Active Vision via Extremum Seeking for Robots in Unstructured Environments: Applications in Object Recognition and Manipulation

Berk Calli, Member, IEEE, Wouter Caarls, Member, IEEE, Martijn Wisse, Member, IEEE, and Pieter P. Jonker, Member, IEEE

Abstract—In this paper, a novel active vision strategy is proposed for optimizing the viewpoint of a robot’s vision sensor for a given success criterion. The strategy is based on extremum seeking control (ESC), which introduces two main advantages: 1) Our approach is model free: It does not require an explicit objective function or any other task model to calculate the gradient direction for viewpoint optimization. This brings new possibilities for the use of active vision in unstructured environments, since a priori knowledge of the surroundings and the target objects is not required. 2) ESC conducts continuous optimization backed up with mechanisms to escape from local maxima. This enables an efficient execution of an active vision task. We demonstrate our approach with two applications in the object recognition and manipulation fields, where the model-free approach brings various benefits: for object recognition, our framework removes the dependence on offline training data for viewpoint optimization, and provides robustness of the system to occlusions and changing lighting conditions. In object manipulation, the model-free approach allows us to increase the success rate of a grasp synthesis algorithm without the need of an object model; the algorithm only uses continuous measurements of the objective value, i.e., the grasp quality. Our experiments show that continuous viewpoint optimization can efficiently increase the data quality for the underlying algorithm, while maintaining the robustness.

Note to Practitioners—Vision sensors provide robots flexibility and robustness both in industrial and domestic settings by supplying required data to analyze the surroundings and the state of the task. However, the quality of these data can be very high or poor depending on the viewing angle of the vision sensor. For example, if the robot aims to recognize an object, images taken from certain angles (e.g., feature rich surfaces) can be more descriptive than the others, or if the robot’s goal is to manipulate an object, observing it from a viewpoint that reveals easy-to-grasp “handles” makes the task simpler to execute. The algorithm presented in this paper aims to provide the robot high quality visual data relative to the task at hand by changing vision sensors’ viewpoint. Different from other methods in the literature, our method does not require any task models (therefore, it is model free), and only utilizes a quality value that can be measured from the current viewpoint (e.g., object recognition success rate for the current image). The viewpoint of the sensor is changed continuously for increasing the quality value until the robot is confident enough about the success of the execution. We demonstrate the application of the algorithm in the object recognition and manipulation domains. Nevertheless, it can be applied to many other robotics tasks, where viewing angle of the scene affects the robot’s performance.

Index Terms—Active vision, extremum seeking control (ESC), object recognition, grasping, manipulation.

I. INTRODUCTION

ACTIVE vision algorithms are utilized in robotics to supply better/more visual data for the execution of a given task by systematically changing the viewpoint and settings of the robot’s sensor [1]. Especially in unstructured environments, these algorithms boost the abilities of robots by making them more robust to variations in data quality. They have various applications in object recognition [2], [3], saliency maximization [4], navigation [5], [6], object modeling [7], [8], and manipulation [9], where the vision sensor viewpoint affects the algorithm performance significantly.

In this paper, we propose a new viewpoint optimization methodology for increasing the efficiency and success rate of an underlying algorithm that utilizes the visual data. Our methodology is unique in the sense that it does not require an explicit objective function for viewpoint optimization; it maximizes a success criterion supplied by the underlying algorithm without the need of a task or environment model. This is achieved by adopting extremum seeking control (ESC) methods [10], which utilize the success criterion in a continuous optimization loop. The algorithms in the literature generally assume that the observation probability distribution is known given the action and current state (either being explicitly available, e.g., [2] and [11] or encoded via a learning process, e.g., [12] and [13]). In other words, these algorithms require a
decent estimate of the outcome for a given camera viewpoint even before visiting it. On the other hand, ESC algorithms do not require such knowledge (therefore are model-free) as they estimate the objective value gradient with continuous measurements and lead the camera accordingly.

