Integration of Visual Cues for Robotic Grasping

Niklas Bergström, Jeannette Bohg, and Danica Kragic

Computer Vision and Active Vision Laboratory,
Centre for Autonomous System,
Royal Institute of Technology, Stockholm, Sweden
{nbergst, bohg, danik}@csc.kth.se

Abstract. In this paper, we propose a method that generates grasping actions for novel objects based on visual input from a stereo camera. We are integrating two methods that are advantageous either in predicting how to grasp an object or where to apply a grasp. The first one reconstructs a wire frame object model through curve matching. Elementary grasping actions can be associated to parts of this model. The second method predicts grasping points in a 2D contour image of an object. By integrating the information from the two approaches, we can generate a sparse set of full grasp configurations that are of a good quality. We demonstrate our approach integrated in a vision system for complex shaped objects as well as in cluttered scenes.

1 Introduction

Robotic grasping remains a challenging problem in the robotics community. Given an object, the embodiment of the robot and a specific task, the amount of potential grasps that can be applied to that object is huge. There exist numerous *analytical* methods based on the theory of contact-level grasping [1]. Even though these approaches work very well in simulation, they cannot simply be applied to object models reconstructed from typically sparse, incomplete and noisy sensor measurements.How to choose a feasible grasp from incomplete information about the object's geometry poses an additional challenge. This paper introduces a vision based grasping system that infers *where* and *how* to grasp an object under these circumstances. This involves a decision about where the hand is applied on the object and how it is orientated and configured.

Current state of the art methods usually approach this problem by concentrating on one of the two questions. The first group of systems, e.g. [2, 3] typically infers grasps based on 3D features resulting in many hypotheses where to apply the grasp. For each hypothesis, a hand orientation is determined. Heuristics are then applied to prune the number of grasp hypotheses. A drawback of these approaches is the high dependency on the quality of the reconstructed data. The second group of approaches, e.g. [4, 5] relies on 2D data and thus avoids the difficulty of 3D reconstruction. Grasp positions are inferred from a monocular image of an object. The difficulty here is the inference of a full grasp configuration from 2D data only. Additional 3D cues are required to infer the final grasp.

In this paper, we propose a method that aims at integrating 2D and 3D based methods to determine both *where* and *how* to grasp a novel, previously unseen object. The first part of the system matches contour segments in a stereo image to reconstruct a 3D wire frame representation of the object. An edge image containing only successfully matched contour segments serves as the input to the second part of the system. Hypotheses about where a grasp can be applied on the 2D contours are generated. By augmenting the 3D model with this 2D based information, we can direct the search for planar object regions. Plane hypotheses that are supported by contour points with a high grasping point probability will carry a high weight. The normal of these planes then define the approach vectors of the associated grasps. In that way both methods complement one another to achieve a robust 3D object representation targeted at full grasp inference.

This paper is structured as follows. In the next chapter we review different grasp inference systems that are applied in real world scenarios. In Sec. 3 we give an overview of the whole system. Section 4 describes the contour matching approach and Sec. 5 the grasp point inference system. This is followed by Sec. 6 where the integration of these two models is described. An experimental evaluation is given in Sec. 7 and the paper is concluded in Sec. 8.

2 Related Work

The work by [2] is related to our system in several aspects. A stereo camera is used to extract a sparse 3D model consisting of local contour descriptors. *Elementary grasping actions* (EGAs) are associated to specific constellations of small groups of features. With the help of heuristics the huge number of resulting grasp hypotheses is reduced. In our system however, the number of hypotheses is kept small from the beginning by globally searching for planar regions of the object model. [3] decompose a point cloud derived from a stereo camera into a constellation of boxes. The simple geometry of a box and reachability constraints due to occlusions reduce the number of potential grasps. A prediction of the grasp quality of a specific grasp can be made with a neural network applied to every reachable box face. In contrast to that, we drive the search for a suitable grasp through information about 2D grasping cues. These have been shown to work remarkably for grasping point detection in [4, 5].

In [4] an object is represented by a composition of prehensile parts. Grasping point hypotheses for a new object are inferred by matching local features of it against a codebook of learnt *affordance cues* that are stored along with relative object position and scale. How to orientate the robotic hand to grasp these parts is not solved. In [5] a system is proposed that infers a point at which to grasp an object directly as a function of its image. The authors apply machine learning techniques to train a grasping point model from labelled synthetic images of a number of different objects. Since no information about the approach vector can be inferred, the possible grasps are restricted to downward or outward grasps. In this paper, we solve the problem of inferring a full grasp configuration from 2D data by relating the 2D grasping cues to a 3D representation generated on-line.

