Visual attention and active vision
Kootstra, Geert Willem

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2010

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Active Object Recognition by Exploration
Abstract

Object recognition is a challenging problem for artificial systems. This is especially true for objects that are placed in cluttered and uncontrolled environments. To solve this problem, we discuss an active approach to object recognition in this chapter. Instead of passively observing objects, we use a robot to actively explore the objects. This enables the system to learn objects from different viewpoints and to select viewpoints for optimal recognition. Active vision furthermore simplifies the segmentation of the object from its background. As the basis for object recognition we use the Scale-Invariant Feature Transform (SIFT). SIFT has been a successful method for image representation. However, a known drawback of SIFT is that the computational complexity of the algorithm increases with the number of interest points. We discuss a growing-when-required (GWR) network for efficient clustering of interest points to reduce the size of the database. The results show successful learning of three-dimensional objects in real-world environments. The active approach is successful in separating the object from its cluttered background, and the active selection of viewpoint further increases the performance. Moreover, the GWR network strongly reduces the number of interest points.

This chapter is based on:


7.1 Introduction

The real world poses many challenging problems for artificial systems. Consider for instance the problem of recognizing objects in the real world. Many object recognition systems that are successful in controlled laboratory environments have problems with the uncontrolled and unpredictable properties of the real world. Whereas, for instance, illumination and background can be controlled in an artificial setting, this is not true for real-world environments. Visual perception, therefore, becomes more challenging in the real world. Natural systems deal with these challenges by using active perception (Gibson, 1979). Instead of passively observing an object, many animals, including humans, explore the object to control the visual input (see figure 6.6). The use of active perception is also very important for artificial systems (Ballard, 1991; Pfeifer & Scheier, 1999). This chapter, discusses an active approach to three-dimensional (3D) object recognition in the real world by an autonomous robot. By actively changing its viewpoint, the robot observes an object from different angles, making it possible to learn to recognize the object from any given viewpoint. Moreover, the system selects the viewpoint that is expected to be most informative for recognition. Furthermore, exploration of the object makes it possible to separate the object from its background, something that is non-trivial when passively observing an object on a highly cluttered background (Metta & Fitzpatrick, 2003).

Like many current approaches to object recognition, our model describes the objects by a set of local interest points (Harris & Stephens, 1988; Lowe, 1999; Schmid & Mohr, 1997). Description in terms of local interest points has the advantage that the representation is more robust to occlusions, clutter and noise. It is also less sensitive to changes in viewpoint. In our method we use the Scale-Invariant Feature Transform (SIFT) for the detection and description of interest points (Lowe, 2004). The choice for SIFT is motivated by the fact that it is a well-known and widely-used interest-point method and the focus of this chapter lies on active vision and not on interest-point methods. Our approach is, however, not restricted to SIFT, but can also be used with other local image detectors and descriptors, like those discussed in Chapter 8 and 9.

Interest points have been successfully used for three-dimensional (3D) object recognition (Ferrari, Tuytelaars, & Van Gool, 2006; Lowe, 2001; Moreels & Perona, 2007; Rothganger, Lazebnik, Schmid, & Ponce, 2006). These studies have demonstrated the ability to learn to recognize objects from multiple viewpoints and subsequently recog-
nize these objects in cluttered scenes. However, learning in these studies takes place in well-controlled environments: the object is usually put on a turntable which carefully rotates the object, while taking pictures of the object with fixed lighting conditions — with the exception of (Moreels & Perona, 2007) — and against a uniform background. This setup reduces the amount of noise and uncertainty and makes it trivial to separate the foreground from the background. It is therefore not representative for real-world environments. A real-world environment is usually highly uncertain and cluttered with many distracting features. In this chapter, we present a method to learn objects in uncontrolled real-world environments, using active vision. We use a mobile robot to actively explore the objects and their environment.

