Exploring Objects for Recognition in the Real World

Gert Kootstra, Jelmer Ypma and Bart de Boer
Artificial Intelligence Department
University of Groningen
Grote Kruisstraat 2/1, 9712 TS, Groningen, The Netherlands
{g.kootstra, jelmer, b.de.boer}@ai.rug.nl

***Abstract* – Perception in natural systems is a highly active process. In this paper, we adopt the strategy of natural systems to explore objects for 3D object recognition using robots. The exploration of objects enables the system to learn objects from different viewpoints, which is essential for 3D object recognition. Exploration furthermore simplifies the segmentation of the object from its background, which is important for object learning in real-world environments, which are usually highly cluttered. We use the Scale Invariant Feature Transform (SIFT) as the basis for our object recognition system. We discuss our active vision approach to learn and recognize 3D objects in cluttered and uncontrolled environments. Furthermore, we propose a model to reduce the number of SIFT keypoints stored in the object database. It is a known drawback of SIFT that the computational complexity of the algorithm increases rapidly with the number of keypoints. We discuss the use of a growing-when-required (GWR) network, which is based on the Kohonen Self Organizing Feature Map, for efficient clustering of the keypoints. The results show successful learning of 3D objects in a cluttered and uncontrolled environment. Moreover, the GWR-network strongly reduces the number of keypoints.**

Index Terms* – *active vision, object exploration, object recognition, SIFT, clustering.

I. Introduction

Perception is essentially a very active process [1]. This is not only true for natural systems, but also for artificial systems sensory-motor coordination plays an important role in the perception of the world [2, 3]. Take for instance the recognition of objects. Instead of passively observing the object, many animals, including humans, explore the object (see fig. 1). This strategy enables us to actively change viewpoint and observe the object from different angles, which makes it possible to learn to recognize the object from any given viewpoint. Moreover, 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 [4]. In this paper, we adopt the strategy to explore objects. We use an active vision approach to achieve 3D object recognition with a robot in uncontrolled real-world environments.

Fig. 1 A 18 months old child exploring an unknown object. This enables the child to observe the object from different viewpoints. In the meanwhile, it makes it possible to discriminate the object from the background.

Like most current approaches to object recognition, our model describes the objects by a set of local interest points [5-7]. 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. To be more specific, we use the Scale Invariant Feature Transform (SIFT) for the detection and description of interest points [8]. Our approach is, however, not restricted to SIFT, but can also usable with other local image detectors and descriptors. In the rest of the paper, we use the term *keypoints* to refer to points of interest detected by SIFT.

Interest points have been successfully used for three-dimensional (3D) object recognition [9-12]. These studies have demonstrated the ability to learn to recognize objects from multiple viewpoints and subsequently recognize these objects in cluttered scenes. However, these methods have not been shown to be able to recognize 3D objects in actual real-world environments, since learning of the objects takes place in a well-controlled environment: 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 [9] – and against a uniform background. This setup makes it trivial to separate the foreground from the background, but is not representative for real-world environments. A real-world environment is usually highly cluttered, which makes the segmentation of the object from its background non-trivial. In this paper, we present a method to learn objects in uncontrolled real-world environments, using active vision to achieve foreground-background segmentation. We use a mobile robot to actively explore the objects and their environment.

The use of active vision to simplify perceptual tasks has been advocated by Ballard, who referred to it with the term *animate vision* [3]. In our approach, we make use of active vision in multiple ways. Firstly, we use active vision to separate the object from its background, similar to [4, 13]. We use a method that can be described as *what-moves-together-belongs-together*. We let our robot circle round the object. Interest points belonging to the object show little displacement, while the robot moves, since the object is near the center of rotation. Interest points in the background, on the

contrary, show relatively large displacements. The amount of displacement of an interest point from one viewpoint to the next is used to classify whether it belongs to the object or to the background.

Secondly, we use active vision to find stable interest points. By changing viewpoint, we can actively test whether an interest point is recognizable from nearby viewpoints. Doing so, we can select stable interest points. Furthermore, we can filter out points that are sensitive to rotation, translation and other affine transformations. This does away with the necessity to use affine invariant interest point detectors (e.g., [14]), which are not only computationally expensive, but have also been shown to perform worse on recognizing non-planar 3D objects than SIFT [9]. A similar approach was taken in [15], where a behavior, inspired by insects, was adopted to find reliable visual landmarks.

