Paradigmatic Revolutions in Computer Vision

Andreas Kriegler[1,2]
[1]AIT Austrian Institute of Technology [2]TU Wien
krieglerandreas@gmail.com

Abstract

Kuhn's groundbreaking Structure divides scientific progress into four phases, the pre-paradigm period, normal science, scientific crisis and revolution. Most of the time a field advances incrementally, constrained and guided by a currently agreed upon paradigm. Creative phases emerge when phenomena occur which lack satisfactory explanation within the current paradigm (the crisis) until a new one replaces it (the revolution). This model of science was mainly laid out by exemplars from natural science, while we want to show that Kuhn's work is also applicable for information sciences. We analyze the state of one field in particular, computer vision, using Kuhn's vocabulary. Following significant technology-driven advances of machine learning methods in the age of deep learning, researchers in computer vision were eager to accept the models that now dominate the state of the art. We discuss the current state of the field especially in light of the deep learning revolution and conclude that current deep learning methods cannot fully constitute a paradigm for computer vision in the Kuhnian sense.

1 Introduction

Kuhn's seminal Structure of Scientific Revolutions (Kuhn, 1970), which he himself simply referred to as Structure, marked a shift on how science is perceived not just by the larger public sphere but by scientists themselves. From seeing science as the ethereal pursuit of knowledge and truth, undertaken by scientists who uphold themselves to strict principles, follow rational judgements and are untouched by trends or personal benefits, to a more honest and realistic description, that sees scientists as human beings in a much larger societal context and less as infallible truth-seekers. Kuhn's Structure also introduced and cemented the concept of paradigms into scientific discourse across the fields. Scientific fields and practitioners working under a paradigm, or in normal science, have a clear world-view through which they perceive their field, a set mode of operation with a number of proven models available to them and engage in puzzle-solving activities for minuscule advances.

The recent revolution in artificial intelligence (AI) driven by deep learning (DL) has already brought a significant number of new technologies for humankind. But DL has not only introduced self-driving cars, enabled mobile robots, advanced speech translation and recognition algorithms, it is also beginning to fundamentally change the way scientists operate. In this work we study the strong impact DL has had on the computing field of computer vision in recent years. Particularly we observe how the field is at a paradigmatic crossroads right now due to the DL revolution.

The remainder of the paper is structured as follows: We first briefly explain Kuhn's idea of paradigmatic science. We then shed light on the recent revolution in AI through deep learning. We show how this impacted the computer vision field, before giving concluding remarks.

I Can't Believe It's Not Better Workshop at NeurIPS 2022.

2 Kuhn's Structure and Paradigms

Kuhn's Structure (Kuhn, 1970) introduced an entirely new vocabulary to the scientific community. His concept of paradigms was especially present — it has been noted that he used paradigm in over twenty distinct ways (Masterman, 1970). He later in his postscript explicitly gave two meanings, firstly as an exemplar or concrete scientific achievement of such magnitude that it defines all subsequent research in a scientific discipline. In the other sense as a disciplinary matrix, i.e. a cluster or constellation of problems, assumptions, beliefs, values, techniques and methods shared by researchers of any given community (Kuhn, 1996). Kuhn argued that science consists of 95% normal or boring science, where scientists tackle puzzles that they know have an attainable solution, can envision the solution and then set out to solve it. They work following a paradigm, the mode of operating of that field at that time. A number of distinct stages can then be observed in the cycle of the sciences according to Kuhn:

1. **Pre-Science**: There is no real consensus and (mathematical) fundamentals are often debated. Activities are diverse, a large number of theories exist, many theories are tailored to fit a certain subset of observations.
2. **Normal Science**: The most prevalent stage, a paradigm has been established, consensus reached, there is little criticism of the theory, some anomalies occur but are brushed aside as such.
3. **Crisis**: The number of anomalies can no longer be simply explained-away. The prevailing theory is under attack from all sides. It requires extraordinary science of gifted individuals, often young or from other fields, to resolve.
4. **Revolution**: A new paradigm has been shown to have merit and its slow adoption begins, but largely not because of logically sound justifications but more so psychological whims.

