Learning Operators for View Independent Object Recognition

Josef Pauli
Christian–Albrechts–Universität
Institut für Informatik
Preusserstrasse 1-9, 24105 Kiel, Germany
jpa@informatik.uni-kiel.de

Abstract

In the context of vision based robotics our work focusses on the recognition of target objects for object grasping. The objects are arbitrarily shaped, and the viewing position and orientation of the camera is arbitrary as well. Due to various imponderables it is hard to geometrical model all relevant 3D object shapes and all effects of perspective projection. Therefore 3D model based approaches for object recognition are unfavorable in our robot application. Rather we use typical 2D appearance patterns of the target object which will be learned in a training phase. To acquire training patterns the turning angle of the object and the focal length of the camera lens must be changed systematically. A multidimensional Gaussian is defined for each typical pattern and used as basis function (GBF) for computing similarities to certain image patches. By appropriate linear combination of the GBFs we get a smart operator for recognizing the target object (regardless of object turn or distance). Using various GBF configurations we expound the trade–off between invariance and reliability of recognition.

1 Introduction

We have equiped a robot manipulator with a vision system for autonomously grasping and arranging target objects. The vision system deals with two sub–tasks, recognizing the relevant object in the image and evaluating the stability of grasping situations. It is trivially accepted, that vision systems have to be purposive for solving a certain task (Aloimonos [1]) in order to avoid useless expenditure of work. Taking this principle into account, we must recover the kind of partial scene information indispensable needed to recognize and manipulate an object.

For example, object recognition has to be grounded on features which discriminate between the target object and other objects and furthermore are easy to extract from the image. Obviously, geometric features describing the exact 3D shape would be sufficient for object recognition. However, in order to compute the geometric features one has to bridge the problematic gap between photometric gray level edges and geometric surface discontinuities (Maxwell and Shafer [4]).

Thus it is advantageous to discover basic features which are computed directly from raw gray levels or filter responses and avoiding image segmentation.

Similarily, for the second sub–task which consists of manipulating an object the 3D shape is not required in full detail. Rather having parallel–jaw grippers mounted we only must determine opposite grasping areas which are nearly parallel. Furthermore for object arranging only the alignment areas of the relevant objects have to be detected. Therefore specific operators are needed for evaluating the relationship between the target object and the grasping fingers (or other objects). Both for the recognition and manipulation task we don't have to reconstruct the target object in detail.

This work focusses on the recognition of target objects by taking the mentioned principle into account. We implemented a vision system which is constructional in the sense that operators for recognizing target objects can be learned for the actual environment. The operators are based on *2D appearance patterns* of the relevant objects or *response patterns* resulting from specific filter operations. The most closely related work is from (Murase and Nayar [5]) who describe a method for *appearance based object recognition*. An appearance manifold of the target will be acquired by systematically changing the object turning angle and taking images in discrete steps. Based on the most important eigenvectors of the covariance matrix the *Karhunen–Loeve transform* is used for compressing the manifold. Because a linear reduction of the dimension takes place the approach is suitable only when the data are sufficiently linearly distributed (Joutsensalo and Miettinen [3]).

Alternatively, we show the use of networks of gaussian basis functions for appearance based object recognition. The approach allows a *nonlinear dimension reduction* and does not assume linear constrained appearance manifolds. By carefully spreading and configuring basis functions an optimal operator can be learned which carries out a compromise between the invariance and reliability criterion.

2 Regularization networks for recognition

Recognition of an object in a certain image area is executable by applying a specific function to the signal structure of the area. The output of that *recognition function* can be defined as a real value between 0 and 1 which encodes the confidence that a certain object is depicted in the image area. Unfortunately, by changing the pose (position or viewing angle) of the cameras the appearance of the objects changes. Regardless of variable size or variable graylevel structure of the 2D pattern of a target object the recognition function should invariantly compute values near to 1. On the other hand, the recognition function should compute values near to 0 for image areas depicting any other object or situation.