Finally, we use active vision to gather more evidence when recognition is problematic. This is especially important in ambiguous situations. If from one viewpoint it is not possible to recognize and object, a more promising viewpoint can be selected. Although, as a human observer, we 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 [16]. In 3D object recognition, an active approach improves recognition [17]. Although not used in the present study, recognition can be even more successful if the system learns which viewpoints give the best information for disambiguation [18, 19]. All together, we exploit the use of active vision in a number of ways in order to improve 3D object recognition.

In addition to the use of active vision to improve the recognition of 3D objects in real-world situations, we propose a method to reduce the number of keypoints in the keypoint database. One of the reasons that SIFT is so successful in object recognition is that it uses a large number of keypoints to represent one object [8]. This makes the system very tolerant to noise, and solves 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 keypoints with the keypoint database. Nearest-neighbor search methods like *k*d-tree search [20] that result in efficient performance in low-dimensional spaces, do not do better than exhaustive search in the high-dimensional space of the SIFT features. To improve computation time, an approximate best-bin-first method has been proposed [21]. But for this method, too, goes that computation time increases with the number of stored keypoints. At the same time, the success in finding the nearest neighbor decreases. It is therefore very useful for 3D object recognition to reduce the number of stored keypoints in an efficient way.

In this paper, we use a growing-when-required (GWR) network [22] for efficient clustering of keypoints. When performing 3D object recognition, many of the acquired keypoints look very similar. There are several reasons for this. First of all, these are keypoints 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, we see ambiguous keypoints on different objects. Our GWR-network clusters these similar keypoints to attain efficient database use.

II. Methods

We used the SIFT detection and description, as described in [8], as the basis for the 3D object recognition. We use a prior smoothing of each octave of $\sigma = 1.2$ instead of $\sigma = 1.6$ as proposed in [8], since this yields better performance in our experiments. Our method for matching the observed keypoints with the database is somewhat different. First, we focused solely on the individual matching of keypoints, and therefore did not use the geometric matching of sets of keypoints as used in [8]. Second, we used an exhaustive search through the database to find the nearest neighbor, since we are interested in the best possible match, to emphasize the performance of our developed methods. Third, we used a threshold on the distance to the nearest neighbor, instead of a best to second-best ratio to determine a match, since this yielded better performance in our experiments. And last, we used a slightly different probabilistic model for recognition than the one described in [10]. The matching and recognition process that we used is described in the next section.

A. Active Vision

One of the contributions of our study to improve 3D object recognition in real-world environments is the use of active vision. We make use of a mobile robot that explores objects by circling around them, observing them every 10 degrees. By actively changing viewpoint, the robot gathers new information that we use in two different ways: to detect stable keypoints and classify them as object or background, and to explore the object in order to gather more evidence to resolve ambiguous situations. Both methods are described in the following paragraphs.

A keypoint is considered stable if it is originally observed at an angle of θ degrees, and subsequently matched in the previous or next image, at $\theta \pm 10$ degrees. A keypoint $\mathbf{k}_i$ 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 keypoint. This filters out all keypoints that are only recognizable from one specific angle.

In the next step, we segment the stable keypoints belonging to the background from those belonging to the object. Each keypoint $\mathbf{k}_i$ 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 keypoints will move little when the robot is exploring the object, whereas the displacement will be relatively large for keypoints in the background. Furthermore, since the robot moves on a flat surface, keypoints will only move in the horizontal direction. Allowing some fluctuations, we classify a stable keypoint as an object point when

$$\left(\left|x_i - x_n\right| < x_T\right) \wedge \left(\left|y_i - y_n\right| < y_T\right) \quad (1)$$

where we use $x_T = 12$ and $y_T = 4$. Otherwise, the stable keypoint 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 keypoints 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 keypoints, $\mathcal{O}$, and the keypoint database, $\mathcal{D}$, we determine the activation of every model, $m_{ID,\theta}$, for object *ID*, and pose θ. The activation of a model is based on the set of observations, $\mathcal{M}$, that support the model, where $\mathcal{M} \subset \mathcal{O}$, and

