A Required Paradigm Shift in Today's Vision Research

Krüger, Norbert

Published in:
Kuenstliche Intelligenz

DOI:
10.1007/s13218-014-0347-7

Publication date:
2015

Document version
Peer reviewed version

Document license
Unspecified

Citation for pulished version (APA):
Krüger, N. (2015). A Required Paradigm Shift in Today's Vision Research. Kuenstliche Intelligenz, 29(1), 89-94. DOI: 10.1007/s13218-014-0347-7

Noname manuscript No.
(will be inserted by the editor)

A Required Paradigm Shift in Todays Vision Research

Interview with Prof. Christoph von der Malsburg

Norbert Krüger

Received: date / Accepted: date

Christoph von der Malsburg obtained his Ph.D. in Heidelberg on a subject of particle physics. He worked at the Max-Planck-Institut für biophysikalische Chemie in Göttingen from 1970 until 1987. Since 1988 he was professor of Computer Science at the University of Southern California (USC) and from 1990 he also co-directed the Institute für Neuroinformatik at the Ruhr-Univer-sität Bochum. Since 2006 he is Senior Fellow at the Frankfurt Institute for Advanced Studies (FIAS). In his work, he has made important contributions in the areas of pattern recognition and computational neuroscience.

KI: You led the Institut fuer Neuroinformatik from 1990 until 2006 at the Ruhr-Universität Bochum. I remember from the time I was a PhD student in your group that you were traveling between Los Angeles and Bochum back and forth twice a year. Now you are working at the Frankfurt Institute for Advanced Studies. Do you miss Bochum and Los Angeles in any way?

Of course I miss them. At that time I had a large group of co-workers in both groups. I had a lot of resources that I could use to develop systems. I now essentially work without any resources apart from a computer and an office. But I must say, I am enjoying both types of activity very much.

N. Krüger
The Maersk Mc-Kinney Moller Institute
Campusvej 55
DK-5230 Odense M
Denmark
+4527787483
E-mail: norbert@mmmi.sdu.dk

KI: What are you concerned with in your current work?

I think I have to collect my thoughts. I staged my life as an attempt to understand the brain and now is cash time. I have to write down my thoughts in a coherent fashion and see how far I have gotten in understanding the brain.

KI: Your group was focusing in the first place on the vision problem with the aim to take the human brain as a model. What kind of knowledge about or which aspects of the brain were useful to derive vision algorithms?

I think it is very important to know what the natural visual system can do and what it cannot do and, by this, to avoid putting up problems that cannot be solved at all. This is certainly one important kind of influence. The second one is that the brain is different from all of computer work in that it does not have a separate programmer. It starts with an initial condition, which has been developed by evolution, and from then on it is on its own resources.

There are two ways it has to solve problems. One is by learning from examples and the other one is by some kind of self-organization or self-interaction in order to create structures that are more self-consistent. And then, of course, studying the brain has to come up with the perspective of a very simple basic data structure, which in the brain is in the form of neurons and their connections and their activity, which can express whatever is to be expressed. This is a drive towards homogeneity and simplicity. I think that is the most important lesson we can learn from the brain.

KI: How would you judge the success of your group in retrospective?

Well, it is the path and not the goal that is important. I think the group has achieved a lot – number one, of course, in founding careers. There are a number of professorships that arose from my institute and also a large number of students who have found good positions in industry. Another aspect is that the group worked very consistently towards a conceptual framework (see Box). Time will tell to which extent this framework will prevail in the scientific medium.

KI: What was the driving force of this framework?

I think the two questions you need to pose when you want to understand the brain and want to create artificial vision systems are, first, what is the basic data structure with which you can represent all the items that you want to express and, second, what are the algorithms – or rather dynamic processes – by which this data structure is put in shape. That is what I would call a cognitive architecture or a visual architecture. We need a data structure and a process of organization, both of which have various time scales: One, concerning the current state, which changes within tenths of a second and, two, the time-scale of memory. Accordingly, the process of organization has one component that brings the brain state into shape and another, which organizes memory. So these are the questions to be answered. It is interesting to see that on the computer vision side, there are no coherent answers to these questions of generic data structures and generic processes of organization on these two time-scales, and there is not even the drive to find those. When focusing there on a particular problem, one usually addresses two issues: A specialized data structure and a specialized algorithm. As I said earlier, in the brain it is different: There is only one basic data structure and one basic mechanism of organization on each time scale. So these questions only arise in the brain context.

