To appear in J. L. van Hemmen and T. J. Sejnowski (Eds.),
Problems in Systems Neuroscience, Oxford University Press (2004).
Submitted and accepted for publication January 2001.
Also: Cognitive Sciences EPrint Archive (CogPrints) 3321, (2003),
http://cogprints.ecs.soton.ac.uk/archive/00003321/

How Does Our Visual System Achieve Shift and Size Invariance?

Laurenz Wiskott

Innovationskolleg Theoretische Biologie
Humboldt-Universität zu Berlin
Invalidenstraße 43, D-10115 Berlin
http://itb.biologie.hu-berlin.de/
wiskott@itb.biologie.hu-berlin.de

Abstract

The question of shift and size invariance in the primate visual system is discussed. After a short review of the relevant neurobiology and psychophysics, a more detailed analysis of computational models is given. The two main types of networks considered are the dynamic routing circuit model and invariant feature networks, such as the neocognitron. Some specific open questions in context of these models are raised and possible solutions discussed.

1 Introduction

The ease with which we recognize common objects from different distances and perspectives and under different illuminations gives the impression that invariant object recognition is a trivial task. However, an apparently small change in the stimulus can cause a dramatic change in the retinal activity pattern. Assume, for example, you are looking at a zebra and change your gaze by just one width of the zebra stripes. Many responses of retinal sensors will be inverted (change from low to high or vice versa) causing a dramatic change of the neural activity pattern, but you still perceive the same zebra. Thus, in order to achieve invariant recognition, our visual system has to be insensitive to these kinds of dramatic changes in the visual input while being still sensitive to more subtle changes that are perceptually relevant, such as when the zebra turns its head.

The question presented and discussed here is confined to the two apparently simplest geometrical invariances, namely shift and size invariance. The discussion is mainly based on computational models of the visual system (not including the large body of literature on application-oriented systems). This simplifies the discussion in some respects, but it also means that many of the experimental details that have not been modeled yet are left aside. One should also keep in mind that even if a computational model is consistent with all known experimental data, it does not mean that it reveals the actual mechanisms used by the biological system. Despite these and other limitations of the discussion I hope it may nevertheless be useful in addressing the question: How does our visual system achieve shift and size invariance?

The text is structured as follows. Section 2 briefly summarizes psychophysical and neurophysiological evidence for shift and size invariance and its limitations. Section 3 provides a list of constraints from neuroanatomy and -physiology that need to be taken into account when developing biologically plausible models. From a computational point of view there are two basic approaches by which shift and size invariance can be achieved: normalizing the image or extracting invariant features. These two approaches are briefly discussed in Section 4. Sections 5 and 6 present models following these two computational approaches and discuss to which extent they can account for different aspects of the visual system. Open questions in modeling shift and size invariance in the visual system are then discussed in Section 7. Section 8 gives a short account of some alternative models not discussed here. A conclusion is given in Section 9.

2 Evidence for Shift and Size Invariance in the Visual System

Evidence for shift and size invariance comes from two sides: psychophysics and neurophysiology. Psychophysical experiments with human subjects indicate that recognition of common visual objects is, within the range tested, independent of location and size. For example, in priming experiments Biederman and Cooper (1991, 1992) tested subjects on line drawings of ordinary objects. They found that naming reaction times did not change significantly if the stimuli (4° in diameter) were shifted by 4.8° visual angle. They also found no significant change if stimuli were changed in size from 3.5° to 6.2° in diameter or vice versa. In both experiments stimuli were presented in or near the center of the visual field. In a perceptual learning task with gray-scale images of common objects Furmanski and Engel (2000) found no significant performance difference if size changed unexpectedly from 16.4°×12.7° to 8.2°×6.4° or vice versa.

Supporting evidence for shift and size invariance also comes from neurophysiological experiments. Tovée et al. (1994), for instance, recorded from face sensitive neurons in the inferotemporal cortex and the cortex in the banks of the anterior part of the superior temporal sulcus of macaque monkeys. They found the firing rates of the neurons to be largely invariant to a change of fixation point up to the edge of the presented faces and only a small reduction up to 4° beyond the edge. Ito et al. (1995) measured the receptive field properties of neurons in the anterior part of the inferotemporal cortex. They found large receptive fields up to about 50° in diameter for critical features of much smaller size, which indicates large shift invariance. 57% of the neurons tested had a stable response within stimulus size ranges larger than 2 octaves. See (Oram and Perrett, 1994) for a more detailed overview.

However, psychophysical results also point to some limitations of shift invariance. Nazir and O'Regan (1990), for instance, showed that recognition of random dot patterns of size 0.97°×0.86° degraded significantly if patterns were trained 2.4° to one side of the visual field and then tested in the center or 2.4° to the opposite side of the visual field. Dill and Fahle (1998) found similar results in a same-different task for random dot patterns of size 0.5°×0.5° shifted up to 2°. The degradation with shift even held when subjects knew the new location beforehand. One possible explanation for these apparently contradicting results might be that our visual system is more invariant for familiar and meaningful objects than for unfamiliar and abstract patterns, like the dot patterns. However, there are also a number of other differences between the experiments pro and contra shift invariance reported here, such as the size of the patterns or the type of task.

3 Neurobiological Constraints

When developing computational models for shift and size invariant recognition in the visual system, one has to account not only for the invariance properties but also for a number of other aspects of the visual system; see Figure 1 and (Oram and Perrett, 1994).

Two pathways: It is generally accepted that the visual system can be divided into two pathways dedicated to different types of processing (Ungerleider and Mishkin, 1982), at least from MT and V4 upwards (see Merigan and Maunsell, 1993, for a review). The ventral pathway, also called form- or *what*-pathway, is dedicated to form perception and object recognition. The dorsal pathway, also called motion- or *where*-pathway, processes motion and other spatial information. Even if the dissociation of *what*- and *where*-processing in the two pathways may not be as strict as generally thought (Merigan and Maunsell, 1993), it seems to be clear that computational models have to provide mechanisms for processing these two aspects of visual information in an explicitly accessible fashion.