Our model-free approach brings several advantages over the methods in the literature, especially for conducting viewpoint optimization in unstructured environments. We choose two domains in robotics to demonstrate and discuss these benefits: active object recognition and object manipulation. In the active object recognition, the duty of the active vision algorithm is to alter the sensor viewpoint for maximizing the recognition rate. The algorithms in the literature (e.g., [3] and [14]–[16]) rely on offline training data for conducting viewpoint optimization, which makes them sensitive to structured noise (e.g., occlusions, extreme lighting conditions). In addition, most of these algorithms employ discrete search techniques, and do not utilize the data between the way points. Our algorithm does not rely on offline training data for viewpoint optimization since it only uses the recognition rate value for continuous viewpoint optimization. In this way, robustness to structured noise is achieved while utilizing all the data acquired by the vision system in the optimization process.

In the object manipulation domain, the goal of the active vision system is to increase the success rate of grasp synthesis algorithms for grasping objects whose models are not available a priori (e.g., [17]–[21]). These algorithms utilize an image of the target object and aim to provide the poses of the robotic fingers on the object surface that achieve a successful grasp. In this case, the viewpoint optimization algorithm maximizes the quality of the grasp (therefore, its probability to be successful) by supplying better images of the object to the grasp synthesis algorithm. Compared to the active object recognition field, there are fewer strategies proposed for utilizing active vision for robotic manipulation. The majority of the strategies integrate viewpoint optimization into the manipulation pipeline, assuming that the object model is known, or a model is generated prior to manipulation (e.g., [22]–[24]). The methods for unknown objects in [25] and [26] boost the performance of a grasping algorithm by minimizing occlusions and measurement uncertainties, respectively. Our strategy, on the other hand, aids the grasp synthesis directly by providing images of the target object for which the algorithm is more confident in generating a good grasp. To the best of our knowledge, our methodology is the only one in the literature that can achieve this, thanks to its model-free nature.

Closest to our work, Zhang et al. [27] design an ESC algorithm for maximizing a saliency metric to help robots detecting interesting things in an environment. Nevertheless, experimental results are only presented for a single execution. In our work, we propose a general active vision methodology with an extensive experimental study.

In this paper, we first explain our ESC-based methodology in Section II. Following that, in Section III, we apply it to an active object recognition problem, discuss the advantages of our strategy in detail with respect to other algorithms in the literature, and present experimental validation. In Section IV, we apply the methodology to a grasp synthesis problem for unknown objects and also provide experimental evaluation. Section V provides the conclusion with discussions and future work.

II. EXTREMUM SEEKING CONTROL-BASED ACTIVE VISION

In our methodology, we use ESC optimizers [10] in a continuous viewpoint optimization loop. ESC algorithms address the problem of objective value optimization when the objective function, its gradient and the optimum value are unknown. They utilize the objective value continuously, estimate its gradient and supply a search direction accordingly. They also have mechanisms to avoid local optima (e.g., with hysteresis functions, in addition, for a strategy to further improve the global convergence performance of a particular ESC algorithm, the reader can refer to [28]). Some widely known applications of ESC are: ignition time selection for combustion engines [29], bioreactor optimization [30], and anti-lock braking system control [31].

In the literature, there are four main types of continuous ESC algorithms: sliding mode ESC [32], neural network ESC (NN-ESC) [33], approximation-based ESC [34], and perturbation-based ESC [35]. The performance of these algorithms are compared in [10]. In this paper, we utilize NN-ESC as it is robust to objective value noise, and provides efficient results in multidimensional optimization problems. However, depending on the characteristics of the problem at hand, other ESC methods may also be preferable and can be used in the exact same methodology presented in this paper.

For our applications, the model-free nature of ESC algorithms (not requiring an explicit objective function) allows us to conduct viewpoint optimization without offline training or a priori information about the task or the target object. The application specific advantages of the method will be covered in detail in Sections III and IV.

A general overview of our method can be seen from Fig. 1. Here, an image is captured by the robot’s camera and supplied to the target algorithm, which would benefit from viewpoint optimization (e.g., object recognition, grasp synthesis). The algorithm processes this image and generates a success rate value (e.g., probability of recognizing an object, the quality
A. 2-D Neural Network ESC

The block diagram of NN-ESC can be seen from Fig. 2. The algorithm generates its own reference \( f_r \) by using the peak detector and the \( W \) switch
\[
f_r = f_p + f_w
\]
(1)

where \( f_p \) and \( f_w \) are the outputs of the peak detector and \( W \) switch, respectively. These switches use the error obtained by the difference between the current objective value and the reference value
\[
e = f_r - f.
\]
(2)

