

SPRINGER BRIEFS IN COMPUTER SCIENCE

Alessandro Betti

Marco Gori

Stefano Melacci

Deep Learning
to See
Towards New
Foundations
of Computer
Vision



Springer

SpringerBriefs in Computer Science

Series Editors

Stan Zdonik, Brown University, Providence, RI, USA

Shashi Shekhar, University of Minnesota, Minneapolis, MN, USA

Xindong Wu, University of Vermont, Burlington, VT, USA

Lakshmi C. Jain, University of South Australia, Adelaide, SA, Australia

David Padua, University of Illinois Urbana-Champaign, Urbana, IL, USA

Xuemin Sherman Shen, University of Waterloo, Waterloo, ON, Canada

Borko Furht, Florida Atlantic University, Boca Raton, FL, USA

V. S. Subrahmanian, University of Maryland, College Park, MD, USA

Martial Hebert, Carnegie Mellon University, Pittsburgh, PA, USA

Katsushi Ikeuchi, University of Tokyo, Tokyo, Japan

Bruno Siciliano, Università di Napoli Federico II, Napoli, Italy

Sushil Jajodia, George Mason University, Fairfax, VA, USA

Newton Lee, Institute for Education, Research and Scholarships, Los Angeles, CA, USA

SpringerBriefs present concise summaries of cutting-edge research and practical applications across a wide spectrum of fields. Featuring compact volumes of 50 to 125 pages, the series covers a range of content from professional to academic.

Typical topics might include:

- A timely report of state-of-the art analytical techniques
- A bridge between new research results, as published in journal articles, and a contextual literature review
- A snapshot of a hot or emerging topic
- An in-depth case study or clinical example
- A presentation of core concepts that students must understand in order to make independent contributions

Briefs allow authors to present their ideas and readers to absorb them with minimal time investment. Briefs will be published as part of Springer's eBook collection, with millions of users worldwide. In addition, Briefs will be available for individual print and electronic purchase. Briefs are characterized by fast, global electronic dissemination, standard publishing contracts, easy-to-use manuscript preparation and formatting guidelines, and expedited production schedules. We aim for publication 8–12 weeks after acceptance. Both solicited and unsolicited manuscripts are considered for publication in this series.

**Indexing: This series is indexed in Scopus, Ei-Compendex, and zbMATH **

More information about this series at <https://link.springer.com/bookseries/10028>


Alessandro Betti · Marco Gori · Stefano Melacci

Deep Learning to See

Towards New Foundations of Computer
Vision

 Springer

Alessandro Betti
Université Côte d'Azur, Inria, CNRS,
Laboratoire I3S, Maasai team
Nice, France

Marco Gori 
Department of Information Engineering
and Mathematics (DIISM)
University of Siena
Siena, Italy

Stefano Melacci
Department of Information Engineering
and Mathematics (DIISM)
University of Siena
Siena, Italy

ISSN 2191-5768 ISSN 2191-5776 (electronic)
SpringerBriefs in Computer Science
ISBN 978-3-030-90986-4 ISBN 978-3-030-90987-1 (eBook)
<https://doi.org/10.1007/978-3-030-90987-1>

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To All People Who Love To Ask Questions

Preface

DEEP learning has revolutionized computer vision and visual perception. Among others, the great representational power of convolutional neural networks and the elegance and efficiency of backpropagation have played a crucial role. By and large, there is a strong scientific recognition of their popularity, which is very well deserved. However, as yet, most significant results are still based on the truly artificial supervised learning communication protocol, which sets in fact a battlefield for computers, but it is far from being natural. In this book, we argue that, when relying on supervised learning, we have been working on a problem that is—from a computational point of view—remarkably different and like more difficult with respect to the one offered by nature, where motion is in fact in charge for generating visual information. Could not be the case that motion is fact nearly all what we need for learning to see? Otherwise, how could eagles acquire such a spectacular visual skills? What else could they grasp from a video to extract precious information for learning? For sure, just like other animals, they do not undergo a massive supervision, but only a reinforcement signal due to their natural interactions with the environment. Current deep learning approaches based on supervised images mostly neglect the crucial role of temporal coherence. It looks like nature did a nice job by using time to sew all the video frames. When computer scientists began to cultivate the idea of interpreting natural video, in order to simplify the problem they remove time, the connecting wire between frames. As a consequence, video turned into huge collections of images, where temporal coherence was lost, which means that we are neglecting a fundamental clue to interpret visual information, and that we have ended up into problems where the extraction of the visual concepts can only be based on spatial regularities.