Layered structure: The visual system has a rich hierarchy of different areas. The ventral pathway, for instance, processes visual information in the following sequence of areas: Retina → lateral geniculate nucleus (LGN) → visual area 1 (V1) → visual area 2 (V2) → visual area 4 (V4) → posterior inferotemporal area (PIT) → central inferotemporal area (CIT) → anterior inferotemporal area (AIT) (→

anterior superior temporal polysensory area (STPa)), with a clear hierarchy given by the connectivity (for reviews see Felleman and Van Essen, 1991; Merigan and Maunsell, 1993; Oram and Perrett, 1994). Computational models should reflect this layered structure.

Feedback connections: The majority of connections in the visual system are reciprocal, i.e. feedforward connections are mirrored by corresponding feedback connections (Felleman and Van Essen, 1991). The role of these feedback connections is fairly unclear. However, a model of the visual system should at least offer a hypothesis as to what the feedback connections are good for.

Feature hierarchy: Single unit recordings from different areas of the ventral stream indicate that the complexity of the critical features causing a neuron to fire gradually increases from bottom to top (Kobatake and Tanaka, 1994; Oram and Perrett, 1994).

Invariance hierarchy: Single unit recordings also show that the receptive field sizes gradually increase from bottom to top and with it the amount of shift invariance (Kobatake and Tanaka, 1994; Oram and Perrett, 1994).

Fast recognition: Measurements of response latencies (see Nowak and Bullier, 1997, for a review) as well as EEG recordings in psychophysical experiments (Thorpe et al., 1996) show that the visual system performs object recognition very rapidly within about 150 ms.

Attention: It is known from psychophysical as well as neurophysiological studies that visual processing can be modified by attention in various ways (see Desimone and Duncan, 1995, for a review). A computational model should offer mechanisms by which attentional selection or biases can be imposed on the processing based on cues such as location, features, or novelty.

Learning: It is infeasible to assume that a complex hierarchical network for invariant object recognition could be genetically predetermined in detail. It is much more likely that the visual system develops through self-organization and unsupervised learning mechanisms from a relatively simple basic structure. A computational model of the visual system thus has to offer ideas about these mechanisms.

4 Two Computational Approaches

From a computational point of view there are two basic approaches known by which invariances can be achieved, firstly by normalization and secondly by extracting invariant features. Some principal pros and cons of these approaches are summarized in this section. A discussion as to how consistent these approaches are with what we know about the visual system is given in the succeeding sections in the context of existing neural models.

4.1 Normalization

In this approach the image of an object in the visual field is normalized to a standard position and size by an internal transformation. Invariant recognition can then be based on this normalized view. This approach has the following principal advantages (+) and disadvantage (–).

+ ***Where*-information is made explicit.** Since an explicit normalization is applied, the information about size and location of the object under consideration is available to the system at any time.

– **Recognition requires normalization.** A normalization requires a rough segmentation to determine which part of the visual field contains the object of interest. This is typically a difficult task in natural environments. Thus a sophisticated mechanism of iterating crude recognition, segmentation, normalization, and verification is required to find the correct normalization and recognize the object with certainty. This costs valuable time.

+ **Minimal information loss.** Since shifting and rescaling of a portion of an image is a simple operation, the normalization can be achieved with minimal information loss and great generality. Thus also unfamiliar and unnatural stimuli can be easily represented in an invariant way.

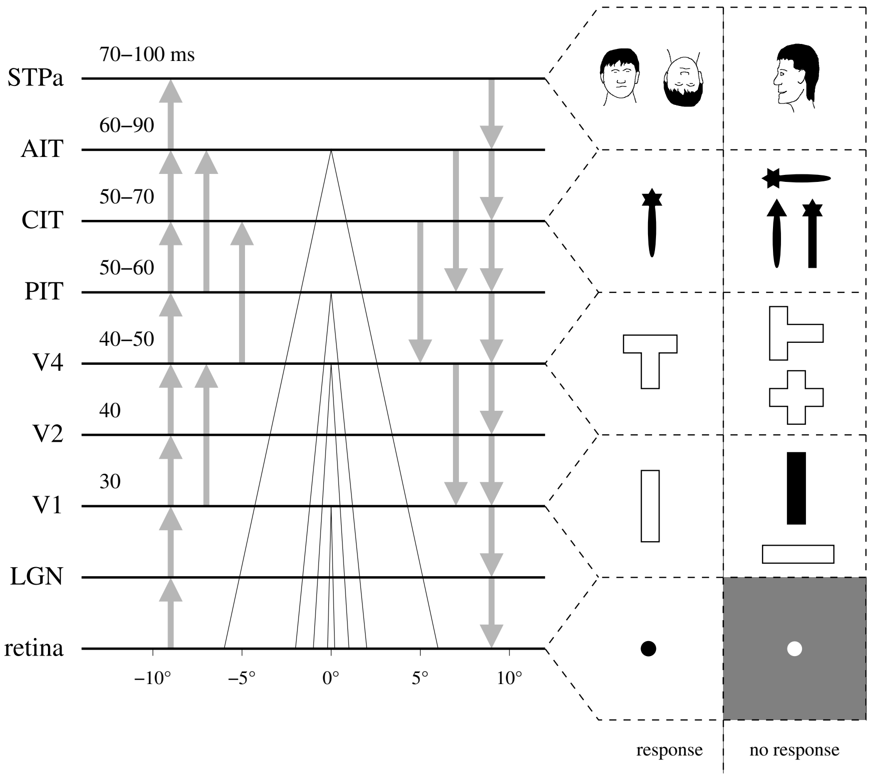

Figure 1: Basic properties of the ventral pathway of the visual system (adapted from Oram and Perrett, 1994). The pathway has a layered structure (from retina to STPa) with layer to layer connections (gray upward arrows). Forward connections are usually mirrored by feedback connections (gray downward arrows). Neurons in different layers respond to features of increasing complexity from bottom to top (sample stimuli to the right). Receptive field sizes and with it the shift invariance also increase from bottom to top (triangles in the center, tip indicating a neuron and base indicating its receptive field size). Response latencies are very short (numbers on the left).