As the algorithm is initialized with the initial objective value \( f_r = f \), the \( W \) switch is not active; the only contribution to the \( f_r \) value comes from the peak detector. The switching mechanism of the peak detector is as follows:
\[
\dot{f}_p = \begin{cases} M, & (e < 0) \\ 0, & (e \geq 0) \end{cases}
\]
(3)

where \( M \) is a positive constant. By this mechanism, the peak detector converges to the value of the maximum objective

value that is measured during the process. In order to maintain this convergence, the rate of change of the objective value should be smaller than \( M \)
\[
|\dot{f}| < M.
\]
(4)

The control law \( u \), which is fed to the robot as velocity reference, tries to minimize the error by the following switching functions with hysteresis:
\[
u_1 = \begin{cases} -U, & e < -\delta_1 \\ U, & e > \delta_1 \end{cases}
\]
(5)

\[
u_2 = \begin{cases} 0, & e < -\delta_2 \\ -U, & e > \delta_2 \end{cases}
\]
(6)

\[
u_3 = \begin{cases} 0, & e < -\delta_3 \\ 2U, & e > \delta_3 \end{cases}
\]
(7)

\[u = \begin{cases} u_1 + u_2, & \text{[previous state] otherwise} \\ u_2 + u_3, & \text{[previous state] otherwise} \end{cases}
\]
(8)

Here, \( U \) is a constant that specifies the magnitude of the velocity reference, and \( \delta_1 \) are the hysteresis widths. The control law is initialized with the following values:
\[
(u_1, u_2, u_3) = (-U, 0, 0).
\]
(9)

The switching mechanism of the control law works as follows: if the direction of the system makes the objective value converge to the optimum value, then this direction is kept. Otherwise, the error will increase, and when it is greater than the hysteresis values, the direction will be changed. The move structure of hysteresis functions brings the following constraint to the hysteresis widths:
\[
\delta_1 < \delta_2 < \delta_3.
\]
(10)

When \( e \) becomes greater than \( \delta_3 \), the control law enters its last state. This state ends as the error is larger than the threshold \( \Delta \) where
\[
\Delta > \delta_3.
\]
(11)

As the \( \Delta \) threshold is exceeded, the reference value \( f_r \) should be reset, so that a new cycle can start. The \( f_w \) function performs this by the following switching function:
\[
f_w = \begin{cases} -W, & e > \Delta \\ 0, & e < -\Delta \end{cases}
\]
(12)

where \( W \) is a positive constant. By this function, when the error is greater than the \( \Delta \) threshold, the \( f_w \) will decrease until the error is less than \( -\Delta \). This makes the control law go back to its first state.

The control signal \( u \) is fed to the systems as velocity input
\[
\dot{x}_r = u.
\]
(13)

While applying an NN ESC algorithm, the parameters are needed to be tuned considering the noise on the objective value...
and the hysteresis that is wished to be allowed to escape from local minima.

In the next section, a way of forming the objective value within the framework is explained.

B. Objective Value Calculation

The success rate of the underlying algorithm form the main component of the objective value that is aimed to be maximized. Nonetheless, we add two other components to impose workspace constraints and to make the optimization process more efficient as follows. The objective value is defined as

\[ f(v, r_s, w_s, x_o) = f_g(v) + f_s(r_s, w_s) + f_o(o_s) \]  

where \( o \) is the success rate value, \( r_s \) is the pose of the robot, \( w_s \) are the workspace constraints, \( o_s \) is the state of the optimizer, and \( f_g, f_s, \) and \( f_o \) are the success rate, workspace and optimizer specific components, respectively. \( f_g \) is calculated specific to the algorithm, and examples of its design will be covered for object recognition and manipulation in the Sections III and IV, respectively. The workspace constraints component is defined as a barrier function [37] as follows:

\[ f_s(r_s, w_s) = \mu_w \sum_{n=1}^{k} \ln(r_{sn} - w_{sn}). \]

Here, \( \mu_w \) is the barrier parameter, \( k \) is the number of degrees of freedom of the robot, and \( r_{sn} \) and \( w_{sn} \) are the components of the robot’s state vector and the workspace constraints. This function drops rapidly as the robot approaches to the close vicinity of the workspace constraints. This makes the objective value violate hysteresis values of the NN-ESC and forces it to change direction.

