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. 55-60

Size Invariance By Dynamic Scaling in Neural Vision Systems

Götz Meierfrankenfeld and Klaus Kopecz

Department of Neurophysics, University of Marburg
Renthof 7, 35032 Marburg, Germany

Abstract. Vision systems in complex environments are faced with the problem of analyzing visual information on multiple scales to support segmentation and size invariant object recognition. We propose a system which detects relevant spatial scales in images and constructs an explicit size invariant representation. Further, it provides a means of navigating through scale space in the presence of ambiguous scale information, thus adds a behaving component to image analysis. The architecture is based on biologically realistic neural networks like neural fields and neurons with bandpass receptive field characteristics.

1. Introduction

The human nervous system shows an intriguing generalization performance with respect to size in object recognition tasks. Size invariant recognition of objects does not require to view the object in several sizes during training. One view of a certain size may be enough to recognize the same object presented at a different distance. Thus, scale invariant recognition cannot be accounted for by a "view based" approach proposed, e.g., for 3D invariant recognition [2]. It is more plausible to think in terms of explicit size invariant representations (ESIR) which are constructed by the nervous system. An ideal ESIR is invariant to spatial scaling of an arbitrary object and thus provides adequate input to a subsequent recognition stage. To build the ESIR, images must be analyzed to extract information about prominent spatial scales. Whenever, this information is available, it might be used for rescaling the image. How rescaling can be done in a neurally plausible way has been described in [5]. There, attentional dynamics has been introduced, which modulates neural weights in a way that they provide the required scaling transformation ("dynamic routing").

In general the problem of selecting a scale of interest cannot be seen in isolation. Without having separated an image into object candidates and background, scale information cannot be object specific. On the other hand, the outcome of a figure/ground segregation depends in general on the scale of observation. However, the selection of scales of interest is one essential ingredient in a complete segmentation system. This implies that whenever images contain

Supported by grant no. 01M3013D from the German Ministry of Science (BMBF).

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. 55-60

more than one prominent scale, mechanisms must be introduced to uniquely select one spatial scale and to sequentially visit structures in scale space.

We address two problems which we solve in a neurally plausible way:
(1) How to define and extract prominent scales within grey-value images?
(2) How to organize an attentional network dynamics which can be used to select prominent scales for steering a desired scaling transformation?

2. Detecting Scales of Prominent Structures

An intuitive way of analyzing spatial scales of a distribution is to consider the power spectrum. A simple example shows that this does not provide complete perceptually relevant information in a way which can be used by a neurally plausible (and technical) system. Consider the one-dimensional gaussian wave packet $g(x) = \cos(\omega_0 x)\ \exp(-x^2/\sigma^2)$. The power spectrum consists of a peak centered around $\omega = \omega_0$ with a width determined by $1/\sigma$. Thus it is relatively easy to detect the characteristic scale ω_0 by picking the local maximum, but it is much more difficult to detect the second characteristic scale σ related to the envelope of the wave packet. This information is "hidden" in the width of the power spectrum peak. Further problems appear if the image information gives rise to higher harmonics in the spectrum, which cannot easily be recognized as such and thus cannot be attributed to size information.

The solution we propose consists of a two-stage nonlinear transformation. The transformation stages are indicated in Fig. 1 by the labels "I" and "II". In the first step we define bandpass filters (receptive fields, RF), $h_1(x - x', \omega)$, with center frequencies ω, which are centered at the image location $x = x'$. In the following, we parameterize these filters by the scale parameter $\sigma \sim 1/\omega$, thus $h_1 = h_1(x - x', \sigma)$. We then perform the nonlinear transformation

$$F(x, \sigma) = \left| \int_I f(x')\, h_1(x - x', \sigma)\, d^2x' \right| \tag{1}$$

where $f(x)$ denotes the image and I the image domain. The result is an edge-like representation at various scales σ (cf. Fig. 1). As bandpass filters we use Laplacians of Gaussians, with four different center frequencies separated by octaves and with mutual overlapping bandwidths. Thus, the outcome of the first processing stage are distributions $F(x, \sigma_i)$ with $i = 1, \ldots, 4$.