Our approach uses active vision in multiple ways. Firstly, we use it to separate the object from its background, similar to (Fitzpatrick, 2003; Metta & Fitzpatrick, 2003). We use a method that can be described as what-moves-together-belongs-together. The robot observes the object while circling around it, a behavior that is comparable to rotating an object in your hand (see figure 6.6). By doing so, the observer learns the appearance of the object from different viewpoints. This solves the object-constancy problem, the problem that objects appear very different from different viewpoints (see Figure 7.1). The exploration of the object enables the system to link the different perspectives and build a representation of the full 3D object.

Secondly, while performing the explorative behavior, interest points belonging to the object will show little displacement on the camera image, since the object is near the center of rotation. Interest points in the background, on the contrary, show relatively large displacements, with the possible exception of points on the floor close to the object. The amount of displacement of an interest point is used to classify whether it belongs to the object or to the background.

Thirdly, active vision is used to find stable interest points. By changing viewpoints, the robot can actively test whether an interest point is recognizable from nearby viewpoints. If so, the interest point can be classified as stable. This process will filter out points that are sensitive to rotation, translation, and other affine transformations. This reduces the necessity to use affine-invariant interest point detectors (e.g., Mikolajczyk & Schmid, 2002), which are not only computationally expensive, but have also been shown to perform worse on recognizing non-planar 3D objects than SIFT (Moreels & Perona, 2007). A similar approach to the one presented here was taken in (Lehrer & Bianco, 2000), where a behavior, inspired by insects, was adopted to find reliable
7.1. Introduction

Figure 7.1: The object-constancy problem. An object appears very differently from different viewpoints. By using an active method for learning and recognition, the different perspectives can be integrated.

Finally, active vision is used to gather more evidence for recognition. This is especially important in ambiguous situations. If from one viewpoint it is not possible to recognize an object, a more promising viewpoint can be selected. Although human observers might have the impression that ambiguous situations are quite rare, we must remember that ambiguity strongly depends on the quality of the sensory system, as can be seen in (Nolfi, 1996). In 3D object recognition, gathering more evidence improves recognition (Roy, Chaudhury, & Banerjee, 2004). Some viewpoints will be more informative than others. We therefore propose a probabilistic method to select the viewpoint that is expected to be most informative as the next viewpoint. Viewpoint selection is also used by (Borotschnig, Paletta, Prantl, & Pinz, 2000; Paletta & Pinz, 2000). The difference with the model presented here is that our model uses a one-shot learning method and perform recognition in a real-world environment.

The proposed method to learn objects by circling around it requires information about the position of the object. This chicken-and-egg problem can be solved by human guidance, like done in (Van Hoof, 2008; Zwinderman, Rybski, & Kootstra, submitted). By letting a human teacher initially denote the object, the position of the object can be determined. Subsequently, the robot can explore the object. However, since we are mainly interested in the possibilities of the explorative behavior, the objects are placed at a fixed position from the robot in the presented research.

In addition to the use of active vision, we propose a method to reduce the number of points in the interest-point database. One of the reasons that SIFT is so successful in object recognition is that it uses a large number of interest points to represent one object (Lowe, 2004). This makes the system very tolerant to noise, and reduces the problem of occlusions. There is, however, an important drawback, namely that a significant
amount of computation in the recognition process is devoted to matching the observed points with the interest-points database. Nearest-neighbor search methods like kd-tree search (Friedman, Bentley, & Finkel, 1977) that are efficient in low-dimensional spaces, do not do better than exhaustive search in the high-dimensional space of the SIFT features. An improvement in computation time can be achieved by an approximate best-bin-first method (Beis & Lowe, 1997). But even then, the computation time increases with the number of stored interest points, while the success in finding the nearest neighbor decreases. It is therefore very useful for 3D object recognition to reduce the number of stored interest points in an efficient way.

In this chapter, growing-when-required (GWR) network (Marsland et al., 2002) is used for efficient clustering of interest points. When performing 3D object recognition, many of the acquired interest points look very similar. There are several reasons for this. First of all, these are interest points belonging to the same point on the object seen from different angles. Secondly, there are similarly looking points on repeating structures on the same object, and finally, different objects can have ambiguous interest points. The GWR network clusters these similar interest points to attain efficient database use.