The optimizer specific component aids the NN-ESC algorithm as follows: while conducting viewpoint optimization, the NN-ESC can get into a direction that has a relatively small gradient. If the gradient is negative, the camera will need to travel long distances before the hysteresis thresholds of the NN-ESC algorithm are violated. For these cases, imposing a time constraint via a barrier function helps to increase the efficiency

\[ f_o(o_s) = \mu_o \ln(\tau_o - \tau_p(o_s)). \]

In (16), \( \tau_p \) is the elapsed time since the maximum value of \( f_g \) is measured in the current state of the optimizer. The value of this component will drop rapidly as the state of the optimizer does not cause the objective value to increase until the allowed time \( \tau_o \). This will act as a penalty component in the objective function, and the state change for the viewpoint optimizer will be triggered. \( \mu_o \) is the barrier parameter of this component.

In the next two sections, we present the application of our methodology to object recognition and grasp synthesis fields, respectively. The advantages of adopting the proposed strategy are given in detail.
maintains system robustness. These points are detailed as search and imperfect training data, the proposed approach from structured noise, the inefficiency caused by discrete search. Therefore, while the algorithms in the literature suffer from occluded data and conducts continuous viewpoint optimization that can adapt/respond to the current state of the environment. On the other hand, the model-free active vision strategy proposed in this paper does not utilize training data set. It can be seen that occlusion affects the recognition rate dramatically.

In the next section, we discuss the advantages of our proposed methodology over the abovementioned algorithms.

A. Advantages of the Proposed Strategy

The active object recognition algorithms in the literature rely on offline training data, which make it harder to maintain robustness for the cases that the data cannot capture the current situation. On the other hand, the model-free active vision strategy proposed in this paper does not utilize training data and conducts continuous viewpoint optimization that can adapt/respond to the current state of the environment. Therefore, while the algorithms in the literature suffer from structured noise, the inefficiency caused by discrete search and imperfect training data, the proposed approach maintains system robustness. These points are detailed as follows.

1) Effect of Structured Noise: The methods in the literature rely on offline information that is mostly assumed to be obtained in the training phase: In all these cases, it is assumed that the observation probability distribution is known for a given action and current class belief. More specifically, if an action, an observation and an object class are denoted by $a$,$o$, and $c$, these algorithms need to know the probability distribution $P(o|c,a)$ in order to decide on the next action (the way that this decision is made varies depending on the algorithm). The learning-based methods and offline methods like [12] and [13] do not explicitly need this information in the decision phase, but it is encoded in the policy that is learned during the training process. Knowing $P(o|c,a)$ implies that, given the class of the object that we are looking for, we need to know what kind of measurement that we are going to have, if we take action $a$. This information can be obtained from position-labeled off-line training images. However, counting on this observation probability distribution for selecting an action relies on a very strong assumption: the data that is acquired offline should be valid for the current environmental conditions. In other words, if the system does not see what it expects to see in the destination viewpoint, this will affect the belief distribution on object class, and the system will tend to believe that the initial reasoning about the object class and pose was wrong. However, unexpected observations at the destination viewpoint may occur due to the occlusions, extreme lighting conditions or any kind of structured noise.

One of the rare active object recognition systems that address the occlusion problem is presented in [51], where a solution is provided by adding synthetically occluded images of the objects to the training data set. However, such a system can only be robust to the trained occlusions. Moreover, their system includes a turntable and two movable cameras with 90° viewing angle difference, which makes the job of the algorithm much easier compared to many realistic robotics scenarios. In addition to that work, a method for calculating a visibility metric for a given viewpoint is proposed in [52], which is aimed to be utilized in cluttered environments, but is not integrated to any viewpoint selection strategy.

2) Inefficiency Caused by Discrete Search: Another disadvantage of the presented active object recognition methods is that, while moving to the next best viewpoint, the data that can be acquired on the way to the destination is not used. However, utilization of this data can be very useful in reducing the execution time of the algorithms.

Very few algorithms use the data continuously in the active recognition literature. Huber et al. [2] and Eidenberger and Scharinger [53] use continuous viewpoint space, but the resultant decisions are discrete, so they do not utilize the data on the way to the destination. Robbel and Roy [54] use Fourier descriptors to model the silhouette of the object and recognize the object by the continuous change of the modeled shape. Although successful and efficient results are obtained by this method, the algorithm is vulnerable to acquired shape changes due to occlusions. A recent work in [47] acknowledges the importance of utilizing the data between way points and optimizes the camera trajectory to benefit from these data. Nevertheless, their viewpoint optimization strategy is not designed to handle structured noise-related issues that are explained in Section III-A1.