In a second step we detect prominent scales within each $F(x, \sigma_i)$ by applying a family of RFs, $h_2(x, s)$, the cross-section of which is illustrated in Fig. 1, part II. Through the inhibitory side-bands, we found these RFs to be very specific when applied to the edge-like representations $F(x, \sigma_i)$. All the RFs $h_2(x, s)$ are only located around the central image location, so that we determine prominent spatial structures with respect to a specific viewpoint. As the final outcome $f_s(s)$ of the two-stage process, we define:

$$f_s(s) = \int_I h_2(x, s) \sum_i F(x, \sigma_i)\, d^2x \tag{2}$$

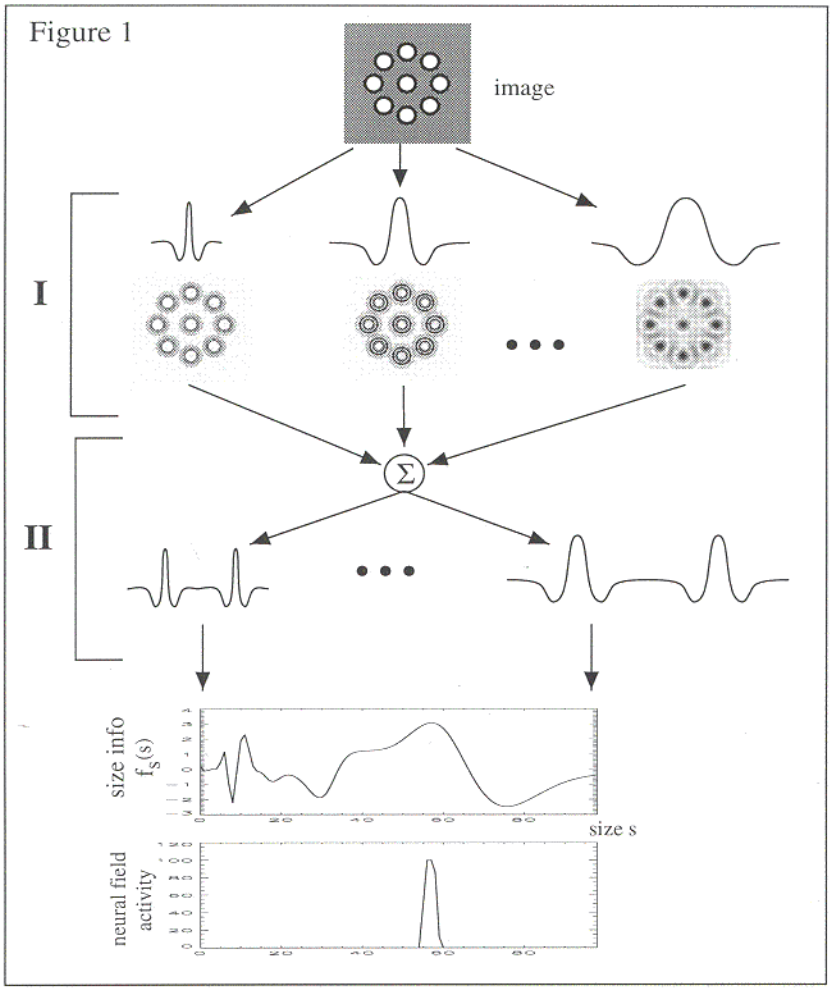

Figure 1

$f_s(s)$ now represents the evidence for a prominent spatial structure at the scale s. This is demonstrated in Figure 1, where the processing of a sample texture is shown. As obvious by the two dominating maxima of $f_s(s)$, the system detects two prominent scales, one related to the texture elements, the other to their collection, which is similar to the envelope in the wave packet example discussed above. The sample texture has by construction the particular property of having no low frequency components, because each local texture element has a vanishing mean intensity [3]. Hence, other approaches to size detection in images like the one given in [5] will hardly detect any structure.