### 7.2 Object Recognition by Exploration

In this section the active approach to object recognition is discussed first. Then the method to select the next viewpoint. The section ends with a description of the method to cluster the SIFT interest points.

#### 7.2.1 The Scale-Invariant Feature Transform

The SIFT detector and descriptor (Lowe, 2004), described in Appendix B, are used as the basis for the 3D object recognition. The method used for matching the observed interest points with the database is somewhat different. Firstly, the method focuses solely on the individual matching of interest points, and therefore does not use the geometric matching of sets of points as used by Lowe. Secondly, a threshold on the distance to the nearest neighbor is used, instead of a best to second-best ratio to determine a match, since this yielded better performance in the experiments. Details on the matching and recognition processes used are described further on.
Figure 7.2: The active vision model to select stable object interest points. While
the robot explores the object, SIFT interest points are detected in the current cam-
era image. The interest points are stored in a temporal database. The current in-
terest points are compared with the previous ones. An interest point is considered
stable if it is observed in the current and in the previous and next observation. The
stable interest points are then divided into background and foreground by compar-
ing their image motion with knowledge about the expected motion. After clustering
the interest points, the robust interest points belonging to the object are stored in
the database.

7.2.2 Active Vision

By actively changing viewpoint, the robot gathers new information that we use in two
different ways: to detect stable interest points and classify them as object or back-
ground, and to explore the object in order to gather more evidence to resolve ambigu-
ous situations. Both methods are described in the following paragraphs. The general
architecture is depicted in Figure 7.2.

An interest point is considered stable if it is originally observed at an angle of $\theta$
degrees, and subsequently matched in the previous or next image, at $\theta \pm 10$ degrees. An
interest point \( \mathbf{k}_i \) is matched to its nearest neighbor in the previous image, \( \mathbf{k}_n \), if the Euclidean distance between both is less than 0.6, where \( \mathbf{k} \) is the 128 dimensional feature vector of the interest point. This threshold is established experimentally. This filters out all interest points that are only recognizable from one specific angle.

In the next step, we segment the stable interest points belonging to the background from those belonging to the object. Each interest point \( \mathbf{k}_i \) has a position \((x_i, y_i)\) at which it is observed in the image. Since the object is in the center of rotation, object points will move little when the robot is exploring the object, whereas the displacement will be relatively large for interest points in the background. Furthermore, since the robot moves on a flat surface, points will only move in the horizontal direction. Allowing some fluctuations, we classify a stable interest point as an object point when

\[
(|x_i - x_n| < x_T) \land (|y_i - y_n| < y_T)
\]  

(7.1)

where we use \( x_T = 12 \) and \( y_T = 4 \) pixels (the resolution of the camera image is \( 360 \times 240 \)). Otherwise, the stable interest point is classified as background. The successful use of this simple classification model nicely illustrates the power of active vision to simplify perceptual tasks. The robot explores the objects from 36 different angles and stores the stable object points along with the object ID and pose. Doing so, the appearances of objects in a cluttered environment are learned.

Once the object database is in place, objects can be recognized. Based on the set of observed interest points, \( \mathcal{O} \), and the interest-point database, \( \mathcal{D} \), we determine the activation of every model, \( m_{b, \theta} \), for object \( b \), and pose \( \theta \). The activation of a model is based on the set of observations, \( M_{b, \theta} \subseteq \mathcal{O} \), that supports the model.

\[
M_{b, \theta} = \bigcup (p_i \in \mathcal{O}| o_n = b \land \alpha_n = \theta)
\]  

(7.2)

where \( o_n \) and \( \alpha_n \) are respectively the object ID and pose of the nearest neighbor \( \mathbf{k}_n \) of \( \mathbf{p}_i \) in the interest-point database. Every supporting observation \( \mathbf{p}_i \) in \( M_{b, \theta} \) gives an activation \( a_i \):

\[
a_i = \exp(-|\mathbf{p}_i - \mathbf{k}_n|)
\]  