It is also interesting to see that even on the neural side, these two issues 'data structure' and 'process of organization' are hardly ever discussed. This is because everybody believes to have concrete answers to them. The generally accepted answer to the data structure question is that the brain is filled with elementary symbols which correspond to neurons or groups of neurons. You can probe them with the electrode and find out what the symbolic meaning of a particular neuron is. A very large industry has been established around that paradigm in the last 40 years, with great success in a way. The prejudice is here, that this is all there is: the brain state being thought to be fully described by stating which neurons are 'on' and which are 'off'. That however would be a very poor data structure, which lacks any means of composing complex representations in a structured way from simple elements. A long time ago I have coined for this weakness the term 'binding problem'. My proposal to solve the binding problem is based on the idea that in reality the data structure of brain state has the form of active nets (see Box).

BOX About here

KI: Do you think that necessarily the same algorithm works at each stage of the processing of the brain?

Yes, I do. There are of course differences in rate of change across the brain. For instance, the primary sensory areas have to deal with rapid changes of state whereas in the prefrontal area you have slower state changes. Correspondingly, when talking about computer vision with a moving camera, you have quickly moving images on the one hand and you have stationary representations of the scene on the other hand. So maybe details of mechanisms of organization of brain state change from here to there, but they all will be of the same style and that style has the form of short term stabilization of attractor states. So the visual system is formulated as a dynamic system, where the interactions are regulated such that certain preferred states are stabilized. As the direct visual input gives rise to many alternate hypotheses, the main effect of state dynamics is to achieve the perceptual collapse: restricting the state to those hypotheses that are compatible with each other in terms of learned constraints. So I do believe that on the level of brain state organization there is a dominating type of mechanism and similarly on the memory level.

KI: When you look back to the development of the last 30 years in the field of computer vision, the term 'biologically motivated vision' has had different meanings and also got different degrees of attention within the computer vision community. In particular Marr's theory is seldom mentioned nowadays, although his ideas had a large impact on computer vision in the 80's and 90's. How would you describe the development of computer vision and in particular the area of biologically motivated vision?

I find it impressive that computer vision (which has been developed with little regard to the brain), although functionally not yet being as mature as biological vision, is far superior to anything that the neural community has been able to present. In that sense 'biologically inspired vision' – at least if you ask for a concrete neural formulation – is a complete failure. Computer vision has had tremendous success with a number of particular problems, but it would be very difficult to translate

those systems into neural terms, given the current prejudices about what the neural data structure is.

KI: That is kind of an frustrating insight, isn't it?

Indeed, a very frustrating insight which I, however, see as a motivation to go beyond the present ideas about neural representation.

KI: Do you see a way these two - nowadays pretty much separated - disciplines could profit from each other? Computer vision does not seem to require any input from the neural community.

When I joined the department of computer science at USC in 1988, computer vision was entirely based on the idea that a human engineer constructs an algorithm and this algorithm performs vision. The field was very averse to any idea of learning. This has profoundly changed; computer vision is now very much driven by learning mechanisms. I think this may be attributed to the influence from the biological side.

KI: When you look at computer vision today, what do you think are the main challenges in vision research?

I think one can see computer vision as a development in which a dozen or two dozen defined problems have been at the focus, and each of these problems has been brought to a certain maturity in terms of algorithmic formulation. I think now is the time to integrate all these different functionalities into one coherent system. One cannot solve these individual problems such as edge detection, motion detection or surface shape extraction without simultaneously solving all the others as well in one framework of mutually supportive subsystems. Hence the main task I see nowadays in computer vision is the formulation of a coherent architecture in which these different functionalities can be linked as a coherent system. I think the time has come for that and we have the computing power to realize it. It is by now clear that a pedestrian solution to the vision problem is way beyond anything we can afford economically. If we do not come up with a framework that makes system construction easy and efficient, we will never get artificial vision. This is the problem, I would suggest, to focus on.

KI: The progress in computer vision came together with insights into the problems that had to be solved, a lot of explicit engineering making use of the regularities that are underlying these problems. This is just the opposite of saying: there is one generic algorithm and one generic data structure to solve all problems. How do you bring these two sides together?

I think it is time to find commonalities between these different sub-systems, find the regularity behind them. Again, it is the brain that proves to us that there is such regularity. If you look through a microscope you find a very homogeneous underlying structure. One regularity of vision is that much of it can be formulated in terms of two-dimensionally extended fields of local variables, like depth profile, motion pattern, illumination, local texture and so on. Another one is that interactions between these different features are mainly in terms of point correspondences – those features that talk about the same point in the outside world should interact most intensely. Another common aspect is that on the basis of direct input each of those variables – depth, illumination, albedo and so on – is subject to great ambiguity. So you need for all of them a kind of probabilistic machinery that can handle this uncertainty and reduce it through constraints acting between them, constraints for example in terms of the kinematics of surface point motion of a rotating solid object.