There exist several other approaches that try to solve the problem of inferring a full grasp configuration for novel objects by cue integration. In [6], a stereo camera and a laser range scanner are applied in conjunction to obtain a dense point cloud of a scene with several non-textured and lightly textured objects. The authors extend their previous work to infer initial grasping point hypotheses by analysing the shape of the point cloud within a sphere centred around an hypothesis. This allows for the inference of approach vector and finger spread. In our approach however, we apply a stereo camera only and are not dependent on dense stereo matching. Due to the application of contour matching, we can obtain sparse 3D models of non-textured and lightly textured objects. [7] showed that their earlier 2D based approach is applicable when considering arbitrarily shaped 3D objects. For this purpose, several views of the object are analysed in terms of potential grasps. While the approach vector is fixed to be either from the top or from the side, the fingertip positions are dependent on the object shape and the kinematics of the manipulator. The best ranked grasp hypothesis is then executed. In our approach, we are not restricted to specific approach vectors whereas our grasp type is assumed to be one of the EGAs defined in [2]. Additionally determining the fingertip positions with the method proposed by [7] is regarded as future work. Finally, in [8] a framework is introduced in which grasp hypotheses coming from different sources e.g. from [2] are collected and modelled as *grasp hypothesis densities*. The grasp hypotheses are strongly dependent on the quality of the 3D object model. The density will therefore contain numerous potential grasps that may not be applicable at all. The authors propose to build a *grasp empirical density* by sampling from the hypotheses that are then grasped with the robot hand. In our case, we are also inferring potential grasps that may not be applicable in practice. However, we are not enumerating hypotheses from different sources but are integrating the information to infer fewer and better hypotheses that are ranked according to their support of 2D grasping cues.

3 System Overview

In our approach the process of grasp inference involves several steps: i) identification, ii) feature extraction, iii) cue integration and iv) grasping. A flow chart of the system is given in Fig. 1 and also shows the utilised hardware.

The first step involves figure-ground segmentation by means of fixation on salient points in the visible scene [9]. A combination of peripheral and foveal cameras is used that are mounted on a kinematic head. Figure 1 (b) and (c) show the left peripheral and foveal views of the head and (d) shows the segmented object.

In this paper, we focus on the feature extraction and cue integration. Full 3D reconstruction of objects with little or no texture from stereo vision is a difficult problem. However, it is debatable if a complete object model is always needed for grasping [7]. We propose a representation that is extractable from real world sensors and rich enough to infer how and where to grasp the considered

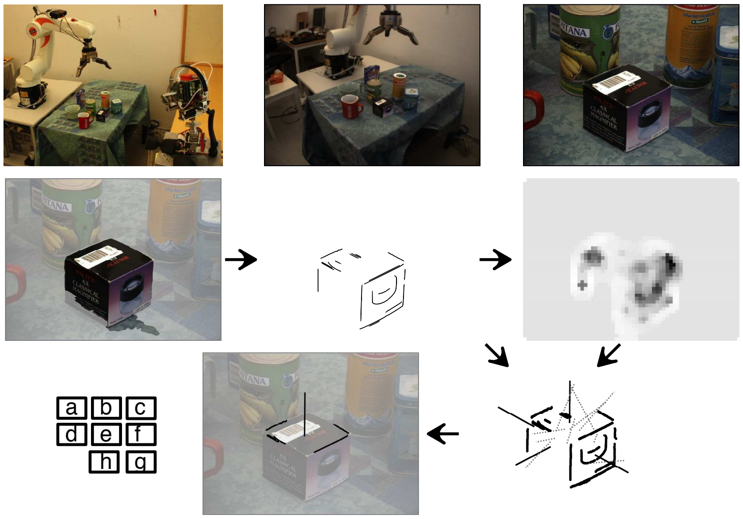

Fig. 1. (a): System setup with 6 DoF KUKA arm, a 7 DoF SCHUNK hand and the ARMAR 3 stereo head. (b,c): Left peripheral and foveal views. d-h: The steps of the grasping system.

object. A general observation that has driven our choice of representation is that many objects in a household scenario, including cups, plates, trays and boxes have planar regions. According to [2] these regions along with their coplanar relationships afford different EGAs. These grasps represent the simplest possible two fingered grasps humans commonly use.