– **No processing towards recognition.** The simplicity of the normalization transformation also implies that no processing towards recognition is achieved. This may be a disadvantage because the time used for the normalization cannot be used for recognition.

4.2 Invariant Features

In this alternative approach some features are extracted from the image that are invariant to the location and size of an object in the visual field. Invariant recognition can then be based on these invariant features. This approach has the following principal advantages (+) and disadvantage (–).

– ***Where*-information may be difficult to extract.** Since the point of extracting invariant features is to ignore any positional and size information, this *where*-information may actually be difficult to extract if needed. In any case, it will require additional machinery and probably additional time to do so, with the possible exception of some simple cases (Wiskott, 1999).

\+ **Recognition does not require knowing where the object is.** A great advantage of this approach is that objects can be recognized without knowing where they are and which size they have. Thus object recognition can be potentially faster than if normalization were required.

– **Usually information is lost.** The invariant features being extracted will be typically tailored to the natural visual environment. Thus if unnatural visual stimuli are presented, the object representation may be insufficient and invariant recognition may degrade. Another type of information loss results from the lack of spatial information which can potentially cause confusion between objects composed of the same local features in different spatial arrangements (but see Ullman and Soloviev, 1999; Mel and Fiser, 2000) and interference between different objects in the visual field.

+ **Processing towards recognition.** Extracting invariant features is already an important step towards object recognition. Thus invariance and recognition are achieved largely simultaneously.

Notice that the four pros and cons listed in this section are complementary to those listed in the previous section and that the first two and the last two pros and cons in each list form pairs related to the same property. Because of this complementarity it may be an appealing idea to combine these two approaches; this issue will be touched upon in Section 7.1.

5 Dynamic Routing Circuit Model

The most prominent neural model for shift and size invariant recognition in the visual system based on a normalization is probably the dynamic routing circuit model by Olshausen et al. (1993). This model implements the normalization in a rather direct fashion; see Figure 2. The connectivity between two successive layers is controlled by routing control units, which can turn on or off certain subsets of connections. If the appropriate connections are activated, a region in the input layer, referred to as the window of attention, is projected to the output layer in a standardized size. This provides a normalized representation of the attended region, based on which recognition can be performed. The connections between the different layers are organized such that small shifts and rescalings can be realized at the lower stages while larger shifts and rescalings are realized at the higher stages. The input layer is associated with the LGN and the output layer is associated with AIT. A closely related model has been presented by (Postma et al., 1997).

Several aspects of the visual system listed in Section 3 can be easily accounted for by the routing circuit model. The network has a layered structure and achieves shift and size invariance (although in principle the invariances do not depend on the type of stimulus; cf. Sec. 2). Units have increasing receptive field sizes and potentially increasing invariances from bottom to top. There is a clear split between *what*- and *where*-information represented in the main network and the routing control units, respectively. Finally, attention is naturally implemented in the network (although only in its spatial form; cf. Sec. 3). However, there are also some open issues here. Feedback is only used in an indirect fashion through the routing control units; there is no role for direct feedback yet. The network does not show a feature hierarchy. So far only local light intensities are represented in the network from bottom to top. The network could easily be adapted to other features such as Gabor-wavelet responses, but this would not introduce a feature hierarchy. Speed is also an open issue. Since the network in its present form requires the control of the routing prior to recognition, it is unclear how the model could account for rapid recognition within 150 ms. Finally, no ideas as to how a routing circuit network could be setup by self-organization have been worked out yet. Some of these open issues will be discussed in greater detail in Section 7.

6 Invariant Feature Networks

There are a great number of neural models based on the idea of extracting invariant features. Prominent examples are the neocognitron (Fukushima et al., 1983), higher-order neural networks (e.g. Reid et al., 1989), and the weight-sharing backpropagation network (LeCun et al., 1989). Invariant features are typically extracted in two steps. First, features are extracted, then invariance is achieved by spatial pooling. For any feature being extracted a set of units with identical receptive fields is distributed over the input layer. In neural network models this is often achieved by a weight-sharing constraint. Pooling over a neighborhood of units sensitive to the same local feature at different locations yields a feature specific response that is invariant to local shifts. A neural module extracting invariant features is illustrated in Figure 3. The extracting and pooling might also be performed on a dendritic tree rather than by a layer of units (Mel et al., 1998). Size invariance can be treated analogously to shift invariance if in the first step common features of different size are extracted (Gochin, 1994). By combining many of these modules at different locations and different levels, one obtains a hierarchical network extracting invariant features of increasing