And then you need, of course, memory structure – stored patterns that are familiar to you: you know what a circle is, you know what a plane surface is – lots of structure that you can recognize in the raw feature distributions and which help to disambiguate the features. So I think what we are up to is some kind of a dynamic framework, which talks about features, about ambiguity, about constraints and high level patterns which can be described in a coherent fashion. I think the task at hand is to find this coherent framework.

KI: So you want to find somehow a more simple description of what so far has been designed in an engineering way?

Yes, the present game in computer vision is for an individual student to find an alternative to existing algorithms and to invent his or her own new data structure and algorithm. Hence a premium is very much in variety. And I think we now need to emphasize homogeneity: We have to look at arrays of different algorithms and try to find and formulate their commonalities.

KI: A recent review paper [4] on the human visual system which is actually an update of earlier papers with similar intention (see, e.g., [7]) outlines the idea of icons of increasing abstraction and spatial extent within different stages of a hierarchical processing scheme (see also Figure 2 in [5] in this issue). Do you think such a concept of the visual cortex is in any way helpful, or is it maybe even misleading, since it simplifies processes too much? In particular V1 and V2 are very large areas compared to other areas in the visual cortex. In the above-mentioned review, they are described as some

kind of 'Zoo of Features'. But does that make sense for these areas?

The expression zoo is appropriate for V1 and V2 already for the reason that these areas form a kind of gateway, through which all visual sub-modalities (color, texture, motion etc.) are to be channeled. And then there is the remarkable fact that the number of neurons increases by a large factor from the geniculate body to V1 (factors of 30 or 50 are quoted sometimes), creating enormous redundancy. This redundancy is very likely used to compress visual information with the help of codebook elements for local texture. These encode statistically dominant patterns in the visual input.

The miracle of vision is that as a baby you are exposed to a number of scenes and then, at a later time, you are exposed to a new environment and you can describe that in all detail with the machinery you have learned in the past. This is, of course, only possible by restricting the range of configurations of basic elements to certain dominating patterns, which need to be represented in the system, like depth discontinuities and so on and so on – in one word, the zoo of which you speak.

KI: Do you think these descriptors are to a large degree genetically coded or do you think they are learned from the statistics of the visual input.

I think what evolution has encoded in genetic terms is an initial state of the system, which is surprisingly complex already. There are these precocious animals, that have to run in the first minutes of their life and have to be able to see the ground and their mother, so that they can follow her. So a lot is already present at birth, probably not due to a process that is engineered in technical terms, by blueprint, but rather by controlling a process of self-organization. What we know about ontogenesis is very much in a style in which there is a growing dynamic system with the genes just handling control parameters.

There is an initial state, but we know from infant vision, that a lot of complicated things develop over a period of months or years, like for example handling depth discontinuities. What do you make out of this strip of texture that you see with the one eye but not the other, since it is occluded? How does that tell you depth ordering? These things seem all to be learned. So the correct view is probably that there is an underlying architecture, an architecture that is set to an initial state which has already some regularity in it, and under the influence of visual input it differentiates over the first ten years of our experience.

KI: I would like to become a bit more concrete here. Would you for example say that the area sizes and the receptive field sizes in the different areas are basically predetermined?

Yes, I would think so. What we know is that the different areas are genetically programmed. The basic topological connections between them seem to be programmed. The properties of the processes put out by neurons, dendrites and axons and their sizes seem to be programmed and that pretty much determines the visual areas, the gross connectivity and the size of receptive fields.

KI: Do you think that the content of the receptive fields, the actual features, are learned or do you think they are also hardwired? When you talk about, for example, edge detection and depth extraction and so on?

The basic layout of the areas is ready at birth. I would not use the term hardwired. It has been developed without external input, even including orientation specificity. This is a fact that has unnerved me for years in my life and only recently I have come to terms with it (see [3]). But then the precise shapes of receptive fields – let them be Gabor-like – seem to very quickly develop under the influence of visual input.

KI: Recently deep hierarchical networks have received a big deal of attention. The wave of neural networks seemed to have lost its power in the early 90's by the awareness of the bias/variance dilemma pointed out by Geman, Bienenstock and Doursat in their paper from 1995 [2], who showed that 'exorbitant data' would be needed to train these networks once they exceed a certain size. Nowadays 'Big Data' is available. Hence what maybe was imagined as 'exorbitant data' in the 80s and 90's which would have been required for training ANNs is available today and is also useable with modern computers. Do you think that 'Big Data' will solve the vision problem?

