Spatial Vision, Vol. 13, No. 2,3, pp. 129–135 (2000)

Editorial

Object recognition and image understanding: Theories of Everything?

Theories of Everything is the title of a recent book that summarises current attempts made in physics to achieve a unified theory of the universe (Barrow, 1991). In the history of epistemology, the question of how humans come to recognise objects is as old as the one about the nature of the universe. The quest for a unified theory of visual object recognition, would, therefore, seem to be a reasonable enterprise. Taking this for granted, one then might guess which perspective could produce such a unified theory. One possibility is a neural-computing approach involving an algorithmic view of brain function. This is, indeed, a good candidate for modeling the basic properties of the human brain, i.e. the ability to learn from past errors and correct itself quasi-automatically. However, the organization of the brain is extremely complex, and it is unlikely that neural nets without recurrent coupling offer more than descriptions of certain aspects of this complexity. This raises the question of how such micro-theories could be combined to produce a theory of the whole. In physics, statistical mechanics was confronted with the same problem, but its task was easier because the thermodynamical macro-quantities, such as temperature and entropy, were already known.

In brain theory, we have to assume that there is not just one level of organisation where macro-quantities are to be searched for. Instead it seems there is a hierarchy of levels of functional organisation, with macro-quantities being attached to each of them. Hence we do not know more about macro-quantities, characterizing cognition and conscious behaviour, than we know about micro-theories. The same considerations probably apply to the neurosciences. Micro-theories abound there, but what brain function *per se* are we going to model by integrating the outputs of a population of cortical simple cells, for example? We are forced to conclude that neither computational theory nor anatomy and physiology can do without the behavioural disciplines, which provide constraints for the quantities the former are supposed to predict. It is equally clear that visual psychophysics would still be restricted to measuring detection thresholds if physiology and formal approaches to visual recognition had not raised further questions about the perceiving and acting brain. All this suggests that there is little chance that any of these disciplines will be able to provide a *unified theory* of visual object recognition on its own. The future success of brain research, and of the quest for understanding object recognition in particular, will come from the close cooperation among the *three complementary disciplines*, anatomy/physiology, behavioural research, and theoretical concepts.

There is ready agreement among disciplines that object recognition and image understanding, i.e. the generation of some nonpictorial description of an image, are fundamental problems both for biological vision and for machine vision. It is equally clear that sensed data have to be related to what is known (visual models) in order to arrive at solutions for these problems. The approaches actually used, however, are something between template matching, or matched filtering, and syntactical image analysis (e.g. Pratt, 1978). One reason for this is the difference in criteria according to which solutions are judged in different disciplines. The neurosciences, for obvious reasons, put emphasis on the biological feasibility of solutions. Given the large body of data on receptive field characteristics in the visual system, they are therefore biased towards template matching. To avoid the inherent ambiguities of verbal communication, those concerned with analysing behaviour have to ask their subjects simple questions upon which, ideally, yes–no answers are given. Consequently, experimental paradigms of the delayed-matching-to-sample type are popular in visual psychophysics. The challenge in the engineering approach, by contrast, is in the analysis of reasonably complex real-world scenes, thus giving rise to the employment of structural or syntactical image analysis derived from formal linguistic methods of natural language analysis (Pratt, 1978).