$$\mathcal{M} = \bigcup \left(\mathbf{p}_i \in \mathcal{O} \mid o_n = ID \wedge \alpha_n = \theta\right)$$

where o_n and α_n are respectively the object ID and pose of the nearest neighbor, $\mathbf{k}_n$, of $\mathbf{p}_i$ in the keypoint database. Every supporting observation p_i in $\mathcal{M}$ gives an activation a_i of

$$a_i = \exp\left(-\left\|\mathbf{p}_i - \mathbf{k}_n\right\|\right)$$

The total activation of model $m_{ID,\theta}$ given the observed keypoints $\mathcal{O}$ is given by

$$A\left(m_{ID,\theta} \mid \mathcal{O}, \mathcal{D}\right) = \frac{\sum_{i \in \mathcal{M}} a_i}{\sqrt{\left|\mathcal{D}_{ID,\theta}\right|}} \quad (2)$$

where $\left|\mathcal{D}_{ID,\theta}\right|$ is the number of keypoints in the keypoint database that are associated with object *ID* en pose θ.

Equation (2) gives the activation of a specific pose of an object. This activation increases with the number of supporting observations relative to the number of keypoints 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 keypoints given the same ratio of matched observations to database keypoints.

Finally, the robot can actively gather more evidence for recognition. When driving around the object, we accumulate the evidence by

$$A\left(m_{ID,\theta}\right) = \sum_{\delta \in \mathrm{E}} A\left(m_{ID,\theta} \mid \mathcal{O}_\delta\right) \quad (3)$$

where E is the set of angles from where the extra observations are made. This results in more robust recognition of 3D objects. In our experiments, we determine the activation of an object by summing up all individual poses. The most active object, *obj*, is the output of the recognition process:

$$obj = \arg\max_{ID} \left(\sum_{\theta=0°}^{350°} A\left(m_{ID,\theta}\right) \right) \quad (4)$$

B. Keypoint Clustering

As explained in the introduction, 3D object recognition with SIFT has the main disadvantage that the computational time needed increases with the number of keypoints stored in the database. We therefore use a growing-when-required (GWR) network [22] to efficiently cluster keypoints that are highly similar. A GWR-network is a clustering method, very similar to a growing-neural-gas (GNG) network [23]. Both networks are based on Kohonen's self-organizing maps (SOM) [24]. A SOM is an efficient method to cluster high-dimensional data. The disadvantage, however, is that the number of clusters 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 keypoints. A GNG-network is an adaptation of a SOM which can dynamically change the number of nodes, i.e., clusters, 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 keypoint. A GWR-network does just that, it adds nodes if this is required.

The GWR-network as described in [22] uses edges between nodes. This is based upon the SOM, and results in a topological preservation of the network in the sense that connected nodes in the network correspond to neighboring points in the input space. Usually, the edges in a GWR network are used in learning to move the neighboring nodes of the winning node closer to the presented input. This is undesirable for object learning, since the presentation of an input not only changes the representation of the corresponding keypoint, but also of neighboring keypoints. In the end, this will result in changing keypoints so much that they are not recognizable anymore. We therefore omitted the edges from the GWR-network.

For the description of our implementation of the GWR-network, we follow the notation and description in [22].

Let K be the set of observed keypoints when learning the objects, A be the set of nodes in the network, $\mathbf{w}_n$ be the weight vector of node n (of the same dimensionality as the SIFT keypoints), and t_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 a randomly picked keypoint from K, and $t_1 = 0$. Then, for each keypoint $\mathbf{k}$ from K we do:

1. $\mathbf{k}$ and the object *ID* and pose θ to which the keypoint belongs are input to the network.
2. Select the best matching node $s \in A$, such that

* Slightly better results could be obtained if one takes advantage of the fact that keypoints in the background move in the same direction as the robot, whereas visible object keypoints are in front of the center of rotation, and therefore move in the opposite direction.