The several steps to build such an object model composed of surfaces are shown in Fig. 1 (d-h). In the segmented foveal view (d) edges are detected and matched across the stereo images to form a 3D wire frame model (e). The projection of this wireframe in one of the images is used to predict where to grasp the object (f). The 3D model is then augmented with this information to detect planar regions that are supported by contour points with a high probability of being graspable (g). The four hypotheses with largest support are indicated with black lines, the others with dashed grey lines. The resulting surfaces provide hypotheses for how to grasp the object. The best hypothesis with respect to plane support and kinematic restrictions of the arm-hand configuration is finally shown in (h).

4 Partial 3D Reconstruction of Objects

Dynamic Time Warping (DTW) is a dynamic programming method for aligning two sequences. The method is described in detail in [10]. Below we give a brief overview of the key points of the algorithm, which is an extension to [11]. The different steps of the method are given in Fig. 2. The leftmost image shows the left foveal view of the object. Canny is used to produce an edge image from

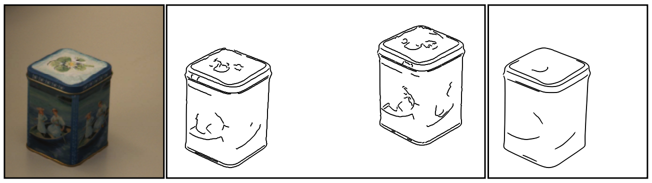

Fig. 2. Left: Left foveal view of object. **Middle**: Contours from left and right foveal views. **Right**: Successfully matched contours.

which connected edge segments (contours) are extracted. Spurious contours are filtered out by restricting their curvature energy and minimum length. The middle image pair shows the contour images from the left and right foveal views. Matching is performed between these two views. DTW is used both for solving the correspondence problem, i.e. which contour that belongs to which, and the matching problem, i.e. which point in the left contour corresponds to which point in the right contour. The latter is performed by calculating dissimilarities between the two contours based on the epipolar geometry, and finding the alignment that minimises the total dissimilarity. The former is performed by integrating the dissimilarity measure with gradient and curvature cues. This is one extension to [11], who could solve the correspondence problem more easily. Another difference is the extension of DTW to handle open and partial contours.

Many contours on the object surface correspond to texture. For 3D reconstruction, as well as 2D grasping point detection as described in Sec. 5, we are only interested in contours belonging to actual edges on the object. As seen in the middle image in Fig. 2, many contours stemming from texture do not have a corresponding contour in the other image and thus will be filtered in the DTW algorithm. Furthermore, shorter contours with higher curvature are less likely to be matched due to a too high total dissimilarity. The resulting matching is used to generate a sparse 3D model of the object.

5 Detecting Grasping Points in Monocular Images

Given the wireframe model reconstructed with the method introduced in the previous section, we search for planar regions that afford EGAs. As it will be shown later, fitting of planes to this raw model will result in many hypotheses stemming from noise and mismatches. In this section, we introduce a method that forms heuristics for searching and weighting of hypotheses according to their *graspability*. We introduce knowledge that comprises how graspable object parts appear in 2D and how these cues are embedded in the global shape of common household objects. Here, we are following a machine learning approach and classify image regions as graspable or not. We briefly describe how our feature vector is constructed and how the training of the model is done. A more detailed description can be found in [12].

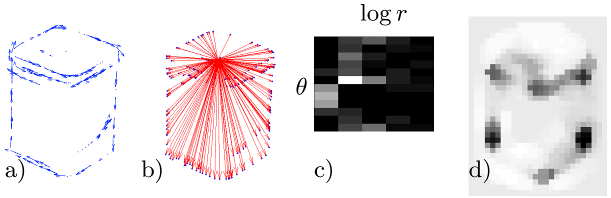

Fig. 3. Example of deriving the shape context descriptor for the matched contours shown in Fig. 2. (a) Sampled points of the contour with tangent direction. (b) All vectors from one point to all the other sample points. (c) Histogram with 12 angle bins and 5 log-radius bins. (d) Classification of the descriptors in each image patch.

