ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 61-66

Object Recognition with Banana Wavelets*

Norbert Krüger, Gabriele Peters

Ruhr-Universität Bochum,
Institut für Neuroinformatik, Germany

Abstract. We introduce an object recognition system, based on generalized Gabor wavelets, called *banana wavelets*. In addition to the qualities frequency and orientation, banana wavelets have the attributes curvature and size. Banana wavelets can be metrically organized, a *sparse* and *efficient* representation of objects is *learned* utilizing this metric.

1. Introduction

In this paper we describe a novel object recognition system in which representations of object classes can be *learned autonomously*. The learned representations allow a fast and effective location and identification of objects in complicated scenes. Our object recognition system is based on three pillars. Firstly, our preprocessing is based on the idea of *sparse coding* [1]. Secondly, effective learning is guided by *a priori* constraints covering fundamental structure of the visual world. Thirdly, we use Elastic Graph Matching (EGM) [6] for the location and identification of objects.

A sparse representation can be defined as a coding of an object by a *small number* of *binary* features taken from a *large feature space*. Sparse coding has biologically motivated advantages like minimizing wiring length and conceptual advantages like increase of associative memory capacity and redundancy reduction (discussed exhaustively in [1]). As an additional advantage in our case sparse coding leads to a siginificant speed–up.

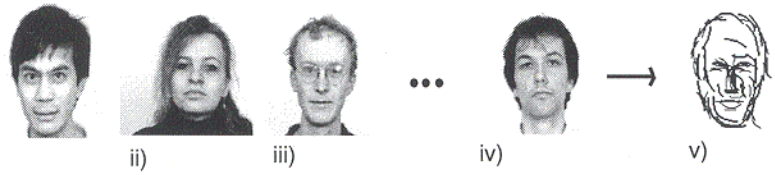

Figure 1: i–iv) Different examples of faces. v) Learned representations.

Our representation of a certain view of an object class comprises only *important features*, learned from different examples (see figure 1, 5 and 6). Learning

*Supported by grants from the German Ministry for Science and Technology 01IN504E9 (NEUROS) and 01M3021A4 (Electronic Eye).

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 61-66

is inherently faced with the bias/variance dilemma [2]: If the starting configuration of the system is very general, it can learn from and specialize to a wide variety of domains, but it will in general have to pay for this advantage by having many internal degrees of freedom —the "variance" problem. On the other hand, if the initial system has few degrees of freedom it may be able to learn efficiently but there is great danger that the structural domain spanned by those degrees of freedom does not cover the given domain of application at all —the "bias" problem. We propose that *a priori* knowledge is needed to cope with the bias–variance dilemma. We have formulated a number of *a priori* principles to reduce the dimension of the search space and to guide learning, i.e., to handle the variance–problem. We expect to avoid the bias–problem because of the general applicability of those principles. In [4] we have already made use of the principles Invariance Maximization (P1) and Redundancy Reduction (P2). Here we introduce an important additional constraint **P3**: Significant features of a local area of the two–dimensional projection of the visual world *are localized curved lines.*

We formalize P3 by extending the concept of Gabor wavelets to banana wavelets (section 2.). To the parameters frequency and orientation we add curvature and size (see figure 2). The space of banana wavelet reponses is much larger compared to the space of Gabor wavelet responses used in [6]. An object can be represented as a configuration of a few of these features (figure 1v), therefore it can be coded sparsely. The space of banana wavelet responses can be understood as a metric space, its metric representing the similarity of features. This metric is essential for our learning algorithm (section 3.). The banana wavelet responses can be derived from Gabor wavelet responses by hierarchical processing to gain speed and reduce memory requirements. The sparse representation combined with our hierarchical feature processing allows a fast and effective locating (section 4.) using EGM.

Our system has certain analogies to the visual system of vertebrates. There is evidence for curvature sensitive features processed in a hierarchical manner in early stages [3]; sparse coding is discussed as a coding scheme used in the visual system [1]; and metric organization of features seems to play an important role for information processing in the brain [3]. We aim to apply these concepts in our artificial object recognition system.

