Hierarchical Neural Networks for Image Interpretation

Sven Behnke

Freie Universität Berlin
Institute for Computer Science

This document contains an extended abstract of my dissertation thesis. The thesis was supervised by Prof. Dr. Raúl Rojas (FU Berlin). My profound gratitude goes to him for guidance, contribution of ideas, and encouragement. The thesis was submitted in October 2002 and defended in November of the same year. Prof. Dr. Volker Sperschneider (Osnabrück) was its second referee.

1 Introduction

Human performance in visual perception by far exceeds the performance of contemporary computer vision systems. While humans are able to perceive their environment almost instantly and reliably under a wide range of conditions, computer vision systems work well only under controlled conditions in limited domains.

The thesis addresses the differences in data structures and algorithms underlying the differences in performance. The interface problem between symbolic data manipulated in high-level vision and signals processed by low-level operations is identified as one of the mayor issues of today's computer vision systems.

The thesis aims at reproducing the robustness and speed of human perception by proposing a hierarchical architecture for iterative image interpretation. It is divided into two parts that cover theoretical aspects and applications of the proposed architecture.

2 Theory

First, I give some background information on the human visual system. The visual pathways, the organization of feature maps, computation in layers, neurons as processing units, and synapses as adaptable elements are covered. Some open questions are discussed, including the binding problem and the role of recurrent connections.

The coverage of related work focusses on two aspects of the proposed architecture: hierarchy and recurrence. Generic signal decompositions, neural networks, and generative statistical models are reviewed as examples of hierarchical systems for image analysis. The discussion of recurrence pays special attention to models with specific types of recurrent interactions: lateral, vertical, and the combination of both.

Neural Abstraction Pyramid Architecture: In the thesis I propose to use hierarchical neural networks for representing images at multiple abstraction levels, as illustrated in Fig. 1. The lowest level represents the image signal. In each new level upwards, the spatial resolution of two-dimensional analog representations decreases while feature

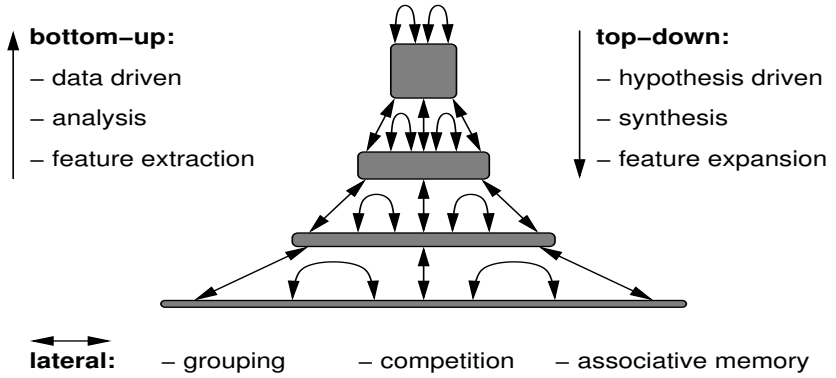

Fig. 1. Integration of bottom-up, lateral, and top-down processing. Images are represented at different levels of abstraction. As spatial resolution decreases, feature diversity and invariance to transformations increase. Simple processing elements interact via local recurrent connections.

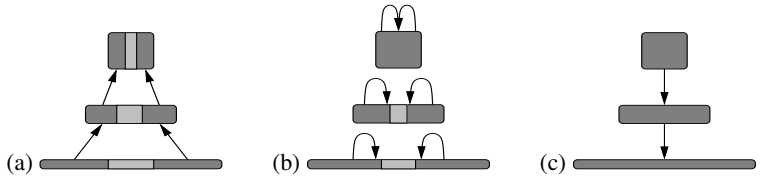

Fig. 2. Iterative image interpretation: (a) interpretation starts where little ambiguity exists; (b) lateral interactions reduce ambiguity; (c) top-down influences bias the low-level decision.

diversity and invariance increase. The representations are obtained using simple processing elements that interact locally. Recurrent horizontal and vertical interactions are mediated by weighted links. Weight sharing keeps the number of free parameters low. Recurrence allows for the integration of bottom-up, lateral, and top-down influences.

Inference in the proposed architecture is performed iteratively. An image is interpreted first at positions where little ambiguity exists. Partial results then bias the interpretation of more ambiguous stimuli. This is a flexible way to incorporate context. Such a refinement is most useful when the image contrast is low, noise and distractors are present, objects are partially occluded, or the interpretation is otherwise complicated.

To illustrate its use, small example networks apply the architecture to local contrast normalization, binarization of handwriting, and shift-invariant feature extraction.

Unsupervised and Supervised Learning: An unsupervised learning algorithm is proposed for the suggested architecture that yields a hierarchy of sparse features, illustrated in Fig. 3. It is applied to a dataset of handwritten digits. The produced abstract features are used as input to a supervised classifier. This classifier outperforms two existing classifiers. Performance increases further when it is used in combination with them.