Based on the underlying representational capabilities of deep architectures and learning algorithms that are still related to backpropagation, in this book we propose that the massive image supervision can in fact be replaced with the natural communication protocol arising from living in a visual environment, just like animals do. This leads to formulate learning regardless of the accumulation of labeled visual databases, but simply by allowing visual agents to live in their own visual environments. We

claim that feature learning arises mostly from motion invariance principles that turns out to be fundamental for detecting the object identity as well as supporting object affordance.

This book introduces two fundamental principles of visual perception. The *first principle* involves consistency issues, namely the preservation of material point identity during motion. Depending on the pose, some of those points are projected onto the retina. Basically, the material points of an object are subject to *motion invariance of the corresponding pixels on the retina*. A moving object clearly does not change its identity, and therefore, imposing an invariance leads to a natural formulation of object recognition. Interestingly, more than the recognition of an object category, this leads to the discovering of its identity.

Motion information does confer not only object identity, but also its affordance, which corresponds with its function in real life. Affordance makes sense for a species of animal, where specific actions take place. A chair, for example, has the affordance of seating a human being, but it can have other potential uses. The *second principle of visual perception* is about its affordance as transmitted by coupled objects—typically humans. The principle states that the affordance is invariant under the coupled object movement. Hence, a chair gains the seating affordance independently of the movement of the person who is sitting (coupled object).

The theory of deep learning to see that is herein proposed is independent of the body of the visual agent since it is only based on information-based principles. In particular, we introduce a vision field theory for expressing those motion invariance principles. The theory enlightens the indissoluble pair of visual features and their conjugated velocities, thus extending the classic brightness invariance principle for the optical flow estimation. The emergence of visual features in the natural framework of visual environments is given a systematic foundation by establishing information-based laws that naturally enable deep learning processes.

The ideas herein presented have been stimulated by a number of questions that we regard of fundamental importance for the construction of a theory of vision. How can animals conquer visual skills without requiring the “intensive supervision” we impose to machines? What is the role of time? More specifically, what is the interplay between the time of the agent and the time of the environment? Can animals see in a world of shuffled frames like computers do? How can we perform semantic pixel labeling by receiving only a few supervisions? Why has the visual cortex evolved toward a hierarchical organization and why did it split into two functionally separated mainstreams? Why top-level visual skills are achieved in nature by animals with foveated eyes thanks to focus of attention? What drives eye movements? Why does it take 8–12 months for newborns to achieve adult visual acuity? How can we develop “linguistic focusing mechanisms” that can drive the process of object recognition?

This book is a humble attempt at addressing these questions, and it is far away from providing a definite answer. However, the proposed theory gives foundations and insights to stimulate future investigations and specific applications to computer

vision. Moreover, the field theory herein proposed might also open the doors to disclose interesting problems in visual perception and capture experimental evidence in neuroscience.

Siena, Italy
August 2021

Alessandro Betti
Marco Gori
Stefano Melacci

Acknowledgements

It is hard not to forget people who have contributed in different ways to shape the ideas that are proposed in this book.

First, the research team in Siena Artificial intelligence Lab (SAILab) has played the most important support for maturing the ideas elaborated in this book. Matteo Tiezzi, Dario Zanca, Enrico Meloni, Lapo Faggi, and Simone Marullo, who are some of the members of the SAILab research team in computer vision, have contributed with their experimental results to shape our ideas and make the theory simpler and more general. There are in fact significant traces of early studies in SAILab from the collaboration with Marco Lippi and Marco Maggini, who contributed to develop the first approaches to learning to see by using motion invariance. In particular, discussions with Marco Maggini, who early discovered a number of slippery issues in the incorporation of motion invariance and clearly identified the major difficulties in carrying out learning in the temporal domain, have been extremely inspiring.

The road that has led to this book certainly crosses some inspiring meeting with Marcello Pelillo and, later on, with Fabio Roli. Marcello ignited my latent passion for unifying and that of looking for invariants in vision. Together with Fabio, years ago, we cultivated the dream of a truly new way of facing computer vision challenges within the “en plein air” framework, which reminds us of painting outdoor—machines which learn directly in their own environment. This is in fact addressed at the end of the book. Most of the comments definitely come from early discussions which began with the Workshop GIRPR 2014.

