RBF Networks for Appearance-Based Object Detection

Josef Pauli, Michael Benkwitz, and Gerald Sommer

Address

Dr. Josef Pauli
Christian–Albrechts–University
Department of Computer Science
Preusserstrasse 1-9
24105 Kiel, Germany

Phone number: +49 431 5604 84
Fax number: +49 431 5604 81
E-mail: jpa@informatik.uni-kiel.de

Abstract

A predominant task occurring in Computer Vision is to recognize and localize the two-dimensional view of an object in the image. E.g., for controlling a vision based grasping robot autonomous object detection is necessary for grasping and assembling desired objects. The work being reported here uses RBF networks for learning and representing object detectors. The primary goal of learning is to generate detectors which are robust with respect to various imponderables, e.g., illumination conditions, projection features of the camera, viewing directions, light reflections on the object. We show the approach exemplary for real world images taken under varying illumination conditions and varying object background.

1 Introduction

Frequently model-based approaches are used for detecting a scene object in the image. First, a three-dimensional model object is assumed to be known a priori as well as the geometric projection features of the camera system [Faugeras, 1993, pp. 33-68]. Second, by taking the projection features into account the model object is transformed into a two-dimensional model and this one is matched with the image [Faugeras, 1993, pp. 483-558]. Third, the place in the image with the highest matching score describes the location of the object. The model object is defined using geometric attributes whereas the image consists of gray levels. Obviously, these are different types of representation. Image segmentation is the usual approach to bridge the gap [Maxwell and Shafer, 1994]. Unfortunately, image segmentation is only useful if (and only if) surface discontinuities of the scene object have corresponding gray level edges in the image. It is a matter of fact, nearly all problems in object recognition can be traced back to the mentioned gap of representation.

Our system bridges the gap of representation by using another category of models dispensing with geometric attributes. Rather the models will be defined directly as two-dimensional views in terms of gray level attributes incorporating implicitly the photometric features of the camera [Murase and Nayar, 1995]. Consequently, we do object detection without image segmentation and apply instead appearance grounded detectors. The detectors will be learned using a Radial Basis Function (RBF) network [Poggio and Girosi, 1990].

2 Learning appearance grounded detectors

2.1 Taking sample images

We take sample images containing the object which has to be detected at a later date. In this work the images differ from each other because of varying illumination conditions and varying object background. Figure 1 shows three exemplary images from an overall collection of 16, the relevant object is outlined by a white rectangle.

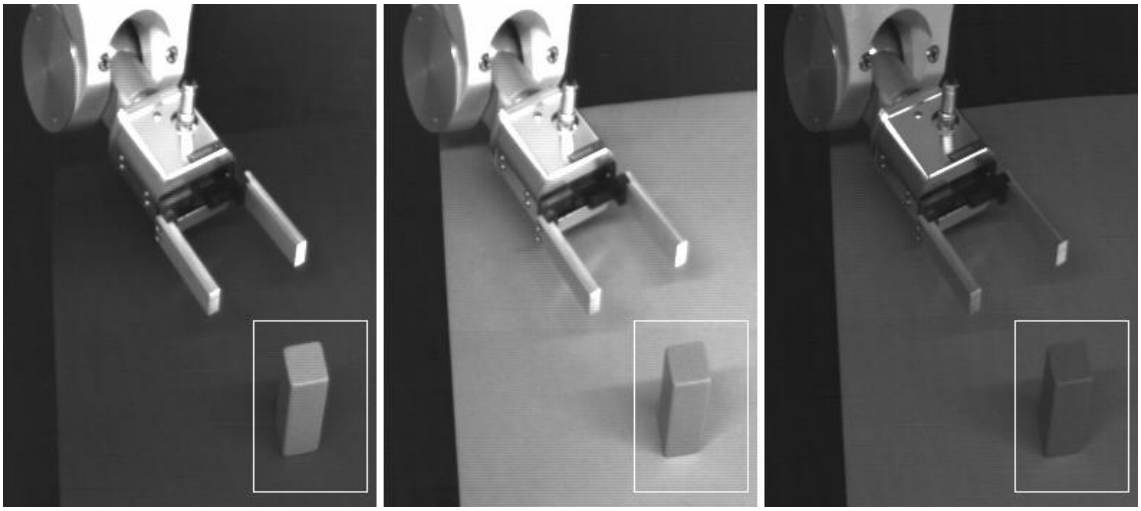

Figure 1: Three sample images with different object background or illumination.