2. The Banana Space

The principle P3 gives us a significant reduction of the search space. Instead of allowing, e.g., all linear filters as possible features, we restrict ourself to a small subset. Considering the risk of a wrong feature selection it is necessary to give good reasons for our decision. We argue that nearly any 2D–view of an object can be composed of localized curved lines. Furthermore, the fact that humans can easiliy handle line drawings of objects strengthens our assumption P3.

Banana Wavelets: A banana wavelet $B^{\vec{b}}$ is a complex–valued function, pa-

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 61-66

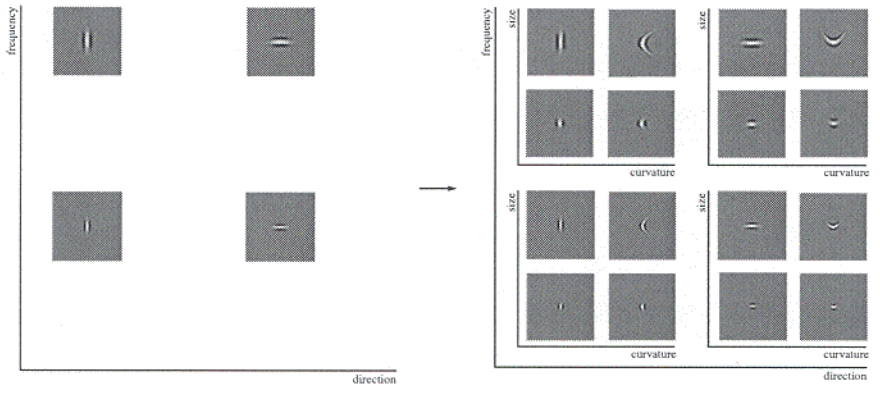

Figure 2: Relation between Gabor wavelets and banana wavelets.

rameterized by a vector $\vec{b}$ of four variables $\vec{b} = (f, \alpha, c, s)$ expressing the attributes frequency (f), orientation (α), curvature (c) and size (s). It can be understood as a product of a curved and rotated complex wave function $F^{\vec{b}}$ and a stretched two–dimensional Gaussian $G^{\vec{b}}$ bent and rotated according to $F^{\vec{b}}$ (see figure 3). Our basic feature is the magnitude of the filter response of a

Figure 3: A banana wavelet is the product of a curved Gaussian $G^{\vec{b}}(x, y)$ and a curved wave function $F^{\vec{b}}(x, y)$ (only the real part of the kernel is shown).

banana wavelet extracted by a convolution with an image. A banana wavelet $B^{\vec{b}}$ causes a strong response at pixel position $\vec{x}$ when the local structure of the image at that pixel position is similar to $B^{\vec{b}}$ (see [5]).

The Banana Space: The six–dimensional space of vectors $\vec{c} = (\vec{x}, \vec{b})$ is called the *banana (coordinate) space* with $\vec{c}$ representing the banana wavelet $B^{\vec{b}}$ with its center at pixel position $\vec{x}$ in an image. In [5] we define a metric $d(\vec{c}_1, \vec{c}_2)$. Two coordinates $\vec{c}_1, \vec{c}_2$ are expected to have a small distance d when their corresponding kernels are similar, i.e., they represent similar features.

Approximation of Banana Wavelets by Gabor Wavelets: The banana response space contains a huge amount of features, their generation requires large cpu–time and memory capacities. In [5] we define an algorithm to derive banana wavelets from Gabor wavelets and banana wavelet responses from Gabor wavelet responses. By this hierarchical processing we achieve a speed up of a factor 15 and a reduction of memory requests by a factor 20. Figure 4 gives the idea of the approximation algorithm.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 61-66

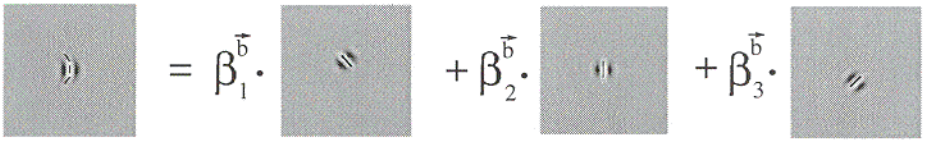

Figure 4: The banana wavelet on the left is approximated by the weighted sum of Gabor wavelets on the right.