I think that the main conclusion that can be drawn from Geman et al's bias/variance dilemma paper is that you do need a bias when you want to learn from extensive input data. High order statistics is a monster you cannot kill by raw learning. You need a potent bias that tunes the brain to the world and I think that the correct bias has not been found yet by the neural community. This comes back to my statement about the data structure accepted by the neural community, which is impotent, is not powerful enough to describe complex situations. So, I do believe that you can do headway with huge processing power and huge numbers of input patterns – I hear that these days there are billions of images drawn into the computer to extract statis-

tics – but it is sort of running up vertical walls. This will not conquer computer vision in its entirety. What has been achieved in these deep layer structures so far is very specialized functionality, mainly classification of different object types, but in other ways it does not compare in any way to human performance. So I think this is almost beating a dead horse. With a proper data structure however, a proper architecture, learning will turn out to be a breeze to the extent that finally - with enough pre-structuring – agents will be able to learn from individual inspection new kinds of objects or patterns.

KI: Do you see there a principal difference to audio processing where – for example in speech recognition – the applied models have become more and more simple and data has just increased while leading to quite significant progress in speech recognition?

You can say that audio is a one dimensional signal in time while vision is a two dimensional signal. I think, that vision is more complicated and it also takes a much larger part of our brain than the auditory system. Vision is much more data intensive than language processing. But I have the naive believe that once correctly formulated, both vision and audition will be put on the same fundamental architectural basis and will look at their basic level equally simple.

KI: At the end of the interview, I would like to pose some more general and open questions with the request of brief answers. More than half of the visual cortex of primates is concerned with vision. Do you think that the understanding of vision is the key to the understanding of cognition?

I don't know whether there could be another pathway to understanding cognition. Looking to the future, I believe that cognition will be understood on the basis of vision.

KI: When will machine vision be better than human vision, or will that never be the case? Do you think there is a fundamental obstacle such that that can never happen?

I think we need to turn a fundamental corner. We have to find a common architecture for vision systems and once that is achieved, then artificial vision will get more efficient or more functional than human vision, simply because new kinds of sensors can be coupled to an artificial vision system.

KI: What would you tell new generations of vision researchers, how to deal with the knowledge we have nowadays about biological models?

I think they should continue to take seriously the basic questions, which they have asked themselves before they went into the field, stay with them and continue pursuing them.

KI: I did not fully understand this answer.

Pay attention to basic questions. What is the architecture, what is the mechanism of brain self-organization – a topic that has completely gone out of fashion – what are the basic mechanisms for learning.

KI: So you say that concrete parameters of the brain, such as receptive field-sizes, the number or size of layers of areas etc. is not worth looking at.

There is a big industry coming up with experimental data about the brain. A fashion that is arriving just now is to get at the connectomics of the brain. This is all very valuable, but will need to be interpreted with the help of a conceptual framework. All this experimental data will not force on us the principle of how the machinery works. We have to come up with our own idea how it might work and eventually test it experimentally, and in order for being able to do that, you need this kind of data.

KI: As a very last question, you talked about the data structures and the algorithms that govern the brain processes, all these structures being rather generic. Do you think that there once will be the paper that explains the brain or do you see vision research more as an evolutionary process in which vision systems become better over time through incremental progress of algorithmic insights and increased computer power?

What I see is that a paradigm shift is required; a change in perspective and that can be achieved with a simple paper. The paper probably needs a lot of discussion of the background and introduction and so on in order to be digestible. But the essence can be very short and then, that would open the door to regular progress in vision, which can then unleash the power of all these 10s of thousands of scientists and engineers working on vision to co-ordinate and to solve the vision problem in a relatively short time.

References

1. Bergmann, U., von der Malsburg, C.: Self-organization of topographic bilinear networks for invariant recognition. Neural Computation (23), 2770–2797 (2011)
2. Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural Computation **4**, 1–58 (1995)

3. Grabska-Barwinska, A., von der Malsburg, C.: Perinatal ontogenesis of orientation specificity and maps in primary visual cortex of higher mammals. J. Neuroscience **28**, 249–257 (2008)
4. Krüger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodríguez-Sánchez, A.J., Wiskott, L.: Deep hierarchies in the primate visual cortex: What can we learn for computer vision? IEEE PAMI **35**(8), 1847–1871 (2013)
5. Krüger, N., Zillich, M., Janssen, P., Buch, A.G.: What we can learn from the primate's visual system. Künstliche Intelligenz - Special Issue on Bio-inspired Vision Systems (2015). DOI 10.1007/s13218-014-0345-9
6. von der Malsburg, C.: A vision architecture. NCTA pp. 1–58 (2014)
7. Oram, M., Perrett, D.: Modeling visual recognition from neurobiological constraints. Neural Networks **7**, 945–972 (1994)
8. Wolfrum, P., Wolff, C., Lcke, J., von der Malsburg, C.: A recurrent dynamic model for correspondence-based face recognition. Journal of Vision (8), 1–18 (2008)

