder, P. K. Allen, C. Lackner, R. Pelossof, Grasp Planning Via Decomposition Trees, in: IEEE International Conference on Robotics and Automation, 4679–4684, 2007.

[27] M. Ciorcarlie, C. Goldfeder, P. Allen, Dexterous Grasping via Eigengrasps: A Low-Dimensional Approach to a High-Complexity Problem, Robotics: Science and Systems Manipulation Workshop .

[28] C. Borst, M. Fischer, G. Hirzinger, Grasping the Dice by Dicing the Grasp, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 3692–3697, 2003.

[29] Y. Li, N. Pollard, A Shape Matching Algorithm for Synthesizing Humanlike Enveloping Grasps, Humanoid Robots, 2005 5th IEEE-RAS International Conference on (2005) 442–449.

[30] T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, R. Dillmann, ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control, in: 6th IEEE-RAS International Conference on Humanoid Robots, 169–175, 2006.

[31] J. Shotton, J. Winn, C. Rother, A. Criminisi, TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation, in: Proceedings of European Conference Computer Vision (ECCV), 2006.

[32] V. Ferrari, L. Fevrier, F. Jurie, C. Schmid, Groups of Adjacent Contour Segments for Object Detection, IEEE Trans. Pattern Anal. Mach. Intell. 30 (1) (2008) 36–51.

[33] C. Dance, J. Willamowski, L. Fan, C. Bray, G. Csurka, Visual categorization with bags of keypoints, in: ECCV International Workshop on Statistical Learning in Computer Vision, 2004.

[34] F.-F. L., P. Perona, A Bayesian Hierarchical Model for Learning Natural Scene Categories, Computer Vision and Pattern Recognition, IEEE Computer Society Conference on 2 (2005) 524–531.

[35] B. Leibe, A. Leonardis, B. Schiele, An Implicit Shape Model for Combined Object Categorization and Segmentation, in: Toward Category-Level Object Recognition, 508–524, 2006.

[36] S. Lazebnik, C. Schmid, J. Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, in: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, 2169–2178, 2006.

[37] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, P. Boyes-Braem, Basic objects in natural categories, Cognitive Psychology 8 (3) (1976) 382–439.

[38] S. El-Khoury, A. Sahbani, Handling Objects By Their Handles, in: IROS-2008 Workshop on Grasp and Task Learning by Imitation, 2008.

[39] R. Pelossof, A. Miller, P. Allen, T. Jebera, An SVM Learning Approach to Robotic Grasping, in: IEEE International Conference on Robotics and Automation, 3512–3518, 2004.

[40] N. Curtis, J. Xiao, Efficient and Effective Grasping of Novel Objects through Learning and Adapting a Knowledge Base, in: IEEE International Conference on Robotics and Automation, 2252–2257, 2008.

[41] J. Grezes, J. Decety, Does Visual Perception of Object Afford Action? Evidence from a Neuroimaging Study., Neuropsychologia 40 (2) (2002) 212–222.

[42] J. Bohg, C.Barck-Holst, K. Hübner, M. Ralph, D. Song, D. Kragic, Towards Grasp-Oriented Visual Perception in Humanoid Robotics, International Journal of Humanoid Robotics 6 (3) (2009) 387–434.

[43] A. Saxena, L. Wong, A. Y. Ng, Learning Grasp Strategies with Partial Shape Information, in: 23rd AAAI Conference on Artificial Intelligence, 1491–1494, 2008.

[44] KUKA, KR 5 sixx R650, `www.kuka-robotics.com`, last visited 2009.

[45] SCHUNK, SDH, `www.schunk.com`, last visited 2009.

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

[47] D. Kragic, M. Bjorkman, H. I. Christensen, J.-O. Eklundh, Vision for robotic object manipulation is domestic settings, Robotics and Autonomous Systems (2005) 85–100.

[48] M. Björkman, J. Eklundh, Foveated Figure-Ground Segmentation and Its Role in Recognition, in: British Machine Vision Conference, 2005.

[49] B. Rasolzadeh, A. T. Targhi, J.-O. Eklundh, An Attentional System Combining Top-Down and Bottom-Up Influences, in: Workshop on Attention and Performance in Computational Vision, 123–140, 2007.

[50] K. Grauman, T. Darrell, The pyramid match kernel: discriminative classification with sets of image features, in: Proceedings of IEEE International Conference on Computer Vision, vol. 2, 1458–1465, 2005.

[51] C. H. Ek, P. H. Torr, N. D. Lawrence, Gaussian Process Latent Variable Models for Human Pose Estimation, in: 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2007), vol. LNCS 4892, Springer-Verlag, Brno, Czech Republic, 132–143, 2007.

[52] G. Mori, S. Belongie, J. Malik, S. Member, Efficient shape matching using shape contexts, IEEE Trans. Pattern Analysis and Machine Intelligence 27 (2005) 1832–1837.

[53] C. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, 2006.

[54] C.-C. Chang, C.-J. Lin, LIBSVM: A Library for Support vector machines, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001.

[55] M. Craven, J. Shavlik, Extracting Tree-Structured Representations of Trained Networks, in: Advances in Neural Information Processing Systems (NIPS-8), MIT Press, 24–30, 1995.

[56] D. Martens, B. Baesens, T. V. Gestel, J. Vanthienen, Comprehensible Credit Scoring Models using Rule Extraction from Support Vector Machines, European Journal of Operational Research 183 (3) (2007) 1466–1476.

[57] D. Kragic, H. I. Christensen, Cue Integration for Visual Servoing, IEEE Transactions on Robotics and Automation 17 (1) (2001) 18–27.

[58] S. Ekvall, D. Kragic, Receptive Field Cooccurrence Histograms for Object Detection, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS'05, 84–89, 2005.

[59] D. Kragic, L. Petersson, H. I. Christensen, Visually guided manipulation tasks, Robotics and Autonomous Systems 40 (2-3) (2001) 193–203.

[60] J. Tegin, S. Ekvall, D. Kragic, B. Iliev, J. Wikander, Demonstration based Learning and Control for Automatic Grasping, Journal of Intelligent Service Robotics 2 (1) (2008) 23–30.

[61] N. Bergström, J. Bohg, D. Kragic, Integration of Visual Cues for Robotic Grasping, in: Computer Vision Systems, vol. 5815 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 245–254, 2009.

Jeannette Bohg holds a M.Sc. in Computer Science from the Technical University Dresden, Germany and a M.Sc. in Applied Information Technology with a focus on Art and Technology from Chalmers in Göteborg, Sweden. Since 2007 she is a Ph.D. candidate in the field of Computer Vision and Robotic Grasping at the department of Computer Science at the Royal Institute of Technology (KTH).

Danica Kragic received her B.S. degree in mechanical engineering from the Technical University of Rijeka, Croatia, and her Ph.D. degree in computer science from the Royal Institute of Technology (KTH), Stockholm, Sweden in 1995 and 2001, respectively. She is currently a professor in computer science at KTH and chairs the IEEE RAS Committee on Computer and Robot Vision. She received the 2007 IEEE Robotics and Automation Society Early Academic Career Award. Her research interests include vision systems, object grasping and manipulation and action learning.