For Kuhn, a scientific field only occupies the pre-science or pre-paradigm stage once, shortly after its inception. Once a paradigm has been established, normal science takes place until there are too many inexplicable observations or anomalies for the current paradigm and the field is at the stage of crisis. Sooner or later a new paradigm is developed and the slow adoption of it begins. During and after a revolution, Kuhn realized that not everyone accepts the new paradigm immediately. In fact he dryly states that older theories or paradigms often only die with their proponents themselves. We can see that for some time in this revolution stage, the field is also marked by strong disagreements over the new theory, perhaps due to insufficient understanding, perhaps due to the new vocabulary introduced. The activity is disorganized and no consensus is reached.

Although Kuhn's ideas have largely stood up to the test of time so far one common criticism of Kuhn's work was his overemphasis on using examples from the natural sciences, specifically physics, as at the time of writing physics had been the leading science for a couple of centuries and Kuhn himself was a trained physicist. Unlike natural sciences where observations of the real world are still the driving factor, in computer science, the object of interest itself has been created by humans. But this also means that all problems that are to be solved were created by humans. This is surely in the spirit of Kuhn treating normal science as "puzzle-solving"—now we are even able to create our own puzzles! As such it could be argued that the entire field of computing always has to be at a stage of normal science. In practice we know that computing is a multifaceted discipline and as such is intimately linked to and partially responsibly for advances in many other engineering and scientific fields. It is then perhaps more appropriate to ask what Kuhnian stage the specific computerized field is in.

Analogously, we construct our main argument of this paper: Deep learning is a strong and successful mathematical and computational paradigm, capable of general function approximation and modeling of probability distributions. Nevertheless, the science of computer vision or the computational perception akin to the human mind, concerned with complex, interdisciplinary problems far beyond a single function approximation, cannot entirely rely upon DL as a paradigm and as such, given the lack of suitable alternatives, is in a state of crisis.

3 Deep Learning for Large-Scale Number Crunching

At a similar point in time as Kuhn's Structure was written, namely the 1950s and 60s, the field of cybernetics or cybernetic devices made its first steps. It promised autonomous agents and general artificial intelligence within just a few years but failed to deliver on this promise. It was soon afterwards downplayed as a less glamorous discipline than symbolic AI (interpretable actions based on rules and knowledge with adequate representations) that is only concerned with the study of self-organizing artificial neural networks (Cariani, 2010): algorithms modeled after our mechanical understanding of the brain. Deep neural networks, or multi-layer perceptrons (algorithms for binary classification) with more than one hidden or intermediate layer, were already shown in 1965 (Ivakhnenko & Lapa, 1965) and the idea of stacking layers spatially for increased receptive fields (region in the input space affecting a feature) à la Convolutional Neural Networks (CNNs) at least as early as 1980 (Fukushima, 1980). Nevertheless, a number of other pieces were required to arrive where we are today at the DL revolution. As time went on following the 1960s, and rule-based systems for AI were shown to be ineffective and too cumbersome (Brooks, 1996), mathematical advances for statistical data analysis continuously contributed ideas and concepts to arrive at machine learning (ML), or deep learning with deep neural network architectures, which is largely synonymous with today's AI. After two AI winters, 2012 was finally the start of the current AI spring also called deep learning revolution (Schmidhuber, 2015). Across multiple fields, namely for biology (Dahl et al., 2014), speech recognition (Dahl et al., 2012), and classification in computer vision (Krizhevsky et al., 2012) the benchmarks were broken by deep neural networks. DNNs have proven to be extremely capable of general function approximation and the idea of gradient learning on large quantities of data powered by parallel-processing hardware is now prominent in many fields of computer science, specifically data science, in natural science (Vanderplas et al., 2012) and social science (Grimmer et al., 2021). Gradient based learning or converging towards some local minimum of a cost function in combination with back-propagation is currently the only possibility to make machine learning algorithms learn that is also tractable in practice although this could change in the future.