(7.3)

This form is in conformity with the activation calculation used in the growing-when-required network discussed in Section 7.3. The total activation of model \( m_{b, \theta} \) given
the observed interest points, $O$, form viewpoint $\delta$, and the interest-point database $D$, is given by:

$$A^\delta_{b, \theta} = \frac{\sum_{i \in M_{b, \theta}} a_i}{\sqrt{|D_{b, \theta}|}}$$  \hspace{1cm} (7.4)

where $|D_{b, \theta}|$ is the number of points in the interest-point database that are associated with object $b$ and pose $\theta$, and $\delta$ is the current viewpoint. Equation (7.4) gives the activation of a specific pose of an object. This activation increases with the number of supporting observations relative to the number of interest points in the database associated with that object/pose. This makes that fewer matched observations are needed for objects that have relatively few interest points. However, the square root in the denominator causes the probability to increase with the number of object points given the same ratio of matched observations to database points. This reflects the idea that there is more confidence when there are more object points.

Finally, the robot can actively gather more evidence for recognition. By rotating around the object, the robot gathers more information about the object under consideration by viewing it from different angles. When driving around the object, we accumulate the evidence by:

$$A_{b, \theta}(t) = \sum_{\delta \in E} A^\delta_{b, \theta}$$  \hspace{1cm} (7.5)

where $A_{b, \theta}(t)$ is the accumulated activation for object $b$ and pose $\theta$ at time $t$ and $E = \{\phi_0, \cdots, \phi_t\}$ is the set of viewpoints from where the observations are made. The change of viewpoint helps to disambiguate object and is therefore expected to result in more robust recognition of 3D objects. In the next section, we will discuss how the next viewpoint is selected by our model.

In the above, the activation of an object in a particular pose is calculated. To obtain the activation of the full object model, the activations of all poses, $\Theta$, for that object are summed:

$$A_b(t) = \sum_{\theta \in \Theta} A_{b, \theta}(t)$$  \hspace{1cm} (7.6)

Figure 7.3 gives an overview of the interest-point database during learning and during
7.2.3 Next-Viewpoint Selection

We use the active capabilities of the robot to explore the objects and gather more information from different viewpoints. In order to select the next viewpoint, we use a probabilistic approach. In this approach, the next viewpoint \( \phi_{t+1} \) is the angle from where we expect the maximum activation of an objects-pose model, that is:

\[
\phi_{t+1} = \arg \max_{\gamma \in (\Theta - \Phi)} E(A_{b,\theta}(t+1))
\]  

(7.7)

where \( \Theta \) is the set of all possible viewpoints, and \( \Phi = \phi_0, \ldots, \phi_t \) is the set of all previous viewpoints. The expected activation of the object-pose model when viewed from viewpoint \( \gamma \) at time \( t+1 \) is given by:

\[
E(A_{b,\theta}(t+1)) = A_{b,\theta}(t) + E(A_{b,\theta}^{\gamma} | O_{b,\theta}) P(O_{b,\theta})
\]  

(7.8)

\[
E(A_{b,\theta}^{\gamma} | O_{b,\theta}) = \sqrt{|D_{b,\theta+\gamma}|}
\]  

(7.9)

\[
P(O_{b,\theta}) = \frac{A_{b,\theta}(t)}{\sum_{i=0}^{N} A_{o_i,\alpha_i}(t)}
\]  

(7.10)