Oswald Lanz has been for us the main reference and source of inspirations for studies in optical flow, which has subsequently given rise to the two principles of perceptual vision. In particular, the conception of the idea of specific velocities associated with visual features has been originated by his clear presentation of the state of the art in optical flow and in tracking.

The studies by Tomaso Poggio on developing visual features under invariance stimulated very fruitful discussion in connection with the Workshop on “Biologically Plausible Learning” at LOD 2020 and have fueled significantly the development of the theory presented in this book.

During the Ph.D. studies of Alessandro Betti, we benefited from a number of constructive criticisms from Stefano Soatto and Michael Bronstein. Among others, they stimulated the importance of setting up an experimental framework adequate to assess the performance. In the same direction, a discussion with Bastian Leibe on the current state of the art in computer vision has contributed to shape and reinforce the idea of “en plein air” discussed in the Epilogue of the book, thus promoting the fundamental principle of replacing the accumulation of visual databases with virtual visual environments. The studies by Ulisse Stefanelli on the reformulation of the principle of least action in physics were a fundamental source of inspiration for the development of the online formulation of learning reported in this book. The same idea used in mechanics gives rise to online gradient-based learning which is nicely related to classic stochastic gradient descent.

We have been inspired by a number of studies in neuroscience, particularly on the mechanisms behind eye movements. Leonardo Chelazzi’s visit to SAILab was very influential concerning the subsequent development of computational models of focus of attention. We strongly benefited from discussions on the different kinds of eye movements and, particularly, on the supposed inhibition of video transmission during saccadic movements. The collaboration with Alessandra Rufa offered the primary support on the formulation of theory of gravitational attraction of attention, which is also at the basis of the local spatiotemporal model reported in the book. Giuseppe Boccignone provided a rich bibliographic support and stimulated many discussions mostly on the joint role of action and perception and on the vision blurring process in newborns and in chicks.

This work has been partially supported by the French government, through the 3IA Côte d’Azur, Investment in the Future, project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002 and by the HumaneAI European research grant Grant agreement ID: 952026 (University of Pisa and University of Siena).

Contents

1	Motion Is the Protagonist of Vision	1
1.1	Introduction	1
1.2	The Big Picture	2
1.3	Supervised Learning Is an Artificial Learning Protocol	4
1.4	Cutting the Umbilical Cord with Pattern Recognition	5
1.5	Dealing with Video Instead of Images	7
1.6	Ten Questions for a Theory of Vision	9
2	Focus of Attention	13
2.1	Introduction	13
2.2	How Can Humans Perform Pixel Semantic Labeling?	14
2.3	Insights from Evolution of the Animal Visual System	15
2.4	Why Focus of Attention?	18
2.5	What Drives Eye Movements?	21
2.6	The Virtuous Loop of Focus of Attention	27
3	Principles of Motion Invariance	31
3.1	Introduction	31
3.2	Computational Models in Spatiotemporal Environments	33
3.3	Object Identity and Affordance	36
3.4	From Material Points to Pixels	38
3.5	The Principle of Material Point Invariance	40
3.6	The Principle of Coupled Motion Invariance	52
3.7	Coupling of Vision Fields	57
4	Foveated Neural Networks	63
4.1	Introduction	63
4.2	Why Receptive Fields and Hierarchical Architectures?	64
4.3	Why Two Different Mainstreams?	66
4.4	Foveated Nets and Variable Resolution	67

- 5 Information-Based Laws of Feature Learning** 73
 - 5.1 The Simplest Case of Feature Conjugation 73
 - 5.2 Neural Network Representation of the Velocity Field 74
 - 5.3 A Dynamic Model for Conjugate Features and Velocities 78
 - 5.4 Online Learning 83
 - 5.5 Online Learning: An Optimal Control Theory Prospective 84
 - 5.6 Why is Baby Vision Blurred? 86

- 6 Non-visual Environmental Interactions** 89
 - 6.1 Object Recognition and Related Visual Skills 89
 - 6.2 What Is the Interplay with Language? 92
 - 6.3 The “en Plein Air” Perspective 95

- Appendix A: Calculus of Variations** 97

- References** 101

- Index** 105