Master Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Visuo-Haptic Uncertainty-Driven Object Shape Completion

Bc. Lukáš Rustler

Supervisor: Mgr. Matěj Hoffmann, Ph.D.
Supervisor–specialist: Jens Lundell
Field of study: Cybernetics and Robotics
January 2022

Acknowledgements

I want to thank everyone who helped me with this work. Firstly, to Mat¥j Hoffmann, who gave me the opportunity to be part of this project, supervised the thesis and supported me during the whole time from technical to formal aspects of this work. I also want to thank Jens Lundell, whose knowledge and experience helped me overcome many problems, and the results would not be possible without him. My gratitude also belongs to Jan Behrens, who provided the base of a simulation environment and gave me an introduction to its functionality. I also thank Martin ’rámek, who helped me create a model of an additional finger for the robot.

I would also like to thank my family for continuous support through the years and my friends for their encouragement. And finally, I thank CTU in Prague for being such a goodalma mater.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prohla²uji, ºe jsem p°edloºenou práci vypracoval samostatn¥ a ºe jsem uvedl ve²keré pouºité informa£ní zdroje v souladu s Metodickým pokynem o dodrºování etických princip· p°i p°íprav¥ vysoko²kolských prací.

V Praze dne 04. ledna 2022

. . . .
Luká² Rustler

Abstract

Shape completion is the task of reconstruction of an object using incomplete information. Recently, advancements in single-view visual-only shape completion have been made. However, the objects can self-occlude or be adversarial for cameras. In this work, we propose a novel active visuo-haptic shape completion method (Act-VH). The method combines a point cloud from a stationary camera and data from haptic exploration with a robotic manipulator. A state-of-the-art implicit surface neural network is used to complete the shapes. The location to explore is actively computed from the reconstruction uncertainty. We evaluate the method both in simulation, for which we present a unique environment, and in the real world. In both worlds, real-time visualizations of current progress are provided. The performance is compared to baseline methods in terms of reconstruction accuracy and exploration efficiency. The results show that Act-VH outperforms all baselines in both criteria. In addition, we perform grasp experiments on real-world objects, reaching a significantly higher grasp success rate than the baseline. Finally, Act-VH is compared to an active visual-only method that can obtain information from occluded parts. Together, this work opens up the door for using active visuo-haptic shape completion in real-world robotic tasks.

Keywords: Perception for Grasping and Manipulation, RGB-D Perception, Deep Learning for Visual Perception, Shape Completion, Shape Reconstruction

Supervisor: Mgr. Mat¥j Hoffmann, Ph.D.

Abstrakt

Modelováni tvaru objektu je úloha zabývající se rekonstrukcí objektu z neúplné informace. V nedávné dob¥ pokro£ila rekonstrukce z neúplné, pouze vizuální informace. Nicmén¥ p°edm¥ty se m·ºou p°ekrývat nebo být nep°átelské pro kameru. Jako °e²ení navrhujeme na²i metodu nazvanou aktivní visuo-haptické modelování tvaru (Act-VH). Tato metoda kombinuje mrak bod· ze stacionární kamery a data z haptické explorace pomocí robotického manipulátoru. Moderní neuronová sít na principu implicitních povrch· je pouºita pro dopl¬ování tvar·. Místo pro exploraci je aktivn¥ po£ítáno z nejistoty rekonstrukce. Metoda je posouzena v simulaci, pro kterou p°edstavujeme unikátní simula£ní prost°edí, a v reálném sv¥t¥. V obou sv¥tech jsou dostupné vizualizace aktuálního pokroku v reálném £ase. P°esnost rekonstrukce a efektivita explorace je porovnána s existujícími metodami. Výsledky ukazují, ºe Act-VH ostatní metody p°ekonává. Krom¥ toho provádíme sérii uchopovacích experiment·, ve kterých jsme dosáhli vy²²í úsp¥²nosti neº existující metody. Nakonec je na²e metoda porovnána s aktivní, pouze vizuální metodou, která je schopná získat informaci i ze zakrytých £ásti. Celkov¥ je tato práce vstupním bodem pro pouºití visuo-haptického dopl¬ování tvar· v robotických úlohách v reálném sv¥t¥.

Klí£ová slova: Vnímání pro uchopování a manipulaci, Vnímání pomocí RGB-D, Hluboké u£ení pro vizuální vnímání, Modelování tvaru objekt·, Rekonstrukce objekt·

P°eklad názvu: Visuo-haptické modelování tvaru objektu °ízené nejistotou

Contents

Chapter 1

Introduction

Figure 1.1: Schematic operation of our active visuo-haptic shape completion method (Act-VH). An initial depth image is used to construct an initial point cloud. Neural network (IGR) generates several possible shape reconstructions, whose discrepancy is used to form a single object reconstruction with uncertainty. The voxel with the highest uncertainty is selected for haptic exploration, and a new points are added to the object representation. This process is repeated, further refining the shape reconstruction. From Rustler et al. [1].

