Advances in Computer Vision and Pattern Recognition

Marius Leordeanu

Unsupervised Learning in Space and Time

A Modern Approach for Computer Vision using Graph-based Techniques and Deep Neural Networks

Springer

Advances in Computer Vision and Pattern Recognition

More information about this series at http://www.springer.com/series/4205

Marius Leordeanu

Unsupervised Learning in Space and Time

A Modern Approach for Computer Vision using Graph-based Techniques and Deep Neural Networks

Springer

Marius Leordeanu
Computer Science and Engineering
Department
Polytechnic University of Bucharest
Bucharest, Romania

ISSN 2191-6586 ISSN 2191-6594 (electronic)
Advances in Computer Vision and Pattern Recognition
ISBN 978-3-030-42127-4 ISBN 978-3-030-42128-1 (eBook)
https://doi.org/10.1007/978-3-030-42128-1

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

Dedicated to my parents who let me be a child, my teachers who let me be a student and my students who give everything a purpose.

In memory of Solomon Marcus (1 March 1925–17 March 2016), Romanian mathematician, writer and guardian angel for many of us. He was a member of the Romanian Academy and Professor of mathematics at the University of Bucharest.

Preface

In *Unsupervised Learning in Space and Time* we address one of the most important and still unsolved problems in artificial intelligence, which is that of learning in an unsupervised manner from very large quantities of spatiotemporal visual data that is often freely available. The book covers our main scientific discoveries and results while focusing on the latest advancements in the field from an original and insightful perspective. The text has a coherent structure and it logically connects, in depth, original mathematical formulations and efficient computational solutions for many different unsupervised learning tasks, such as graph and hypergraph matching and clustering, feature selection, classifier learning, object discovery, recognition, and segmentation in video. The tasks are presented with a unified picture in mind, which puts together and relates at many levels different tasks that converge into the more general unsupervised learning problem. We start from a set of intuitive principles of unsupervised learning, and then gradually build the mathematical models and algorithmic tools that are necessary. We eventually reach a general computational model for unsupervised learning, which brings together graphs and deep neural networks. Overall, the book is deeply grounded in the scientific work we have developed together with our professors, colleagues, and doctoral students at the Robotics Institute of Carnegie Mellon University, Intel and Google Research, the Institute of Mathematics "Simion Stoilow" of the Romanian Academy and the University Politehnica of Bucharest. For our work on unsupervised learning, in 2014 we were awarded the "Gigore Moisil Prize", the highest in mathematics given by the Romanian Academy.

Organization and Features

The book is organized into eight chapters, which take the reader from a set of initial intuitions and common sense principles for unsupervised learning, to different tasks, computational models, and solutions which are introduced and integrated together, chapter by chapter, as follows:

Chapter 1: In the first chapter, we introduce seven principles of unsupervised learning, and then make a brief pass through the subjects covered in the next chapters in strong relation to these basic principles and concepts. Chapter 1 gradually builds a big picture of the book, without covering the very last concepts and models, which are presented in the final chapter.

Chapter 2: In the second chapter, we introduce the problems of graph and hypergraph matching, going from initial motivation and intuition to efficient algorithms for optimization and unsupervised learning. In this chapter, we present the Spectral Graph Matching algorithm, which is later related to the method presented in Chap. 6 for unsupervised object segmentation in video. We also present the Integer Projected Fixed Point (IPFP) method, whose clustering extension (Chap. 3) is later used on hypergraph clustering (Chap. 3), unsupervised feature selection and classifier learning (Chap. 4), and descriptor learning and object discovery in video (Chap. 5). We extensively compare our methods to many other approaches for graph and hypergraph matching and demonstrate a significant boost. We also show how unsupervised learning for graph matching can significantly improve performance.

Chapter 3: In the third chapter, we extend the formulations and algorithms from the second chapter to the task of graph and hypergraph clustering. The two problems are strongly related, and similar models and algorithms can address both. We present an efficient clustering algorithm based on the integer projected fixed-point method from the second chapter. The IPFP-clustering method is then applied to the tasks defined in Chaps. 4 and 5.

Chapter 4: In the fourth chapter, we present an efficient approach to linear classifier learning formulated as a clustering problem. The formulation leads to both feature selection and classification, for which we also provide an unsupervised solution. We introduce the idea of a feature sign and show that by knowing this sign we could learn without knowing the samples' labels. The algorithm used for learning is the same as the clustering-IPFP method from Chap. 3. We compare to many linear classification approaches, including Support Vector Machines (SVM) on the task of video classification and show significantly more powerful generalization power from limited training data.

Chapter 5: In this chapter, we put together all the building blocks presented so far. By following the initial unsupervised learning principles from Chap. 1, we create a fully unsupervised system for object segmentation in video, which learns over several generations of classifiers, using features that start from simple pixels to deep features extracted from the whole image. We show in extensive experiments that our approach is more efficient and accurate than previously published methods on several challenging datasets.