Fig. 2 A number of images from our database.

$$s = \underset{n \in A}{\operatorname{argmax}} \left\| \mathbf{k} - \mathbf{w}_n \right\|$$

3. Calculate the activity of the winning node
   $a_s = \mathrm{e}^{-\left\| \mathbf{k} - \mathbf{w}_s \right\|}$
4. Calculate the firing counter h_s
   $h_s = 1 - \left(1 - \mathrm{e}^{-\alpha_b t_s / \tau}\right) / \alpha_n$
   where $\alpha_b = 1.05$, $\alpha_n = 1.05$ and $\tau = 3.33$
5. if $\left(a_s < a_T\right) \wedge \left(h_s < h_T\right)$, add a new node r
   - $A = A \cup \{r\}$
   - $\mathbf{w}_r = \mathbf{k}$
   - $R_r = \{\langle ID, \theta \rangle\}$
   where $a_T = 0.8$ and $h_T = 0.4$.
6. Else, adapt the weights of the winning node
   - $\mathbf{w}_s = \mathbf{w}_s + \eta \cdot h_s \cdot \left(\mathbf{k} - \mathbf{w}_s\right)$
   - $R_s = R_s \cup \{\langle ID, \theta \rangle\}$
   where $\eta = 0.05$.
7. $t_s = t_s + 1$

When the presented keypoint 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 keypoints. If, on the other hand, the presented keypoint differs from the existing nodes, and the firing counter of the nearest node is below the threshold h_T, the presented keypoint is added as a new node. In this way, the GWR-network clusters similar keypoint, 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 keypoints when such a keypoint is observed. This is in contrast with [8], where important evidence is discarded by choosing only keypoints that match uniquely with one object.

III. Results

We used seven objects placed in a cluttered environment for our image database (see fig. 2). A mobile robot equipped with a CCD camera was used to take images from 36 different viewpoints around the objects. The image database consists of four different sets, where the orientation of the objects is respectively 0°, 90°, 180° and 270° with respect to the environment. This provides a different background for each object/pose combination. In the experiments, training was done on one set, while the other three sets were used for testing the performance. This resulted in 12 different cross-validation tests.

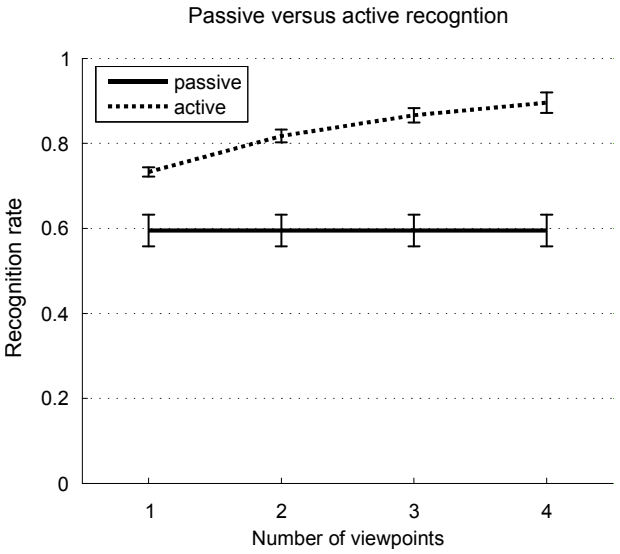

Fig. 3 The mean recognition rates for the passive and active method. The active method is plotted as a function of the number of viewpoints used to explore the object. The plot of the passive method is flat, since it does not use exploration. The error bars give the 95% confidence interval.

In our first experiment, we tested the performance of our active approach to 3D object recognition and compared it with a passive approach. Fig. 3 shows the mean recognition rate of both approaches. For the active approach, the performance is plotted as a function of the amount of accumulated evidence. Since the passive approach does not accumulate evidence from multiple viewpoints, it is drawn as a horizontal line. The error bars show the 95% confidence intervals. The means and confidence intervals are calculated from the 12 cross-validations. The active approach clearly outperforms the passive approach. Already with one viewpoint, the use of active vision to select stable object keypoints gives significantly better performance than passively considering all visible keypoints (respectively 73% and 60% success). With increasing accumulation of evidence, the recognition rate rises from 73% to 90%.