3) Effect of Imperfect Training Data: Using training images for selecting the next best viewpoint may be problematic when the training images are not taken in laboratory conditions. Two common cases are using object images from internet search [55], [56] and a robot learning a new object [57]. This causes a problem that is analogous to the one explained in Section III-A1, but this time it is caused by the training data. Moreover, in the case of obtaining images from the Internet, the relative pose of the camera with respect to the object is unknown. This makes the effect of the action on the current state impossible to calculate, since the current state and the
states of the other online images cannot be known. As a result, the reasoning process of the abovementioned active object recognition algorithms becomes inapplicable.

The implementation using our algorithm presented in the next section does not rely on any offline data and it conducts a continuous viewpoint optimization. Therefore, the advantages caused by imperfect training data and discrete search are avoided.

B. Implementation and Experiments

In this section, we present an implementation of our methodology to the active object recognition problem. We first explain the utilized object recognition algorithm followed by the description of our experimental setup and the experimental results with and without structured noise.

1) Object Recognition: For object recognition, we chose to use the bag of features method [58], since it is a commonly used algorithm in the literature and its advantages and drawbacks are well known. Nonetheless, it is important to note that the active vision strategy proposed in this paper can be used by any other object recognition method which outputs a recognition rate value. The features are detected using speeded up robust features descriptors, and each object in the training set is trained with a one-vs-all scheme via a support vector machine (SVM) using a radial basis function as kernel. In the classification phase, SVM provides signed distances to the separating hyperplane for each object. These distances are converted to probabilities by fitting sigmoid functions using Platt scaling [59]. The parameters of the scaling are obtained by a separate training image set of each object. The probabilities of each class are then compared to find the most probable class. We would like to clarify that the training here is solely for the object recognition algorithm, and no training phase or training images are necessary for the viewpoint optimization purposes unlike the abovementioned methods in the literature.

For calculating the $f_g$ component of (14), we directly use the distance to the separating hyperplane of the most probable class instead of using the classification probability values that are obtained by Platt scaling. There are mainly two reasons for this choice: First, the sigmoids fit by Platt scaling saturate at both ends of the curve which affects the optimizer’s performance; very small changes are observed at the saturated regions which does not provide information that is rich enough for the optimization. Also, in the nonsaturated regions, the increase in the recognition value may be too sudden and steep due to the scaling which also affects the performance of the optimizer negatively. Moreover, the scaling parameters that are obtained by the fitting process are dependent on the quality of the training image set, which we would like to avoid. On the other hand, by minimizing the distance to the separating hyperplane, we are able to optimize the recognition performance regardless to the scaling quality. This choice is validated by our preliminary experiments.

2) Experimental Setup: Our experimental setup is made up of a UR5 type Universal Robots arm, a webcam attached to the end-effector of the robot and a rotating platform (Fig. 5).

Fig. 5. Experimental setup: The UR5 type robot arm, the webcam attached to the tooltip of the robot, and the rotating platform.

Fig. 6. Objects that are used in the experiments.

We have trained the object recognition algorithm with twelve different objects, which can be seen from Fig. 6.

The performance of the algorithm is tested with and without occlusion for each object. For both cases, six experiments are conducted per object with 60° rotation difference (introduced by the rotating platform), 144 experiments in total. The optimization loop is run at 7 Hz. For each run, the initial velocity of the robot is selected randomly, and the NN-ESC algorithm is activated. If the most probable class changes during the operation, the optimizer is restarted. In order to avoid rapid switching of the most probable class, a low pass filter is applied to the class probabilities. The parameters of the NN-ESC are selected as $U = 0.02$, $\delta_1 = 0.08$, $\delta_2 = 0.11$, $\delta_3 = 0.14$, $\Delta = 0.17$, $M = 0.2$, and $W = 0.5$.