Chapter 6: In the sixth chapter, we continue our exploration of unsupervised object discovery in video and present an original formulation of object discovery as segmentation in a space-time graph in which every pixel video is a node. We introduce a novel Feature-Motion matrix, which couples elegantly motion and appearance and demonstrates that the main object of interest can be discovered as a strong cluster in the space-time graph by efficiently computing the eigenvector

of the Feature-Motion matrix. The mathematical formulation and solution is thus directly related to the spectral graph matching approach from Chap. 1. Our spectral clustering approach to object discovery in space and time is fast and completely unsupervised, while also capable to accommodate any type of pretrained features, if needed. We test on three challenging datasets and outperform other unsupervised approaches. We also boost the performance of other supervised methods, when including their outputs into the Feature-Motion formulation.

Chapter 7: In the seventh chapter, we move to the next level of unsupervised learning over multiple generations. We introduce a teacher-student system, which learns in a self-supervised manner, such that the population of student ConvNets trained at one iteration form the teacher at the next generation. The ideas build upon the material presented so far, but the system is original and shows how it can learn from video, without any human supervision, to segment objects into single images. While the previous chapters are more focused on classical graph models than on deep neural networks, in Chap. 7 we change the focus to deep learning.

Chapter 8: In the last chapter, we merge the graph and neural networks models into a new recurrent space-time graph neural network (RSTG) model, which leverages the benefits and features of both, including the ability to learn over deep layers of features and multiple scale, as well as the capacity to send messages iteratively across both space and time. The RSTG model takes full advantage of the space-time domain and proves its effectiveness on high-level tasks, such as learning to recognize complex patterns of motion and human activities. In the last part, we introduce the Visual Story Network concept—a universal unsupervised learning machine, which learns through multiple prediction pathways, between different world interpretations, by optimizing its own, self-consensus.

Target Audiences

The book is written especially for people with exploratory, naturally curious, and passionate minds, who are daring to ask the most challenging questions and accept unconventional solutions. They could be young students or experienced researchers, engineers, and professors, who are studying or already working in the fields of computer vision and machine learning. We hope the book will satisfy their curiosity and convey them a unified big picture of unsupervised learning, starting from some basic, common sense principles, and developing towards the creation of a fully universal unsupervised learning machine. However, in order to grasp in sufficient detail the complex material covered by the book, readers are expected to have a solid background in mathematics and computer science, and already be familiar with most computer vision and machine learning concepts. To fully understand the more technical parts, which bring together many graph algorithms and deep neural

network models, spread across several computer vision problems, readers are encouraged to master fundamental elements of linear algebra, probability and statistics, optimization, and deep learning. *Unsupervised Learning in Space and Time* is ultimately for people who are determined to find the time and space to learn and discover by themselves.

Bucharest, Romania Marius Leordeanu
May 2020

Acknowledgements

This book would not be possible without my dear professors, mentors, colleagues, and students who have devoted a considerable amount of effort to our collaborative work. They gave me their time, trust, and support for which I am and will always be grateful. We shared hopes, passion, and values. Above all, we shared our love for knowledge and for the fascinating world of vision. Trying to make computers learn to see the world as we learn to see it, is a really hard problem. But the challenge is fascinating, the intellectual reward could be immense, and it is worth all our focus, passion, and spark of creativity. Unsupervised learning is probably one of the ultimate quests in science and technology today, with the potential to open doors towards a territory that is beyond our imagination now. We can only hope that the way we will understand unsupervised learning in the natural world will help in the way we will create and use resources, communicate with each other, and ultimately live our lives.

We should hope that a better understanding of learning, by establishing bridges between many domains, from mathematics and computer science to neuroscience, psychology, philosophy, and art, will take us to a better understanding of ourselves. At the end, it will lead to developing a clearer vision of our own meaning and purpose. To this noble goal, I dedicate this book and towards this goal I hope to grow and direct my efforts.

My deepest and most sincere thoughts of gratitude go first to my professors and mentors who gave me a chance to find my own way in research and grow over the years.

I thank deeply my dearest Ph.D. advisor Martial Hebert, a true model as a researcher and human being, during my Ph.D. years. He guided me with much kindness, wisdom, and light through my very first steps in becoming a computer vision and robotics researcher, at Carnegie Mellon University (CMU). I learned a lot from him and enjoyed immensely doing research together with him. Many of the core ideas and results presented in this book are created and developed together with Martial, under his wise and caring guidance. I also want to express my deepest gratitude to my other great advisor during my first Ph.D. years, Prof. Robert

Collins, with whom it was such a wonderful pleasure to work, create, and debate ideas, and write my first papers on object tracking and unsupervised learning.

I thank deeply to one of my dearest friends and mentors throughout my entire career, Rahul Sukthankar, who I met during my Ph.D. years at CMU. Our journey of working together has been truly amazing. I have learned so much from him and together with him, at both scientific and spiritual levels. Some of the best papers I wrote were together with Rahul. We met in a magic space, governed by the love of science and discovery, where we ask the most interesting questions, take the most daring intellectual journeys, and find the most surprising solutions.