2.2 Transforming the sample images

We apply an operator to the sample images in order to transform the gray levels of the depicted object into a specific structure, respectively. Based on this structure of the operator outcome the depicted object can be looked for in the image efficiently (see later in section 3). For example, our operator transforms the gray levels of the depicted object into a unique peak structure with the peak being localized approximately at the center of the depicted object. The operator is designed on the basis of a gauss-modulated cosinus wave which is two-dimensionally extended. Actually, the operator is a special case of the complex Gabor wavelet function (for short, wavelet) taking only the real part into account [Rioul and Vetterli, 1991]. Figure 2 shows the operator response for the left image of Figure 1 and Figure 3 shows especially for the area of the object the operator response in a three-dimensional manner.

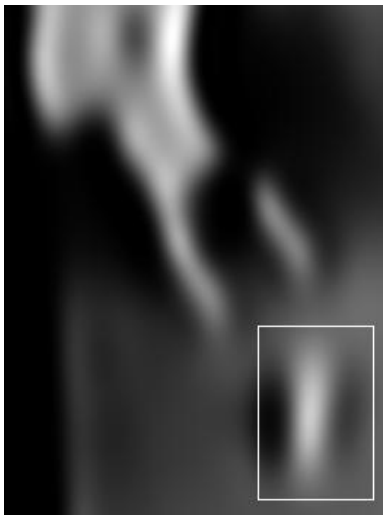

Figure 2: Response of the wavelet operator, shown as gray levels.

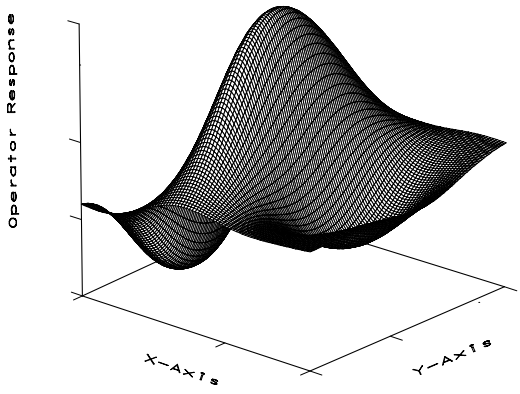

Figure 3: Operator response in the rectangular image patch, shown in three dimensions.

2.3 Clustering specific patterns of the object

From each of the 16 sample images we are interested in the operator response pattern of a small rectangular area having the relevant object inside (see Figure 2). These specific patterns are extracted in order to generate a training set from which to automatically learn the object detector. For example, Figure 4 shows the relevant patterns computed from the three sample images of Figure 1. To gain an insight regarding to the deviations of all 16 patterns we represent the intensity structure along a middle staight line in horizontal direction (X-direction) and along a middle staight line in vertical direction (Y-direction) (see Figure 5). Figure 6 shows the overlay of intensity structures in X-direction and Figure 7 in Y-direction.

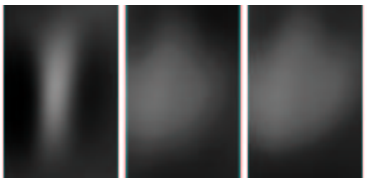

Figure 4: Three exemplary operator response patterns.

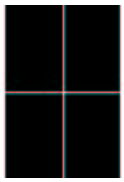

Figure 5: Two orthogonal straight lines for selectively showing the intensity structures.

