Object Localization in 2D Images based on Kohonen's Self-Organization Feature Maps

C. Yuan and H. Niemann

Chair for Pattern Recognition
University of Erlangen–Nuremberg
Martensstraße 3, 91058 Erlangen, Germany
email: (yuan, niemann)@informatik.uni-erlangen.de

ABSTRACT

This paper presents a hybrid approach for neural object localization and recognition in 2D grey level images. The system combines an auto-associative network, two self-organization feature maps (SOMs), and a three layer feed-forward network trained with dynamic learning vector quantization (DLVQ). By using a hidden layer smaller than the input/output layers, the auto-associative network can be expected to find efficient ways of encoding the information contained in the input data set. Thus a dimension reduction of the input image can be achieved. The object localization scheme is then directly based on features which are detected automatically using the Kohonen's SOMs. After preprocessing images are split into small blocks and input to two Kohonen maps. Through training, the first map can detect the object area of the input image, while the second map can detect the object specific features. By integrating the features extracted from the output of the two maps and the DLVQ methods, we can locate different objects and estimate object pose (translation, rotation within the image plane and scale parameter).

1. INTRODUCTION

Object detection and localization has many applications in such areas as automatic target recognition, manufacturing, inspection etc. In the past artificial neural networks (ANNs) have shown their good ability in solving pattern classification problems [1]. Like the representative work of A. Khotanzad [2], objects are usually first segmented from the background, and its size is properly scaled to be fit into the input of ANN. Then some features will be extracted from the segmented objects and input to a properly structured net. Subsequently objects will be identified using a supervised learning method. Following that the pose parameters of recognized object can be estimated using some other ANNs. One limit of this approach is that it cannot estimate the scale parameter of object. Also it is generally very difficult to achieve a robust and automatic object segmentation because of changes of illumination, noise and occlusion. Furthermore only limited precision of pose estimation can be achieved if the object varies only slightly at different pose.

In order to overcome those drawbacks we use an unsupervised learning mechanism, which is based on the self-organization feature maps proposed by T. Kohonen [3, 4]. Our goal is to locate different 2D objects, irrespective of object translation, rotation, scaling and illumination changes. We use the approach of appearance based modeling, whereby blocks cut sequentially out of an image are encoded by the SOM. By using features extracted from the output of the maps, objects can be identified and located efficiently.

The organization of the rest of this paper is as follows. Section 2 deals with the basic theory of the Kohonen's SOM. Section 3 describes in detail our localization and recognition algorithm. Experimental results using our approach to localize five different industrial objects are presented in section 4. Finally Section 5 summarizes the whole paper.

2. THE SELF-ORGANIZATION MAPS

Amongst different ANN paradigms, SOM is one of the most popular unsupervised learning mechanisms depending on competition and a winner-take-all strategy. Application of SOM includes, but is not limited to image segmentation [5], object recognition [6], speech recognition [7], vector quantization [8].

The goal of the SOM is the projection of an input space of n-dimensions into a position in the map, which is made of a one- or two-dimensional lattice of output layer neurons. Figure 1 illustrates the network architecture for the two-dimensional case. The neuron whose weight vector has a minimum of the Euclidean norm distance from the input vector $\vec{x}$ is chosen as the winning neuron. The weight of the winner j and its neighborhood neurons are trained according to

$$\boldsymbol{w}_j(t+1) = \boldsymbol{w}_j(t) + h_{jk}(t)[\boldsymbol{x} - \boldsymbol{w}_j(t)] \qquad (1)$$

where t is the time during learning and $h_{jk}(t)$ is the neighborhood function, a smoothing kernel which is maximum at the winning neuron j. A wide applied neighborhood function is

$$h_{jk}(t) = \eta(t) \exp\left[\frac{-|\boldsymbol{r}_j - \boldsymbol{r}_k|^2}{2\sigma^2(t)}\right] \qquad (2)$$

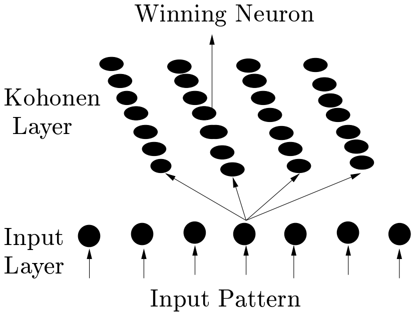


Figure 1. Architecture of the Kohonen network, two-dimensional case.

where $\eta(t)$ is a scalar valued learning rate and $\sigma(t)$ defines the width of the kernel. While $\boldsymbol{r}_j$ represents the location of the winner node, $\boldsymbol{r}_k$ ranges over all nodes. $h_{jk}(t)$ approaches 0 as $|\boldsymbol{r}_j - \boldsymbol{r}_k|$ increases and also as t approaches infinity.