Shape context (SC) [13] is a widely applied descriptor that encodes the property of *relative shape*, i.e. the relation of the global object shape to a local point on it. The descriptor is invariant to 2D rotation, scale and translation. Figure 3 shows an overview on the computation of SC. N samples are taken with a uniform distribution from the contour. For each point we consider the vectors that lead to the remaining $N-1$ sample points. We create a log polar histogram with K angle and radius bins to comprise this information. For the feature vector, we subdivide the image into 10 × 10 pixel patches. A patch descriptor is composed by accumulating the histograms of all those sample points that lie in the patch. We calculate the accumulated histograms at three different spatial scales centred at the current patch and concatenate them to form the final feature descriptor.

This feature vector is then classified by a grasping point model as either graspable or not. This model is an SVM that we trained off-line on the labeled database developed in [5]. An example of the classification results with an SVM trained on a pencil, a martini glass, a whiteboard eraser and two cups is shown in Fig. 3 d). Patches with a high graspability are characterised by rounded and parallel edges which indicate similarity to handles, rims or thin elongated structures. However, the approach direction is not easily inferred.

6 Cue Integration

To generate grasping hypotheses, we are interested in finding planar surfaces, i.e. finding contours that lie in the same plane. The set of plane hypotheses is defined as $\Pi = \{\pi_i\}$, $\pi_i = (\overline{n}_i, \mu_i)$, where $\overline{n}_i$ is the normal and μ_i the centre point on the plane. When searching for hypotheses, we start be selecting a point p_1 on one of the contours and a point p_2 nearby. We assume that these points are likely to lie in the same planar region(s) on the object. Then, there will be a third point p_3 on the remaining contours that defines such a region. By searching over the set of potential p_3, we try to find all these planes. Given p_1, p_2 and p_3, a plane hypothesis $\tilde{\pi}_i$ can be defined. Since the depth is quantised, the three selected points may produce a non optimal plane. Therefore we use RANSAC [14] over small contour regions defined by these points to optimise the plane. The hypothesis is accepted or rejected depending on the amount of contour points neighbouring p_1, p_2 and p_3 that are close enough to $\tilde{\pi}_i$. If accepted a more exact π_i is computed by performing regression on the full set of contour points not exceeding a certain distance to $\tilde{\pi}_i$. After the planes related to p_1 have been found, a new p_1 is selected and the procedure is repeated.

In order to restrict the search, whenever a contour point has been assigned to a plane it will be unavailable when choosing p_1. This will, apart from reducing the computational time, drastically reduce the number of hypotheses and remove most duplicates. This puts requirements on how the selection of p_1 is made. If chosen badly, it is possible to miss good hypotheses if for instance p_1 is not chosen from a contour corresponding to an actual edge. To solve this problem we use the information from the 2D grasping point detection. We start by extracting local maxima from the classification result. Because contour points in these regions are likely to be graspable, we choose p_1 from among these. As we will show in Sec. 7, this will result in a faster and more reliable search than randomly choosing p_1. The search for hypotheses continues until all points from regions with local maxima have been considered. We enforce that the normals are in the direction pointing away from the mean of all contour points.

As a final step planes are ranked according to graspability. For each plane

$$support(\pi_i) = \sum_{j \in \{all\ points\}} w(p_j) * P(p_j)/(\lambda_1 + \lambda_2) \tag{1}$$

where $w(p_j) = 1 - 2\frac{1}{1+e^{-d(p_j,\pi_i)}}$, $d(p_j, \pi_i)$ is the distance of p_j to the plane π_i, $P(p_j)$ is the probability that p_j is a grasping point, and $\lambda_{1,2}$ are the two largest eigenvalues from PCA over the inliers. This gives a support value that favours planes with dense contours whose points have a high graspability. Estimated planes may have a normal that does not correspond perfectly to the normal of the real plane. This plane will still get support from points that are close and are likely to stem from the real plane. Normalising with the sum of the eigenvalues ensures that planes without gaps are favoured over planes formed only from e.g. two sides. It also reduces the support for planes with points from falsely matched contours that will lie far from the actual object. Moreover, by calculating the eigenvalues we are able to filter out degenerate planes that have a small extension in one direction.

The normals of the final plane hypotheses are then defining the approach direction of the grasp and the smallest eigenvector of the related set of contour points the wrist orientation.