My deepest gratitude and warmest thoughts also go towards my first professor of computer vision, Ioannis Stamos, at Hunter College of the City University of New York. He welcomed me into this wonderful world of vision and gave me the chance to do, for the first time in my life, what I have always dreamed about. My very first contact with research in computer vision and my very first papers were written together with him. Meeting him was a true blessing and thus, I embarked on the journey to find how vision and mind works. From that moment on, 18 years ago, my dream became my profession.

My deepest thanks and warmest thoughts also go to my dear mentor and friend Cristian Sminchisescu, who gave me the extraordinary chance to do top computer vision in my own country, Romania. Under his guidance and wisdom, I have started my career as an independent researcher. I thank him greatly for our fruitful collaboration together. He has been a true model of diligence, motivation, and success.

I also send my deepest feelings of hope, admiration and love to my extraordinary friend and mentor, Leon Zagrean, Professor of Medicine and Neuroscience at Carol Davila University in Bucharest. He is a true pioneer in neuroscience research in our country and a man of very high moral and human values. His profound love for life and endless search for harmony in space and time, has shaped and influenced the core ideas proposed in this book in surprising and wonderful ways.

I also thank deeply my dear colleagues, mentors and friends at the University Politehnica of Bucharest, Institute of Mathematics of the Romanian Academy and University of Bucharest, who actively participated in the creation of this book though numerous discussions, exchange of ideas, projects, and papers that we worked on together. Adina Florea, Ioan Dumitrache, Stefan Trausan Matu, Emil Slusanchi, Traian Rebedea, Elena Ovreiu, Nirvana Popescu, Mihai Dascalu, Lucian Beznea, Vasile Brinzanescu, Dan Timotin, Cezar Joita, Cristodor Ionescu, Sergiu Moroianu, Liviu Ignat, Ionel Popescu, Bogdan Ichim, Bogdan Alexe, Radu Ionescu, Viorica Patraucean, Razvan Pascanu, and Gheorghe Stefanescu are the people together with whom we participate every day in growing a young community of AI researchers, professionals, and engineers in Romania and Eastern Europe and bring a wind of hope and change in this part of the world. I must thank them, from the bottom of my mind and heart for being strong and staying together in our endeavor.

Last, but not least, I thank my dear Ph.D. and graduate students. They are the main purpose of my work as a professor and researcher. This book is made possible by them and ultimately dedicated to them and to the young generation, which they

represent. A large part of the material presented in this book is the result of their effort and passion for knowledge and discovery. Ioana Croitoru, Vlad Bogolin, Andrei Zanfir, Mihai Zanfir, Emanuela Haller, Otilia Stretcu, Alexandra Vieru, Iulia Duta, Andrei Nicolicioiu, Elena Burceanu, Alina Marcu, Dragos Costea, Nicolae Cudlenco, Mihai Pirvu, Iulia Paraicu, and Cristina Lazar participated substantially in the research results reported in this book and also gave me constant feedback during the writing process. I must thank them for their full support and great work. They give me one more strong reason to believe that the next generation will surpass the last and the world will be in good hands.

Source Materials: The material presented in most chapters is based in large part on a number of published works, as follows:

- Chapter 2:
  Leordeanu, Marius, and Martial Hebert. "A spectral technique for correspondence problems using pairwise constraints." IEEE International Conference on Computer Vision (ICCV), 2005.
  Leordeanu, Marius, Martial Hebert, and Rahul Sukthankar. "An integer projected fixed point method for graph matching and map inference." In *Advances in Neural Information Processing Systems* (NIPS) 2009.
  Leordeanu, Marius, Rahul Sukthankar, and Martial Hebert. "Unsupervised learning for graph matching." *International Journal of Computer Vision* 96, no. 1 (2012): 28–45.
  Leordeanu, Marius, Andrei Zanfir, and Cristian Sminchisescu. "Semi-supervised learning and optimization for hypergraph matching." *IEEE International Conference on Computer Vision*, 2011.
- Chapter 3:
  Leordeanu, Marius, and Cristian Sminchisescu. "Efficient hypergraph clustering." In *Artificial Intelligence and Statistics* (AISTATS), 2012.
- Chapter 4:
  Leordeanu, Marius, Alexandra Radu, Shumeet Baluja, and Rahul Sukthankar. "Labeling the features not the samples: Efficient video classification with minimal supervision." In Thirtieth AAAI conference on artificial intelligence (AAAI). 2016.
- Chapter 5:
  Stretcu, Otilia, and Marius Leordeanu. "Multiple Frames Matching for Object Discovery in Video." British Machine Vision Conference (BMVC) 2015.
  Haller, Emanuela, and Marius Leordeanu. "Unsupervised object segmentation in video by efficient selection of highly probable positive features." IEEE International Conference on Computer Vision (ICCV) 2017.
- Chapter 6:
  Haller, Emanuela, Adina Magda Florea, and Marius Leordeanu. "Spacetime Graph Optimization for Video Object Segmentation." arXiv preprint arXiv:1907.03326 (2019).