The SOM is unlike most classification or clustering techniques in that it provides a topological ordering of the classes. So the map can catch the specific features of the input data automatically and organize them spatially.

3. LOCALIZATION ALGORITHM

3.1. Preprocessing

Due to the great number of pixels carrying the visual information, it is desirable to reduce the image data to be processed at the first stage. We use an auto-associative network (See figure 2) for extracting the most representative low-dimensional subspace from a high-dimensional pattern vector space. The network is trained through letting the outputs mimic the inputs. After training the output is discarded and the hidden layer is kept. We apply this method on the input image of 256 x 256 pixels two times and get the desired output image of 64 x 64 pixels.

3.2. Object Localization

After preprocessing, we use two Kohonen maps trained successively on the scale reduced images to locate different objects. The first map is trained on all the 256 blocks with the block size of 4x4 pixels cut from the dimension reduced images of the training set. The input of the first map has therefore 16 neurons. There are 12 neurons in the output layer, which can organize themselves according to the spatial properties of the 4x4 pixels into 12 different clusters. After training we examine the correlation of the reference vectors of each output neuron with the input image and can determine whether an output neuron stands for object area (edge, corner and bright area) or background (dark area). We label all the 256 blocks of an image according its output neuron index and can then compute three features: number of blocks belonging to object, number of blocks belonging to background, and the ratio of them. According to these three features we can differentiate objects which have different shapes and also know the scale information of them.

In order to differentiate some objects which look very similar, we use the second map. This map has an input of 4 and an output of 9. It receives blocks of 2 x 2 pixels from the object area and is aimed to automatically detect and locate object center and some specific features of the object. Examples of such features are center position of the object and some other geometric features, which appear at different location when the object pose changes. Therefore object translation parameter can also be achieved immediately from the position of object center.

For estimating the rotation parameter of these objects, we have used two approaches. For objects whose pose can be determined through locating pose specific geometric features, we use the result of the second map directly. We compute the rotation parameter by just computing the direction of the line connecting the object center position and the feature position. In case no pose specific feature exists or it is difficult to differentiate such feature position from other part of the object, we use a three layer feed-forward net trained with dynamic learning vector quantization (DLVQ) for each object class [9], which we will introduce shortly in section 3.3.

3.3. Dynamic LVQ for estimating object rotation parameter

The DLVQ is a new LVQ [10] algorithm, which in turn is related closely with the SOM. The idea of this algorithm is to find a natural grouping in a set of data. The algorithm presupposes that the vector $\boldsymbol{x}_i$ belonging to the same class are distributed normally with a mean vector $\boldsymbol{\mu}_i$. A feature vector $\boldsymbol{x}$ is assigned to the class Ω_i with the smallest Euclidean distance $||\boldsymbol{\mu}_i - \boldsymbol{x}||^2$. However DLVQ is different from SOM in that DLVQ net is a three layer feed-forward network trained with a supervised learning mechanism. Since the algorithm generates the hidden layer dynamically during the learning phase, it was called dynamic LVQ.

The feature vector input to the DLVQ net is gained as follows. We use the auto-associative net again on the 64x64 pixels image and get an 16x16 pixels image. After that a feature vector of 256 dimension is input to the DLVQ net. Since the difference of the feature vectors of an image under different pose is greater than that under the same pose, it is very reasonable to treat the problem of estimating object rotation parameter as a classification problem using DLVQ. In our experiment the sampling interval of the rotation parameter is 10^o. Therefore there are 36 different pose parameter to be classified. At the beginning the number of hidden layer neurons N is the same as the number of different rotation pose of the object (36). During training the number of hidden layer neurons rises until a satisfactory recognition rate is achieved. The output layer consists only one neuron whose value varies from 0 to N-1. Therefore after training the net can compute object rotation parameters through indicating the net output as one of the N different values.

4. EXPERIMENTAL RESULTS

In our experiment we want to demonstrate our method through locating five different industrial objects. Images are taken for the objects under different illumination, in different scale and different translation and rotation positions in the plane (See Figure 3). We show the effect of dimension reduction of the input image using the auto-associative network in Figure 4.

