Active vision for dexterous grasping of novel objects
Arruda, Ermano; Kopicki, Marek; Wyatt, Jeremy

DOI:
10.1109/IROS.2016.7759446

License:
None: All rights reserved

Document Version
Peer reviewed version

Citation for published version (Harvard):

Link to publication on Research at Birmingham portal

Publisher Rights Statement:
(c) 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

Checked 25/8/2016

General rights
Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

• Users may freely distribute the URL that is used to identify this publication.
• Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
• Users may use extracts from the document in line with the concept of ‘fair dealing’ under the Copyright, Designs and Patents Act 1988 (?)
• Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

Take down policy
While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact UBIRA@lists.bham.ac.uk providing details and we will remove access to the work immediately and investigate.
Active vision for dexterous grasping of novel objects

Ermano Arruda  Marek Kopicki  Jeremy L. Wyatt

Abstract—How should a robot direct active vision so as to ensure reliable grasping? We answer this question for the case of dexterous grasping of unfamiliar objects. When an object is unfamiliar, much of its shape is by definition unknown. An initial view will recover only some surfaces, leaving most of the object’s surface unmodelled, and also leaving shadow regions which may or may not contain obstacles. These two features make it difficult both to select reliable grasps, and to plan safe reach-to-grasp trajectories. Grasps typically fail in one of two ways, either unmodelled objects in the scene cause collisions, or object reconstruction is insufficient to ensure that the grasp points provide a stable force closure. These problems can be solved more easily if active sensing is guided by the anticipated actions. Our approach has three stages. First, we take a single view and generate candidate grasps from the resulting partial object reconstruction. Second, we drive active vision to maximise surface reconstruction quality around the planned contact points. During this phase the anticipated grasp is continually refined. Third, we direct gaze to unmodelled regions that will affect the planned reach to grasp trajectory, so as to confirm that this trajectory is safe. We show, on a dexterous manipulator with camera on wrist, that our approach (85.7% success rate) outperforms a randomised algorithm (64.2% success rate). Our approach also matches the grasp success of our original method, but with fewer views to pick the grasp.

I. INTRODUCTION

Grasping of novel objects is a hard problem on which there has been steady progress [10], [11], [8], [14], [7], [16], [6], [3], [15], [4]. We now possess methods that are able to generate dexterous grasps for unfamiliar objects, using incomplete object reconstructions. Nonetheless, the reliability of grasping rises with the quality and completeness of the reconstruction available. Given an active vision system, we would like to minimise the number of views taken, while maximising grasping reliability.

At the root of the difficulties is a chicken and egg problem. On the one hand, given that the initial point cloud can be highly incomplete, it is hard to plan a reliable grasp to begin with. On the other hand, if we knew the likely planned grasp then we could direct gaze more efficiently. In this paper we solve this problem by employing a grasp planner that can generate grasps for novel objects in the face of fragmentary reconstructions. We use grasp candidates to guide active vision, and the results of active vision to refine grasp planning.

First we describe related work, and then proceed to describe our active vision method. This has two parts, a routine driven by the planned contact points, and a routine driven by the need to ensure a safe reach to grasp trajectory. We then present experimental results on 14 novel objects, comparing our method with a randomised view planner.

II. RELATED WORK

Active vision, or more generally active perception, is defined as the study of modelling, planning and control strategies for perception when the sensor can be actively moved [2]. The field of active perception for robots started with work by [2], [1]. The greatest advantage of active perception is that many problems that are hard to solve in the passive observer paradigm become easier. In the context of manipulation, researchers have focused on devising strategies for view selection based on recovery of the full shape of the object to be grasped [12], [5].