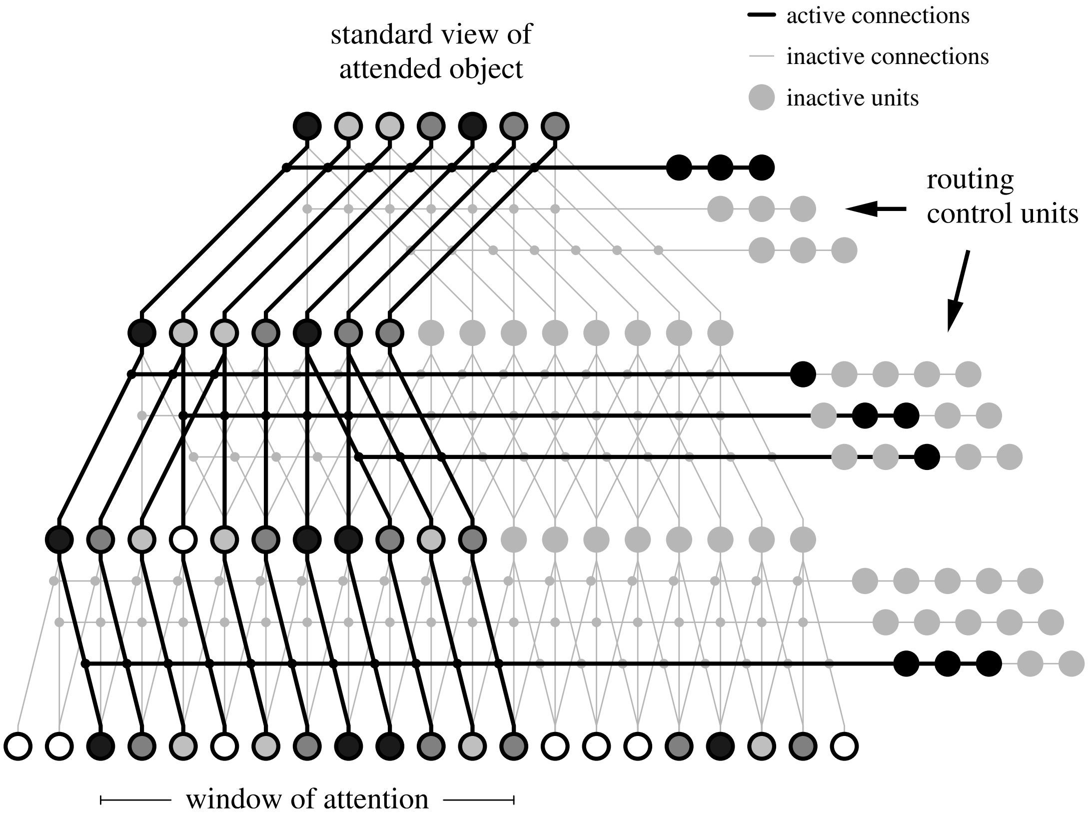

Figure 2: Schematic illustration of a routing circuit. The activity of units represent a feature value, such as local light intensity, and is indicated by different gray values. The same type of feature is used in the whole network (no feature hierarchy). Most of the existing connections between two successive layers are disabled (gray lines) through inhibitory mechanisms by the routing control units. The remaining active connections (black lines) establish a mapping between a region in the input layer (bottom layer), referred to as window of attention, and the output layer (top layer). This provides a normalized view of the attended object.

complexity and increasing invariance. Figure 4 shows such a hierarchical network, which is similar to a neocognitron (Fukushima et al., 1983).

The basic architecture has been extended in several ways. To include top-down attention to a location Salinas and Abbott (1997) have added so-called gain fields to allow selecting a local region and enable feature extracting units only there. One can also imagine top-down attention to objects or features if the facilitation acts on different sets of units sensitive to a common feature rather than location as illustrated in Figure 5 (cf. Koch and Ullman, 1985). These attentional control mechanisms are similar to those in the routing circuit model in that they work top-down and require indirect feedback. In (Fukushima, 1986) saliency driven attention based on direct feedback was implemented. Assume that at the top level there are units that respond to objects (single units are considered here for simplicity and may actually correspond to larger groups of neurons). If several objects are present in the visual field, several of these units will be active. Through a winner-take-all mechanism, one of these units can be selected and feedback connections can be used to trace back which of the feedforward units and connections gave rise to the response of the winning unit. By facilitating those connections and units and suppressing others, the system can attend to the most salient object. This is also illustrated in Figure 5.

The detailed connectivity of an invariant feature network is too complicated to be determined genetically. It is therefore interesting to see that invariances can be learned in a hierarchical network based on visual experience (Wallis and Rolls, 1997; Wiskott, 1999), leading to a connectivity like in a typical invariant feature network. The respective learning principle is based on the assumption that the external world changes slowly while the primary sensory signals change quickly, e.g. the response of retinal photoreceptors change quickly due to their small receptive field sizes. Unsupervised learning of invariances can then be based on the

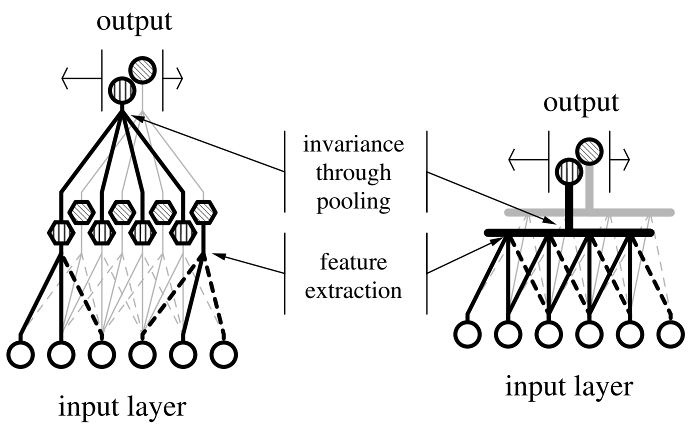

Figure 3: Modules for extracting shift invariant features. Black lines highlight the receptive fields of some selected units (— excitatory connections, - - - inhibitory connections). **Left:** In a first step arrays of units (hexagons) with identical receptive fields extract features (indicated by different textures) at any given location. In a second step the activity of all units sensitive to the same feature is pooled. The pooling units (textured circles) then respond to this feature invariant to local shift within the receptive field (indicated by the arrows ($\mapsto$) left and right the pooling units). **Right:** The same computation could also be performed on the dendritic tree (thick lines) rather than by an explicit neural layer. However, for clarity the left type of illustration will be used in the following.

objective of extracting slowly changing features from the quickly varying sensory input (Földiák, 1991), which then leads to a robust and invariant representation of the environment.

Invariant feature networks can account for many aspects of the visual system listed in Section 3. They have a layered structure with increasing feature complexity as well as increasing shift invariance. Size invariance could be dealt with analogously. Object recognition can be very fast, since without attentional control all processing is purely feedforward. Attention can be implemented in various ways which also provides a function for direct feedback connections. Finally, simulations have shown that the connectivity of invariant feature networks can be learned based on visual experience. Only one of the major issues considered here remains unclear: How can an invariant feature network process *where*-information separately from *what*-information?

7 Open Questions

In the preceding sections a short review of the two main classes of neural models for shift and size invariant recognition was given. It became clear that the routing circuit model leaves more questions open than invariant feature networks. However, this is not surprising, since many more researchers have been working on the latter. Thus at the current stage of the discussion both models should be considered further. I will discuss two open questions regarding the routing circuit model and one open question regarding the invariant feature networks.