3. Learning

Extracting Significant Features Per Instance: Our aim is to extract the local structure in an image I in terms of curved lines expressed by banana wavelets. We define a *significant feature per instance* of an object by two qualities. Firstly it has to cause a strong response (**C1**), secondly it has to represent a local maximum in the banana space (**C2**). Figure 5bi)–iv) show the significant features per instance for a set of cans (each banana wavelet is described by a curve with same orientation, curvature and size). In terms of analogy to the processing in area V1 in the vertebrate visual system C1 may be interpreted as the response of a certain column which indicates the general presence of a feature, whereas C2 represents the intercolumnar competition giving a more specific coding of this feature [3].
Clustering: After extracting the significant features per instance in different pictures we apply an algorithm to extract invariant local features for a *class of objects*. Here the task is the selection of the *relevant features* for the object class from the noisy features extracted from our training examples (see figure 5bi)–iv)) We assume the correspondence problem to be solved, i.e., we assume the position of certain landmarks of an object to be known on pictures of different examples of these objects. In some of our simulations we determined corresponding landmarks manually, for the rest we replaced this manual intervention by motor controlled feedback (see section 5.). In a nutshell the learning algorithm works as follows: For each landmark we divide the significant features per instance of all training examples into clusters. Features which are close according to our metric d are collected in the same cluster (P2: Reduction of redundancy). A significant feature for an object class is defined as a representative of a *large* cluster. That means this or a similar feature (according to our metric d) occurs often in our training set, i.e., has a high invariance (P1). We end up with a graph with its nodes labeled with banana wavelets representing the learned significant features (see figure 5bv) and [5]).

4. Matching

To use our learned representation for location and classification of objects we define a similarity function between a graph labeled with the learned banana wavelets and a certain position in the image. A *total similarity* simply averages *local similarities*. The local similarity expresses the system's confidence whether

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 61-66

a pixel in the image represents a certain landmark. The graph is adapted in position and scale by optimizing the total similarity. The graph with the highest similarity determines the size and position of the objects within the image.

In a nutshell the local similarity is defined as follows (for details see [5]): For each learned feature and pixel position in the image we simply check whether the corresponding banana response is high or low, i.e., the corresponding feature is present or absent. Because of the sparseness of our representation only *a few* of these checks have to be made, therefore the matching is very *fast*. Because we make use only of the *important* features, the matching is very *efficient*.

5. Simulations and Conclusion

Learning of Representation: Firstly we apply the learning algorithm to data consisting of manually provided landmarks. Our training sets consist of a set of approximately 60 examples of an object viewed in a certain pose. As objects we used cans and faces. Corresponding landmarks are defined manually on the different representatives of a class of objects (figure 5a). Figure 5 and 1 show some of the learned representations.