We need a scheme for approximating the recognition function based on sample data of the input–output relation. The function should fit the sample data to meet *closeness* constraints and should generalize over the sample data to meet *smoothness* constraints. Neglecting the aspect of generalizing leads to overfitted functions, otherwise, neglecting the fitting aspect leads to overgeneralized functions. Thus both aspects have to be combined to get a qualified function approximation. The *regularization approach* (Poggio and Girosi [8]) incorporates both constraints and determines such a function by minimizing a functional.

Let $S = \{(\vec{p_i}, r_i) \in (R^m \times R) | i = 1, \cdots, N\}$ be the sample data representing the input–output relation of the function that we want to approximate. The functional (1) consists of a *closeness term* and a *smoothness term* which are combined by a *regularization factor* λ expressing relative importance.

$$H(f) = (\sum_{i=1}^{N} (r_i - f(\vec{p_i}))^2) + \lambda \parallel Pf \parallel^2 \tag{1}$$

The first term computes the sum of squared distances between the desired and the actual outcome of the function. The second term incorporates a differential operator P for describing the smoothness of the function.

Under some pragmatic conditions (see again [8]) the solution of the regularization functional is given by equation (2).

$$f(\vec{p}) = \sum_{i=1}^{N} v_i G_i(\vec{p}, \vec{p_i}) \tag{2}$$

The *basis functions* G_i are specified for a limited range of definition having $\vec{p_i}$ as the center. Based on a general *window function* G we get the N versions G_i by shifting the center of definition through the input space to the places $\vec{p_1}, \cdots, \vec{p_N}$. The several versions are called *support functions* due to the local range of definition. Equation (2) says that the solution of the regularization problem is a linear combination of (typically nonlinear) support functions.

In our application the sample data S consist of image patterns $\vec{p_i}$ and recognition values r_i. Each pattern represents the signal structure of a certain image area and will be formulated by a vector taking a specific order into account. The dimension n of such a vector is equal to the pixel number of a pre–defined pattern size which typically can be *a few hundreds up to a few thousands*. The recognition value r_i is a real scalar between 0 and 1. Based on the sample data S a recognition function f_r has to be acquired which is based on support functions and combination factors. We define the support functions as multidimensional symmetric gaussians (*Gaussian Basis Functions, GBFs*) in which the dimension is again equal to the pixel number of the patterns.

The number of GBFs could (!) be equal to the number of samples in S. In that case the center of each GBF would be defined by the patterns $\vec{p_i}$. On account of applying principles of the *minimum description length* (MDL, Rissanen [10]) it is of interest to discover the minimum number of support functions needed to reach a critical quality of the function approximation. The actual motivation for incorporating the MDL principle is to get a recognition function which is efficiently applicable (see below). Rather than using the patterns $\vec{p_1}, \cdots, \vec{p_N}$ for defining GBFs we cluster the patterns into M groups – with $M \leq N$ – according to similarity and compute typical patterns $\vec{c_1}, \cdots, \vec{c_M}$. Each typical pattern $\vec{c_j}$ is used for defining a GBF by taking the pattern as the center of the definition range.

$$G_j(\vec{p}, \vec{c_j}) = exp(-\frac{\parallel \vec{p} - \vec{c_j} \parallel^2}{2\sigma_j^2}) \tag{3}$$

The function G_j computes a value of similarity between the typical pattern $\vec{c_j}$ and a new pattern $\vec{p}$. The specification of similarity is affected by the pre–specified

parameter σ_j which determines the support size and shape of the GBF. It is intuitively clear that the ranges of definition of the support functions G_j must overlap to a certain degree in order to approximate the recognition function appropriately. The overlap between the GBFs is just determined by the parameters σ_j. The combination of the GBFs is defined by the factors w_j.

$$f_r(\vec{p}) = \sum_{j=1}^{M} w_j G_j(\vec{p}, \vec{c_j}) \tag{4}$$

Equations (3) and (4) define the scheme for all recognition functions to be discussed in this work.