Shape completion is the task of reconstruction of an object using incomplete information. The output object is usually a triangular mesh or a voxel grid. Shape completion is an active research problem with potential in a wide range of applications and influences robotic tasks, such as grasping. Nowadays, the information is often acquired by visual sensors, which provide RGB images with or without depth data. However, using visual-only data is complicated. For example, when only a stationary camera is available, the object self-occludes,i.e., information from only one side of the object is available. One can collect more samples, but it requires more cameras, a movable camera, or a turntable under the object. Additionally, even with current cameras, a lot of noise is present when capturing adversarial objects such as transparent or dark items. A straightforward solution is to explore the unseen parts by touch and reconstruct the shape with visuo-haptic data. However, current visuo-haptic methods use a heuristic to choose where to touch [2] or require an impractical number of touches with enormous computational time [3, 4].

We provide a solution to these issues with our Act-VH method. Act-VH is an active visuo-haptic shape completion method, which iteratively refines the output shape. The method is depicted in Fig. 1.1. Firstly, depth information of the scene is used to construct an initial point cloud. Then, a state-of-the-art deep implicit surface neural network (Implicit Geometric Regularization for Learning Shapes (IGR) [5]) is used to generate a set of possible shapes, which is utilized to obtain a single shape with uncertainty. The segment of the object with the highest uncertainty is selected for haptic exploration, providing new information which is merged to the point cloud. The entire process is repeated, further refining the reconstruction.

A manuscript based on the visuo-haptic shape completion part of this work is currently under review for The IEEE International Conference on Robotics and Automation (ICRA) 2022 with Robotics and Automation Letters (RA-L) option [1]. Figures that were used in this manuscript as well as here include a reference to [1] in their captions. An accompanying video is here: https://youtu.be/Ft1PUYRNFHw. The simulation environment and the data from experiments pertaining to [1] are available at https://github.com/ctu-vras/visuo-haptic-shape-completion. Additional internal code is available in the attachment of this work. Data collected during the experiments (point clouds and ROSbags) can be downloaded from https://drive.google.com/drive/folders/1Du8hVDbsFYqEvb-hwS6ISpnaE4RGJW8c?usp=sharing.

Chapter 2

Related Work

Data for shape completion can be obtained from several inputs. The most obvious is purely visual input from RGB or depth sensors. In combination with robotic arms and hands, the data can be also collected by haptic exploration. In many recent works, these two inputs are combined. The data can be collected only once or gathered actively usually known as active perception [6].

2.1 Visual-only Shape Completion

Visual-only approach to shape completion is the most common one. Usually sensors with depth perception are used. This approach is beneficial because of the global nature of the information gained a single sample contains a lot of data. The first visual approaches were geometry- or template-based. The geometry ones assume the objects are symmetric and complete the reconstruction by mirroring it through its axis of symmetry, as done by Bohg *et al.* [7]. Other work from Schnabel *et al.* [8] uses primitive shapes detected in the input point cloud. Template-based methods just search in a database of known objects to find the most similar one. Example of template-based method was proposed by Pauly *et al.* [9]. It is clear that these approaches will not work in the general case, *e.g.*, the geometry-based method will perform poorly for non-symmetric objects and the template-based will fail if there is incorrect or no match in the database.

With progress in Machine Learning (ML), approaches based on ML were proposed [2,5,10–15], using mainly Gaussian Processes (GPs). The early approach by Li *et al.* trained a Gaussian Process Implicit Surface (GPIS) to reconstruct objects [10]. The problem with GPIS is in its poor scaling over a lot of points in the input point cloud. Thus, the input needs to be down-sampled which results in loss of information and details of the reconstruction. More recent works utilize Deep Learning (DL) methods, mainly Convolutional Neural Networks (CNNs) [2,11–13]. Using CNN, the objects are represented as voxel grids. Using voxel grids is useful, for example, to represent probabilistic attributes of the reconstruction. However, to preserve fine details, the grids would have to be sampled with very high resolution. And it is, even with modern computers, very computationally demanding because the requirements for memory grow cubically with the resolution.

However, in the world of shape reconstruction, new approaches based on implicit surfaces were proposed [5,14,15]. These methods work directly with point clouds (with different needs of point cloud preprocessing), returning triangle meshes. The methods produce reconstructions with higher quality, while preserving better efficiency than voxel grid methods. They are not originally meant for shape completion from incomplete input, but can be also utilized that way.

Even though shape completion results of the mentioned methods are impressive, it will probably never reach the quality of reconstruction created from a full point cloud. However, to capture the complete point cloud, the camera or the object must be moveable. It can be done by choosing the next-best-view [16–18], using a turntable under the objects [19–22] or using a coordinate measuring machine with a laser scanner [23].