In the second experiment we compared the performance of the GWR-network with the standard SIFT method, both using active vision. The learned keypoints are presented to the GWR-network in random order. We therefore performed ten different experimental runs to test the performance of the GWR-network. The standard SIFT method has on average 6429 keypoints in the database. The GWR-network has 2396 keypoint clusters (37% of standard SIFT), making it 2.7 times

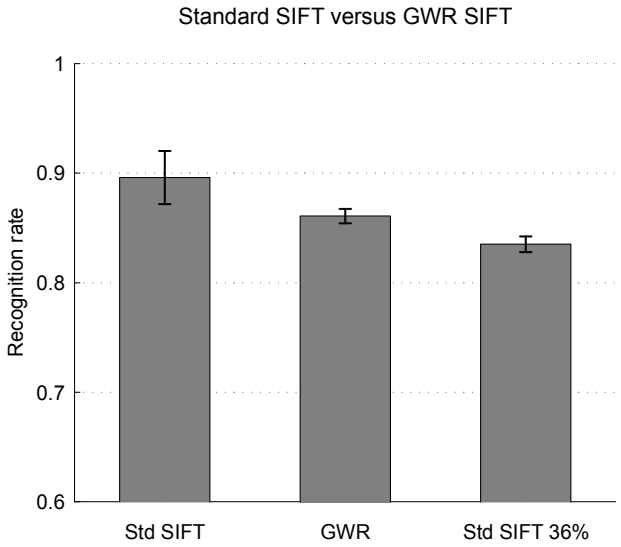

Fig. 4 The mean recognition rates for standard SIFT, the GWR-network and standard SIFT using 36% of the keypoint database. Standard SIFT uses on average 6429 keypoints. The GWR-network has on average 2396 keypoint clusters (37% of standard SIFT). For the GWR-network and SIFT using 36% keypoints, the data is acquired from 10 experimental runs. The error bars give the 95% confidence intervals.

faster than standard SIFT. We also compared the GWR-network with standard SIFT that used only 36% of the keypoints, approximately the same amount of keypoints as the GWR-network uses. These keypoints are selected randomly from the keypoint database. Again we performed 10 experimental runs. Fig. 4 shows the result of this experiment. We see that the standard SIFT method performs best, with 90% recognition rate. The GWR-network performs slightly (but significantly) worse. On the other hand, the GWR-network performs significantly better than standard SIFT using approximately the same number of keypoints.

IV. Discussion

Our experiments show the successful use of object exploration for 3D object recognition. Exploration is used for (1) stable keypoint detection, (2) object segmentation, and (3) exploration. The active vision approach performs significantly better than when passively learning to recognize objects. The problem of the passive method is that it not only learns to associate the keypoints of the object itself with the object, but also the background keypoints. Because we want the objects to be recognized on different backgrounds, the passive approach performs badly in recognizing the objects.

Our second experiment tested the use of GWR-networks for clustering keypoints. Although the performance of the GWR-network in terms of recognition rates is worse than standard SIFT, it is computationally more efficient. The GWR-network uses 2.7 times fewer keypoints, while performing significantly better than SIFT using the same number of keypoints. Reducing the amount of keypoints is important for object recognition using SIFT, especially with a growing number of objects. Our results using the GWR-network are promising, although there is room for improvements.

One of the problems with the GWR-network is that keypoints that are presented early to the network are badly represented in the final network, in contrast to keypoints that are presented later. This might account for the fact that the GWR-network performs worse than standard SIFT using the complete keypoint database. The keypoints presented first are simply lost. However, the fact that the network performs significantly better than SIFT using the same number of keypoints, shows that it is capable to effectively cluster the keypoints. Further research needs to be done to deal with this problem.

In this study, we did not use additional methods to improve the recognition rate. A good way to boost recognition is to use a geometric fit between sets of keypoints, for instance the geometric verification method described in [8]. This method can be used both with our active vision method and with the GWR-network. We expect a similar increase in performance for our methods. A further improvement in the computational efficiency for all proposed methods can also be achieved by applying, for instance, the best-bin-first search method [21] for matching keypoints in addition with the methods presented in this paper.