7.1 How can Routing Circuits have a Feature Hierarchy?

Routing circuits only achieve a normalization. No feature extraction or other kind of processing towards object recognition is performed. Thus routing circuits do not have a feature hierarchy. If the input layer represents gray values, the output layer also represents gray values. This is in conflict with neurophysiological results which indicate that neurons in higher layers are selective for more complex features than those in lower layers. Thus, the question is: How can routing circuits have a feature hierarchy?

The first naive approach would be to strictly alternate a feature extraction stage like the first stage in Figure 4 and a routing stage. The feature extraction stage would increase the number of units at any location. This is no problem in the invariant feature network since there the density of locations that need to be represented

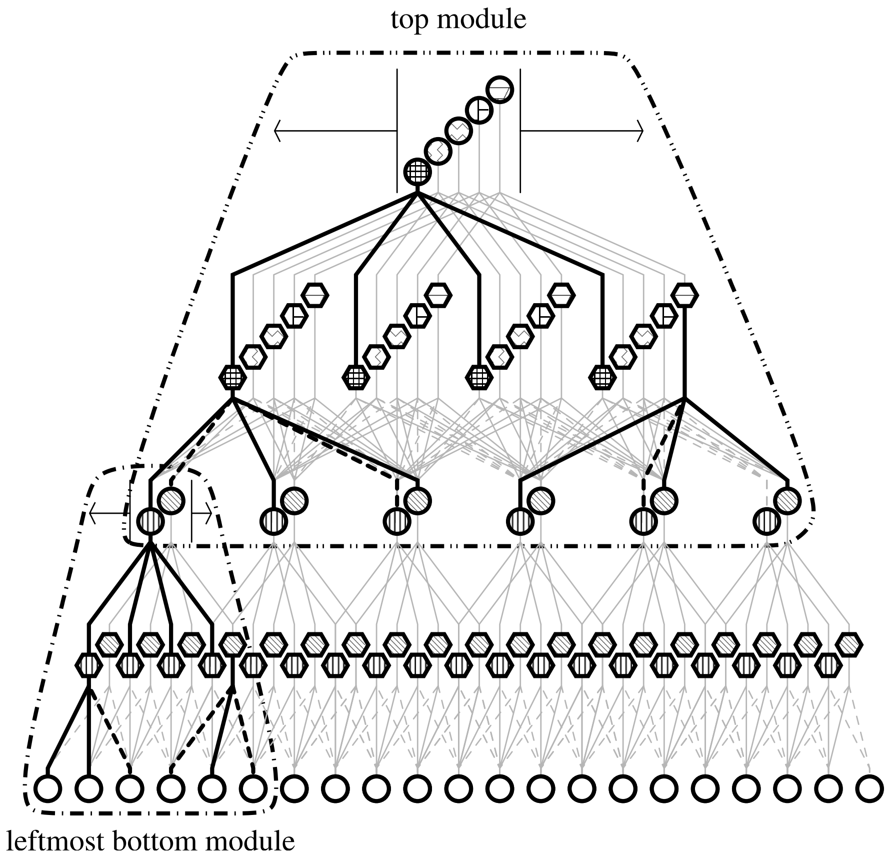

Figure 4: A hierarchical network for extracting invariant features can be built by replicating the module depicted in Figure 3 at different locations and different levels. The top module has the same structure as the bottom modules except that more complex features in greater number are combined and larger invariance is achieved. The bottom modules have some overlap. As one proceeds up the hierarchy spatial specificity is traded for feature specificity.

is reduced in the pooling stages. However, this is not the case in the routing stages. In the standard routing circuit model the total number of units is only decreased by reducing the overall size of the represented field but the density of complex features would be the same as that of the simple features at the bottom layer, which would lead to an explosion of the number of units and a very redundant representation.

Thus, it is clear that the spatial density of the representation has to decrease from bottom to top. A second approach would therefore be to include the dynamic routing mechanism into the standard invariant feature network. This is illustrated in Figure 6. A problem with this solution, however, is that local features can only be represented with a fixed spacing or, alternatively, local distortions and deviations from the regular spacing have to be explicitly represented by the routing control units and possibly also memorized. In any case, this architecture looks very similar to the standard invariant feature network with attentional control. It is therefore an open question whether a feature hierarchy could be included in the routing circuit model in a sensible way while preserving properties that differ significantly from the invariant feature network.

7.2 How can Dynamic Routing Circuits be Fast Enough?

Another open question with the dynamic routing circuit is its speed. We know that humans can do visual classification tasks, e.g. animal vs. no animal, within about 150 ms regardless of the accurate location of the objects in the visual field. In the general case the standard procedure for invariant object recognition in a dynamic routing circuit would probably be as follows: (i) The window of attention is widely open and the presented image gets propagated through the routing circuit. (ii) Some mechanism determines the region in the visual field the system has to attend to. (iii) The routing control units are appropriately activated and