2.2 Haptic-Only Shape Completion

Current robots are already capable of haptic exploration and can be also used for shape completion. With modern force and tactile sensors, haptic exploration can be more accurate than the visual methods. Moreover, the robots are able to explore all parts of the object, even from behind. Most recent haptic-only approaches use ML-based methods similar to the ones used in visual-only approaches. Ottenhaus et al. [24] proposed a method based on implicit shape potentials, Yi et al. [25] on GPs, Driess et al. [26] on GPIS's and Dragiev et al. [27] on Gaussian Process Implicit Shape Potentials (GPISPs). The advantage of Gaussian-based methods is the ability to compute uncertainty and actively gather new information to reduce it. Even though touch is usually more precise than visual sensors, it also captures more local information explores smaller regions. Consequently, haptic-only completion requires a lot of touches (tens to hundreds), which is both impractical and time-demanding.

2.3 Visuo-Haptic Shape Completion

Combining the previous approaches could theoretically preserve the advantages and address the limitations at the same time. Again, methods based on GPIS's [28, 29] and GPs [4] are proposed. These methods require dense information all around the object. With a fixed camera capturing only one side of the object, a lot of surface remains to be explored by touch. CNN-based methods [3, 30] can be used with fewer touches, but suffer from low resolution of the reconstruction. Smith et al. proposed approaches [31, 32] based on Graph Neural Network (GNN). Reconstructions by these methods have a higher resolution, but are nonsmooth and, for now, evaluated only in a simulation.

An important part of haptic exploration is the decision where to touch. One can touch the object randomly as done by Smith et al. [31], or always select a position opposite of the camera (from behind) as Watkins-Vall et al. [30]. These are, however, not so effective. Uncertainty from Gaussian distribution can be used to explore more effectively [3, 4, 28, 29], or it can be learned where to touch as in Smith et al. recent work [32].

The main contributions of the work presented in this thesis are: (i) a novel active visuo-haptic shape completion method; (ii) a visuo-haptic simulation environment; (iii) real time visualizations suitable for benchmarking; (iv) an empirical evaluation of the proposed method against state of the art, presenting, both in simulation and on real hardware, improvements in terms of reconstruction accuracy and grasp success rates.

Chapter 3

Shape Completion Methods

This chapter gives an overview of state-of-the-art methods for shape reconstruction. Our experiments to determine the most suitable one for this work are presented. Examples of two groups of shape completion approaches are given, together with experimental results and comparison.

3.1 Neural Network Types

Shape completion method is an essential part of the pipeline. We tested and compared multiple state-of-the-art networks based on two principles: (i) voxel-based; and (ii) implicit surface based.

For each network, a new dataset was created from the same training samples in a format needed by the given network. The tested networks are described below, and a table with the properties of each network can be found in Section 3.2.

3.1.1 Voxel-based Networks

The network introduced by Lundell et al. [13] was selected as a representative of this type of networks. The authors have already proved that the network can be used for shape completion from an incomplete point cloud. However, as our experiments showed, the method is impractical for iterative shape refining. The network is trained from multiple incomplete views of the same object. To allow incorporation of tactile information, the network would have to be trained evenly with simulated tactile information all around the objects (a similar thing was done in [30]), which is unfeasible in the general case.

The network utilizes a voxel grid as input. The main disadvantage of this approach is the cubically growing computation and memory requirements which result in low object resolution. As such, fine object details are not preserved, which is important when, for instance, sampling grasps.

3.1.2 Implicit surface networks

These networks, in contrast to voxel-based methods, can reconstruct visuo-haptic data without explicit training. Firstly, we evaluated the DeepSDF [14] network, which learns the Signed Distance Function (SDF) defined as:

$$SDF(x) = s; \tag{3.1}$$

where $x \in \mathbb{R}^3$ is a three-dimensional position of a point in a point cloud and $s \in \mathbb{R}$ is a signed distance to the surface. The signed distance is used to determine which points are inside (negative) and outside (positive) of the surface. The desired triangle mesh can be obtained with Marching Cubes algorithm [33] from an isosurface of $SDF(\cdot) = 0$, i.e., from all points with zero distance to the surface. Implicit surface principle shown on a ball can be seen in Fig. 3.1.

(a) : 2D depiction of the signed distance field.

(b) : Zoomed cross-section with implicit surface SDF = 0 (black semicircle), red points outside (SDF > 0) and blue point inside (SDF < 0).

(c) : Triangle mesh reconstructed from implicit surface SDF = 0.

Figure 3.1: 2D signed distance field (a) with a cross-section showing outside and inside points (b). Triangle mesh reconstructed from isosurface SDF = 0. Inspired by Park et al. [14].

The principle of SDF allows to train the network with the whole point cloud and then use just part of it to reconstruct the mesh, which enables integration of tactile information. However, DeepSDF is heavily dependent on preprocessing of the input. The network expects, for both training and prediction, a point cloud with around 500 000 points randomly sampled with Normal distribution around the surface. Each sample consists of a point location in 3D space and a signed distance to the surface. See Fig. 3.2a for an example.

Computation of signed distances to the surface is easily achievable when sampling the whole mesh for training but more problematic when creating the input from incomplete point cloud captured with a camera, as it is not clear what is inside and outside. For this reason, we started to experiment with IGR [5]. This work uses the same network structure as DeepSDF [14] but works with raw point clouds only 3D position and 3D normal for each point are needed. See Fig. 3.2 for comparison with DeepSDF.