Nonetheless, for most practical manipulation purposes, full object reconstruction is either too costly or simply infeasible. It is also typically redundant, since most of the time only a limited portion of the object surface is in contact during a grasp. These practical considerations were taken into account by a number of works. For instance, the approach proposed by [10], [11] is able to transfer previously demonstrated grasps to new objects without the need for grasp force analysis, and is able to cope with an incomplete point clouds of the target object. Additional efforts were made by [8], focusing on task and grasp transferability from limited...
training data, i.e. demonstration and partial object point clouds. The work done by [7], focused on learning grasps by letting the robot autonomously explore and try grasps while at the same time being able to transfer those self-discovered grasps to novel objects. In [15], efforts were made towards finding stable grasps given limited visibility of object shape from cluttered scenes. The problem of shape incompleteness is dealt with by Bohg et al. [4] by trying to fill the gap between the missing parts of the objects using symmetry assumptions.

Although there has been progress in grasping in the face of partial reconstruction and novel objects, there exists a clear need for active perception so as to decide when and how much to fill in the missing information. In addition, we wish to ensure robot safety, avoiding hardware damage due to unexpected collisions. We now proceed to describe our proposed approach to tackle these issues.

III. VIEW SELECTION

We first sketch our method, and then proceed to the details. The robot begins by taking a single view from a fixed location of the scene. A depth camera mounted on the robot’s wrist is used. The robot is then able to choose views, which in turn provide incomplete point clouds of the object. A dexterous grasp planning algorithm is then run, which generates a large number of candidate grasps on the partial point cloud for the object. These grasps will typically assume the existence of graspable surfaces on both sides of the surface defined by the point cloud. The predicted contact locations are then used to drive the next view. The next view is chosen to maximise the quality of the point cloud at the planned contact locations. If a grasp cannot be found, we employ information gain view planning, using a 3D occupancy map. This is used to calculate the probability of a collision free trajectory. Active views for safety are driven to reduce the average entropy in cells through which the candidate reach-to-grasp trajectory passes. This ensures a safe grasp. We now proceed to describe the representations, and the three criteria used to drive active vision at different stages (contact based, information gain, and safety based).

A. Representations

We start by describing the underlying representations used to define our approach. Let $\Xi = [\xi_1, \xi_2, \ldots, \xi_N]$ be a list of possible camera poses, where $\xi_i \in SE(3)$, and $V \subset \Xi$ is the set of already visited camera poses. This list must be finite, and should provide good coverage of the workspace. In addition, let $\gamma$ be a point cloud obtained from a certain camera pose $\xi$. We define $\Gamma_t$ as the combined object point cloud, segmented from the table plane, after $t$ views have been taken,

$$\Gamma_t = \Gamma_{t-1} \cup segmented(\gamma), \quad \text{(1)}$$

i.e., $\Gamma_t$ is the result of segmenting the object point cloud from the table plane in $\gamma$ and integrating this result with our previous obtained object point $\Gamma_{t-1}$.

In addition to the object point cloud, we also maintain a representation of the full robot workspace as a 3D occupancy grid, implemented with an octree. We shall refer to this 3D occupancy grid as $\Lambda$, which is updated after each view and observation ($\xi, \gamma$). The implementation we use [9] allows us to easily represent known and unknown parts of the robot workspace $\Lambda$ and thus to define the information gain and safety based view planning strategies.
It is possible to find a grasp trajectory $\tau$ by transferring a learnt grasp $G$ to the given object represented by $\Gamma_i$, using the method of Kopicki, Wyatt et al. [10]. Using the same method, we are able to extract contact points from $\Gamma_i$, yielding a list of contacts $C = [c_1, \ldots, c_M]$, where $c_i = (w_i, p_i, n_i)$ is composed of a weight $w_i \in \mathcal{R}$, indicating its relevance to the grasp, the contact location $p_i \in \mathcal{R}^3$, and the surface normal at that point $n_i \in \mathcal{R}^3$. Points are relevant to a grasp if they are close to planned contact point for that grasp. The weight $w_i$ falls off exponentially as the distance from the planned finger position increases. Let us also define $\Omega = [(c_1, z_1), \ldots, (c_M, z_M)]$ as a list of contact points expanded to include the current quality $z_i$ of the observation of each point from the best view $\xi$ to date. Contact driven vision prioritizes looking at the planned contact points for which there is currently low quality reconstruction, rather than elsewhere on the object. We now describe contact-based view selection in detail.