Currently, the mentioned approximation scheme is popular in the neural network literature under the terms *GBF network* or *regularization network* (Bishop [2, pp. 164–191]). A regularization network consists of an input layer, a layer of hidden nodes and an output layer. In our application the input layer and the output layer consist of one node respectively, and the number of hidden node is equal to the number of GBFs. The layer of hidden nodes approximates the appearance manifold of the target object and therefore the whole network can be used as recognition function. The input node represents the input pattern of the recognition function. The hidden nodes are defined by the M support functions and all these will be applied to the input pattern. The output node computes the recognition value by a weighted combination of results coming from the support functions. The input space of the regularization network is the set of all possible patterns of the pre–defined size but each hidden node only responds significant for a certain subset of these patterns. Unlike simple applications of regularization networks, in our application (of object recognition) the dimension of the input space is extremely high.

3 Learning operators for recognition

The approach for learning a recognition function is as follows:
(i) We take sample images containing the object which has to be recognized at a later date. The samples differ from each other due to a systematic change of the view conditions. (ii) Optionally, we apply specific filters to the image in order to enhance or express certain properties (see section 6). (iii) From each of the (filtered) sample images we extract a small rectangular area having the relevant object inside. The generated set of training patterns is the basis for learning the recognition function (that is, the GBF network). (iv) According to the approach for learning a GBF network we first have to cluster the training patterns with regard to similarity. (v) Finally, we determine appropriate combination factors of the GBFs by least squares fitting using the *pseudo inverse technique*.

Steps (i), (ii), and (iii) will be illustrated in the sections below. Step (iv) is implemented as follows. We use a clustering approach which is similar to the error–based ISODATA clustering algorithm in Schalkoff [11, pp. 109-125]. The algorithm initially groups the patterns using the standard *K–means* method. Then, clusters exhibiting large variances are split in two, and clusters that are too close together are merged. Next, K–means is reiterated taking the new clusters into account.

This sequence is repeated until no more clusters are split or merged. The allowed variances within a cluster can be controlled by specific parameters.

Also step (v) needs further explanation. First, a set of M support functions will be applied to each pattern $\vec{p_i}$ of a set of N training patterns. This results in a matrix A of similarity values with N rows and M columns. Second, we define an N–dimensional vector $\vec{h}$ of desired output values. E.g., the GBF network has to compute constantly the recognition value 1 for each training pattern. Third, we define a vector $\vec{w}$ which comprises the unknown combination factors $w_1, \cdots, w_M$ of the support functions. Finally, the problem is to solve the equation $A\vec{w} = \vec{h}$ for the vector $\vec{w}$. According to Press et al. [9, pp. 671–675] we compute the pseudo inverse of A and determine the vector $\vec{w}$ directly.

$$A^{\#} = (A^T A)^{-1} A^T, \qquad \vec{w} = A^{\#}\vec{h} \tag{5}$$

The learned operator for the recognition of an object is defined by a GBF network. The collection of GBFs is based on a set of typical patterns (appearance patterns). The support of the GBFs specifies the generalizing ability for applying the operator to new patterns of the object (not included in the training set). The question of interest is: *How many GBFs are needed and what size of the support is appropriate for robust object recognition?* The robustness will be defined by incorporating an *invariance* criterion and a *reliability* criterion. The invariance criterion strives for an operator which responds nearly equal for any appearance pattern of the target object. The reliability criterion aims at an operator which clearly discriminates between the target object and any other object or situation. Regions of the appearance space which represent views of objects other than the target object or any background area should be given low confidence values.

We will experimentally demonstrate a conflict in trying to maximize both criterions simultaneously. Therefore related to the overfitting/overgeneralizing dilemma (discussed above) a compromise is needed. By changing the number and the support size of the GBFs we show the invariance and reliability performance of recognition functions. Section 4 presents experiments on object recognition under arbitrary view angle, section 5 deals with object recognition under arbitrary view distance. Finally, section 6 discusses the approach and mentions future work.