- Chapter 7:
  Croitoru, Ioana, Simion-Vlad Bogolin, and Marius Leordeanu. "Unsupervised Learning of Foreground Object Segmentation." *International Journal of Computer Vision* (IJCV) 2019: 1–24.
  Croitoru, Ioana, Simion-Vlad Bogolin, and Marius Leordeanu. "Unsupervised learning from video to detect foreground objects in single images." IEEE International Conference on Computer Vision (ICCV) 2017.
- Chapter 8:
  Nicolicioiu, Andrei, Iulia Duta, and Marius Leordeanu. "Recurrent Space-time Graph Neural Networks." In Advances in Neural Information Processing Systems (NeurIPS) 2019.

Funding Sources: Writing of the book and a good part of the research work presented was supported through UEFISCDI, from EEA Grant EEA-RO-2018-0496 and projects PN-III-P1-1.1-TE-2016-2182 and PN-III-P1-1.2-PCCDI-2017-0734.

Endorsements

The book is a pleasure to read. It is timely, in the all-important quest for effective strategies for unsupervised learning. The author describes his work on the past decade, with many new additions, and an interesting philosophical outlook in the last chapter through his Visual Story Graph Neural Network. A key mechanism in the book is the use of his "Integer Projected Fixed Point" (IPFP) method with first-order Taylor series expansion, so the problem can be solved in a cascade of linear programs. Another returning key mechanism is the spectral clustering by "learning" the fit to the principal eigenvector of the adjacency matrix or feature-motion matrix. Many books on deep learning and the quest for unsupervised strategies lack a focus on video analysis, and this book fully compensates this. The author has for many years been a pioneer in this spatiotemporal AI domain. His work has significantly influenced many other works and patents. The chapters are built up in a logical order, increasing in complexity from graph/hypergraph matching and clustering to appearance and motion, and to exploiting large numbers of networks forming "students" creating "teacher" over a number of generations. A realistic approach is offered by allowing a little bit of supervised information at the start of the process, like assigning a small number of "Highly Probable Positive Features" (HPP) and "feature signs". It's impressive to learn how excellent the implementations of the described theories almost always outperform the current state of the art, often by a significant margin, and very often explicitly in the speed domain. The author takes great care in comparing the proposed methods with many current models and implementations. Also, each chapter gives a deep and complete overview of the current literature. The theory is described well, with both a solid mathematical theory, an intuitive story, and with critical discussions, many parameter variations, discussion of pitfalls, and extensive quantitative results. As

the author remarks, there is a lot of work to do. But this book is a significant step forward, with proven effectivity of the many ideas. All in all, an important new book on stage.

Bart M. ter Haar Romeny
Emeritus Professor in Computer Vision and Biomedical Image Analysis
Department of Biomedical Engineering, Eindhoven University of Technology
Eindhoven, The Netherlands

I very much enjoyed the systematization and the logical sequencing of the questions posed and addressed, in analyzing the unsupervised learning process. Marius Leordeanu is a wonderful critical thinker, who takes us on an exciting journey, sharing original insights and drawing us into the story of a beautiful adventure from imaging to seeing things in images and videos. I found his approach appealing and inviting, his enthusiasm contagious, and his arguments solid and well presented and I am sure this book will become a standard reference for researchers in the field of Computer Vision.

Alfred M. Bruckstein
Technion Ollendorff Chair in Science
Technion Israel Institute of Technology
Haifa, Israel

I thoroughly enjoyed reading *Unsupervised Learning in Space and Time*! This is a complex topic of very active research and is challenging to capture in a single volume. Rather than presenting a dry review of recent work, Marius Leordeanu guides the reader along a fascinating journey using a handful of well-formulated but intuitive principles that motivate this perspective on the research space. The authors' infectious love for the subject is evident throughout the book—it will inspire the next generation of researchers to dream bigger dreams!

Rahul Sukthankar
Distinguished Scientist, Google Research
Mountain View, California, USA

Unsupervised Learning in Space and Time outlines a pathway to solving one of the most complex open issues in today's Computer Vision. Just as the completion of big puzzle game should start by sorting the pieces by color, form, size, and orientation before actually trying to fit the pieces together, Marius puts into perspective all the complementary tasks and processing steps involved in unsupervised learning before letting us enjoy the fully assembled picture. We are then left to dream with our minds' eye of the possibilities just opened by the acquired insights.

Emil Slusanschi
Professor of Computer Science
University Politehnica of Bucharest
Bucharest, Romania

In a captivating storytelling fashion, Marius captures the reader's attention by displaying the pieces of a puzzle whose completion leads to a result that seems to be the answer to a key question in the field of Computer Vision: how can we learn in an unsupervised manner.