3. Scale Attention and Dynamic Scaling

Having determined the scale information $f_s(s)$, local maxima must be selected and used for rescaling the original image. The process of selecting a specific scale of interest can be viewed as attending to a specific size. Due to the expected noisy structure of the scale information, local maximum selection should be conditioned in the sense that not every small peak is attended to and that the selection is stable with equally competing candidates. A dynamic neural network which accomplish this task reliably is the *neural field* [1]. With respect to target selection it has been applied and discussed in detail in [4]. In brief, the dynamics of the neural field $u(s,t)$ is defined by

$$\dot{u}(s,t) = -u(s,t) \; - \; h \; + \int w(s-s')\,T[u(s',t)]\,ds' \; + \; f_s(s) \tag{3}$$

$T(\cdot)$ is a positive threshold function. The distribution $T(s) = T[u(s)]$ will be referred to as *activity*. h is a parameter, which can be viewed as the threshold of T. If the connectivity $w(s-s')$ is chosen to be composed of local excitation and global inhibition, the dynamics exhibit a stationary solution composed of one unique localized cluster of activity which will be located at one of the local maxima of the input scale information $f_s(s)$ (see Fig. 1, bottom). Importantly, this local maximum has to exceed a certain amplitude and width criterion to be selected. The activity cluster can now be used to scale image information appropriately, to construct the size invariant representation.

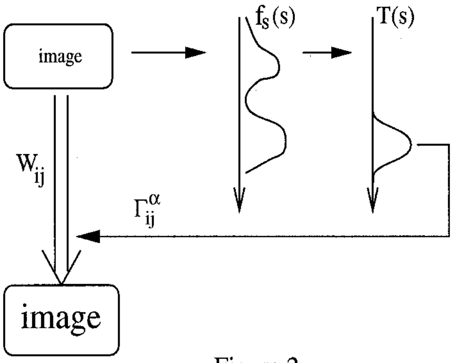

Figure 2

In the framework of neural connectivities, information can be scaled by "dynamic routing" (see [5] for details). We briefly describe a modified version here. Fig. 2 illustrates the scaling mechanism. Let W_{ij} the weights connecting image location x_j with the location y_i on the size invariant representation. Assume that the image contains a structure of a certain size s which defines the scaling factor $\alpha = s/s_0$, where s_0 is the desired size of the structure on the invariant representation. The scaling can then be done by setting:

$$W_{ij} \sim \exp\left[-\frac{(\alpha\, i \, - \, j)^2}{2\alpha^2}\right] \tag{4}$$

This form fulfills the need for low-pass filtering the image (or target) when shrinking (or magnifying) the image to avoid aliasing. To extract the desired

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. 55-60

scaling which is coded by the location of the activity cluster on the neural field, we set

$$W_{ij} \sim \int \Gamma_{ij}^{\alpha}\, T(\alpha\, s_0)\, d\alpha \quad \text{with:} \quad \Gamma_{ij}^{\alpha} = \exp\left[-\frac{(\alpha i - j)^2}{2\alpha^2}\right] \tag{5}$$

which determines the effective scaling by an average of all α's weighted by the activity $T(s)$.