B. Contact Based View Selection

We let the viewing direction of a certain view pose $\xi_k \in \Xi$ be the vector $v_k$, which we always constrain to point towards the object $\Gamma_i$. We define the quality of observation of a contact point $c_i$ from a given $\xi_k$ as

$$\theta_{ki} = \theta(\xi_k, c_i) = \arccos(\min(0, v_k^T n_i)). \quad (2)$$

This models the fact that the depth errors rise as the surface becomes perpendicular to the image plane. Thus when looking at contact point with surface normal $n_i$, we assign higher values to views that are square on, where the image plane normal and the surface normal directly oppose one another, i.e. $n_i$ and $v_k$ form an angle of 180 degrees, according to our convention that $v_k$ always looks towards the object.

Thus, for each element $(c_i, z_i) \in \Omega$ we store the contact points $c_i$, and define $z_i$ the best quality of observation to date with respect to all visited poses as

$$z_i = \arg \max_{\xi_j \in \mathcal{V}} \theta_{ji}. \quad (3)$$

Finally, let $F_\tau = [f_1, \ldots, f_R]$ be the list of finger link normals for the finger surfaces that involved in the grasp. These are calculated for the last time step of the trajectory $\tau$. We then define the value of a particular (untired) view with respect to a particular contact $c_i$ as

$$\sigma(\xi_k, F_\tau, c_i) = w_i \sum_{r=1}^{R} \max_{\theta_{ki}}(\theta_{ki}, z_i) \frac{1 - \text{sign}(f_r^T n_i)}{2}. \quad (4)$$

This defines high value views as being those views which gaze head on at contact points. Note that when looking at a certain contact point $c_i$ we are able either to improve our previous best viewing quality if $\theta_{ki} > z_i$, or leave it as it is. Note also that the multiplying term $\frac{1 - \text{sign}(f_r^T n_i)}{2}$ serves as a switch that yields 0 or 1. This simply ensures that the view must be of the side of the point cloud where the finger will contact. In other words it models the geometric constraint that a link must have with a contact point, i.e. the surface normal $f_r$ of a given finger link must point in the opposite direction of the contact normal $n_i$, otherwise this contact point is meaningless with respect to this given finger link, meaning that viewing it is not useful. Finally, the normalised weight $w_i$ scales this value according to its overall relevance to the grasp as defined by the approach of Kopicki, Wyatt et al. [10]. It follows that the total utility of a given view $\xi$ is given by

$$u_1(\xi, \Omega, \tau) = \sum_{i=1}^{N} \sigma(\xi, F_\tau, c_i). \quad (5)$$

We are then able to rank the potential views by calculating the total value of a view with respect to all contact points, and picking the view that has the maximum value according to Eq 6.
Algorithm 2 Select Next Best View Contact Based
1: function NBVCOUNTACTBASED(Ξ, Γ, Λ, V, Ω, τ)
2: ξ = None
3: if |V| = 0 then
4: ξ = head(Ξ)
5: else if Ω ≠ ∅ then
6: Let ξ be selected according to Eq 6
7: else
8: Let ξ be selected according to Eq 14
9: end if
10: Return ξ
11: end function

C. Information Gain View Selection

Of course, if no grasp can be found, then grasp driven view selection cannot run. In this case, the robot should look at the workspace around the recovered point cloud. To support this we define an information-gain based utility function for view selection. Intuitively, this strategy makes sense, since no contacts were found with the knowledge we have about the object shape so far, represented by Γi. Therefore, one should ideally adopt an exploratory behaviour to seek for new parts of the object.

For this purpose, let bmin(Γt), bmax(Γt) ∈ R3 be the respective minimum and maximum limits of the bounding box that circumscribes the object point cloud Γt. We are then able to extract the set of voxels Λobject ⊂ Λ inside this bounding box. If we assume the surface of this voxelised solid box Λobject is visible from all cameras, as shown in Fig 3. Then we can define a simple strategy to minimise the entropy about the object’s shape, by selecting views that are going to have maximum predicted information gain about the voxels in Λobject. Intuitively, our goal is to select views such that we gradually sculpt the solid cube, such that we will eventually reach a constant entropy value for this cube, due to self-occluding parts of the object, from which point no views are going to bring any more information gain.