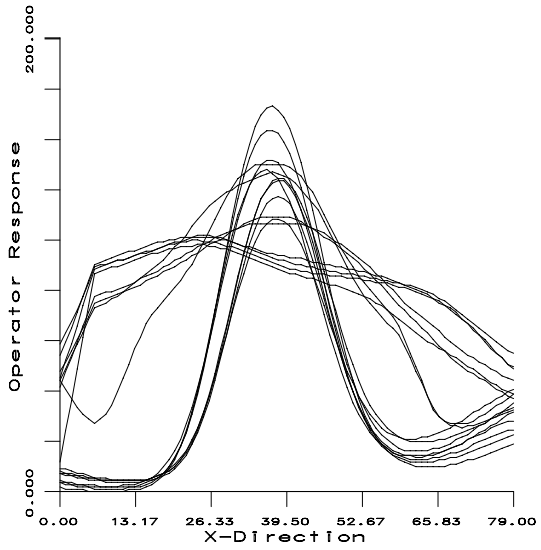

Figure 6: Overlay of 16 intensity structures in X-direction.

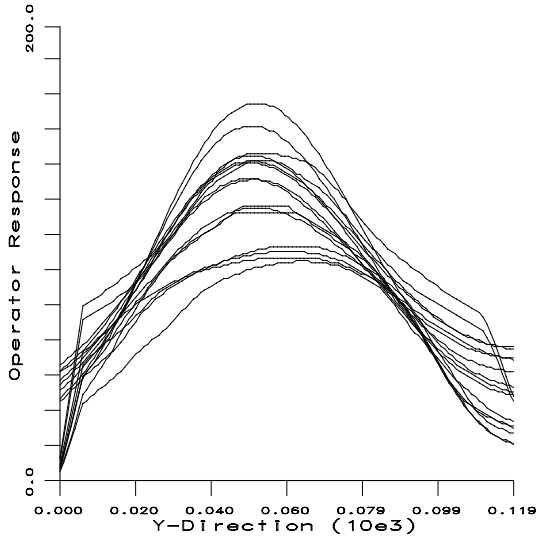

Figure 7: Overlay of 16 intensity structures in Y-direction.

According to the approach for learning an RBF network [Hush and Horne, 1993] we first have to group the 16 patterns for getting a smaller number of typical patterns. In this work we used a clustering approach which is similar to the error-based ISODATA clustering algorithm in [Schalkoff, 1992, pp. 109-125]. The algorithm initially groups the patterns using a standard K-means clustering approach. 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.

In the underlying situation the algorithm groups the 16 patterns into four clusters. For each cluster a typical pattern is created by computing point for point the mean value based on all patterns of the cluster. The typical patterns of the four clusters are shown in Figure 8, respectively. Additionally, with each typical pattern a second pattern is associated consisting of variance values representing point for point the mean deviation of the pattern values in the cluster. Figure 9 shows for the second cluster of Figure 8 the intensity structure of the typical pattern (middle curved line) together with the deviations for each point (upper and lower curved lines).[1] The four typical patterns of Figure 8 are used as centers of four Gaussians

[1]Only the intensity structure and the deviations along the middle straight line in X-direction are depicted.

buildung up the hidden layer of the RBF network. The number of dimensions of the Gaussian is set equal to the number of pixels in the pattern. For example, in order to prepare the experiments for this report the pattern size did vary between 100 and 10000 pixels. To specify the extension of a Gaussian in each dimension the σ-values (which fix the turning-points) have to be chosen appropriately. We define these σ-values respectively as the product of each variance value in the variance pattern with a user specified parameter (for each point in the pattern).

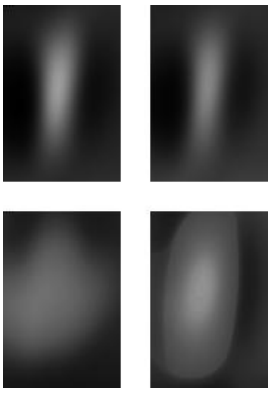

Figure 8: Typical patterns of four clusters.

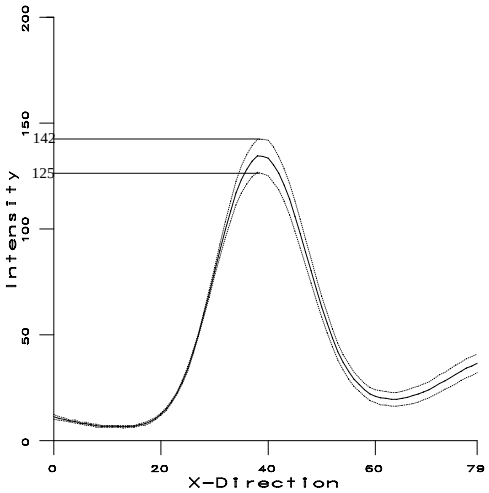

Figure 9: Deviations between the patterns of the second cluster.