In words, the expected new activation of the model is based on the old activation. This value is increased with the expected extra activation gained from the new viewpoint given that we are looking at object \( b \) and pose \( \theta \), \( E(A_{b,\theta}^{\gamma} | O_{b,\theta}) \), multiplied by the probability that we are actually looking at object \( b \) at pose \( \theta \), \( P(O_{b,\theta}) \). Equation (7.9) can be inferred from equation (7.4) when we assume that we observe all interest points belonging to object \( b \) at pose \( \theta + \gamma \). Although this is not the case in reality, we can assume that a constant proportion of the interest points are observed. The next-viewpoint choice is not influenced by this constant factor. And the probability \( P(O_{b,\theta}) \) is the activation of the model divided by the total activation of all object-pose models. By selecting the viewpoint that optimizes the expected activation of one of the object-pose models, we select the most informative viewpoint as the next.
7.3 Clustering Interest Points: Growing When Required

As explained in the introduction of this chapter, 3D object recognition with SIFT has the main disadvantage that the computational time needed increases with the number of interest points stored in the database. We therefore use a growing-when-required (GWR) network (Marsland et al., 2002) to efficiently cluster interest points that are highly similar. A GWR network is a clustering method, very similar to a growing-neural-gas (GNG) network (Fritzke, 1995). Both networks are based on Kohonen’s self-organizing maps (SOM) (Kohonen, 1990). A SOM is an efficient and unsupervised method to cluster high-dimensional data. The disadvantage, however, is that the number of clusters (i.e., nodes in the map) needs to be set in advance. This makes a SOM highly inappropriate for object recognition with SIFT, since the number of clusters depends on the number of unique interest points. A GNG-network is an adaptation of a SOM which can dynamically change the number of nodes in the network. However, the drawback of a GNG-network is that new nodes are only added after a number of inputs. This is not desirable for object recognition, since we would like to add a node in the network when we observe a completely new interest point. A GWR network does just that, it adds nodes if this is required.

The GWR network as described in (Marsland et al., 2002) uses edges between nodes. This is based on the SOM, and results in a topology preserving network in the sense that connected nodes in the network correspond to neighboring points in the input space. Usually, the connections between nodes in a GWR network are used in the learning process to move the neighboring nodes of the winning node closer to the presented input. Although this provides a better distribution of the nodes over the input data, it is undesirable for object learning, since in that case the presentation of an input not only changes the representation of the corresponding interest point, but also of neighboring interest points. This will result in changing the interest points so much that they do no longer correspond with the original input. Since this will impair recognition, we omitted the edges from the GWR network.

For the description of our implementation of the GWR network, we follow the notation and description in (Marsland et al., 2002). Let \( K \) be the set of observed interest points when learning the objects, \( A \) be the set of nodes in the network, \( w_n \) be the weight vector of node \( n \) (of the same dimensionality as the SIFT interest points), and \( r_n \) be the activation counter. Furthermore, each node holds a record, \( R_n \), of all associated objects.
and poses. We initialize the network with $A = n_1$, where the weight vector of $n_1$ is initialized with a randomly picked interest point from $K$, and $t_1 = 0$. Then, for each interest point $k$ from $K$, we do:

1. $k$ and the object $b$ in pose $\theta$ to which the interest point belongs are input to the network.

2. Select the best matching node $s \in A$, such that
   
   $$s = \arg \max_{n \in A} |k - w_n|.$$  

3. Calculate the activity of the winning node
   
   $$a_s = \exp(-|k - w_n|)$$

4. Calculate the firing counter is calculated by the decaying function:
   
   $$h_s = 1 - (1 - \exp(-\alpha_b t_s / \tau)) / \alpha_n$$

   where the shape of the function can be set using the parameters. In the experiments, we used: $\alpha_b = 1.05$, $\alpha_n = 1.05$, and $\tau = 3.33$.

5. If the activation of the winning node is low ($a_s < a_T$), that is when the input does not really match any of the existing clusters, and when the winning node is mature ($h_s < h_T$), create a new node $r$. Add $r$ to the network, initialize its weight with $k$, and set $\langle b, \theta \rangle$ as the reference list:

   $$A = A \cup r$$
   $$w_r = k$$
   $$R_r = \{\langle b, \theta \rangle\}$$

   where the thresholds are $a_T = 0.8$ and $h_T = 0.4$.

