Preface

Summary of the Thesis

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.

This 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 major issues of today's computer vision systems. This thesis aims at reproducing the robustness and speed of human perception by proposing a hierarchical architecture for iterative image interpretation.

I propose the use of hierarchical neural networks for representing images at multiple abstraction levels. The lowest level represents the image signal. As one ascends these levels of abstraction, the spatial resolution of two-dimensional feature maps decreases while feature 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 integration of bottom-up, lateral, and top-down influences.

Image interpretation 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.

The proposed architecture can be trained using unsupervised and supervised learning techniques. This allows one to replace manual design of application-specific computer vision systems with the automatic adaptation of a generic network. The task to be solved is then described using a dataset of input/output examples.

Applications of the proposed architecture are illustrated using small networks. Furthermore, several larger networks were trained to perform non-trivial computer vision tasks, such as the recognition of the value of postage meter marks and the binarization of matrixcodes. It is shown that image reconstruction problems, such as super-resolution, filling-in of occlusions, and contrast enhancement/noise removal, can be learned as well. Finally, the architecture was applied successfully to localize faces in complex office scenes. The network is also able to track moving faces.

Acknowledgements