2.4 Configuring the object detector

According to the approach for learning an RBF network [Hush and Horne, 1993] we finally assign to each training pattern a desired output vector. Based on the difference between the actual outcome and the desired outcome the weights for combining the nodes of the hidden layer are trained. The number of hidden layer nodes is equal to the number of cluster centers (four in our illustration) and the number of output nodes is equal to the number of objects which must be detected (one in our illustration). The output scalar should take a real value between 0 and 1 in order to encode a confidence as to whether the object of interest is inside a certain image area. For learning the weights (four in our illustration) we assign the value 1 to each of the 16 training patterns and apply the α-LMS rule [Widrow and Lehr, 1990]. Depending on the value for α in the interval $0.1 \leq \alpha \leq 1.0$, up to 100 learning cycles are enough for stabilizing the adaptation. By applying the learned RBF network to image areas confidence values are computed which fall after normalization into the unit interval.

3 Using the learned object detector

The learned detector is defined by an RBF network and is mainly composed of a number of typical patterns together with possible deviations from each typical pattern. Based on this specific arrangement of the detector we implemented an efficient method to look for the depicted object in the image. The efficiency arises from three factors. First, the typical patterns of the detector can be used in parallel for matching with the image. Second, the unique peak type of the operator response (of the depicted object) is to a certain degree invariant with respect to shrinking the image. In fact, all the experiments done for this report can been carried out with a shrinking factor of 8 leading to similar results. Third, by making use of the unique peak structure of the typical patterns we can efficiently select small areas of the image as candidate places for possible locations of the object.

The approach for detecting objects is as follows. First, we take a new image containing the

desired object using any illumination and object background. Second, we apply the operator used during the training phase. Third, for each typical pattern of the learned detector we take the point having the maximal intensity value and compute for this point an interval of possible intensity deviations. For example, in Figure 9 the two horizontal lines define variations of the maximal response in the interval beginning at the intensity of 125 and ending at 142. Fourth, we look for local maxima in the filtered image and only those local maxima being inside the relevant interval will be considered in detail. Fifth, only the image areas which surround a relevant local maximum must be put into the RBF network. The network computes a confidence value for the existence of the object in this areas. Sixth, the image area for which the RBF network computes a global maximal value is the location of the desired object.

4 Experiments

To illustrate the robustness of the approach the learned detector has been applied to 30 test images. Every image shows a part of the robot hand, a scene object for grasping and a background plane. All together three different objects and 10 different background planes have been taken into account. Of course, one object is actually the one for which the detector has been learned. The other two objects differ from the first one with regard to shape (see the three objects in the images of Figure 10). The background planes are nearly homogeneous and differ with regard to the mean gray level. Therefore the 10 planes can be organized with regard to increasing values of contrast between object and background graylevels.

First, the detector has been applied to the 10 images containing the well known scene object. As a result, the procedure has localized the object with a precision of plus or minus two pixel (see in the left image of Figure 10 the square patch at the center of the object silhouette). The confidence values measured on the basis of the different background planes are shown in Figure 11 (see the curve highlighted with square signs). Second, the detector has been applied to the 20 images containing one of the other two objects with varying background planes, respectively. Image by image the maximum confidence value within the object silhouette has been determined (see the square patches in the middle and right image of Figure 10). As a result, these confidence values are significantly lower compared to the values for the first object as shown in Figure 11. The curve highlighted with triangle (plus) signs shows the confidence values for images containing the second (third) object. Therefore, if the three objects all together are arranged in a scene the detector would correctly localize the desired object despite of varying background and illumination.