If one desires aforementioned exemplars to also have explanatory capabilities, that is theories that not only describe but explain the phenomena they are applied to, current learning-based methods cannot constitute a paradigm, although the emerging field of explainable AI (Samek et al., 2019) seeks to develop and study methods to make black box models interpretable to the user and advocates the use of explainable learning algorithms. In the following sections we show how gradient-based data-hungry deep learning has pushed computer vision into this limbo of relying heavily on those models yet they cannot be considered a paradigm on their own right.

4 A Brief History of Computer Vision

Visual perception enables humans to take in large amounts of information from their environment and process this information to understand their surroundings and form decisions for actions. The field of computer vision is dedicated to artificially reproducing this perception capability, a necessity for embodied AI such as mobile robots. It was originally perceived to be a simple problem but over the following decades it was realized that computational perception is a very complex field with many connections to neuropsychology, gestalt principles, optics, colour theory and other fields. Marr's 1982 posthumously published book *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (Marr, 1982) is regarded as a classic for cognitive scientists and computational visionists alike. Any theory would need to work on different levels of analysis, the computational, algorithmic and hardware implementation level independently. As mentioned in the afterword, if the book were to be rewritten today Marr would certainly dedicate a chapter to learning approaches. Marr's bold computational approach to studying vision, with his idea of using tokens or low level features such as oriented edges for what he called the primal sketch, and using these tokens for subsequent processes such as object recognition excited a number of researchers in the field. For a long time hand-crafted features were the driving force of research, that is expert-built algorithms to extract specific features such as corners and edges from images and much effort was put into designing the optimal filters. Interest only gradually shifted away, with the realization that this approach is too rigid for real world images. Let us now observe some more concrete examples how the word paradigm has appeared in Computer Vision (CV) literature up to now.

The seminal work of RANCSAC (random sample consensus) from 1981 promised and delivered a "paradigm for model fitting with applications to image analysis [...]" (Fischler & Bolles, 1981). It is still used today as a way to discard outliers in a set of observations to more accurately fit a parameterized mathematical model to a set of observations. In the ten years later published review paper for robust regression methods in CV (Meer et al., 1991), RANSAC is also continuously referred to as a paradigm. It stands to reason, that RANSAC really is the only work in CV that could constitute a paradigm in Kuhn's sense. At the very least as an exemplar and possible also as a disciplinary matrix.

Ten further years later, De La Torre & Black, 2001 argued that "the automated learning of low-dimensional linear models from training data has become a standard paradigm in CV". It has always been noted that images are very high-dimensional objects (a FullHD image for example is just a 2D grid-based ordering of data points in 1920×1080 or 2073600 dimensional space). Many of those dimensions strongly correlate and methods of learning low-dimensional embeddings via dimensionality reduction such as principal component analysis have proven successful. Here, paradigm is more used in the sense of a disciplinary matrix and less as an exemplar because no single work or achievement possibly encompasses the entirety of linear models.

In Klette & Reulke (2005), a number of paradigm shifts are discussed for modeling 3D scenes, from the related fields of photogrammetry, remote sensing and computer vision. They use Kuhn's original definition of paradigm from the 1962 version of Structure but quickly go on to derive their own definition as a "paradigm shift characterized by gradual transitions in a period of several years, leaving basic knowledge unchanged, but adding completely new opportunities" (Klette & Reulke, 2005). This describes an evolutionary rather than revolutionary sequence, which is different from Kuhn's original notion. Finally, post DL revolution in the year 2015, Ros et al. propose an offline-online perception paradigm for autonomous driving. Their work, while certainly useful at the time, cannot be understood as a paradigm in either of Kuhn's definitions, since it is has not lead to major changes in research direction or methodology.

So as time goes on, the rigor with which Kuhn's concept and vocabulary is applied seems to diminish. Today, although other sensing modalities exist, images are still the main workhorse of many CV algorithms with the general motivation that lower-dimensional representations of an image still hold much information of the images content. CNNs as variants of DNNs provide such feature representations and this can be seen as a natural continuation from earlier feature engineering or learning of lower-dimensional representations.