An experiment is considered successful, if the algorithm can increase the recognition probability of the correct object up to 95%. If this success rate cannot be achieved within 30 s, the process is terminated, and the most probable class
### TABLE I

**Summary of the 144 Experiments: Success Rate and Average Execution Time of the Algorithm With and Without Occlusion for Different Termination Levels**

<table>
<thead>
<tr>
<th>Sites/occl.</th>
<th>Initial view</th>
<th>Reached %95 (ESC)</th>
<th>Terminated after 30 s (ESC)</th>
<th>Avg. time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3w/o occl.</td>
<td>Initial view</td>
<td>36.1%</td>
<td>83.3%</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>Reached %95 (ESC)</td>
<td>91.6%</td>
<td>95.8%</td>
<td></td>
</tr>
<tr>
<td>3w/ occl.</td>
<td>Initial view</td>
<td>36.1%</td>
<td>97.2%</td>
<td>10.4</td>
</tr>
<tr>
<td></td>
<td>Reached %95 (ESC)</td>
<td>91.6%</td>
<td>95.8%</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 7. Experimental results with “hagel slag” box. (a) Snapshots from an experiment without occlusion. Initial view (left). Final view (right). (b) Recognition probability versus time. The recognition rate increases over time and hits 95%.

3) **Experimental Results:** All the experimental results are summarized in Table I. For the experiments without occlusions, the object is correctly recognized directly from the initial view (before active vision is initiated) with 95% recognition rate for 36.1% of the cases. For the other cases, when the proposed active recognition algorithm is run, 95% recognition rate is achieved by changing the viewpoint of the camera via ESC algorithm, and the success rate increases to 91.6%. For the 4.2% of the cases, the algorithm fails to reach 95% recognition rate in 30 s, and the algorithm is terminated. However, the most probable class is still correct at the end of the process (which sums up to 95.8% success). For the other 4.2% of the cases, the object is guessed wrong as the highest measured probability during the process belongs to a wrong object.

Two samples from the experiments are presented in Figs. 7 and 8. When the distance to the separation hyperplane is raised above the hysteresis threshold values, the ESC algorithm changes velocity reference and forces the distance to decrease. This causes an increase in the recognition rate of the object and 95% recognition rate is achieved eventually. The average time of obtaining the correct guess is 9.4 s.

In order to evaluate the performance of the algorithm under structured noise, occlusion is applied. As the results are compared with the nonoccluded case, the significant difference appears to be the decrease while obtaining 95% recognition rate. Although the object is guessed correctly in 97.2% of the cases, the recognition rate cannot be raised to 95% in some
runs due to the occluded regions of the object. Also, a slight increase in average convergence time is observed since it takes more time to find a good viewpoint due to occlusions.

Two samples from the experiments with occlusion are given in Figs. 9 and 10. As can be seen from Figs. 9 and 10, the algorithm successfully avoid occlusions and leads the robot to a good viewpoint for recognition. However, the process takes more time and the recognition rate varies more than the nonoccluded case. These large variations are due to the steep scaled values, and might have been a problem for the ESC algorithm, however, thanks to the better-behaved hyperplane distance that we used in the optimization process, the algorithm is able to optimize the viewpoint successfully in the majority of the cases (as explained in Section III-B1).

For the cases in which the algorithm fails to recognize the object, it is observed that the algorithm gets stuck in a region where the number of features is low. As a result, a healthy guess cannot be made by the recognition algorithm, and the optimizer cannot go out of local maxima and fails.

Next, we present the application of our active vision methodology for the grasp synthesis problem.

IV. VIEWPOINT OPTIMIZATION FOR GRASP SYNTHESIS

In unstructured environments, the model of the target object is usually unavailable to the robot prior to manipulation. Therefore, grasp synthesis algorithms need to rely on the incomplete information acquired by their vision sensors while deciding on a grasping strategy. The algorithms in the literature generally optimize a grasp using one single image of the target object either using learning methods [17]–[19], [60] or heuristics [20], [21]. The fact that they synthesize a grasp using a single viewpoint makes the pose of the object with respect to the sensor a crucial factor: When the object is observed from a “good” viewpoint (which can be different for each method and object), these algorithms provide a very high success rate for a huge variety of objects. However, for many other viewpoints, the success rates of the algorithms drop significantly. This problem can be solved by utilizing an active vision strategy. As it is indicated by Spaan [61], this kind of visual exploration is a crucial element in order to fill the gap between neuroscience models and robotic implementations of grasping.