Our first step to fulfill this goal is to define a rule with which we can determine the set of visible voxels in Λobject visible from a camera pose ξ. The visibility test is performed using a typical frustum culling graphics procedure, with a few slight modifications. First, we transform the set of voxels Λobject into the camera coordinate system. During the projection phase of the pipeline, we allow many free voxels along the line of sight to be projected onto identical image coordinates, but we do not allow either unknown voxels, nor occupied voxels to be projected on top of one another. As a consequence, we find a border in our initial solid cube Λvisible(ξ) ⊂ Λobject, which contains all free voxels visible on the image plane, together with boundary voxels that might be either unknown or occupied, as shown in Fig 5. Thus,

\[ ξ = \arg \max_{ξ ∈ Υ - V} u_1(ξ, Ω, τ). \] (6)

Algorithm 3 Safety Exploration

1: function SAFETYEXPLORATION(Ξ, Γ, Λ, τ, V)
2: stop = false
3: while not stop do
4: ξ, value = safetyNBV(Ξ, Λ, τ)
5: V = append(T, ξ)
6: γ = capture(ξ) ⊳ Point cloud from pose ξ.
7: Γ = Γ ∪ segmented(γ)
8: Λ = updateOctTree(Λ, γ, Ω, ξ)
9: T = append(T, τ)
10: stop = CHECKSTOPSAFETY(value)
11: end while
12: p = p(τ|Λ)
13: Return (V, p, Γ, Λ)
14: end function

Algorithm 4 Safety Exploration View Selection

1: function SAFETYNBV(Ξ, Λ, τ)
2: Λc = findPassingVoxels(Λ, τ) ⊳ Finding voxels through which the hand is passing
3: Using Λc, let ξ be selected according to Eq 14
4: value = u2(ξ, Λc)
5: Return (ξ, value)
6: end function

\[ Λ_{visible}(ξ) = \{s_1, \ldots, s_D\} \] is defined as our set of voxels of interest for information gain prediction. We then follow to describe the information gain prediction for the set of voxels Λvisible(ξ).

1) Information Gain Prediction: Let the occupancy probability of a voxel \( s_d \in Λ_{visible} \) up to our most recent observations \( o_{1:t} \) be \( p_{sd} = p(s_d|o_{1:t}) \). We can write the entropy of this voxels by viewing it as a Bernoulli random variable with entropy

\[ H(s_d) = -p_{sd} \log(p_{sd}) - (1 - p_{sd}) \log(1 - p_{sd}), \] (7)

By using a log-odds representation of our occupancy probability such as in [9], [13], we can then define future predicted occupancy probability of \( s_d \) as

\[ L(s_d|o_{1:t}, o'_{t+1}) = L(s_d|o_{1:t}) + L(s_d|o'_{t+1}), \] (8)

where \( o'_{t+1} \in O = \{occupied, free\} \) is an imaginary future measurement and \( L(s_d|o) \) is also called inverse sensor model [17]. The inverse sensor model is defined likewise as in [9] as
measurement has uniform distribution, i.e., \( p \) value resulting from an imaginary measurement as the expected voxel results on the bottom image are defined as being visible after the execution of this occupied voxels occlude each other. As a result, only the voxels coloured in which free voxels are assumed to be transparent, whereas unknown or the real depth camera specifications, frustum culling procedure is performed occupied voxels are represented in red, free voxels are green and unknown Fig. 5. Cross-section view of a typical visibility check. In the picture, odds is then \( p_{sd} = p(s_d|o_{t:t+1}) = 1 - \frac{1}{1 + \exp(L(s_d|o_{t:t+1}))} \). 