BOX: THE DYNAMIC LINK ARCHITECTURE

Data Structure

System: a very large network expresses both system structure and specific memories

State: activity of a sparse subset of the system's nodes and links.

Organization

State Dynamics: activation and deactivation of nodes and links under signal exchange, weakly influenced by input.

System Evolution: plastic change of connections under the influence of state activity, turning the system into a superposition of regular network structures (see Figure).

The active networks that constitute system state are formed as attractors of network dynamics and have "regular net structure": sparse networks with an abundance of cooperative loops (alternate pathways between nodes), supporting prediction of one signal by others. Active nodes are interpreted as active features or feature values, while active links serve to bind nodes into structured representations.

Application to Vision

Visual structure is represented by two types of net structures (see [6] for somewhat more detail):

Planes: two-dimensionally extended nets with nodes representing local features and links representing their neighborhood relationships in the image plane. A combinatorially rich multiplicity of 2D nets are embedded in the system, locally amounting to codebooks of feature patterns (textures). Separate planes represent different modalities (gray-level, texture, color, depth, motion, reflectance, illumination, ..). Different modality planes are connected by links that implement consistency constraints between feature values. Visual input usually activates mutually exclusive alternate feature value units ("hypotheses"), which inhibit each other. This and the consistency constraint connections drive the perceptual collapse towards a consistent interpretation of the input.

Projections: nets that form homeomorphic fiber projections between planes (linked nodes in one plane connect to linked nodes in another). When visual input patterns move or shift, projections connect shifted versions with each other. This is the basis for motion tracking and for invariant representation.

For a face recognition system exemplifying state dynamics see [8]. For a model of the growth of control structures for dynamic projections see [1].

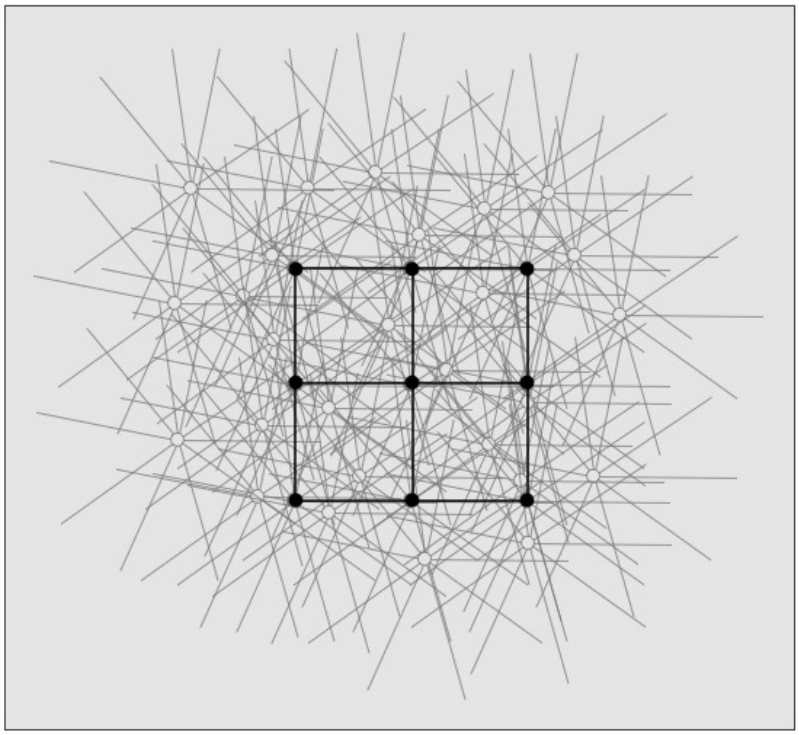

Fig. 1 The data structure of the brain as overlay of structured nets (highly schematic). **Gray Background:** System network, a seemingly random tangle of connections, although in reality a superposition of sub-networks of regular net structure. **Solid Foreground:** System state, active nodes and links, forming a regular net structure with alternate pathways between nodes. Two-dimensional planes with local connections and homeomorphic fiber projections between such planes are examples of regular net structures.