Most grasp synthesis algorithms have an internal optimization process in order to synthesize the best grasp with the
data at hand; for a given single image, they utilize a grasp quality metric and maximize it by changing the grasping pose. The goal of the viewpoint optimization in this case is to maximize the grasp quality by changing the viewpoint of the sensor and supplying “better” images to the grasp synthesis algorithm. The definition of a “better” image naturally depends on the grasp synthesis algorithm. Nevertheless, our method does not require any information about how the grasp synthesis algorithm works; it treats the algorithm as a black box and only utilizes the grasp quality value.

In the manipulation literature, active vision is often utilized for known object models [22], [23]. In this case, the role of viewpoint optimization is to localize the target object. In [24], a complete 3-D model of the object is generated prior to manipulation. On the other hand, very few works utilize active vision for grasping unknown objects. In [26], active vision is used to refine the surface reconstruction of the grasp location candidates, which results in a more reliable grasp execution. Kahn et al. [25] address the problem of grasping unknown objects in the presence of occlusion. In this case, the active vision system aims to minimize occlusions on the grasp handles of the objects. Similar to our case, this approach also uses a local optimizer that generates continuous trajectories rather than selecting discrete next best views. Nevertheless, their approach is specifically designed for scenarios with occlusions.

Our methodology, being the only model-free approach applied to object manipulation, aids the grasp synthesis algorithm directly by supplying better views: since the model of the object is unknown, the objective function based on the grasp quality cannot be calculated for an unvisited viewpoint. From the optimization perspective, this means the objective function that we want to optimize is unknown. Moreover, the gradient of this function is not calculable or cannot be measured directly. We can only calculate the value of the objective function for our current state, or in other words, our objective value. Our method utilizes this value for viewpoint optimization, and does not require an objective function. The next section presents our implementation and experimental results.

A. Implementation and Experiments

Within the abovementioned framework, for grasp synthesis, we used a silhouette-based algorithm in which the grasping point locations are optimized on a parametric model of the silhouette contour considering force closure and curvature of the contact surface (maximizing concavity) based on the method presented in [62]. In a nutshell, the silhouette of the object is segmented and modeled using elliptic Fourier descriptors. By taking the first and second derivatives of these parametric models, explicit equations can be obtained for calculating tangent and normal vectors for a given point on the contour. These expressions are used to optimize curvature values and force application directions. The algorithm searches for the grasping point locations that exhibit maximum concavity, while also satisfying the force closure property. Therefore, for this algorithm, the quality of the grasp is measured by the sum of curvature of the stable grasping points, and this measure is optimized to find the best grasp.

For viewpoint optimization, the quality value of the best grasp that is synthesized using the last captured image is used as the $f_g$ value of (14), and combined with workspace and optimizer specific components as explained in Section II-B. Similar to the active object recognition case, our methodology can also be utilized with any other grasp synthesis algorithm that can output a grasp quality value.

The experiments are conducted using the same UR5 setup with an addition of the Delft Hand 3 gripper [63] attached to the end-effector of the robot (Fig. 11). This gripper can apply a power grasp as well as a precision grasp. A distance sensor, which is used to close the gripper when the object is close enough, is placed at the palm of the gripper.

Six household objects in Fig. 12 are positioned in front of the robot with three different orientations, and the robot is initialized in two different poses as presented in Fig. 13 (six experiments per object). The viewpoint optimization is terminated after one full circle of NN-ESC (when $W$ threshold is exceeded). This last pose of the gripper is the approach vector, as the gripper follows a straight line toward the object and closes when the grasping distance is achieved. The distance threshold is set to power grasp for all the objects except for the table tennis racket. Note that with the chosen grasp
This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 13. Illustration of the experiments: The three different orientations for the object with 120° rotation difference and two different initial positions of the robot, which have 60° rotation difference around the x-axes of the camera frame.