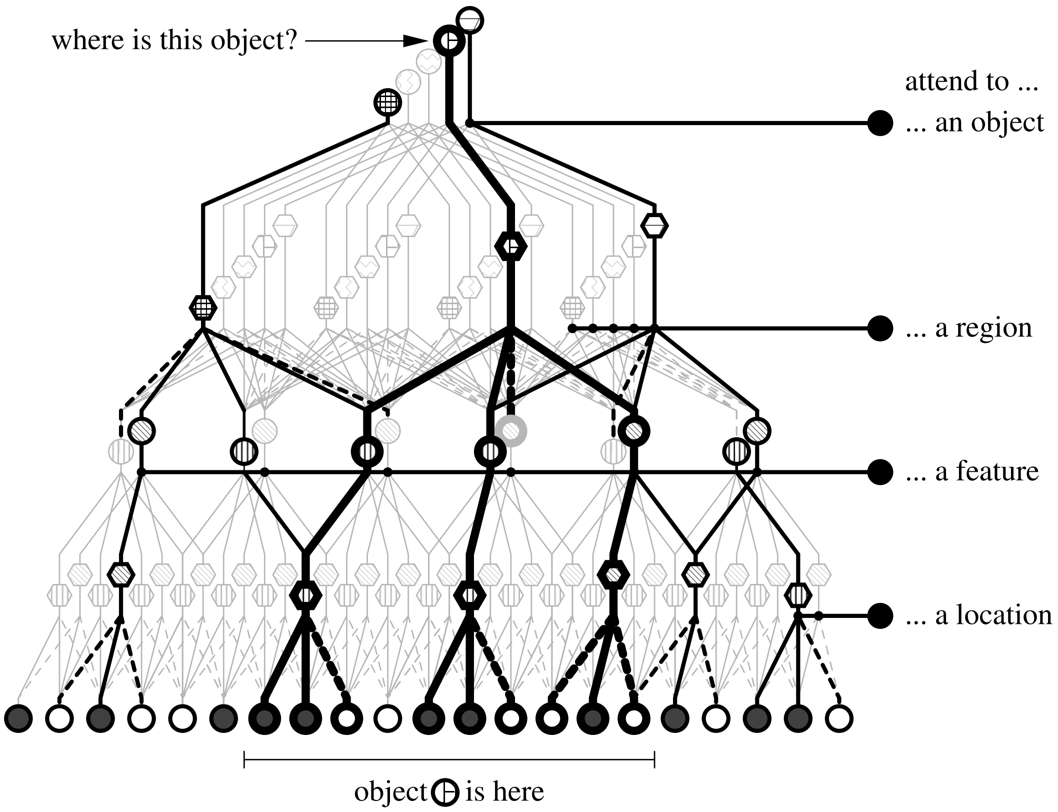

Figure 5: The hierarchical network for extracting invariant features with two types of attentional control. Firstly, the control units to the right can facilitate or enable (thin black lines) and disable (connections not shown) certain sets of units and connections so that only information about a particular object, region, feature, or location is being processed. This is a form of top down attentional control. Secondly, those connections and units which gave rise to the activity of a particular unit at the top can be facilitated through direct feedback connections (thick black lines and units). This is a form of saliency driven attention. If only the facilitated part of the network remains active, the location of the attended object can be easily determined by some mechanisms not illustrated here.

the dynamic routing is established. (iv) The image gets propagated through the routing circuit once again such that the attended object is normalized to standard size and location. (v) Object recognition can take place based on the normalized view. Even if we use very optimistic estimates for the times used by these five steps, it is clear that the whole process cannot be finished within 150 ms. Thus, the question is: How can fast object recognition be explained by dynamic routing?

One simple solution would be to assume that the rapid recognition experiments done so far do not require dynamic routing and can be explained simply with the window of attention widely open (Bruno Olshausen, 2000, personal communication). The task of distinguishing animal pictures from non-animal pictures (Thorpe et al., 1996) might be solved just based on simple feature detectors, such as eye- and fur-detectors, which do not require a normalization of the input image. If this solution turns out to be true, it would indicate that the routing circuit can initially work in an invariant feature network mode.

Another solution might be to assume that in simple cases the window of attention can be determined very rapidly on a low level. By modifying an earlier experiment (Subramaniam et al., 1995) Biederman and coworkers have performed rapid recognition experiments in which black line drawings were presented on a white background every 70 ms at one of four different locations in the visual field and the subjects had to respond when they recognized a particular object. Performance was nearly perfect despite the fact that the location of stimulus presentation changed randomly from image to image, i.e. every 70 ms (Irving Biederman, 2000, personal communication). In this experiment the location of the objects could have been detected purely on the basis of low-level cues (black lines on white background). Such a low-level based