5 The State of Modern Computer Vision

As mentioned earlier, DNNs have taken loose inspiration from neuroscience but our current understanding of the brain is too limited to create a computational copy. For example, neuroscientists indicate that real neurons do not calculate their output by simply summing up weighted inputs. Similarly in the opposite direction, CNNs have done little to explain how our brain processes percepts, because our technical representation via discrete 2D images is nowhere near how human perception, with the inclusion of memory, works. As such, since its inception there were no generally accepted theories in vision to answer most cognitive problems although the ideas of (Marr, 1982) provided some answers to very simple perception phenomena. Today there are no accepted theories either. So it seems that the ideas of deep learning are no use in helping us explain human vision from a computational point of view, the original goal of CV research.

One might argue that modern CV research is not concerned to explain human perception. Perhaps CV is more of an engineering field, because so many useful things are directly enabled by it, think of mobile robots, self-driving cars or motion capture systems. Perhaps gradient-based learning and CNNs can be considered a paradigm in this application-focused, pragmatic and industry-driven scope? But even here the discourse is too heated, there is too little consensus, there are too many problems pointed out by people all across the field. There is the problem of brittleness, that is CNN models can be easily tricked with nonsense patterns that the machines recognize as familiar objects or minor perturbations that alter the classification (Zhou & Firestone, 2019). The problem of data dependency which naturally means models as theories are only partially transferable to other observations, for example, most machine learning models applied for self-driving scenarios struggle with differences in appearance due to seasonal changes, when everything is blanked by a snow-white cover. There are ecological and economical problems due to energy consumption since we are receiving diminishing returns from training very large-scale models, the training of which produces carbon-dioxide emissions comparable to some U.S. cities (Thompson et al., 2021). The number of theorists and theories (models) is seemingly similar with new architectures populating the zoo every minute - there are 2000 different models available on the popular website "Papers with code" at the time of writing. There are ethical considerations which stand in the way of adoption in many real-world cases, for example the question of accountability for the decisions of a self-driving car in case of accidents. There is the problem of model opaqueness which could result in discoveries that the scientist will have a hard time understanding (Boge, 2022). Finally the natural world is inherently uncertain and many of these models struggle with the high variance in real world scenes, which is particularly noticeable in robotic applications (Sünderhauf, 2018). All this means that the algorithms might perform well on laboratory test datasets, but fall short in many applications in the real world. Surely all these are symptoms a mature vision paradigm should not have. Thus, learning-based methods alone are also insufficient in creating a perception system for the real world.

Nevertheless, one argument can be made that DL models constitute a paradigm in CV, at least in one sense of the word. The day-to-day problems in CV show many characteristics of puzzles in the Kuhnian sense: these problems are solvable given the current paradigm, the solution can be envisioned and attained by the researcher. The largely accepted way for puzzle-solving is training back-propagation driven deep neural networks on labeled training data – this also constitutes the main activities CV researchers conduct on a daily basis. Because it has been so successful in delivering models that break many visual dataset benchmarks, which is still the dominant way to mark progress in the field and thus the primary incentive for new publications, it has found widespread adoption amongst researchers and acts as a disciplinary matrix. It is now largely impossible to do computer vision research without including some concepts from DL but these neural network architectures should be seen only as tools to an end and never as the goal itself. In this way, CV research will be able to progress towards proper computational perception for intelligent systems.

In summary, while the application of deep neural networks for computer vision tasks has yielded substantial improvements on laboratory benchmarks, this cannot realistically be considered a paradigm for computer vision especially in light of its original goals of providing computational explanations for the processes of visual perception in cognitive sciences but also for the more application oriented side of modern CV.

6 Conclusions

We analyzed a potential beginning paradigmatic shift in the Kuhnian sense for the computing field of computer vision induced by the deep learning revolution of the last decade. While computer vision was part of this revolution from the beginning, there are still many unresolved fundamental issues—brittleness, explainability, and generalization of the used neural networks—which we believe are necessary to be satisfactorily addressed by a mature paradigm. As such Computer Vision is at a paradigmatic crossroads: If one does not accept DL as a paradigm, Computer Vision is still in the Pre-Science stage in Kuhnian terms. If we do accept gradient-based learning as a paradigm, it would seem that Computer Vision is heading towards a Crisis. The question is whether deep learning alone can really solve the challenge of perceiving and understanding vision or whether it must be extended by other, yet to be discovered, means.