In recent years, however, a number of common denominators between these seemingly divergent research interests has become apparent. One example of this is attention, a classical subject of psychology. It has now become a hot topic both for the neurosciences and for computer science. The need to solve problems of competition among concurrent processes of data analysis, task requirements, and the necessity for economic allocation of processing resources constitutes the driving force for interdisciplinary exchange (e.g. Felleman and Van Essen, 1991; Desimone and Duncan, 1995). Of common concern is also the idea of visual perception and recognition being action-oriented. This led to a renewed interest in the concept of *ecological optics* as advanced by the psychologist Gibson (1966), which has its counterparts in the research on sensory–motor integration in cognitive science and in the approach of *active vision* in robotics (e.g. Aloimonos, 1993). In neurophysiology, the analysis of single unit responses in the inferotemporal cortex of the monkey brain has revealed a degree of invariance against changes in stimulus conditions that is not found at earlier stages of visual processing. Invariant pattern and object recognition has always been a central issue for computational vision, but it is not yet clear whether such characteristics can be described best in terms of implicit processing models (e.g. Riesenhuber and Poggio, 1999; Zetzsche and Caelli, 1989) or whether structural schemes of analysis (Jüttner *et al.*, 1997) have to be invoked. Related to this, face recognition, traditionally the domain of neuropsychologists, has received considerable attention in the pattern recognition community with regard to technical applications (see Wechsler *et al.*, 1998). More generally, it has become clear in cognitive science that perceptual categorisation is the main way that organisms make sense of experience and is highly adaptive to context and task demands (Rosch, 1978; Staddon, 1983; Edelman, 1987). For technical object recognition this opens the possibility of a significant reduction of processing capacities, since models or templates for many objects can be replaced by models for single object classes. However, such class concepts do not exist *a priori* but have to be developed (Caelli and Bischof, 1997). This suggests that learning is the *conditio sine qua non* for object recognition by the living brain as well (Staddon, 1983; Edelman, 1987).

This brief and accidental list of interdisciplinary research interests in object recognition and image understanding highlights some of the topics that were dealt with in May 1999 at a Workshop on *Object Recognition and Image Understanding by Brain and Machines*. The host of this Workshop was the *Werner Reimers Stiftung* in Bad Homburg v.d.H., Germany, and the meeting was supported by the *Deutsche Forschungsgemeinschaft*, DFG. The present collection of articles is the result of this attempt to promote the exchange between sub-disciplines in brain research.

More systematically, the 18 papers collected in this Special Issue of *Spatial Vision* may be divided into five categories — neuroanatomy/physiology, neuropsychology, psychophysics, computational vision, and machine vision. This order suggests the existence of a systematic transition from purely biological to purely computational accounts of vision. However, many of our contributions actually defy a clear classification within that taxonomy. Their scope is much broader and often combines behavioural and theoretical aspects of vision research. This will become evident in the following brief summaries according to the above scheme.

Young first reviews recent neurophysiological and neuroanatomical evidence of the structure of the visual system. He then focuses on the problem of relating these structural results to function. Thereby, the traditional view of vision as information processing is questioned. An alternative perspective is proposed, which seeks to provide a neurobiological basis for the notion of vision as knowledge-rich inference.

Tanaka reviews recent findings concerning the functional architecture of area TE in the inferotemporal cortex of the monkey. He relates the flexible properties of visual object recognition, such as tolerance to illumination, viewing angle and object pose, to response invariances of TE cells and their arrangement in columnar modules within which cells respond to similar features.

Logothetis addresses the neurophysiological basis of face recognition. Research findings concerning properties of human face recognition are related to results from learning studies involving monkeys. It is argued that selectivity properties similar to those for faces may be elicited with other homogenous object classes after extensive training. This suggests that faces may not be a *special type* of stimulus *per se* but are by default a *special class* of the primate recognition system.

Landis reviews clinical evidence concerning various forms of disruption of space perception due to cortical lesions. He points out that such lesions may differentially affect retinotopic, egocentric, and allocentric frames of reference, as well as selective attention to far or near space and to global or local features of space. Together, these findings suggest the existence of disseminated functional modules for space control, rather than a unified space representation restricted to the parietal cortex.

Barth argues that important aspects of early- and middle-level visual coding may result from an efficient processing of the visual input. He demonstrates this concept by showing how functional properties of MT neurons, such as orientation and direction selectivity, can be predicted from a spatio-temporal representation in terms of differential geometry. Related findings are also readily employed for solving technical problems of flow-field analysis.

Krieger, Rentschler, Hauske, Schill and *Zetzsche* investigate saccadic mechanisms of feature selection in object and scene analysis from a perspective of higher-order statistical analysis. Evidence is presented that nonredundant, intrinsically two-dimensional image features, like curved lines and edges, play an important role in the saccadic selection process. Such feature extraction may form the bottom-up component of a mechanism which, in conjunction with a knowledge-driven top-down component, provides an efficient control of saccadic scanning.