4 Object recognition under arbitrary view angle

For learning an appropriate operator we actually must take sample images of the target object under several view angles. Due to a momentary lack of camera mobility we instead turn the object and acquire turn–dependent appearance patterns (size of the object patterns: 15x15 = 225 pixel). Figure 1 shows a subset of eight patterns from an overall collection of 32. The collection is devided into a training and a testing set comprising 16 patterns each. The training set has been taken by equidistant turning angles of 22.5 degree and the testing set differs by an offset of 10 degree. Therefore, both in the training and testing set the turn of the object changes in discrete steps over the range of 360 degree.

Fig. 1. Varying turning angle of the target object.

The collection of GBFs and their combination factors will be learned according to the approach of section 3. By modifying the number and/or the support of the GBFs we get specific operators. In the first experiment a small support has been chosen which results in a spare overlap of the GBFs. Four variants of operators will be defined for recognizing the target by choosing 2, 4, 8 and 16 GBFs respectively. Figure 2 shows four curves ((a), (b), (c), (d)) of confidence values computed by applying the operators to the target object of the test images. The more support functions are used, the higher the confidence values for recognizing the target. The confidence values vary significant when changing the turning angle of the object, and therefore the operators are hardly invariant.

The second experiment differs from the first in that a large support of the GBFs has been used leading to a broad overlap. Figure 3 shows once again four curves of confidence values produced by the new operators. Consistently, the invariance criterion improves and the confidence nearly takes the desired value 1. Taking only the invariance aspect into account, the operator characterized by many GBFs and large support is the best (see curve (d) in Figure 3).

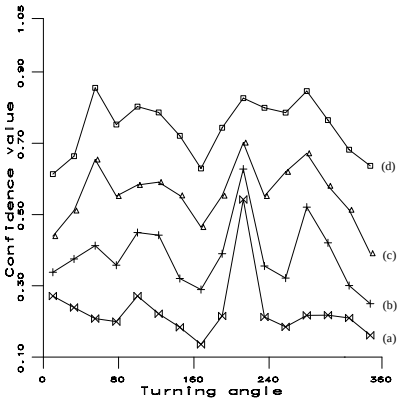

Fig. 2. Confidence of target recognition, four different sizes of GBF network, constant *small* support.

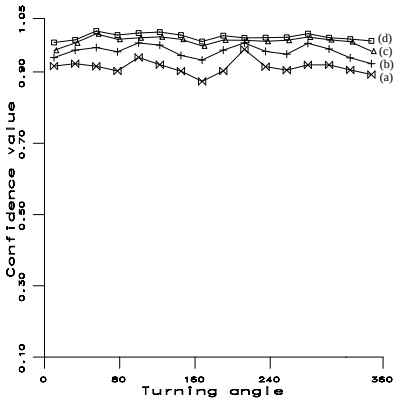

Fig. 3. Confidence of target recognition, four different sizes of GBF network, constant *large* support.

The third experiment incorporates the reliability criterion into object recognition. An operator is reliable, if the recognition value computed for the target object is significant higher than those of other objects. In the experiment we apply the operators to the target object and to three test objects (outlined in Figure 4 by white rectangles). Based on 16 GBFs as support functions we define 6 operators by systematically increasing the support in 6 steps.

Figure 5 shows four curves dedicated to the target and the three test objects. If we enlarge the support of the GBFs and apply the operators to the target object a slight increase of the confidence values occurs (curve (a)). If we enlarge the sup-

port in the same way and apply the operators to the test objects the confidence values increase dramatically (curves (b), (c), (d)). Consequently, the curves for the test objects approach the curve for the target object. Increasing the support of the GBFs makes the operator more and more unreliable (overgeneralization). However, according to the first experiments an increasing support makes the operator more and more invariant with regard to object turn (overfitting). Thus a compromise has to be made in specifying an operator for object recognition.

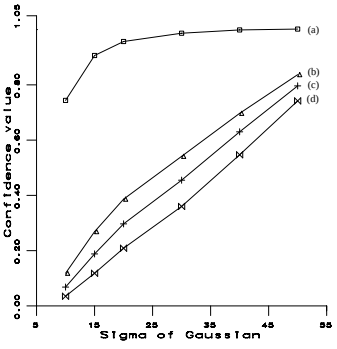

Fig.4. Target object (outlined by a bold rectangle) and test objects (outlined by fine rectangles).