Summarizing, we showed the successful use of active vision to simplify complex recognition tasks. We furthermore, demonstrated the possibility to reduce the number of keypoints for SIFT by using a GWR-network. Both methods make implementing object recognition in the real world more feasible.

References

[1] J. J. Gibson, *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin, 1979.
[2] R. Pfeifer and C. Scheier, *Understanding Intelligence*: MIT Press, 1999.
[3] D. H. Ballard, Animate Vision, *Artificial Intelligence*, vol. 48, pp. 57-86, 1991.
[4] G. Metta and P. Fitzpatrick, Early integration of vision and manipulation, *Adaptive Behavior*, vol. 11, pp. 109-128, 2003.
[5] C. Harris and M. Stephens, A Combined Corner and Edge Detector, presented at The Fourth Alvey Vision Conference, Manchester, UK, 1988.
[6] C. Schmid and R. Mohr, Local greyvalue invariants for image retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 530-535, 1997.
[7] D. G. Lowe, Object recognition from local scale-invariant features, presented at International Conference on Computer Vision, Corfu, Greece, 1999.
[8] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, vol. 60, pp. 91-110, 2004.
[9] P. Moreels and P. Perona, Evaluation of Feautures Detectors and Descriptors based on 3D Objects, *International Journal of Computer Vision*, vol. 73, pp. 263-284, 2007.
[10] D. G. Lowe, Local Feature View Clustering for 3D Object Recognition, presented at IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, 2001.
[11] V. Ferrari, T. Tuytelaars and L. v. Gool, Simultaneous Object Recognition and Segmentation from Single or Multiple Model Views, *International Journal of Computer Vision*, vol. 67, pp. 159-188, 2006.
[12] F. Rothganger, S. Lazebnik, C. Schmid and J. Ponce, 3D Object Modeling and Recognition Using Local Affine-Invariant Image

Descriptors and Multi-View Spatial Constraints, *International Journal of Computer Vision*, vol. 66, pp. 231-259, 2006.
[13] P. Fitzpatrick, First Contact: an active vision approach to segmentation, presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vagas, Nevada, 2003.
[14] K. Mikolajczyk and C. Schmid, An affine invariant interest point detector, presented at the 7th European Conference on Computer Vision, Copenhagen, Denmark, 2002.
[15] M. Lehrer and G. Bianco, The turn-back-and-look behaviour: bee versus robot, *Biological Cybernetics*, vol. 83, pp. 211–229, 2000.
[16] S. Nolvi, Adaptation as a more powerful tool than decomposition and integration, presented at the Workshop on Evolutionary Computing and Machine Learning, 13th International Conference on Machine Learning, Bari, Italy, 1996.
[17] S. D. Roy, S. Chaudhury and S. Banerjee, Active recognition through next view planning: a survey, *Pattern Recognition*, vol. 37, pp. 429-446, 2004.
[18] L. Paletta and A. Pinz, Active object recognition by view integration and reinforcement learning, *Robotics and Autonomous Systems*, vol. 31, pp. 71-86, 2000.
[19] H. Borotschnig, L. Paletta, M. Prantl and A. Pinz, Appearance-based active object recognition, *Image and Vision Computing*, vol. 18, pp. 715-727, 2000.
[20] J. H. Friedman, J. L. Bentley and R. A. Finkel, An algorithm for finding best matches in logarithmic expected time, *ACM Transactions on Mathematical Software*, vol. 3, pp. 209-226, 1977.
[21] J. Beis and D. G. Lowe, Shape indexing using approximate nearest-neighbour search in high-dimensional spaces, presented at Conference on Computer Vision and Pattern Recognition, Puerto Rico, 1997.
[22] S. Marsland, J. Shapiro and U. Nehmzow, A self-organising network that grows when required, *Neural Networks*, vol. 15, pp. 1041-1058, 2002.
[23] B. Fritzke, A Growing Neural Gas Network Learns Topologies, presented at Advances in Neural Information Processing Systems (NIPS'94), Denver, 1995.
[24] T. Kohonen, The self-organizing map, *Proceedings of the IEEE*, vol. 78, pp. 1464-1480, 1990.