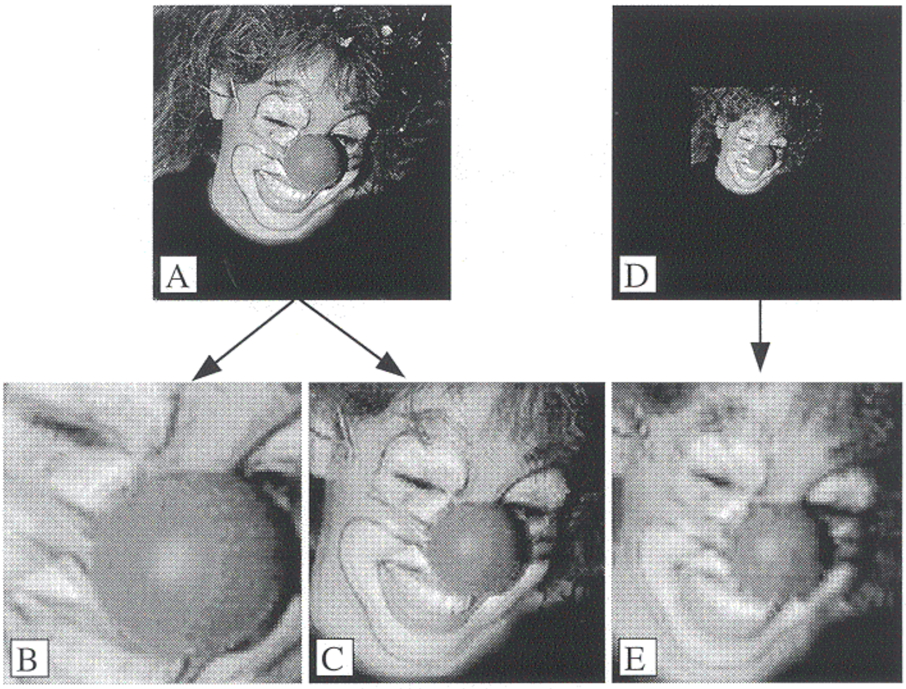

Figure 3

Figure 3 shows scaling and invariance properties of our system, by applying it to an image of a face viewed from two distances (panels A and D). Our system detects two prominent scales in the image. One is related to the nose/eye region, the other to the entire face. The images scaled according to these two detected sizes are depicted in panels B and C, respectively. Panel D shows the original image at a reduced size. Here, again the same two prominent structures are detected. The scaling onto the face region is depicted in Panel E which is identical to panel C (except for the information loss due to the size reduction). In this sense the projection of the face region is invariant against scalings of the original image.

4. Discussion and Conclusion

We presented a neurally motivated system which is able to detect perceptually relevant size information in grey-value images. Further, we showed how this scale information can be used the rescale the image to construct size invariant representations. In general, scale information will be ambiguous. Neural field dynamics is introduced which establish *attention for scale*, i.e. which select prominent maxima of ambiguous scale information.

In the case of having non-unique scale information, the problem appears how to deal with this ambiguity. The solution we propose is to unfold the ambiguity in time, thus to sequentially visit prominent structures in scale space. For constructing panels B and C in Fig. 3, this has been accomplished by introducing an additional slow inhibitory degree of freedom, which destablizes a selection after some time, so that the system can switch to a different scale. This adds a behaving component to image analysis. The principle of *global precedence* [3] may be used for further constraining the dynamics of scale selection. According to this principle, scenes are in general analyzed from global to local scales, which could be considered in our model by assigning larger weights to global scale information.

From the viewpoint of image segmentation one can consider the here presented transformation as one processing step within an iterative segmentation loop. Based on initially scaled information, segmentation can proceed, e.g., based on contour information, which in turn would provide input by feedback into the scaling system to refine the initial guess of a relevant scale. Similarly, the initial scaling transformation might affect the locus of spatial attention (the gaze direction of the camera or eye), which in turn can modify scale information. As a future line of research, this integrative behaving scenario is expected to be a step towards the understanding of human performance in image analysis.

References

[1] S. Amari. Dynamics of pattern formation in lateral–inhibition type neural fields. *Biol. Cybern.*, 27:77–87, 1977.

[2] S Edelman and D Weinshall. A self-organizing multiple view representation of 3d objects. *Biol. Cybern.*, 64:209–219, 1991.

[3] HC Hughes, G Nozawa, and F Kitterle. Global precedence, spatial frequency channels and the statistics of natural images. *J. Cog. Neurosci.*, 8:197–230, 1996.

[4] K. Kopecz and G. Schöner. Saccadic motor planning by integrating visual information and pre-information on neural, dynamic fields. *Biol. Cybern.*, 73:49–60, 1995.

[5] BA Olshausen, CH Anderson, and DC van Essen. A multiscale dynamic routing circuit for forming size- and position invariant object representations. *J. Computational Neurosci.*, 2:45–62, 1995.