7 Experiments

The goal of the proposed method is to generate good grasping hypotheses for unknown objects in a robust and stable manner. Furthermore, as few false positives as possible should be generated. In this section, we will show that this is achieved for objects and scenes of varying geometrical and contextual complexity.

Figure 4 shows different objects used for the experiments. The corresponding matched contours are shown on the row below. The upper right of the figure contains the output of the grasping point detection. Finally, the last row shows the five planes with best support for each object. These four objects are selected to pose different challenges to our system: The hole puncher has a *complex geometric structure*, but with easily detectable edges. Due to many close parallel contours on the tape roll, we get some *false matches*. The tea canister object is *highly textured*, and its lid has many *parallel edges* which causes problems when finding the top plane. The magnifier box resides in a more complex scene in which Canny produces more *broken edges* that complicate the matching problem.

In all cases the two best hypotheses (red and green) shown in the bottom row are graspable, and correspond to how a human probably would have picked up the objects under the same conditions. For the puncher, the hypotheses give the choice of picking up from the object's front or top. This is an example of one of the benefits of our method: we do not need to constrain the approach direction. In the tape roll case there are several severe mismatches (marked in the figure). These correspond to a depth error of up to 50 cm, and are actually part of three plane hypotheses. Here the normalisation makes sure they get low support. Because of the parallel edges on the tea canister's lid, several hypotheses with good support are found on the top. The red hypothesis gets more support though, as it has more contour points close to the plane. In the case of the magnifier box, matching is harder, and we get much fewer and shorter edges. The longest contour is actually the one corresponding to the image of the magnifier. This affects the results from the support computations since the contours from the sides are not complete. The hypothesis from the right side clearly gets largest support. When finally choosing a grasp configuration kinematic constraints or other preferences will guide which of them to choose.

As mentioned in the previous section, the choice of the starting point is crucial to the performance of plane detection. We compared the method described in Sec. 6 to other approaches like random choice or a systematic search from the longest to the shortest contour. The assumption behind the latter method is that longer contours are more likely to originate from an actual edge of the object rather than from texture. We have performed an extensive evaluation of each method on the data in Fig. 4 to estimate their robustness, and will show how the proposed method outperforms the random and sequential method. Given the same input, all three methods will result in different plane hypotheses for each run due to the application of RANSAC in the plane estimation phase. The quality of a detected plane is measured by Eq. 1.

Figure 5 shows three representative examples for each of the three methods applied to the magnifier box. The two plane hypotheses that have the highest support are red and green. The best results for each method are shown in the leftmost column. Our method produced results similar to the top left example in Fig. 5 most times. The best result for the random selection only contains two hypotheses corresponding to real planes. The other two examples contain cases of missed planes (e.g. the top plane in the middle figure) and wrong planes being preferred over hypotheses corresponding to real planes. As with our method, the sequential selection produces more stable results. However, the problem of missed planes and ranking wrong planes higher than real ones persists.

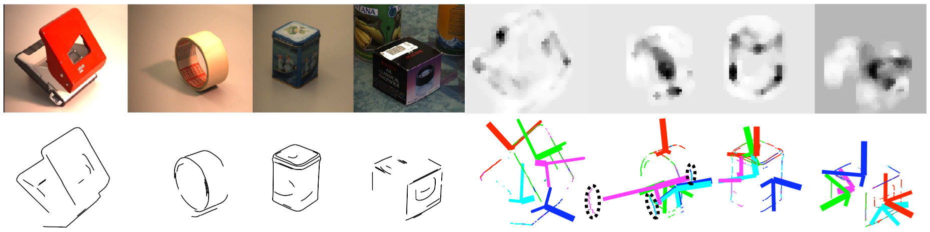

Fig. 4. Four objects, their matched contours, grasping point probabilities and finally the five best hypotheses for each object. The hypotheses are coloured, from best to worst, red, green, blue, cyan, magenta. False matches are circled in black. (Best viewed in colour)