6. Else, adapt the weights of the winning node and add $\langle b, \theta \rangle$ to the reference list:

   $$w_s = w_s + \eta \cdot h_s \cdot (k - w_s)$$
   $$R_s = R_s \cup \{\langle b, \theta \rangle\}$$
where the parameter that determines the adaptivity of the node is $\eta = 0.05$.

7. Proceed to the next cycle:

$$t_s = t_s + 1$$

When the presented interest point is sufficiently similar to the winning node, it is clustered with that node, and the description of the node is altered to better represent all associated interest points. If, on the other hand, the presented interest point differs from the existing nodes causing the activation of the winning node to be below the threshold $a_T$, and the firing counter of the nearest node is below the threshold $h_T$, the presented interest point is added as a new node. In this way, the GWR network clusters similar interest points, thus creating a smaller database for recognition.

A record is kept of all objects and poses that correspond to the nodes in the network. This allows for supporting all objects containing similarly looking interest points when such a point is observed. This is in contrast with (Lowe, 2004), where important evidence is discarded by choosing only interest points that match uniquely with one object. Figure 7.3 gives an overview of the interest-point database both with and without the use of the GWR network.

### 7.4 Experiments and Results

We used seven objects placed in a cluttered environment for our image database (see figure 7.4). A mobile robot equipped with a CCD camera was used to take images from 36 different viewpoints around the objects with intervals of $10^\circ$. The image database consists of four different sets, where the orientation of the objects is respectively $0^\circ$, $90^\circ$, $180^\circ$, and $270^\circ$ with respect to the environment, resulting in a different background for the objects. In the experiments, training was done on one single set, while the other three sets were used to test the performance. This resulted in 12 different cross-validation tests.

#### 7.4.1 Active Segmentation

Figure 7.4 shows a number of examples of interest points that are filtered using active vision. The white squares are the interest points that are both stable and move accord-
Chapter 7. Active Object Recognition by Exploration

Figure 7.3: The interest-point database. a) During learning, the SIFT interest-point descriptors (squares) are stored along with the associated object and pose. b) When using the GWR network, the interest points are clustered (ellipses), and the cluster descriptor (related to the average of the clustered points) is stored with the associated object(s) and pose(s). c) During recognition, the observed interest points (gray-filled squares) are matched with the database in feature space. The nearest neighbors (black-filled squares) active the associated object-pose models. d) In case of GWR-clustering, the observations are matched with the clusters, and the nearest clusters in feature space (black-filled ellipses) support the associated object-pose models.

ing to the foreground. A point is stable when it is also observed in the previous or in the next observation. A point moves according to the foreground if it satisfies equation 7.1. The gray squares are stable interest points that move according to the background. The black squares are the interest points that are either unstable, or move according to the background. It can be appreciated from the figure that the majority of object in-
terest points lie indeed on the object and most background interest points are correctly filtered out. However, quite some interest points on the objects are also filtered out due to instability. It is a know drawback of SIFT that it produces many interest points that are not observable from slightly different viewing angles. The filtering of interest points will result in better object models. The object models will neither include background points, nor will they include unstable points that are not reobservable from small deviating points of view.

The numbers on the bottom left of the images in Figure 7.4 indicate the number of true positives, that is the number of interest points that are correctly classified as object versus the number of false positives, the number of incorrectly classified object points. The numbers on the bottom right give the precision, where the precision is calculated as: $\text{precision} = \frac{tp}{tp + fp}$.