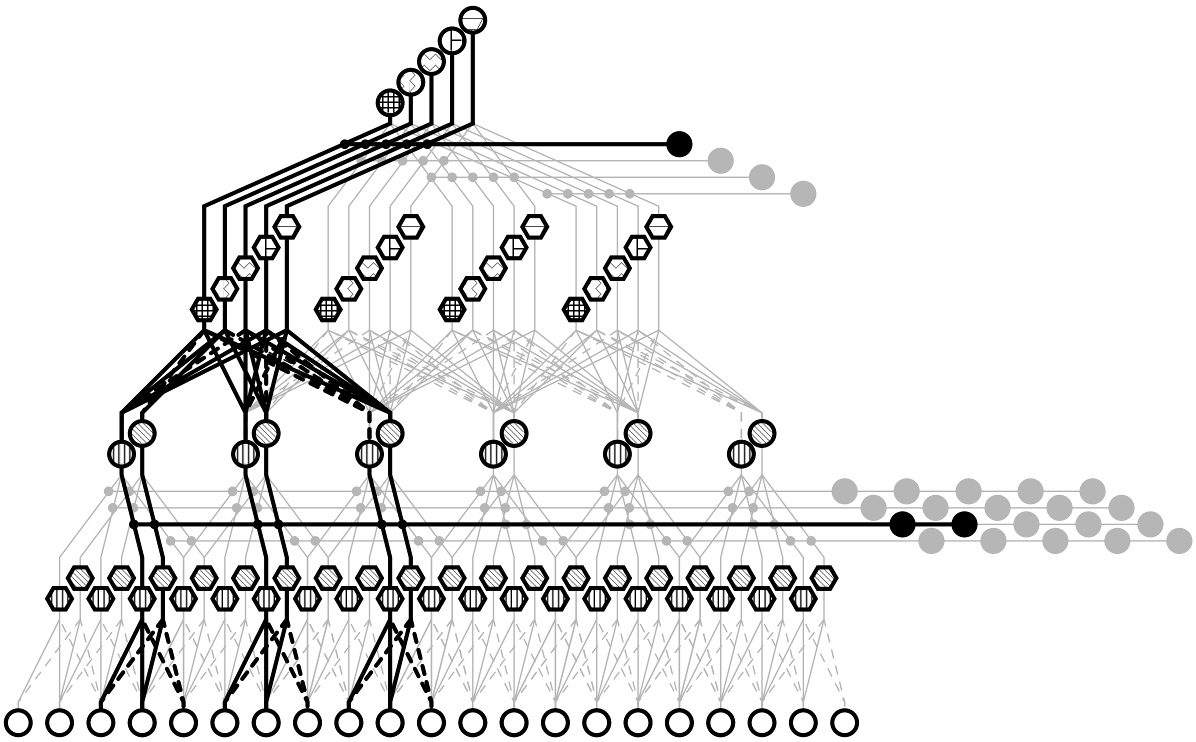

Figure 6: A hierarchical network that is a combination of an invariant feature network and a dynamic routing circuit. The pooling step is replaced by dynamic routing, thereby preserving control over the spatial normalization transformation. With a rigid routing scheme, like the one indicated in this figure, local features could be represented only with a fixed spacing. This problem could be solved by a more flexible routing scheme that would also include local distortions. This however, would make the control more demanding and the architecture similar to the original invariant feature network with top-down attentional control, which does not depend on the detailed control of the distortions.

attentional control would bypass the ventral pathway and permit to skip step (i) above and reduce the time used for step (ii). The control mechanism proposed in (Olshausen et al., 1993) is of this type, but the control dynamics is recurrent and based on gradient ascent, which makes the system slow. However, if a more rapid version of this control mechanism could be developed, it might be possible that the system would be fast enough even if dynamic routing were required for recognition.

A more detailed analysis of existing experimental data and possibly further experiments are required to validate or rule out one or both of these solutions.

7.3 How can Invariant Feature Networks Process *Where*-Information?

The major open question for the invariant feature networks is that of how they process *where*-information. There are some models which address this issue. Hummel and Biederman (1992), for instance, have built a network model called JIM (John and Irv's model) for recognizing 3D-objects made of simple geometric shapes. Part of JIM extracts invariant features and other parts process information about the location and size of the features. However, this model is incomplete in that the *where*-information is not extracted from the input image but provided separately by hand. Learning to extract invariant features and *where*-information from input images has been demonstrated in (Jacobs et al., 1991) and (Wiskott, 1999), in the former case through supervised learning and in the latter case through unsupervised learning. However, in both cases only one object in front of a blank background was visible at a time so that extracting *where*-information was greatly simplified.

Thus, processing *where*-information in an invariant feature network remains an open issue. However, in context of attentional control some of the necessary machinery has already been developed. Two basic mechanisms are needed. Firstly, communication from the *where*- to the *what*-system requires a mechanism for focusing on a particular location to answer the question: What do I see here? This can be done by the top down attentional mechanism described in Section 6. Secondly, communication from the *what*- to the *where*-system requires a mechanism for determining the location of a recognized object to answer the question: Where do I see this? This can be done by the saliency driven attentional mechanism also described in Section 6. Thus, we see that some basic machinery for communication between the *what*- and *where*-system has been developed. What remains to be done is to link together these existing components and to develop a readout mechanism for object location in case of saliency driven attention. A detailed comparison with psychophysical and neurophysiological results would then have to show how plausible such a model would be.

8 Other Models

There are a number of models for shift and size invariant recognition that have not been considered in the discussion above. Some of these models will now be briefly mentioned.

Several models are based on a fixed global invariance transformation, such as a combination of the log polar and Fourier transform (Cavanagh, 1978) or the R-transform (Reitböck and Altmann, 1984). These are an extreme case of an invariant feature network. They are not considered here for two reasons. Firstly, since the transformation is global, recognition requires a segmented object; cluttered scenes cannot be processed well. Secondly, the transformation requires a very precise connectivity and no mechanisms as to how these structures could be learned or self-organize have been presented.

Dynamic link matching (Bienenstock and von der Malsburg, 1987; Konen et al., 1994; Wiskott and von der Malsburg, 1996) is a dynamics which establishes a topology-preserving mapping between similar objects in two different layers. This can be used to find a normalizing transformation to achieve invariant recognition. Since the dynamics requires an interplay between synaptic changes and induced correlations, dynamic link matching is too slow to account for the rapid invariant recognition capabilities of humans and is not considered here. However, dynamic link matching may play an important role in the self-organization of the visual system (Wiskott and von der Malsburg, 1996).

An interesting mechanism for extracting invariant features has been proposed by Buonomano and Merzenich (1999). It is based on local interactions generating a temporal code that is shift invariant. It would be interesting to investigate to what extent this mechanism could substitute for the invariant feature module described in Section 6.

9 Conclusion

How does our visual system achieve shift and size invariance? I have discussed this question here from a theoretician's point of view by giving an overview over computational models and pointing out some open issues in developing these models further. At the level of the discussion presented here the invariant feature networks seem to be consistent with most of the neurobiological constraints listed in Section 3, while there are more open issues for the dynamic routing circuit model, which represents here the computational approach of normalization. However, this picture might change if the discussion is carried further and more experimental details are taken into account. It may also be that both types of networks need to be combined as briefly considered in Section 7. The visual system could work in an invariant feature mode in the beginning and then use mechanisms of dynamic routing for a refined perception. Finally, it may also turn out that neither of the two network models discussed here is realized in the visual system. Some alternative models were mentioned in Section 8.

The question of shift and size invariance may appear to be too specific to be worth being raised as one of the important questions in systems neuroscience. Wouldn't be the question of how our brain builds invariant representations in general be much more suitable? I think it depends on the answer. Either the brain solves all invariance problems in a similar way based on a few basic principles or it solves each invariance problem in a specific way that is different from all others. In the former case the more general question would be appropriate and one could consider the more specific question of shift and size invariance as a representative example. Solving the problem of shift and size invariance would then provide the key to all other invariance problems. In the latter case, i.e. if all invariance problems have their specific solution, the more general question would indeed be a set of questions and as such not appropriate to be raised and discussed here. There is, of course, a third and most likely alternative and that is that the truth lies somewhere between these two extremes.

Acknowledgments

I am grateful to Bruno Olshausen and Irving Biederman for fruitful discussions and valuable feedback and to Raphael Ritz for critically reading the manuscript. I also wish to thank Andreas Herz for providing excellent working conditions. Financial support was given by HFSP (RG 35-97) and the Deutsche Forschungsgemeinschaft (DFG).