Bogdan Alexe
Pro-Dean and Associate Professor of Computer Vision
University of Bucharest
Bucharest, Romania

Unsupervised learning is such an important topic in machine learning and computer vision, which has not been fully explored yet. The book takes us in an in-depth exploration of recent techniques and methods proposed for unsupervised learning for several tasks in computer vision. And it is a wonderful learning experience, not only for more experienced researchers, but also for beginners in machine learning. The various proposed models and theories converge at the end, when Marius introduces a new model for unsupervised learning in computer vision, the Visual Story Graph Neural Network, which makes use of classifiers based on weak signals, trained in a teacher-student fashion, and reinforcing each other on several layers of interpretation for an image in time. This model also opens new research opportunities for visual-language tasks, but this will probably be the topic of a different book!

Traian Rebedea
Associate Professor of Computer Science
University Politehnica of Bucharest
Bucharest, Romania

Marius Leordeanu, one of the most prolific and creative researchers of his generation, describes in this book fundamental elements of unsupervised learning for vision, spanning topics from graph matching, clustering, feature selection, and applications, to neural networks. These topics are a necessary read for all who want to acquire a deep understanding of the field. The book follows Leordeanu's research path, making it not only current, but also essential for researchers. Marius' passion, enthusiasm, along with his intuition and insights are all reflected here. This is an important book for computer vision.

Ioannis Stamos
Professor of Computer Science
Hunter College
City University of New York
New York City, USA

Contents

Chapter 1
Unsupervised Visual Learning: From Pixels to Seeing

1.1 What Does It Mean to See?

I am trying to imagine how the world looked like the first time I opened my eyes. What did I see? I am pretty sure that I saw every pixel of color very clearly. However, did I see objects or people? When I looked at my mother, what did I see? Of course, there was that warm, bright presence keeping me close and making me feel good and safe, but what did I know about her? Could I see her beautiful, deep eyes or her long dark hair and her most wonderful smile? I am afraid I might have missed all that, since I did not really know back then what "eyes", "nose", "mouth", and "hair" were. How could I see her as a person, a human being just like myself when I did not even know what I am or what a human being is.

As I could not relate anything to past experiences, there was nothing that I could "recognize" or see as something. There were no "things to see" yet because there was no relationship yet built between the different parts of the image. Pixels were just pixels, and I was probably blind to everything else. By just observing that some groups of pixels have similar colors did not mean that I could see them as being something. There was no past behind anything, to connect things to other things and give them meaning and their "reality" as things.

Now when I look at my mother I can see who she is: there are so many experiences which bind us together. There are many images of her at different stages of my life as well as many images of myself at those times. All those memories are strongly connected together and interact within a story that is coherent in both space and time. All those images of mother, linked through her unique trajectory in my life, make her who she is in my universe. They make her what I see when I look at her.

Now, when I see a single pixel on her face, in fact I see so many things at the same time. I see a skin pixel, a face, and a human pixel. I also see a pixel of my own mother—that special and uniquely important person in my life. Then, at a higher level in space and time, that pixel is also part of us, myself and her, mother and son, which connects all the way back to that very first moment. I also see a pixel belonging to the human race and to life and Earth, as it travels around the Sun. It is only now that I can see all those layers of reality, which required years to form.

My vision is now deeper than it was then, at the very first moment, when it barely existed. All these visual layers containing objects and parts of objects, interactions,

M. Leordeanu, *Unsupervised Learning in Space and Time*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-42128-1_1

and activities exist now and are real at the same time. Back when I opened my eyes for the very first time, the world just started to move. I must have felt a deep, strong urge pulling me towards the unexpected lights, with their surprising and seductive dance of patterns. But what did I know then about what would follow? Everything I see now is our story. It is the visual story of myself as part of the world, as I am growing to see it even better and hopefully become able to imagine how might have been then and what followed next.

1.2 What Is Unsupervised Visual Learning?

While my mother has always been next to me when it mattered most and thought me some of the most important lessons that I needed to know, during the first years of my life, she was definitely not the one who taught me how to see. She did not take every pixel in my visual field to give it meaning at so many levels. That would have been simply impossible. It was my brain who taught me how to see, after learning, mostly in a natural and unsupervised way from many experiences I had in the world. It is my brain that takes every pixel that I perceive and gives it meaning and value. It is my brain that makes it all real for me, copying all these pixels and arranging them on different higher or lower, simultaneous space and time layers of seeing. All those layers, present and past, find consensus into a coherent story. It is in that story that I give a meaning to here and now, it is in that space and time world that I see.

From this fundamental point of view, understanding unsupervised visual learning will help us understand that there is so much more to learn about the world and so much more to see in order to better take care of our world and improve our lives. Unsupervised visual learning in the natural world is, for the same reasons, also fundamental for understanding intelligence, how the mind works, and how we can build truly intelligent systems that will learn to see as humans do and then learn to be in harmony with what we are.

Unsupervised learning is one of the most intriguing problems that a researcher in artificial intelligence might ever address. How is it possible to learn about the world, with all its properties and so many different types of objects interacting in simple or very complex ways? And, on top of that, how is it possible to learn all this without access to the truth? Is that even possible? We do not have an answer to that yet, but what we do know is that children can learn about the world with minimal interventions from their parents. And when parents or teachers do intervene in our educations, they do not explicitly tell us everything. They definitely do not start marking every pixel in our visual field with a semantic label. From the first months of our lives, our brain starts learning, by itself, to arrange and group pixels into regions, to which we also begin to give meanings as we continuously gain experience. At least, at the very beginning, our experiences do not involve complex physical interactions with the world and our visual learning is mostly based on observation.