The result of using two SOMs on the dimension reduced image is as follows. After training of the first map, the map self-organizes itself and its output neurons map to different parts of the image (object area including edges, corners, etc. and the dark background). The reference vectors of three typical neurons representing object area are displayed in Figure 5 (a, b and c), where we can see how the pixel gray values change in each situation. All the 4 x 4 blocks of the image can then be labeled either as object area or as background according to the index of the output neuron. With those three features extracted from the output of the map, we can differentiate the three lids from the case and the fan using a simple three layer feed-forward network. Knowing the number of blocks belonging to object area, the scale information of each object class is then determined. Here 100% of recognition rate of the case and the fan is achieved using a set of new images different from the training set.

Subsequently the second map was trained in order to differentiate the three kinds of lids. Specific features of the object, which are in our case the center of lids and the hole(s) can be detected based on the output of the second map. The reference vector of the neuron in the second map representing the hole(s) is displayed in Figure 5d. According to the number of holes detected, we can differentiate the three kinds of lids. Here we achieve an average of 93.3% recognition rate on the three lids, which is an increase of 6.4% as compared with 86.9% using a three-layer feed-forward network trained by backpropagation.

Table 1 summarizes the recognition results with a disjoint testing set of 70 images/object for the five objects under different illumination and noise condition, after training the network with a training set of 50 images/object. For lids with one hole and two holes, we compute the rotation parameter directly based on the location of hole(s). For lids without hole, case and fan, we use DLVQ methods. Localization precision in translation is 3 pixels. The localization result in rotation is 5^o using DLVQ and 3.5^o with the SOM method.

Objects	# Total	# Correct	Percentage(%)
lid0	70	69	98.6
lid1	70	63	90.0
lid2	70	64	91.4
case	70	70	100
fan	70	70	100

Table 1. Recognition results on the test images

5. CONCLUSION

A new hybrid approach for 2D object localization which has invariance against translation, rotation and scale changes is presented. With a relative small set of training images, the proposed network based algorithm has the advantage of self-organization and can produce outputs indicating the class and pose information of objects. Future work will focus on object localization under varying background using the same approach and compare the localization precision with other approaches.

REFERENCES

[1] A. Ravichandran and B. Yegnanarayana. Studies on object recognition from degraded images using neural networks. *Neural Networks*, 8(3):481-488, 1995.

[2] A. Khotanzad. Recognition and pose estimation of unoccluded three-dimensional objects from a two-dimensional perspective view by banks of neural networks. *IEEE Trans. on Neural Networks*, 7(4):897-905, July 1996.

[3] Teuvo Kohonen. Self-organization and Associative Memory. Spring Verlag, Berlin-Heidelberg-New York-Tokio, 1989.

[4] Teuvo Kohonen. The self-organizing map. In *Proceedings of the IEEE*, 78(9):1464-1478, September 1990.

[5] S. M. Bhandarkar and J. Koh and Minsoo Suk. Multiscale image segmentation using a hierarchical self-organizing map. *Neurocomputing*, 14(3):241-272, March 1997.

[6] H. M. Lakany and E.-G. Schukat-Talamazzini and H. Niemann and M. Ghoneimy. Object Recognition from 2D images using Kohonen Self-Organised Feature Maps. In *Pattern Recognition and Image Analysis*, 7(3):301-308, March 1997.

[7] S. Albeverio and N. Kruger and B. Tirozzi. An extended Kohonen phonetic map. In *Mathematical and Computer Modelling*, 25(2):69-73, February 1997.

[8] Hao Bi and Guangguo Bi and Yimin Mao. Globally Optimal Vector Quantizer Design Using Stochastically Competitive Learning Algorithm. In *Proc. Int. Symp. on Speech, Image Processing and Neural Networks*, 650-653, Hong Kong, 1994.

[9] Andreas Zell. Simulation Neuronaler Netze. Addison-Wesley, Bonn-Paris-New York-Tokyo-Singapore, 1994.

[10] T. Kohonen and J. Hynninen and J. Kangas and J. Laaksonen and K. Torkkola. Lvq_pak: the learning vector quantization program package. Technical report A30, Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland, 1996.