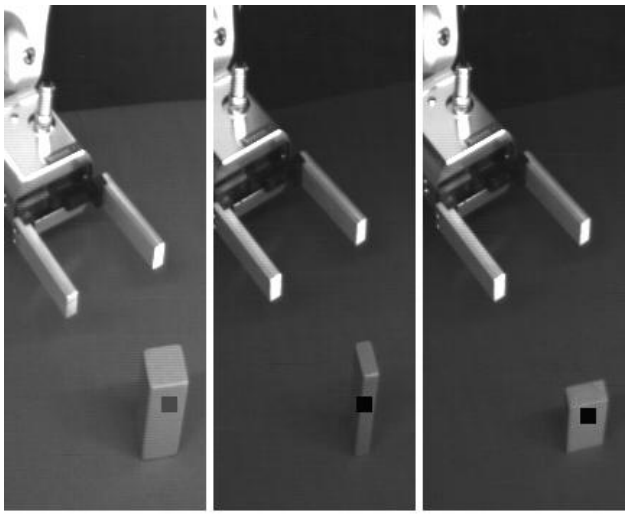

Figure 10: Applying the detector to test images, only three from a collection of 30 are shown.

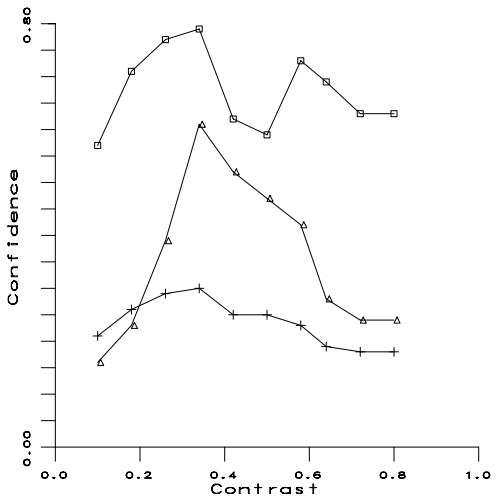

Figure 11: Confidence values for the existence of the desired object in an image.

5 Conclusion

We presented an approach for object detection which does not require explicitly encoded knowledge of the three-dimensional geometric shape. Rather the meaning of an object is grounded in photometric appearance [Cottrell *et al.*, 1990]. That is, the detector for the object must be learned online on the basis of the actual graylevel structure and elementary filter operators. In this way knowledge of the object appearance is encoded in the detector implicitly. For representing and learning the detector an RBF network is used. Object detection has been performed successfully in images taken under arbitrary illumination and arbitrary homogeneous object background.

In principle the RBF network together with the learning procedure constitute a shell for object detection. That is, specific image operators have be incorporated for building desired detectors. Currently we implement operators for separating textures. Based on those an RBF network can be created for detecting textured objects in a structured scene. The number of hidden layer nodes in the RBF network is variable depending on variance of imaging conditions and varying inhomogeneous object background. Due of this flexibility, functions for object detection can be created with high reliability. Finally, the number of output layer nodes is variable depending on the number of objects which have to be detected in the image. In this way a single RBF network represents object detectors for several objects all together.

References

[Cottrell *et al.*, 1990] G. Cottrell, B. Bartell, and C. Haupt. Grounding Meaning in Perception. In H. Marburger, editor, *14th German Workshop on Artificial Intelligence*, volume 251 of *Informatik Fachberichte*, pages 307–321, 1990.

[Faugeras, 1993] O. Faugeras. *Three-Dimensional Computer Vision.* The MIT Press, Massachusetts, 1993.

[Hush and Horne, 1993] D. Hush and B. Horne. Progress in Supervised Neural Networks. *IEEE Signal Processing Magazine*, 10:8–39, January 1993.

[Maxwell and Shafer, 1994] 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.

[Murase and Nayar, 1995] H. Murase and S. Nayar. Visual Learning and Recognition of 3D Objects from Appearance. *International Journal of Computer Vision*, 14:5–24, 1995.

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

[Rioul and Vetterli, 1991] O. Rioul and M. Vetterli. Wavelets and Signal Processing. *IEEE Signal Processing Magazine*, pages 14–38, October 1991.

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

[Widrow and Lehr, 1990] B. Widrow and M. Lehr. 30 Years of Adaptive Neural Networks - Perceptron, Madaline, and Backpropagation. *Proceedings of the IEEE*, 78:1415–1442, 1990.