Briscoe makes a case for vision as a sequential analysis of image information. He presents a neural model based on recurrent self-organising feature maps which operates upon a temporal trace of local features rather than upon a detailed internal iconic representation. Thus, recognition becomes a process spread over multiple time intervals and occurs as a result of the learning and identification of temporal sequences.

Rentschler and *Jüttner* explore the dynamics and the context dependence of visual category learning from a psychophysical perspective and of cognitive modelling. They demonstrate how machine-vision techniques may be used to visualise and to analyse the learning process and how they account for variations in learning speed imposed by contextual manipulations of the set of learning patterns.

Biederman summarises recent research in psychology and neurophysiology concerning the format of mental object representations. He concludes that the available evidence clearly supports the notion of a structural object description in terms of certain three-dimensional primitiva, the so-called geons, rather than a view-based description in terms of multiple two-dimensional templates.

Edelman and *Intrator* introduce a new model of object recognition based on a former view-based account. Their so-called chorus-of-fragments approach combines *what* and *where* information in terms of units which are tuned both to specific shapes and coarse location information. The authors describe a pilot implementation of their model and review supporting evidence concerning its theoretical foundation.

Christou and *Bülthoff* describe new approaches for studying the development of mental representations of scenes and objects in virtual-reality environments. Concerning scene recognition, they demonstrate that familiar views of a scene are more easily recognised than depth-rotated views of the same scene. Concerning the recognition of shapes, they find a significant contribution of the visual background to identification performance and a pronounced viewpoint dependency of the latter.

Osman, Pearce, Jüttner and *Rentschler* present a new approach to human recognition of 3D objects adopted from machine vision. Their technique aims at reconstructing visual representations underlying object recognition and is applied to behavioural data from an experiment exploring the role of haptic exploration in object learning. The analysis clearly reveals that haptic experience supports the evolution of object representations, increasing the degree of attribute differentiation and deepening the relational depth of structural encoding.

Poggio and *Shelton* summarise recent research with respect to object-recognition models based on regularisation theory. Such models typically are trained in learning-from-examples paradigms. The authors show how one particular form of this type of model works with complex scenes and discuss the relevance of their approach for neural computing in the cerebral cortex.

Bischof provides a review of recent work on rule-based pattern and object recognition in machine intelligence. Such techniques imply a decomposition of an object into its constituent parts, which then are characterised in terms of part-specific and relational features. Rule regions are defined for these features in the associated feature spaces. The paper discusses aspects of generation, application and evaluation of such rules.

Caelli discusses the learning of scene interpretations from the perspective of machine intelligence. As an example, he outlines the architecture of a recent image- annotation system, called CITE. The system generates multiple hypotheses concerning the labelling and grouping of image regions. Ambiguities are resolved by a process of relaxation labelling and constraint propagation. This process triggers a knowledge-driven resegmentation of the input image, thus leading to a closed top-down control loop.

Lazarescu, Venkatesh and *West* address the issue of machine learning in dynamic scenes. Here, learning is considered as an incremental process where not all data are known before the start of the training. Important aspects of incremental learning, such as memory size, forgetting and concept drift, are discussed and illustrated by an application to spatio-temporal tracking.

Ballard et al. emphasise the task-specific nature of vision and visual attention. They propose a hierarchical framework of so-called visual routines that serve as building blocks for the spatio-temporal organisation of behaviour. Their approach is illustrated by a system which simulates elements of car-driving behaviour, for instance when encountering traffic lights and stop signs.

The paper by *Bunke* finally provides an introduction to graph theory as applied to syntactical pattern recognition. Graphs provide a versatile and flexible formalism to build structured object representations with well-defined invariance properties. A similarity measure for graphs is derived and used to introduce the novel concept of a mean graph, which enables a set of graphs to be represented by its most typical member.