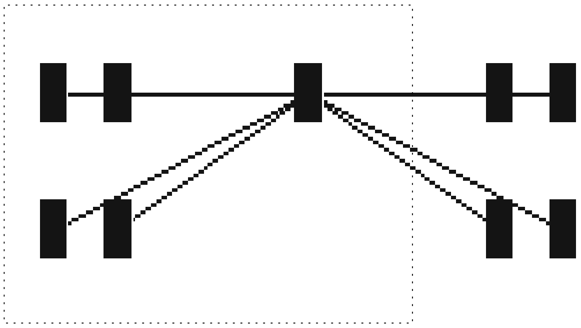
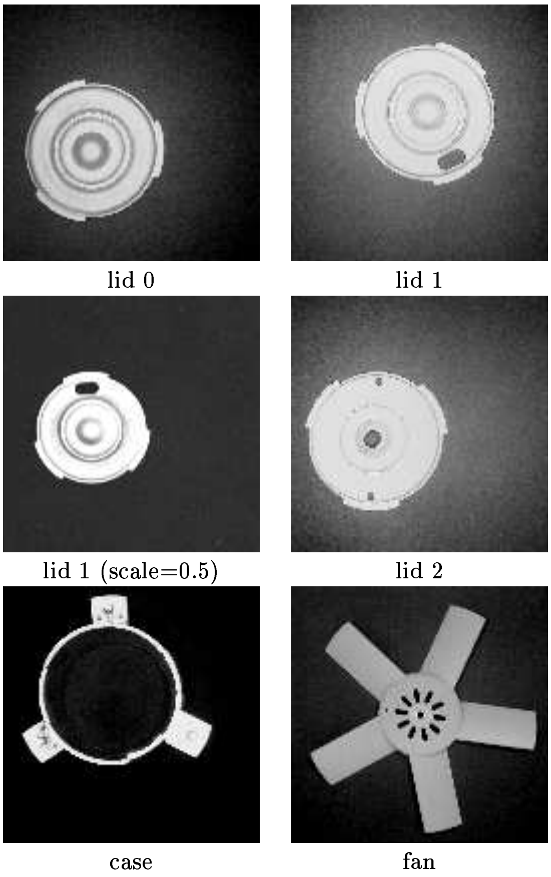


Figure 2. Auto-associative network for image dimension reduction



lid 0

lid 1

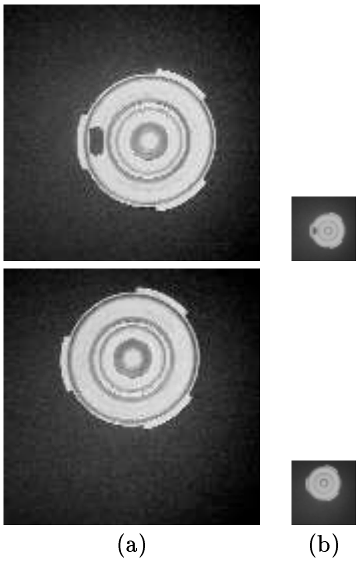
lid 1 (scale=0.5)

lid 2

case

fan

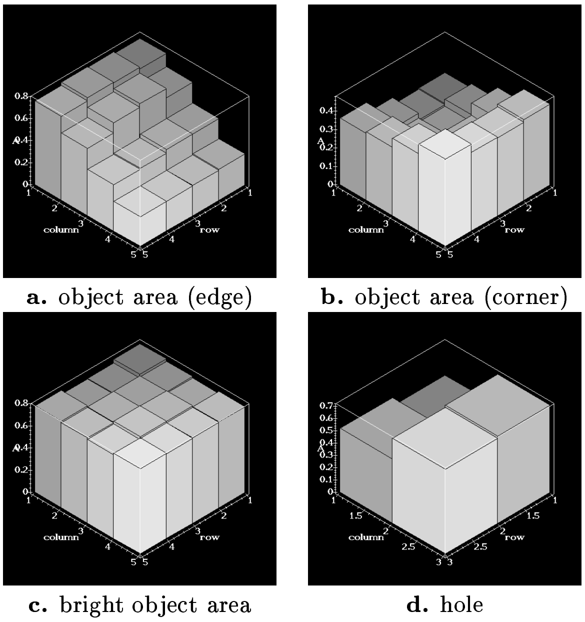
Figure 3. Objects used for experiments.



(a)

(b)

Figure 4. Image scale reduction using the trained auto-associative network. (a) is the original image with size 256 x 256 pixels, (b) is the image scaled down to size 64 x 64 pixels



a. object area (edge)

b. object area (corner)

c. bright object area

d. hole

Figure 5. Reference vectors of the neurons representing object area and hole(s). x direction: column, y direction: row, z direction: pixel gray value