Fig.5. *Discriminating* the target and other objects, constant size of GBF network, increasing support.

5 Object recognition for arbitrary view distance

For learning an appropriate operator we actually must take sample images of the target object under several spatial distances between object and camera. Due to the lack of camera mobility we instead change the focal length of the lens in order to reach similar projection effects. A small (large) focal length leads to a large (small) size of the object pattern in the image. Figure 6 shows on the left the image of an experimental scene (depicting the target object and other objects) taken under a mean focal length. On the right a collection of 11 training patterns depicts the target object which has been taken under a systematic decrease of the focal length in 11 discrete steps. The size of the object pattern changes from 15x15 pixel to 65x65 pixel. Each training pattern encodes essential information and therefore we don't have to build clusters. Accordingly, for each training pattern a single GBF is defined. The combination factors of the GBFs are determined as before.

A further collection of 10 test images has been acquired which differs from the training set by using intermediate values of focal length. We constructed three operators for object recognition taking small, middle, and large support of the GBFs. In the first experiment these operators have been applied to the target of the test images. Figure 7 shows in curve (a) the confidence values for recognizing the target object taking a small support into account. The confidence value differ significantly by changing the focal length and is far away from the desired value 1. Alternatively, if we use a middle support the confidence values have approached to 1 and the smoothness of the curve improved (curve (b)). Finally, using a large support will lead to invariant recognition values near to 1 (curve (c)).

Fig. 6. Scene for object recognition (left), target object under *varying focal length* (right).

In the second experiment we investigate the reliability criterion for the three operators from above. The operators will be applied to all objects of the test image (image on the left in Figure 6), and the highest confidence value of recognition has to be selected. Of course, it is expected to get the highest recognition value from the target object. Figure 8 once again depicts (equal to Figure 7) the confidence values of applying the three operators to the target object (curves (a), (b), (c)). Furthermore, if we use the operator of large support for all objects of the test images we frequently get higher confidence values for objects other than the target object (see curve (c1)). In those cases the operator fails to localize the target object. Alternatively, applying the operator of middle support will improve the realibility criterion (curve (b1) rarely surpasses curve (b)). Finally, the operator of small support localizes the target object in all test images. The highest confidence value will be computed just for the target object (curve (a) and curve (a1) are identical).

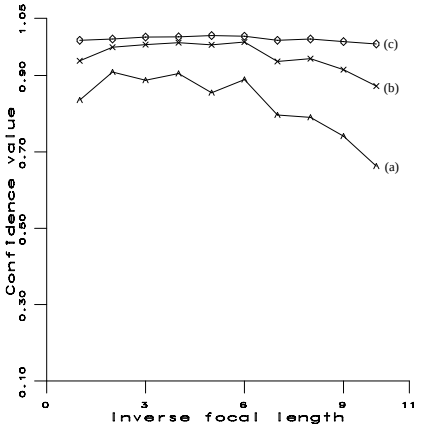

Fig. 7. *Invariance aspect* of target recognition, constant size of GBF network, small, middle, and large support.

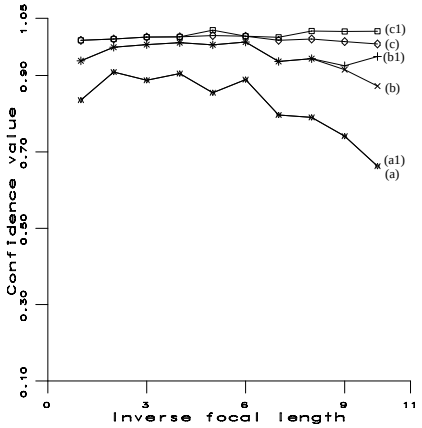

Fig. 8. *Reliability aspect* of target recognition, constant size of GBF network, small, middle, and large support.