TABLE II
SUMMARY OF THE EXPERIMENTAL RESULTS FOR THE GRASPING SCENARIO

<table>
<thead>
<tr>
<th>Object Name</th>
<th>Successful Exp. / Total Exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Box</td>
<td>6/6</td>
</tr>
<tr>
<td>Spray Bottle</td>
<td>5/6</td>
</tr>
<tr>
<td>Glass</td>
<td>2/2</td>
</tr>
<tr>
<td>Tennis Racket</td>
<td>2/6</td>
</tr>
<tr>
<td>Beer Bottle</td>
<td>2/2</td>
</tr>
<tr>
<td>Water Bottle</td>
<td>2/2</td>
</tr>
<tr>
<td>Total</td>
<td>19/24</td>
</tr>
</tbody>
</table>

The results are summarized in Table II, and object initial and final poses are presented in Fig. 14. In all the experiments, the viewpoint optimization was successful in increasing the viewpoint quality with respect to the given measures, i.e., force closure and model curvature. This aids the grasp synthesis algorithm since the grasping points are placed to much better locations on the object. The new viewpoints are also better approach vectors for grasping. These advantages bring successful grasps for the box, wine glass, beer bottle, and water bottle for all different initial viewpoints. For the spray bottle, only one initial view is failed. This makes up 94% success for the objects that are grasped with power grasps. However, for the table tennis racket, four failed out of six initial viewpoints. Fig. 15 presents the objective value evolving in time and the generated velocity references for the spray bottle experiment whose snapshots are given in Fig. 14. These results are summarized in Table II.
In addition, the underlying algorithm should be fast enough to be run continuously (in 5–10 Hz) to maintain the efficiency. We also observed that having a high signal-to-noise ratio for the objective value boosts the performance of the viewpoint optimization procedure significantly as, e.g., for the case of NN-ESC, low hysteresis thresholds can be chosen.

ESC methods are local optimizers, and even though they have mechanisms for escaping the local optima, they may still get stuck in suboptimal states, e.g., some of the failed cases of our experiments. Combining the methodology with a higher level reasoning strategy, e.g., partially observable Markov decision processes [64], can increase its success significantly. In a recent study [28], a simplex-based strategy is also proposed to improve the global convergence performance of a particular ESC algorithm. While integrating this strategy would increase the convergence time, it can certainly be preferable in critical tasks.

**REFERENCES**


**Fig. 16.** Snapshots from the failed experiments. First column presents the initial views, and second column is the final views (both with EFD models). Green points are the EFD models, yellow points are the model centers, and blue points are the grasping points.


[40] C. Zhang and R. Ordonez, “Non-derivative optimiz-
Berk Calli (M’15) received the Bachelor of Science and Master of Science degrees in Mechatronics from Sabanci University, Istanbul, Turkey, with his thesis on integrated visual servoing and force control for robotic manipulation, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2015 with his thesis on viewpoint optimization for improving grasp synthesis performance.

He is currently with the GRAB Laboratory, Yale University, New Haven, CT, USA, working on vision-based manipulation, dexterous manipulation, and manipulation benchmarking.

Wouter Caarls (S’03–M’12) received the M.Sc. (Hons.) degree in artificial intelligence from the University of Amsterdam, Amsterdam, The Netherlands, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, on the subject of the automatic optimization of a parallel computer architecture for smart cameras.

He is currently an Assistant Professor with the Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil, investigating the applications of reinforcement learning in robotics. His research interests include robotics, machine learning, optimization, parallel algorithms, and image processing.

Martijn Wisse (S’02–M’04) received the M.Sc. and Ph.D. degrees in mechanical engineering from the Delft University of Technology, Delft, The Netherlands.

He is currently a Full Professor with the Delft University of Technology. His previous research interests included passive dynamic walking robots. His current research interests include the field of robot manipulators for agile manufacturing, underactuated grasping, open-loop stable manipulator control, design of robotic arms and robotic systems, agile manufacturing, and the creation of start-up companies.

Pieter P. Jonker (M’91) received the M.Sc. degree in electrical engineering from the Twente University of Technology, Enschede, The Netherlands, in 1979, and the Ph.D. degree in physics from the Delft University of Technology (TUDelft), Delft, The Netherlands, in 1992.

He is currently a Full Professor of Vision-Based Robotics with the Bio-Mechanical Engineering Group, TUDelft. With Dr. M. Wisse, he runs the Dutch Bio-Robotics Laboratory, TUDelft. His current research interests include bioinspired real-time embedded vision systems for robotics, surveillance, and augmented reality, and on hierarchical reinforcement learning for walking robots.

Dr. Jonker is a Fellow of the IAPR.

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, 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 component of this work in other works.