References

[1] Kuhn, T.S. (1970) *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

[2] Masterman, M. (1965) Criticism and the Growth of Knowledge: The Nature of a Paradigm. *Proceedings of the International Colloquium in the Philosophy of Science*, London.

[3] Kuhn, T.S. (1996) *The Structure of Scientific Revolutions (3rd edition).* Chicago: University of Chicago Press.

[4] Cariani, P. (2010) On the Importance of Being Emergent. *Constructivist Foundations* **5**(2):86-91.

[5] Ivakhnenko, A. & Lapa, V. (1965) *Cybernetic predicting devices.* New York: CCM Information Corp..

[6] Fukushima, K. (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* **36**(4):193-202.

[7] Brooks, F.P. (1996) The Computer Scientist as Toolsmith II. *Communications of the ACM* **39**(3):61-68.

[8] Schmidhuber, J. (2015) Deep learning in neural networks: An overview. *Neural Networks* **61**:85-117.

[9] Dahl, G.E. & Jaitly, N. & Salatkhutdinov, R. (2014) Multi-task Neural Networks for QSAR Predictions. *arXiv:1406.1231.*

[10] Dahl, G.E. & Yu, D. & Deng, L. & Acero, A. (2012) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing* **20**(1):30-42.

[11] Krizhevsky, A. & Sutskever, I. & Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* **60**:84-90.

[12] Vanderplas, J. & Connolly, A.J. & Ivezic, Z. & Gray, A. (2012) Introduction to astroML: Machine Learning for astrophysics. *Conference on Intelligent Data Understanding (CIDU)*, pp. 47-54.

[13] Grimmer, J. & Roberts, M.E. & Stewart, B.M. (2021) Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science* **24**:395-419.

[14] Samek, W. & Montavon, G. & Vedaldi, A. & Hansen, L.K. & Müller, K.R. (2019) *Explainable AI: interpreting, explaining and visualizing deep learning.* Springer Nature.

[15] Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Cambridge: The MIT Press.

[16] Fischler, M.A. & Bolles, R.C. (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* **24**(6):381-395.

[17] Meer, P. & Mintz, D. & Rosenfeld, A. & Kim, D.Y. (1991) Robust regression methods for computer vision: A review *International Journal of Computer Vision* **6**(1):59-70.

[18] De La Torre, F. & Black, M.J. (2001) Robust principal component analysis for computer vision. *International Conference on Computer Vision (ICCV)*, pp. 361-369.

[19] Klette, R. & Reulke, R. (2005) Modeling 3D Scenes: Paradigm Shfits in Photogrammetry, Remote Sensing and Computer Vision. *Communication and Information Technology Research Technical Report 155.*

[20] Ros, G. & Ramos, S. & Granados, M. & Bakhtiary, A. & Vazquez, D. & Lopez, A.M. (2015) Vision-based offline-online perception paradigm for autonomous driving. *Winter Conference on Applications of Computer Vision (WACV)*, pp. 231-238.

[21] Zhou, Z. & Firestone, C. (2019) Humans can decipher adversarial images, *Nature communications* **10**(1334).

[22] Thompson, N.C. & Greenewald, K. & Lee, K. & Manso, G.F. (2021) Deep Learning's Diminishing Returns, *IEEE Spectrum.*

[23] Boge, F.C. (2022) Two Dimensions of Opacity and the Deep Learning Predicament, *Minds and Machines* **32**, pp. 43-75.

[24] Sünderhauf, N. & Brock, O. & Scheirer, W. & Hadsell, R. & Fox, D. & Leitner, J. & Upcroft, B. & Abbeel, P. & Burgard, W. & Milford, M. & Corke, P. (2018) The limits and potentials of deep learning for robotics. *International Journal of Robotics Research* **37**(4-5).