6 Discussion

We have presented an approach for object recognition which does not require a priori knowledge of the three–dimensional geometric shapes. Rather the knowledge about objects is grounded in photometric appearance. As a consequence, the operator for object recognition must be learned on the basis of *raw gray levels or elementary filter responses*. The vision system for object recognition has to be adjusted to the actual environment in order to reach autonomy. By doing experiments prior to the application stage (like the one presented in this work) we can later make use of the learned recognition functions.

For representing and learning the operator a regularization network is used. In our application it is implemented with gaussian basis functions (GBFs) but any other bell–shaped parabolic function is possible as well. The regularization factor (see equation (1)) is controlled by the number and the support size of the basis functions. Various configurations reflect the well–known *invariance/reliability conflict* in object recognition (see Figures 3 versus 5, and 7 versus 8).

Increasing the support (and/or increasing the number) of GBFs makes the operator for object recognition invariant but unreliable. In order to reach a *certain degree of discriminability* between the target object and other objects the claim for strict invariance has to be reduced to *appoximate invariance*. Therefore, a further goal of doing experiments prior to application stage is to discover an appropriate compromise between invariance and reliability of object recognition.

The greatest strength of the approach is the ability to learn (approximate) invariants under *real world changes*. Usual methods for invariant pattern recognition (Wood [12]) have the constraint that the permitted transformations are acting directly on the patterns. As opposed to that in the recognition of three–dimensional objects one has to deal with changing view directions, view distances, object background, illuminations and maybe further imponderable changes. Therefore the pattern transformations are much more complicated because they originate from real world changes. Fortunately, our experiments proved, that approximate invariants can be learned with regularization networks.

The approach is generic in several aspects. First, invariants can be learned for any imaginable real world changes. For example, in (Pauli et al. [7]) we demonstrated the robust object recognition under varying object background and varying illumination. Second, the approach can also be applied to filtered images rather than raw images. For example, by using the output of a bandlimited filter we can learn a reliable recognition function which is specific to certain signal frequencies (to recognize certain shapes or textures). Third, the appearance based approach can not only be applied to object recognition but also to *situation recognition*. For example, in (Pauli [6]) we evaluated the stability of grasping situations based on the recognition of the relationship between the target object and the robot fingers.

Usually the problem of object recognition comes in combination with object localization in the image. Due to place limitation we don't mention the approach and therefore the interested reader is recommended to (Pauli et al. [7]).

In future work we will develop strategies for combining several invariants. Recognition functions have to be learned which are robust with regard to more complicated real world changes.

7 Facilities

Sun SPARCstation 10/40, TRC–Bisight active binocular head, Stäubli–Unimation RX–90 robotic arm, computer vision system KHOROS.

Acknowledgement
I am grateful to G. Sommer for new insights and useful discussions.

References

[1] Y. Aloimonos. Active vision revisited. In Y. Aloimonos, editor, *Active Perception*, pages 1–18. Lawrence Erlbaum Associates Publishers, New Jersey, 1993.

[2] Ch. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, London, 1995.

[3] J. Joutsensalo and A. Miettinen. Self–organizing operator map for nonlinear dimension reduction. In *International Conference on Neural Networks*. Perth, Australia, November, 1995.

[4] B. Maxwell and S. Shafer. A framework for segmentation using physical models of image formation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 361–368, 1994.

[5] H. Murase and S. Nayar. Visual learning and recognition of 3D objects from appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

[6] J. Pauli. GBF network architectures for robot vision. In *International Conference on Artificial Neural Networks in Engineering*. Missouri, November, 1996. submitted.

[7] J. Pauli, M. Benkwitz, and G. Sommer. RBF networks for object recognition. In B. Krieg-Brueckner and Ch. Herwig, editors, *Workshop Kognitive Robotik*, ZKW-Bericht 3/95, Center for Cognitive Sciences, Bremen University, 1995.

[8] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1990.

[9] W. Press, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C*. Cambridge University Press, Massachusetts, 1992.

[10] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, 1984.

[11] R. Schalkoff. *Pattern Recognition - Statistical, Structural, and Neural Approaches*. John Wiley and Sons, New York, 1992.

[12] J. Wood. Invariant pattern recognition - a review. *Pattern Recognition*, 29(1):1–17, 1996.