After a general discussion of supervised learning problems, gradient descent techniques for feed-forward and recurrent neural networks are reviewed separately. Improvements to the backpropagation technique and regularization methods are discussed, as well as the difficulty of learning long-term dependencies in recurrent networks. It is suggested to combine the RPROP algorithm with backpropagation through time to achieve stable and fast learning in the proposed recurrent hierarchical architecture.

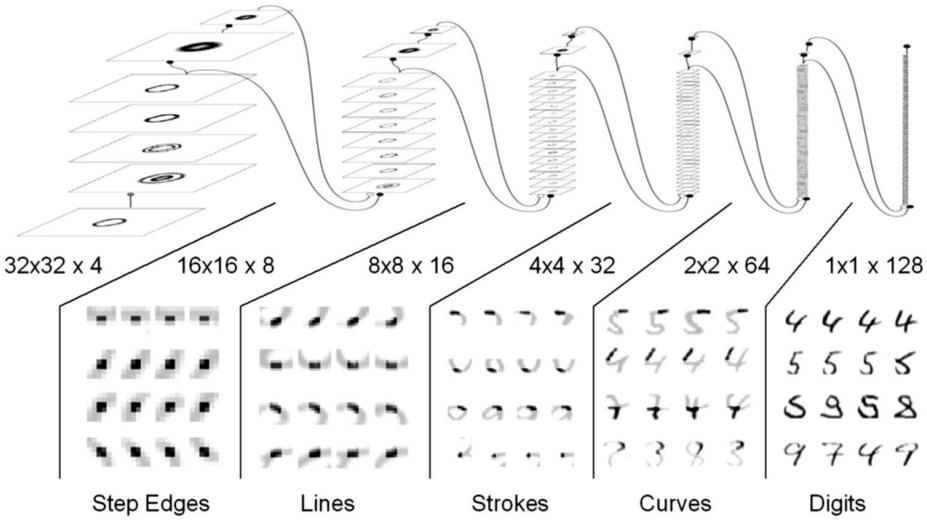

Fig. 3. Unsupervised learning a hierarchy of sparse digit features. The network architecture is sketched in the upper half. It consists of several layers of specific excitatory feature arrays with decreasing resolution. Unspecific inhibition is mediated by subsampled smoothed sums of local excitation. The lower part shows the four best stimuli for four of the feature detectors of each layer. One can observe an increasing degree of abstraction with distance from the input signal.

3 Applications

Recognition of Meter Values: The proposed architecture is applied to recognize the value of postage meter marks. After describing the problem, the dataset, and some preprocessing steps, two classifiers are detailed. The first one is a hierarchical block classifier that recognizes meter values without prior digit segmentation. The second one is a neural classifier for isolated digits that is employed when the block classifier cannot produce a confident decision. It uses the output of the block classifier for a neighboring digit as contextual input. This system can read meter values reliably.

Binarizaton of Matrix Codes: The second application deals with the binarization of matrix codes. After the introduction to the problem, an adaptive thresholding algorithm is proposed that is employed to produce outputs for undegraded images. A hierarchical recurrent network is trained to produce these outputs even when the input images are degraded with typical noise. This allows to read low-quality code images for which adaptive thresholding fails.

Image Reconstruction: The proposed architecture is also applied to image reconstruction problems. Super-resolution, the filling-in of occlusions, and noise removal/contrast enhancement are learned by hierarchical recurrent networks. Images are degraded and networks are trained to reproduce the originals iteratively. The same method is also applied to the reconstruction from image sequences, as shown in Fig. 4.

Face Localization: The last application deals with a problem of human-computer interaction: face localization. A hierarchical recurrent network, shown in Fig. 5, is trained

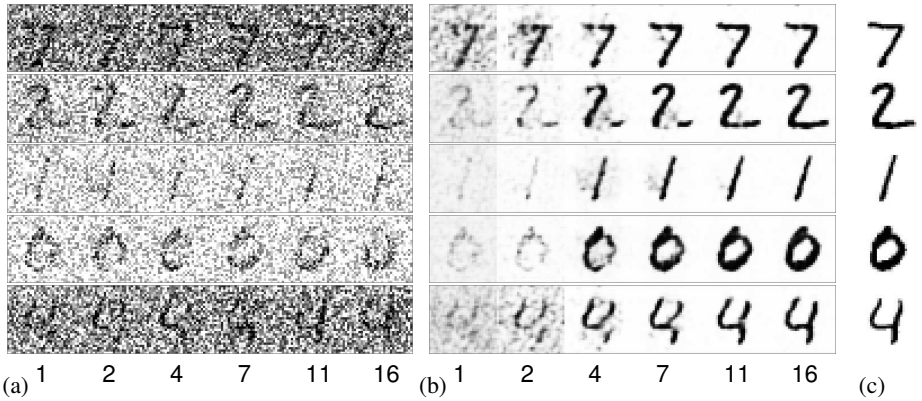

Fig. 4. Image reconstruction: (a) degraded input sequence; (b) output over time; (c) target.