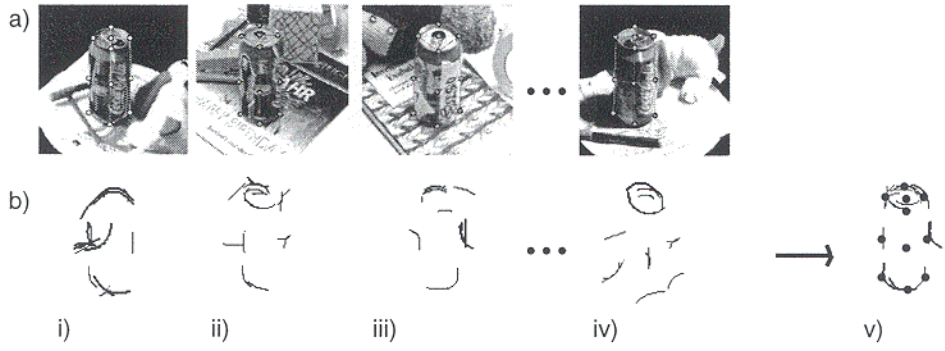

Figure 5: a: Pictures for training. bi)–iv): Significant features per instance describing beside relevant information also accidential features like background, shadow or surface textures. c: the learned Representation.

To avoid the manual generation of ground truth we made use of motor controlled feedback. Our aim is the construction of training data in which a certain object is shown under changing background and illumination but without changing of the position of the landmarks. Then we can simply apply our learning algorithm to this data using a rectangular grid. For the learning of a representation for cans we put a can on a rotating plate and changed background and lighting conditions in a sequence of pictures (see figure 6). For the generation of ground truth for frontal faces we recorded a sequence of pictures in which a person is sitting fixed on a chair. Illumination and background is changed as for cans. To extract representations for different scales we simply apply the learning algorithm to the very same pictures of the different sequences scaled accordingly (see figure 7).

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 61-66

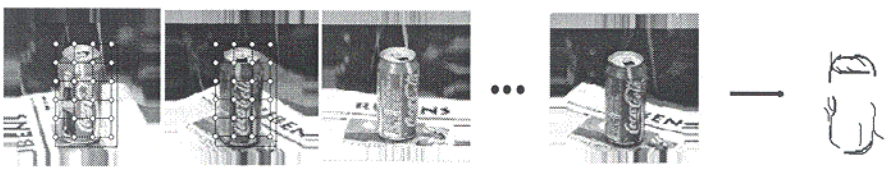

Figure 6: Automatical generation of ground truth for cans.

Matching: For the problem of face finding in complex scenes with large size variation a significant improvement in terms of performance and speed compared to the older system [6, 4] (which is based on Gabor wavelets) could be achieved. (Figure 7) shows some examples of matches and mismatches. The object finding in one picture approximately requires 1.5 seconds on a Sparc Ultra. In [5] we also performed successfully matching with cans and other objects, as well as various discrimination tasks.
Conclusion: We showed that our object recognition system is able to learn autonomously an efficient representation from noisy data applicable to a wide range of problems.

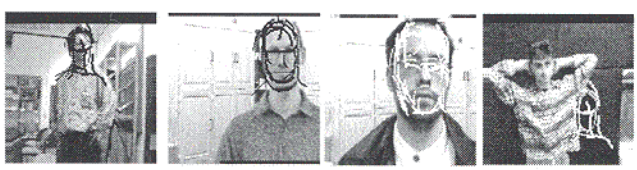

Figure 7: Face finding with autonomously learned representations for three scales. The mismatch (right) is caused by the person's unusual arm position.

References

[1] D. Field, "What is the Goal of Sensory Coding?," *Neural Computation*, vol. 6, no. 4, pp. 561–601, 1994.

[2] S. Geman and R. Doursat, "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, vol. 4, pp. 1–58, 1995.

[3] M.W. Oram and David I. Perrett, "Modeling Visual Recognition from Neurobiological Constraints," *Neural Networks*, vol. 7, pp. 945–972, 1994.

[4] N. Krüger, M. Pötzsch, T. Maurer, M. Rinne, "Estimation of Face Position and Pose with Labeled Graphs," BMVC96, pp:735–743.

[5] N. Krüger, G. Peters, C. v.d. Malsburg, "Object Recognition with an Autonomously Learned Representation", Technical Report, 1996.

[6] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R.P. Würtz, W. Konen, "Distortion Invariant Object Recognition in the Dynamik Link Architecture," *IEEE Transactions on Computers*, vol. 42, no. 3, pp. 300–311, 1992.