Using the definitions aforementioned, when no contacts are available, we are finally able to select next best views according to a maximum information gain strategy via

\[
\xi = \arg \max_{\xi \in \Xi} \frac{u_2(\xi_k, \Lambda_{\text{visible}}(\xi), \Gamma_t)}{D},
\]

where \( D = |\Lambda_{\text{visible}}| \) is the number of visible voxels. Note that we refer to the average information gain per voxel, since different views have different numbers of visible voxels in their field of view after frustum culling. This makes the predicted information different gain for different views comparable.

2) Information Gain View Selection: Using the definitions aforementioned, when no contacts are available, we are finally able to select next best views according to a maximum information gain strategy via

\[
\xi = \arg \max_{\xi \in \Xi} \frac{u_2(\xi_k, \Lambda_{\text{visible}}(\xi), \Gamma_t)}{D},
\]

\[
\xi = \arg \max_{\xi \in \Xi} u_2(\xi, \Lambda_{\text{visible}}(\xi), \Gamma_t).
\]
Having retrieved those voxels, let $p_{s_c}$ be the probability of occupancy of a given voxel $s_c \in \Lambda_c$. The probability of collision can be calculated by

$$p_{\text{collision}}(\tau, \Lambda_c) = 1 - \prod_{s_c \in \Lambda_c} (1 - p_{s_c}). \tag{15}$$

For numerical reasons, we prefer to refer to Eq 15 using only the product term, representing the probability that all voxels along the trajectory $\tau$ are free, in its logarithmic form as

$$\kappa(\tau, \Lambda_c) = \ln \prod_{s_c \in \Lambda_c} (1 - p_{s_c}) = \sum_{s_c \in \Lambda_c} \ln(1 - p_{s_c}), \tag{16}$$

note that $p_{\text{collision}}(\tau, \Lambda_c) = 1 - \exp(\kappa(\tau, \Lambda_c))$.

Finally, to select views in order to get better estimations for Eq 15, we use the same utility function defined in 13. Thus if we let $\Lambda_{c_{\text{visible}}}^{\xi}(\xi) \subset \Lambda_c$ be the set of visible voxels by a certain view pose $\xi$. Next best views are then selected according to

$$\xi^* = \arg \max_{\xi \in \mathcal{V}} u_2(\xi_k, \Lambda_{c_{\text{visible}}}^{\xi}(\xi), \Gamma_t). \tag{17}$$

In practice, we allow safety exploration to run while the information gain is above a predefined threshold, i.e. $u_2(\xi_k, \Lambda_{c_{\text{visible}}}^{\xi}(\xi), \Gamma_t) > \beta$. If this criteria is not met, the final probability of collision is reported according to Eq 15. The trajectory $\tau$ is therefore executed or not based on the probability of collision.

IV. EXPERIMENTS

The following section outlines the experiments we conducted to test our view selection approach. General pseudocode of our implementation is described in Alg 5, which is divided into two sub-phases. First a contact-based next best view exploration procedure is run as outlined by Alg 1. In this first phase, at least two views are selected, up to a maximum of 7 views if after the second view no grasp trajectory and contacts are found. The first view is fixed, only subsequent views after this fixed view are selected according to the criteria for contact-based view selection. The second phase of Alg 5 is run in order to estimate the probability of collision of the best promising candidate trajectory $\tau^*$ selected as the trajectory with the lowest probability of collision prior to the safety view exploration phase, given our current knowledge of the object $\Gamma_t$ and workspace $\Lambda$. This second phase is outlined in Alg 3. Note that the safety exploration phase stops if the current selected view predicts information gain below a certain threshold $\beta$. If, at the end of the safety exploration phase, we discover that this trajectory $\tau^*$ has collision probability above a certain threshold $\alpha$, we reject $\tau^*$ and cycle back to phase 1, i.e. Alg 1.