References

Biederman, I. and E. E. Cooper (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception 20*, 585–593.

Biederman, I. and E. E. Cooper (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance 18*(1), 121–133.

Bienenstock, E. and C. von der Malsburg (1987). A neural network for invariant pattern recognition. *Europhysics Letters 4*(1), 121–126.

Buonomano, D. V. and M. Merzenich (1999, January). A neural network model of temporal code generation and position-invariant pattern recognition. *Neural Computation 11*(1), 103–116.

Cavanagh, P. (1978). Size and position invariance in the visual system. *Perception 7*, 167–177.

Desimone, R. and J. Duncan (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience 18*, 193–222.

Dill, M. and M. Fahle (1998, January). Limited translation invariance of human visual pattern recognition. *Perception and Psychophysics 60*(1), 65–81.

Felleman, D. J. and D. C. Van Essen (1991, January). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex 1*, 1–47.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation 3*, 194–200.

Fukushima, K. (1986). A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics 55*, 5–15.

Fukushima, K., S. Miyake, and T. Ito (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Trans. on Systems, Man, and Cybernetics 13*, 826–834. Reprinted in *Neurocomputing*, J. A. Anderson and E. Rosenfeld, Eds., MIT Press, Massachusetts, pp. 526–534.

Furmanski, C. S. and S. A. Engel (2000, March). Perceptual learning in object recognition: Object specificity and size invariance. *Vision Research 40*(5), 473–484.

Gochin, P. M. (1994, September). Properties of simulated neurons from a model of primate inferior temporal cortex. *Cerebral Cortex 5*, 532–543.

Hummel, J. E. and I. Biederman (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review 99*(3), 480–517.

Ito, M., H. Tamura, I. Fujita, and K. Tanaka (1995). Size and position invariance of neural responses in monkey inferotemporal cortex. *J. of Neurophysiology 73*(1), 218–226.

Jacobs, R. A., M. I. Jordan, and A. G. Barto (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision task. *Cognitive Science 15*, 219–250.

Kobatake, E. and K. Tanaka (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. of Neurophysiology 71*(3), 856–867.

Koch, C. and S. Ullman (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology 4*(4), 219–227.

Konen, W., T. Maurer, and C. von der Malsburg (1994). A fast dynamic link matching algorithm for invariant pattern recognition. *Neural Networks 7*(6/7), 1019–1030.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation 1*(4), 541–551.

Mel, B. W. and J. Fiser (2000). Minimizing binding errors using learned conjunctive features. *Neural Computation 12*(2), 247–278.

Mel, B. W., D. L. Ruderman, and K. A. Archie (1998). Translation-invariant orientation tuning in visual "complex" cells could derive from intradendritic computations. *The Journal of Neuroscience 18*(11), 4325–4334.

Merigan, W. H. and J. H. R. Maunsell (1993). How parallel are the primate visual pathways? *Annual Review of Neuroscience 16*, 369–402.

Nazir, T. A. and J. K. O'Regan (1990). Some results on translation invariance in the human visual system. *Spatial Vision 5*(2), 81–100.

Nowak, L. G. and J. Bullier (1997). The timing of information transfer in the visual system. In Rockland et al. (Eds.), *Cerebral Cortex*, Volume 12, Chapter 5, pp. 205–241. New York: Plenum Press.

Olshausen, B. A., C. H. Anderson, and D. C. Van Essen (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. of Neuroscience 13*(11), 4700–4719.

Oram, M. W. and D. I. Perrett (1994). Modeling visual recognition from neurobiological constraints. *Neural Networks 7*(6/7), 945–972.

Postma, E. O., H. J. van den Herik, and P. T. W. Hudson (1997). SCAN: A scalable neural model of covert attention. *Neural Networks 10*(6), 993–1015.

Reid, M. B., L. Spirkovska, and E. Ochoa (1989). Simultaneous position, scale, and rotation invariant pattern classification using third-order neural networks. *International Journal of Neural Networks - Research & Applications 1*(3), 154–159.

Reitböck, H. J. P. and J. Altmann (1984). A model for size- and rotation-invariant pattern processing in the visual system. *Biological Cybernetics 51*, 113–121.

Salinas, E. and L. F. Abbott (1997, June). Invariant visual responses from attentional gain fields. *Journal of Neurophysiology 77*(6), 3267–3272.

Subramaniam, S., I. Biederman, P. Kalocsai, and S. R. Madigan (1995, May). Accurate identification, but chance forced-choice recognition for RSVP pictures. In *Proc. Association for Research in Vision and Ophtalmology, ARVO'95, Ft. Lauderdale, Florida.*

Thorpe, S., F. Fize, and C. Marlot (1996, June). Speed of processing in the human visual system. *Nature 381*, 520–522.

Tovée, M. J., E. T. Rolls, and P. Azzopardi (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *J. of Neurophysiology 72*(3), 1049–1060.

Ullman, S. and S. Soloviev (1999, October). Computation of pattern invariance in brain-like structures. *Neural Networks 12*(7/8), 1021–1036.

Ungerleider, L. G. and M. Mishkin (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield (Eds.), *Analysis of Visual Behaviour*, Chapter 18, pp. 549–586. Cambridge, MA: MIT Press.

Wallis, G. and E. Rolls (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology 51*, 167–194.

Wiskott, L. (1999). Learning invariance manifolds. In *Proc. Computational Neuroscience Meeting, CNS'98, Santa Barbara.* Special issue of *Neurocomputing 26/27*, 925–932.

Wiskott, L. and C. von der Malsburg (1996). Face recognition by dynamic link matching. In J. Sirosh, R. Miikkulainen, and Y. Choe (Eds.), *Lateral Interactions in the Cortex: Structure and Function*, Chapter 11. http://www.cs.utexas.edu/users/nn/web-pubs/htmlbook96/: The UTCS Neural Networks Research Group, Austin, TX. Electronic book, ISBN 0-9647060-0-8.