While interacting with the world is crucial for learning, in this book we are mainly interested in what we can learn *by watching* the world over space and time, as it

reveals itself in front of our eyes. While we briefly touch the topic of interaction in final chapter and discuss how we could take actions and then learn from the outcome, we believe that we should first focus on the more limiting case of learning from space-time data on which we do not have influence over to better understand what are the fundamental limits of unsupervised learning. What assumptions should we make? What type of data do we need to have access to and how much of it is required? What type of classes and concepts can we learn about? What types of computational models could solve unsupervised learning?

As we show throughout the book, these essential questions can reveal universal answers, which could become practical tools for making possible many real-world applications in computer vision and robotics. In the final part of the book, we will adventure ourselves more into the world of imagination and dare to envision a universal computer vision system that could learn in space and time by itself. From the beginning, we will establish a set of general principles for unsupervised learning, which we will demonstrate chapter by chapter with specific tasks, algorithms, and extensive experimental evaluations. At the end, we will show how one could use these basic principles to build a general system that learns by itself multiple layers of interpretation of the space-time world within a single, unified Visual Story Network.

By the end of the book, we will better understand that unsupervised learning is ultimately about learning in the natural world and it cannot happen by itself without input from the vast ocean of data, which obeys physical and empirical statistical laws. Unsupervised learning is not just about fast and efficient algorithms or the architecture of a certain computational model, it is also very much about the world in which that model operates. At the end, we have to reach the level of learning in which we interact with the outside world, of which the learning system itself is part of. From this perspective, unsupervised learning in computer vision becomes a wonderful chance to learn about learning and about ourselves and how we came to discover and "see" the world around us.

1.3 Visual Learning in Space and Time

Vision is so rich that a picture can tell a thousand words. Vision is also our first window into the world and our most important sense. Vision happens in space and time, creating everything we see, from an object at rest or in motion to an activity that takes place in the scene and the whole story that puts all actors and their intricate interactions together. Vision almost has it all, from the simplest pixels of color and common physical objects to the wildest and most profound imagination, in an attempt to create and reflect the world in which we live in. As a consequence, vision must take into consideration the physical and empirical statistical laws, which give the natural world coherence and consistency in space and time. Therefore, it has to build upon certain grouping properties and statistical principles, which reflect such consistency. We must understand these principles and use them in building computational models

of learning if we want to have a real chance to learn, in an unsupervised way, in the wild.

There is a certain advantage in thinking in space and time. Everything that exists and happens around us is in both space and time. There is nothing truly static. Today, however, computer vision research is largely dominated by methods that focus on single image processing and recognition. There are very few approaches that start directly with videos in mind. That is due more to historical reasons: it is less expensive to process single images and humans show that recognition in a single image is possible. So, if humans do it and it is less expensive, why not make programs do the same thing? We argue that single-image tasks make more sense in the supervised setting. If we want to learn unsupervised, then we should better consider how things are in the real world: objects and higher level entities, actions and interactions, complex activities and full stories, all exist in both space and time, and the two are deeply linked. Every object changes a little bit, in appearance or position, from one moment to the next. In every single place, there is an element of change, a vibration in both time and space, which we should learn from. Objects usually move differently than their surroundings. They also look different than their background. Thus, changes co-exist in both temporal and spatial dimensions from the start. At the same time, objects are consistent and coherent in both space and time. Their movements usually vary smoothly and their interactions follow certain patterns. Their appearance also varies smoothly and follows certain geometric and color patterns of symmetry. Therefore, considering videos as input at the beginning of our journey seems to be a must.

All these properties of the physical world could be taken advantage of only if we consider the space-time reality. We should definitely take advantage of every piece of information and property that universally applies in the natural world if we want to solve in a fundamental and general way the hard problem of unsupervised learning.

There is a second, practical advantage in considering videos at input and approaching space and time, together, from the beginning. There is a tremendous amount of unlabeled video data freely available and being able to label it automatically would give any solution a huge advantage over the strictly supervised setting that requires very expensive manual annotation.

Thus, the ability to perform unsupervised learning is extremely valuable for both research and industry. Moreover, the increased amount of available video data, as compared to single images, has the potential to greatly improve generalization and would allow learning of objects and their interactions together, from the beginning. Object classes are often defined by their actions and roles they play in the larger story. There should definitely be an agreement between the properties of an object at its local level and its role played in the global spatiotemporal context and we could only take advantage of such agreements if we consider the spatiotemporal domain from the start.