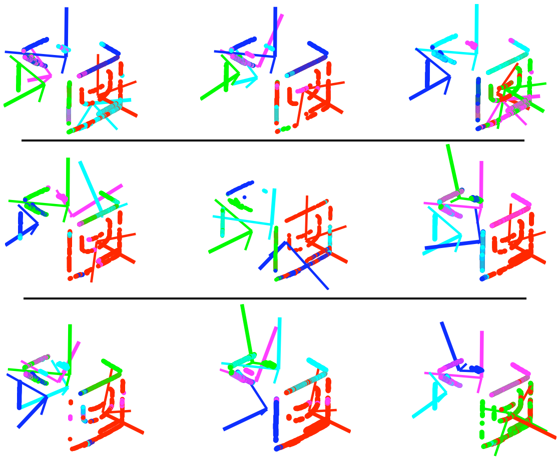

Fig. 5. Top row: Proposed method. Middle row: Random selection. Bottom row: Sequential selection. Colours in the same order as in Fig. 4 (Best viewed in colour)

In cases of simple hardly textured objects in non-cluttered scenes, all three methods have a comparable performance. However, in real world applications we need to deal with objects of arbitrary geometry in complex scenes in which segmentation is hard due to sensory noise, clutter and overlaps.

8 Conclusion

We have presented a method for generating grasping actions for novel objects based on visual input from a stereo camera. Two methods have been integrated. One generates a wire frame object model through curve matching, and associates EGAs to it. The other predicts grasping points in a 2D contour image of the object. The first accurately predicts how to apply a grasp and the other where to apply it. The integration generates a sparse set of good grasp hypotheses. We have demonstrated the approach for complex objects and cluttered scenes.

Our future work will exploit the use of the method in an integrated learning framework. Hypotheses will be generated as proposed and used for picking up objects. The system will then be able to view the object from different directions in order to generate a more detailed model.

Acknowledgments This project has been supported by the EU IST-FP7-IP GRASP (2008-2012) and the Swedish Foundation for Strategic Research through project CORS.

References

1. Nguyen, V.D.: Constructing stable grasps. Int. J. on Robotics Research **8**(1) (1989) 26–37
2. Kraft, D., Pugeault, N., Baseski, E., Popovic, M., Kragic, D., Kalkan, S., Wörgötter, F., Krueger, N.: Birth of the Object: Detection of Objectness and Extraction of Object Shape through Object Action Complexes. Int. J. of Humanoid Robotics (2009)
3. Hübner, K., Kragic, D.: Selection of Robot Pre-Grasps using Box-Based Shape Approximation. In: IEEE Int. Conf. on Intelligent Robots and Systems. (2008) 1765–1770
4. Stark, M., Lies, P., Zillich, M., Wyatt, J., Schiele, B.: Functional Object Class Detection Based on Learned Affordance Cues. In: 6th Int. Conf. on Computer Vision Systems. Volume 5008 of LNAI., Springer-Verlag (2008) 435–444
5. Saxena, A., Driemeyer, J., Kearns, J., Ng, A.Y.: Robotic Grasping of Novel Objects. Neural Information Processing Systems **19** (2006) 1209–1216
6. Saxena, A., Wong, L., Ng, A.Y.: Learning Grasp Strategies with Partial Shape Information. In: AAAI Conf. on Artificial Intelligence. (2008) 1491–1494
7. Speth, J., Morales, A., Sanz, P.J.: Vision-Based Grasp Planning of 3D Objects by Extending 2D Contour Based Algorithms. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems. (2008)
8. Detry, R., Başeski, E., Krüger, N., Popović, M., Touati, Y., Kroemer, O., Peters, J., Piater, J.: Learning object-specific grasp affordance densities. In: Int. Conf. on Development and Learning. (2009)
9. Björkman, M., Eklundh, J.O.: Attending, Foveating and Recognizing Objects in Real World Scenes. In: British Machine Vision Conference. (2004)
10. Bergström, N., Kragic, D.: Partial 3D Reconstruction of Objects for Early Reactive Grasping. Technical report, CAS, KTH Stockholm (2009) `www.csc.kth.se/~nbergst/files/techreport09.pdf`.
11. Romero, J., Kragic, D., Kyrki, V., Argyros, A.: Dynamic Time Warping for Binocular Hand Tracking and Reconstruction. In: IEEE Int. Conf. on Robotics and Automation. (May 2008) 2289–2294
12. Bohg, J., Kragic, D.: Grasping Familiar Objects Using Shape Context. In: Int. Conf. on Advanced Robotics. (June 2009)
13. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Trans. on Pattern Analysis and Machine Intelligence **24**(4) (2002) 509–522
14. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6) (1981) 381–395