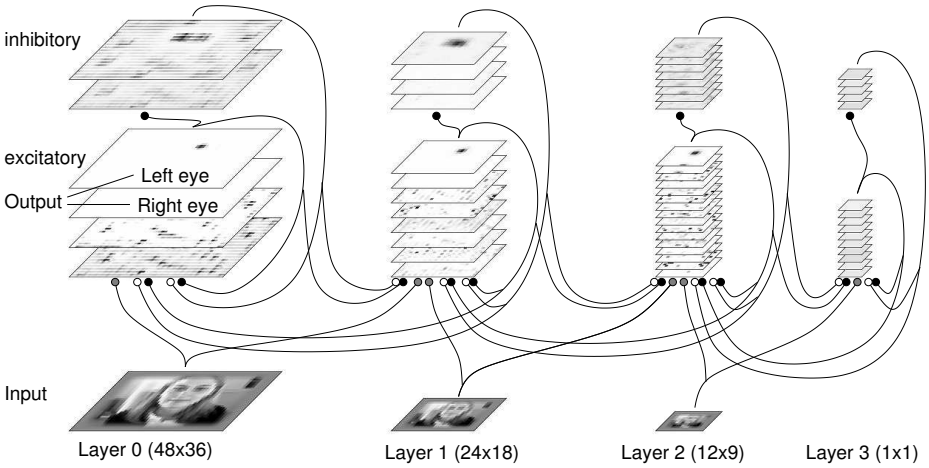

Fig. 5. Sketch of the network used for learning face localization. It consists of four layers with excitatory and inhibitory feature arrays. There is a local recurrent connectivity.

on a database of images that show persons in office environments. The task is to indicate the eye positions by producing a blob for each eye. The network outperforms a hybrid localization system that was proposed by the creators of the database. It is also able to track a moving face.

4 Summary and Conclusions

The successful application of the proposed architecture to several non-trivial computer vision tasks shows that the design patterns used are advantageous for such problems.

The architectural bias of the Neural Abstraction Pyramid facilitates learning of efficient image representations. The networks utilize the two-dimensional nature of images as well as their hierarchical structure. Because the same data structures and algorithms

are used from the lower layers of the pyramid all the way to its top, the interface problem between high-level and low-level representations does not occur.

The use of weight sharing allows to reuse examples seen at one location for the interpretation of other locations. It helps to limit the number of free network parameters and hence facilitates generalization. Restricting the weights to mediate specific excitation and unspecific inhibition constrains the representations used by the networks, since it enforces sparse features. A similar effect can be achieved with a low-activity prior.

The employment of recurrence was motivated by the ubiquitous presence of feedback in the human visual system and by the fact that an iterative solution to a problem is frequently much simpler than direct one. Recurrence allows to integrate bottom-up, lateral, and top-down influences. If local ambiguities exist, the interpretation decision can be deferred, until contextual evidence arrives, yielding a flexible use of context. Parts of the representation that are confident bias the interpretation of less confident parts.

This iterative approach has anytime characteristics. Initial interpretation results are available very early. If necessary, they are refined as the processing proceeds. The advantages of such a strategy are most obvious in situations which are challenging for current computer vision systems. While for the interpretation of unambiguous stimuli no refinement is necessary, the iterative interpretation helps to resolve ambiguities. Hence, its use should be considered when image contrast is low, noise is present, or objects are partially occluded. Since the recurrent networks can integrate information over time, they are suitable for the processing of input sequences, such as video.

The application of learning techniques to the proposed architecture shows a way to overcome the problematic design complexity of current computer vision systems. Supervised learning offers the possibility to specify the task through a set of input/output examples. Automatic optimization of all parts of the system is possible in order to produce the desired results. In this way, a generic network becomes task-specific.

Future Work: Several interesting aspects have not been covered in the thesis. They include implementation options, the use of more complex processing elements, and the integration of the perception network into a complete system.

Implementation options for the proposed architecture include SIMD processor extensions, parallel processors, and massive parallel implementations, including analog VLSI. The less flexible an implementation is, the greater is the potential for improvements in costs, size and power consumption. Dense arrays of vias could allow for tight integration of CMOS image sensors and multiple layers of analog processing arrays in a stack of chips. This has the potential for inexpensive, small, low power devices that have the computational power of today's supercomputers for computer-vision tasks.

The simple processing elements used resemble feed-forward neural networks with a single output-unit. One could investigate the use of biologically more realistic units that generate spikes and have dynamic synapses. Another interesting line of research would be to give the representations a probabilistic interpretation by viewing the abstraction pyramid as graphical belief network and applying generalized belief propagation.

Finally, the investigated system constitutes a perceptual network. It could be complimented by an inverse hierarchical network that expands abstract action decisions to low-level actuator commands. Connected to an embodied agent, such a complete system would allow for the use of reinforcement learning techniques.