### 7.4.2 Active vs Passive Object Recognition

In our next experiment, we tested the performance of our active approach to 3D object recognition and compared it with a passive approach that does not use active vision for robust interest points filtering and multiple viewpoints. The recognition performances are shown in Figure 7.5a. The plot shows the mean recognition rate over the 12 tests as a function of the number of viewpoints. The active methods accumulate evidence as given by Equation (7.5). The choice of the next viewpoint depends on the used next-viewpoint selection method, which is discussed in the next paragraph. Since the passive approach only uses a single viewpoint and does not accumulate evidence it is plotted as one wide bar. The error bars give the 95% confidence intervals on the mean. The active approach clearly outperforms the passive approach. Already with one viewpoint, the use of active vision to select stable object points gives significantly better performance than passively considering all visible interest points, with respectively 73% and 60% success (t-test: $p$-value < $10^{-4}$). With increasing accumulation of evidence, the recognition rate rises from 73% to about 90% (t-test: $p$-value $\ll 10^{-4}$). Additionally, Figure 7.5b shows the decrease of the number of interest points when using the active method. This means that the active recognition system not only gives better recognition performance, it is also faster than a passive system.

We furthermore compared the performance of our next-viewpoint selection method to an approach where the next viewpoint is a simple 30° interval, as well as to random-
Figure 7.4: Examples of filtered interest points using active vision. The white squares are the interest points that are both stable (i.e., found in the previous or next observation) and move according to the foreground. The gray squares are stable points that move according to the background. The black squares are unstable interest points. The numbers on the bottom left give the number of correctly classified object points versus the number of incorrect classified object points. The number on the bottom right gives the precision.
7.4. Experiments and Results

Figure 7.5: Passive versus active object recognition. In (a), the mean recognition rate is shown as a function of the number of accumulated viewpoints. Already for one viewpoint, the active-vision method has a significantly better performance than the passive method. The performance increases when more viewpoints are accumulated. The passive method does not explore new viewpoints and is therefore plotted as one wide bar. In (b), the number of interest points are shown. It can be appreciated that the use of active vision greatly reduces the number of interest points. The error bars give the 95% confidence intervals on the means.

viewpoint selection. Figure 7.6 shows the mean recognition rates over the 12 tests. The error bars show the 95% confidence intervals. The difference of our next-viewpoint selection method with the interval method and the random selection method shows a small significant difference only for the second viewpoint (p-values of respectively 0.03 and 0.02 using a t-test). The difference in performance for more viewpoints is not significant, but is in favour of the ext-viewpoint-selection method. A significant increase in performance for the second viewpoint is important since this enables the system to recognize objects with fewer observations. Although, admittedly, the differences are small.

7.4.3 Interest-Point Clustering using the GWR Network

We first tested the influence of the thresholds $a_T$ and $h_T$ and the adaptation parameter $\eta$ in the GWR network on the recognition performance and the reduction of interest points. The results are shown in Figure 7.7. The plots show a trade off between
Chapter 7. Active Object Recognition by Exploration

<table>
<thead>
<tr>
<th>Number of viewpoints</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The recognition rates for the passive approach and three active methods: a fixed 30° interval, random viewpoint selection and our next viewpoint method. The error bars give the 95% confidence intervals in the means. Our viewpoint-selection method is significantly better when using two viewpoints.

**Figure 7.6:**

The recognition rates for the passive approach and three active methods: a fixed 30° interval, random viewpoint selection and our next viewpoint method. The error bars give the 95% confidence intervals in the means. Our viewpoint-selection method is significantly better when using two viewpoints.

The recognition performance and the number of interest-point clusters for $\eta_T$ and $h_T$. With higher values for these parameters, the recognition rates increase, as well as the numbers of interest points. The best recognition results are obtained with $\eta = 0.15$. Lower values for $\eta$ give worse recognition performance, while the number of resulting interest-point clusters is the same. Since the recognition for $\eta = 0.15$, $a_T = 0.9$ and $h_T = 0.55$ is similar to the performance of standard SIFT without interest-point clustering, we use those parameter settings to compare GWR SIFT with standard SIFT.