1.3.1 Current Trends in Unsupervised Learning

The interest in unsupervised learning is steadily increasing in the machine learning, computer vision, and robotics research. Classical works are based on the observation that real-world data naturally groups into certain classes based on certain core, innate properties, related to color, texture, form, or shape. Thus, elements that are similar based on such properties should belong to the same group or cluster, while those that are dissimilar should be put in different clusters. Consequently, the very vast research field of clustering in machine learning was born, with a plethora of algorithms being proposed during the last fifty years Gan et al. [1], which could be grouped into several main classes: (1) methods related to K-means algorithm Lloyd [2] and Expectation-Maximization (EM) Dempster et al. [3], which have an explicit probabilistic formulation and attempt to maximize the data likelihood conditioned on the class assignments; (2) methods that directly optimize the density of clusters, such as the Mean Shift algorithm Comaniciu and Meer [4], Fukunaga and Hostetler [5] and Density-Based Spatial Clustering (DBSCAN) Ester et al. [6]; (3) hierarchical approaches that form clusters from smaller sub-clusters in a greedy agglomerative fashion Day and Edelsbrunner [7], Ward Jr [8], Sibson [9], Johnson [10] or divisive clustering methods (DIANA) Kaufman and Rousseeuw [11], which start from a large cluster and iteratively divide the larger clusters into smaller ones; (4) spectral clustering algorithms, which are based on the eigenvectors and eigenvalues of the adjacency matrix or the Laplacian of the graph associated with the data points Cheeger [12], Donath and Hoffman [13], Meila and Shi [14], Shi and Malik [15], Ng et al. [16]. The clustering algorithms discussed in the present book and applied to different computer vision problems are mostly related to the class of spectral clustering methods.

Until not so long ago, most unsupervised learning research was focusing on proposing and studying various kinds of clustering algorithms Duda et al. [17] and for a good reason. Most unsupervised learning tasks require, implicitly or explicitly, some sort of clustering. We all researchers in machine learning would hope, even without saying it, that the full structure of the world, with its entities moving, relating, and acting in different ways and being grouped into specific classes, should emerge naturally in a pure unsupervised learning setup. The discovery of such structure with well-formed entities and relations immediately implies some sort of data clustering. Also, the insightful reader will surely observe throughout the book, that the methods proposed here are also based, at their core, on clustering principles.

Current research in unsupervised learning is much more versatile, diversified, and complex than the more general clustering approaches from 20 years ago. However, unsupervised learning is still in its infancy with many pieces missing in the still mysterious unsolved jigsaw puzzle. There are still a lot of unanswered questions and some other questions that have not been even asked yet. At this point, we begin to realize that the space-time domain offers a great advantage over the single-image case, as the temporal dimension brings an important piece of information when it comes to clustering. Objects differ from each other not only in the way they look but also in the way they move. Things that belong together look alike but also move

together, whereas those which do not, separate in time and space. The time dimension, which enforces additional consistency and coherence of the world structure, suddenly becomes a crucial player in the unsupervised learning puzzle.

Therefore, there it comes as no surprise the fact that the initial modern techniques specific to computer vision for unsupervised learning were dedicated to the spatiotemporal, video domain. For example, in a classic pioneering paper Sivic and Zisserman [18], authors propose an algorithm for retrieving objects in videos which is based on discovering a given object in video based on matching keypoints that are stable with respect to appearance and geometry between subsequent frames. Then, such stable clusters are associated with individual physical objects. While the paper is not specifically dedicated to unsupervised learning and clustering, it is in fact heavily relying on it for the task of object retrieval from videos. In our earlier work on object discovery in videos Leordeanu et al. [19], we took a similar approach and discovered objects as clusters of keypoints, matched between video frames, that are geometrically stable for a sufficient amount of time. We noticed an interesting fact in our experiments: when a group of keypoints is geometrically stable for a specific amount of time, the probability that they indeed belong to a single object increases, suddenly from almost 0 to almost 1—this indicates, again, that time can provide very strong cues for what should and what should not belong together in the unsupervised learning game.

Since the first methods that discover objects in videos in an unsupervised manner, other researchers have started to look into that research direction as well Kwak et al. [20], Liu and Chen [21], Wang et al. [22]. The task of object discovery in video is gaining momentum and nowadays, most approaches are formulated in the context of deep learning. There seem to be several directions of research related to learning about objects from video in an unsupervised fashion. One direction crystallizes around the task of discovering object masks from videos. There are already several popular benchmarks for video object segmentation (e.g., DAVIS dataset [23]) with methods that vary from the fully unsupervised case [24–28] to models that are still unsupervised with respect to the given dataset but are heavily pretrained on other datasets [29–42] or having access to the human annotation for the first video frame [23]. While the case when a method is allowed to use powerful pretrained features on different datasets in order to learn on a new dataset is very interesting and has an important insight to give in the future of unsupervised learning, the case itself is definitely not unsupervised. However, once we have powerful pretrained features it really does not matter whether they have been trained in a supervised or unsupervised manner. We should therefore consider this situation, when we are allowed to use pretrained features as a very important one, since in practice it is always the case that we already have a huge number of pretrained models and features available and we should find the best and most efficient way in order to use them, within an unsupervised learning scheme, on novel tasks and datasets, which keep growing in number, diversity, and size.