These synopses are dominated by a number of basic themes, or leitmotivs, that appear in several variations. One prominent leitmotiv is the nature of mental representations for space and objects, as explicitly addressed by Tanaka, Logothetis, Landis, Biederman, Edelman and Intrator, Christou and Bülthoff, and Osman *et al.* Another leitmotiv concerns the problem of interpretation, both of scene and of object information. It is discussed in the contributions of Poggio and Schelton, Bischof, Caelli, Lazarescu *et al.*, Ballard *et al.*, and Bunke. Furthermore, there is a group of papers which deal, on a fundamental level, with our functional understanding of how the brain processes visual information. Several alternatives to the classical representational notion of vision are proposed. These include vision as inference (Young), vision as sequence analysis (Barth, Briscoe, Krieger *et al.*), and the task-specific nature of visual processing (Ballard *et al.*). Finally, learning represents the global leitmotif underlying almost all contributions. Learning processes in general are highly adaptive to contextual information. Three papers specifically endeavour to elucidate the role of context, namely in category learning (Rentschler and Jüttner), in object recognition (Christou and Bülthoff), and in visually guided behaviour (Ballard *et al.*).

Given these shared epistemological interests, we see great potential to intensify the co-operation among the various research areas. This pertains in particular to the prevailing relatively loose connection between researchers working in neurophysiology and in machine vision. The former are in a situation that is reminiscent of the state of affairs of physics in the 20th century, when a plethora of elementary particles awaited systematic ordering before the emergence of a unified theory of electromagnetic and weak processes. There our understanding of inanimate nature has been advanced through intense interaction between experimental and theoretical research. To fulfil in the 21th century the promises made during the past *decade of the brain*, it may well be necessary to arrive at a comparably intense interaction among experimental and theoretical subdisciplines of cognitive science.

Acknowledgement

The organizers of the Symposium, Object Recognition and Image Understanding by Brain and Machines, are grateful to the Werner Reimers Stiftung, Bad Homburg v.d.H., Germany, and the Deutsche Forschungsgemeinschaft, DFG, for generous support. Frau Gertrude Söntgen was responsible for making the meeting also socially most enjoyable.

REFERENCES

Aloimonos, Y. (Ed., 1993). *Active Vision Revisited.* Lawrence Erlbaum, Hillsdale, NJ.

Barrow, J. D. (1991). *Theories of Everything.* Oxford University Press, Oxford.

Caelli, T. and Bischof, W. F. (1997). *Machine Learning and Image Interpretation.* Plenum Press, New York.

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

Edelman, G. M. (1987). *Neural Darwinism. The Theory of Neuronal Group Selection.* Basic Books, New York.

Felleman, D. and Van Essen, D. (1991). Distributed hierarchical processing in primate visual cortex, *Cerebral Cortex* **1**, 1–47.

Gibson, J. J. (1966). *The Senses Considered as Perceptual Systems.* Houghton Mifflin, Boston.

Jüttner, M., Caelli, T. and Rentschler, I. (1997). Evidence-based pattern classification: A structural approach to human perceptual learning and generalization, *Journal of Mathematical Psychology* **41**, 244–259.

Pratt, W. K. (1978). *Digital Image Processing.* John Wiley, New York.

Rosch, E. (1978). Principles of categorization, in: *Cognition and Categorization*, Rosch, E. and Lloyd, B. (Eds), pp. 27–48. Lawrence Erlbaum, Hillsdale, NJ.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex, *Nature Neuroscience* **2**, 1019–1025.

Staddon, J. E. R. (1983). *Adaptive Behavior and Learning.* Cambridge University Press, Cambridge.

Watanabe, S. (1985). *Pattern Recognition. Human and Mechanical.* John Wiley, New York.

Wechsler, H., Phillips, P. J., Bruce, V., Soulie, F. F. and Huang, T. (Eds) (1998). Face recognition: From theory to applications. NATO ASI Series F, Springer, Berlin.

Zetzsche, C. and Caelli, T. (1989). Invariant pattern recognition using multiple filter image representation, *Computer Vision, Graphics, and Image Processing* **45**, 251–262.

INGO RENTSCHLER
TERRY CAELLI
WALTER BISCHOF
MARTIN JÜTTNER