In the next experiment, we compared the performance of the GWR network with the standard SIFT method, both using active vision with a fixed interval of 30°. The learned interest points are presented to the GWR network in random order. We therefore performed ten different experimental runs to test the performance of the GWR network. Figure 7.8a shows the decrease of 50% in the number of interest-point clusters used by the GWR network. This results in a great improvement of the computational speed of the system. The recognition performance is shown in Figure 7.8b. We also compared the GWR network with standard SIFT that uses only 50% of the interest points, the same amount of interest points as used by the GWR network. These interest points
Figure 7.7: Recognition rate and number of interest-point clusters for different parameter settings of the GWR network. For larger $a_T$ and $h_T$ thresholds, the recognition performance improves, but also the number of interest-point clusters. A larger value for $\eta$ results in better recognition without an increase of clusters. To keep up with the performance of the standard method not using clustering, we use $a_T = 0.9$, $h_T = 0.55$, and $\eta = 0.15$ in the subsequent experiment.

are selected randomly from the interest-point database. Again we performed 10 experimental runs. It can be appreciated from the plot that the GWR network performs similar to standard SIFT, but significantly better than standard SIFT using the same number of interest points. This shows that the GWR network effectively clusters the interest points, using only 50% of interest points without loss of performance.
Figure 7.8: Object recognition with and without interest-point clustering using the GWR network. (a) Using GWR clustering reduces the number of interest points in the database by 50%, yielding a speed-up in computation time. (b) The mean recognition rates using the GWR network are compared with standard SIFT using all interest points and using the same amount of interest points as GWR (50%). For the GWR network and SIFT using 50% interest points, the data is acquired from 10 experimental runs. The error bars give the 95% confidence intervals on the mean. The recognition performance using the GWR network is similar to standard SIFT using all interest points, whereas it is significantly better than standard SIFT using 50% of the points.

7.5 Discussion

The experiments show the successful use of object exploration for 3D object recognition. Exploration is used in different ways, (1) to acquire evidence from multiple viewpoints, (2) to detect stable interest points, (3) to segment the object from the background, and (4) to select informative viewpoints. The active vision approach performs significantly better than passively observing. The problem of the passive method is that it not only learns to associate the object with interest points that actually lie on the object itself, but also with points in the background. Moreover, many unstable points pollute the database. These problems are solved by active exploration of the objects. By selecting the next viewpoint based on the optimization of the expected activation of object models, the system further increases its recognition performance with subsequent viewpoints.

The proposed next-viewpoint selection method results in a small but significant im-
7.5. Discussion

Improvement of performance for the second viewpoint. However, the improvement is not significant when more viewpoints are used. There is room for improvement of the method. One of the problems is the calculation of the expected gain in activation for taking a next viewpoint. Since the proposed method is a one-shot-learning method, this expectation cannot be realistically estimated. We therefore propose to learn these probabilities during subsequent encounters with the object, to get realistic figures.

The use of a GWR network for the clustering of interest points has also been analyzed. The GWR network results in a strong reduction in the number of interest points that need to be stored in the database, while maintaining the recognition performance of standard SIFT. The GWR network performs significantly better than SIFT using the same number of interest points. Reducing the amount of interest points is important for object recognition using SIFT, especially with a growing number of objects. The results show that the GWR network is capable of effective clustering of interest points.

The use of the GWR network is very similar to the bag-of-features approach (Csurka et al., 2004). However, in the bag-of-features approach, the clusters are learned beforehand. Since the number of clusters and the position of the clusters cannot change during execution, the approach cannot deal with object types that are never seen before. Our approach constantly adapts to the input, and will create new clusters if necessary in new circumstances. In future work, we would like to investigate these properties of the GWR network and compare our system to the bag-of-feature approach.

In the presented study, no additional methods are used to improve the recognition rate. A good way to boost recognition is to use a geometric fit between sets of interest points, for instance the geometric verification method described in (Lowe, 2004). This method can be used both with the proposed active-vision method and with the GWR network.

Summarizing, this chapter showed the successful use of active vision to simplify complex recognition tasks. The quality of the object representations is improved by exploring the objects and better recognition is obtained by efficiently taking different viewpoints. The GWR network furthermore demonstrated the possibility to reduce the number of interest points. These methods make the implementation of object recognition in the real world more feasible.
Chapter 7. Active Object Recognition by Exploration