Another direction of research on the task of learning about objects from video is that of discovering specific keypoints and features that belong to single objects, along with modeling the dynamics of such object keypoints or object parts

[19, 43–45]. On a complementary side, we also have a limited number of papers that address the problem of unsupervised learning for matching such keypoints of object parts, while taking into consideration both the local appearance of parts and their geometric relationships [46–51]. The core general idea is to optimize the model parameters such that matching process will find the most consistent group of assignments between keypoints or dense object regions, which yield the strongest consistency (or cluster) in terms of both local appearance and topology or geometric relationships.

So far, the approaches discussed have been limited to discovering objects as they are seen in images or videos, without taking into consideration the consistent spatial structure of the 3D world. We should keep in mind that it is precisely this stable structure that yields consistent video, depth, or RGB-D sequences. Once we discover the keypoints or regions that belong to a certain object, we could leverage the geometric and camera constraints, even if only implicitly, in order to improve the object discovery in the image and also infer its 3D structure in the world [52–54]. Moreover, once we make the connection between the image of the world and the 3D world itself, we could also take into consideration the static 3D world, the moving objects as well as the camera motion. In fact, we could consider them simultaneously and make them constrain each other during learning, such that a system that predicts motion could be constrained and "taught" by the one that predicts depth and vice versa, alternatively. We then reach a new level in unsupervised learning, which is currently receiving a growing attention, in which complex systems composed of complementary and collaborative pathways are put together to reinforce and also to constrain each other. Thus, we can learn in an unsupervised way, simultaneously, to predict the depth, the camera motion, and its intrinsic parameters from simple RGB video [55–62].

While the concept and art of combining multiple pathways into a global unsupervised learning system still has to be developed, it immediately leads to our novel concept of a universal unsupervised learning machine, the Visual Story Network (VSN), which we propose in the last chapter. This unsupervised learning machine would learn through self-supervision and consensus among many prediction pathways, such that a unified and coherent story is obtained about "everything" that it can sense, interact with, and predict. For more details regarding our novel VSN concept, we refer the reader to Chap. 8.

Before we discuss the Visual Story Network, we should also bring to the reader's attention a novel trend in unsupervised learning that is focusing on putting together multiple modalities and senses. Once we get the idea of using multiple sources of information as self-supervisory signal and observe that the more such sources we have the better, we immediately want to cross the barrier of vision and include touch, auditory, equilibrium, temperature, smell, or any other type of sense into the unsupervised learning equation. This research direction of unsupervised learning by cross-modal prediction, while it is not new [63] it is currently generating an increasing interest [64–69] and it directly relates to more general principles, which we aim to lay down and substantiate in this book. The more diverse and independent types of input and predictions the better, because the harder it will be to find consistency and consensus, but also the more reliable and robust the final results will be when

that will happen. We could start imagining how, in fact, the unsupervised learning problem could begin solving by itself as we keep adding information and constraints into the game. However, we should keep in mind that we might not be able to learn everything from the start and begin by learning first simpler and fewer tasks, with a limited set of data and predictive pathways. This relates to the idea of curriculum learning [70], which has been researched in machine learning in the past decade. At the same time, we should also expect that learning could take several generations of students and teachers, which become stronger from one iteration to the next as they explore a continuously growing world. However, at this point, we should not jump too far ahead and instead return to the more basic ideas and principles, which we will use to build a stronger case for an unsupervised learning system in the final chapters of the book.

First, we propose to go back to some of the earliest ideas ever proposed for unsupervised learning and grouping in humans. We should be prepared to go back and deep in time if we want to be able to see far ahead into the future.

1.3.2 Relation to Gestalt Psychology

Many of the current approaches in unsupervised learning are strongly related and some even inspired by the ideas laid down by the Gestalt school of psychology which was established in Austria and Germany at the very beginning of the twentieth century Koffka [71], Rock and Palmer [72]. The German word *gestalt* used in Gestalt psychology refers to configurations of elements or patterns. This school of thought introduced and studied the main idea that objects are wholes that emerge from parts and are something else than the simple sum of their parts. That is where the saying "the whole is greater than the sum of its parts" comes from. Therefore, the Gestalt cognitive scientists proposed and studied several "laws of grouping," which the brain uses to form such "whole" entities. Such grouping laws include, for example, the *law of proximity* which states that elements that are nearby are likely to belong together, or the *law of similarity* that states that similar things should also be grouped together. In the same way, elements that display symmetry, geometric continuity, or a common fate (similar motions) are also likely to belong together. Besides these principles of combining smaller things into greater ones, Gestalt psychology studied the way we consciously see things as whole entities and interpret them in relation to past experiences.

Below we will present our own key observations regarding unsupervised learning, which we group into a set of principles, which we expect to be true only in a statistical sense, not necessarily all the time. Our principles are strongly related to the initial Gestalt laws and in some sense they could be seen as a modern re-interpretation of those laws in the context of modern machine learning and computer vision. While the Gestalt principles talk mostly about the initial stages of grouping, we go further and present principles from a computational point of view in order to eventually build artificial systems that learn to see by themselves.