A. Methodology

Using Alg 5 we performed trials on a set of 14 novel objects shown in Fig 7. In our experiments, we compared our active view algorithm with a random view selection strategy. In other words, we substituted all calls of the selection procedures Alg 2 and Alg 4 by a uniform random view selection scheme. Furthermore we limited the two phases of this modified random-based approach to be constrained to the same number of views that our algorithm performed in both phases. It is also important to note that in our experiments we have set the size of the voxels in our 3D occupancy map $\Lambda$ to be 0.0025$m$, for relatively fine precision. Table I shows the final data for this experiment.

B. Results

The results shown in Table I outline the contrast between the two approaches. We first note that the success rate of our proposed view selection approach achieved a success rate of 85.7%, whereas the modified random-based approach showed a success rate of only 64.2%. A closer look at Table I reveals that random exploration tended to yield unsafe grasps, under the same view number constraints as our active view selection approach. This indicates that random view selection would probably need to cycle back to generate new grasp trajectory candidates more times, which seems a natural consequence of its sub-optimal exploratory behaviour.
TABLE I
Trial Results

<table>
<thead>
<tr>
<th>Objects</th>
<th>Grasp Count</th>
<th>HBV &amp; Random</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>bowl</td>
<td>4</td>
<td>4</td>
<td>success</td>
</tr>
<tr>
<td>small</td>
<td>3</td>
<td>4</td>
<td>success</td>
</tr>
<tr>
<td>bucket</td>
<td>5</td>
<td>4</td>
<td>success</td>
</tr>
<tr>
<td>coke</td>
<td>2</td>
<td>5</td>
<td>success</td>
</tr>
<tr>
<td>cup</td>
<td>3</td>
<td>5</td>
<td>success</td>
</tr>
<tr>
<td>dustpan</td>
<td>3</td>
<td>3</td>
<td>success</td>
</tr>
<tr>
<td>glass</td>
<td>4</td>
<td>4</td>
<td>fail</td>
</tr>
<tr>
<td>gurger</td>
<td>3</td>
<td>4</td>
<td>success</td>
</tr>
<tr>
<td>jug</td>
<td>3</td>
<td>5</td>
<td>success</td>
</tr>
<tr>
<td>mrmuse</td>
<td>3</td>
<td>4</td>
<td>success</td>
</tr>
<tr>
<td>mg1</td>
<td>3</td>
<td>4</td>
<td>success</td>
</tr>
<tr>
<td>remise</td>
<td>3</td>
<td>3</td>
<td>success</td>
</tr>
<tr>
<td>stand2</td>
<td>3</td>
<td>3</td>
<td>success</td>
</tr>
<tr>
<td>toothpaste</td>
<td>4</td>
<td>4</td>
<td>fail</td>
</tr>
</tbody>
</table>

Success Rate: 0.857 0.642 Average Views: 4.7

One such example is highlighted by Fig 6, in which the final trajectory executed with probability of collision 1.0 and, indeed, makes the robot hand collide with a part of the mug not involved in the grasp, finally failing for safety reasons. We also note that our collision probability appears to be over-sensitive, the random approach also succeeded for various cases in which the probability of collision was 1.0. Nonetheless, even for successful grasps as the one depicted in Fig 2, grasps with probability 1.0 tended to collide prematurely with different parts of the object. In addition, we also noted that for the case of the toothpaste, the trivial solution of a grasp with as few collisions as possible might yield grasps with very poor grip. This indicates future work towards a middle ground between these two extremes.

Finally, as shown by Table I and in Fig 8, our approach had equivalent success rate to prior work done by Kopicki and Wyatt et al [10]. In our experiments our approach used on average 4.7 views for grasp planning, as compared with 7 in [10]. Additional views were only used to assess safety.

V. CONCLUSIONS

We have proposed an effective approach for view selection comprising two stages. The first stage guides gaze by planned contacts, seeking a low noise point cloud near those contacts. If no contact points are available, we guide gaze so as to minimise the entropy of the 3D occupancy grid map around our object cloud $\Gamma$. After candidate grasps are found, we gaze to assess the safety of the reach-to-grasp trajectory candidate. We showed that this yields a better success rate compared to a random strategy, and that we slightly improve on [10], while using fewer views for grasp planning.

REFERENCES