Technische Universität München
Fakultät für Elektrotechnik und Informationstechnik
Lehrstuhl für Datenverarbeitung

Learning Image and Video Representations Based on Sparsity Priors

Xian Wei

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzende/-r: Prof. Dr. Jörg Conradt

Prüfende/-r der Dissertation:

1. Prof. Dr.-Ing. Klaus Diepold
2. Priv.-Doz. Dr. Martin Kleinsteuber

Die Dissertation wurde am 17.11.2016 bei der Technische Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 07.04.2017 angenommen.

Xian Wei. *Learning Image and Video Representations Based on Sparsity Priors.* Dissertation, Technische Universität München, Munich, Germany, 2017.

Berichte aus der Informatik

Xian Wei

Learning Image and Video Representations Based on Sparsity Priors

Shaker Verlag
Aachen 2017

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: München, Techn. Univ., Diss., 2017

Printed in Germany.

ISBN 978-3-8440-5303-6
ISSN 0945-0807

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

Thanks to my family.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Martin Kleinsteuber, for his continuous support to my Ph.D. study. His deep insights and meticulous guidance helped me through all the time in research and writing of this dissertation. Especially, I want to thank him for his financial support in the last year.

My cordial thanks also go to Prof. Dr.-Ing. Klaus Diepold for his kind support to my doctoral degree program for providing me with facilities and personnels, in particular, for his generosity to take care of the administration of my dissertation examination.

I wish to express my gratitude to my mentor, Dr. Hao Shen, for his insightful comments and encouragement, especially, for his patience and expertise in teaching me research and scientific writing.

I would like to acknowledge the financial support from China Scholarship Council (CSC) for provision of scholarship. They gave me a financial support for four years of studies and researches. I am grateful to TUM graduate school for their support to my international visits and other academic activities.

I thank my colleagues from GOL and LDV. They are Dr. Michael Zwick, Dr. Simon Hawe, Clemens Hage, Martin Kiechle, Dominik Meyer, Matthias Seibert, Alexander Sagel, Julian Wörmann, Sunil Ramgopal Tatavarty, Peter Hausamann and Sundeep Patil. I am lucky to share the happy five years with them. My thanks are also due to Ms. Ricarda Baumhoer for her assistance and advices on administration through my Ph.D. study.

I would like to express my special thanks to Simon Hawe, Clemens Hage and Martin Kiechle, for their advices and software support for my Ph.D. research. My special thanks are also due to my colleagues in the same office, Martin Knopp, Alexander Sagel and Peter Hausamann, I am happy to work with them.

Abstract

Recent development in representation learning shows that appropriate data representations are the key to the success of machine learning algorithms, since different representations can entangle different explanatory information of the data. Among the various methods of learning representations, sparse representations of data have been observed to contain rich distributed information of the data with respect to specific learning tasks, such as image classification, regression, etc. By taking advantage of such a benefit, the focus of this dissertation is on developing algorithmic framework that allows disentangling the underlying explanatory factors hidden in sparse representations of image and video data. For example, explanatory information considered in this dissertation can be an underlying linear system that explains the dynamics of texture videos, or the similarity of image data points that explores the intrinsic structure of data. Moreover, such disentangled factors have shown to conveniently solve various computer vision problems. Specifically in this dissertation, they are dynamic texture modeling and low dimensional image representations. The key concept behind this development is to construct a joint cost function, which combines the criteria for learning sparse representations and the criteria for discovering underlying factors in the learned sparse representations. Since the admissible sets of solutions to our optimization problem are restricted on appropriate matrix manifolds, geometric optimization techniques that exploit the underlying manifold structures of solutions can be employed to efficiently solve such an optimization problem. Finally, we leverage the advantage of differential geometric optimization to develop a collection of efficient algorithms on appropriate differentiable manifolds.

The key difficulty for solving the proposed joint learning problem is the differentiability of sparse representation with respect to a given dictionary. For addressing such a challenge, we consider the sparse coding problem by minimizing a quadratic reconstruction error with appropriate convex sparsity priors, such as elastic net prior and Kullback-Leibler divergence prior. In this way, sparse representation can be shown to be a locally differentiable function with respect to a dictionary, and hence a generic form of the directional derivative of sparse representation with respect to the given dictionary is developed. The ability to compute such a derivative leads to various further learning mechanisms in sparse representations that disentangle different underlying explanatory factors. By leveraging such an algorithmic benefit and geometric optimization techniques, in what follows, we construct joint learning cost functions to study two aforementioned challenging computer vision problems, dynamic textures and image dimensionality reduction.

Modeling Dynamic Textures (DT) is a long standing active research topic in the computer vision community. Study and analysis of DT attracts both theoretical and practical research efforts, such as building a stable DT modeling system, video segmentation, video recognition and video synthesis. However, the continuous change in the shape and appearance of a dynamic texture makes the application of traditional computer vision algorithms very

challenging. Thus, finding an appropriate spatio-temporal generative representation model to explore the evolution of the dynamic textured scenes is the key to many DT studies. One classical technique is to model the dynamical course of DTs as a Markov random process. Following the Markov random process, one typical model is developed and widely applied to the practice, namely, linear dynamical system (LDS). LDS assumes that each observation is correlated to an underlying latent variable, or "state", and the dynamic process of these consecutive states can be captured by a parameter transition operator. In this dissertation, we follow the framework of classical LDS, and present to treat the sparse coefficients over a learned dictionary as the underlying "states". In this way, the dynamical process of dynamic textures exhibits a transition course of corresponding sparse events. Next, our goal is to find a suitable and robust linear transition matrix that captures the dynamics between two adjacent frames of sparse representations in time series. Under several reasonable assumptions, we read this transition as a linear transformation matrix with the constraint of stability. Under this way, a DT sequence is represented by an appropriate sparse transition matrix together with a dictionary, shortly called DT parameters. Such learned DT parameters can be used for various DT applications, such as DT synthesis, recognition and denoising.

The second computer vision problem studied in this dissertation is finding an appropriate low dimensional representations of raw images. It is known that natural images are often very high dimensional, statistically non-Gaussian, and show abundant varying texture patterns. Hence, they are difficult to be explicitly parameterized by a common probabilistic model. Therefore, some machine learning techniques, such as linear smooth regression, may not be directly used to construct the prediction model for such raw images. Finding appropriate low dimensional representations of image data is an efficient way to promote the further prediction models learning. In this dissertation, we present a unified algorithmic framework for learning low dimensional representations of images for the three classic machine learning scenarios of unsupervised, supervised and semi-supervised learning. The core concept of our development is to combine two popular data representation criteria, namely sparsity and trace quotient. The former is known to be a convenient tool to identify underlying factors, and the latter is known for disentangling underlying discriminative factors. We construct a generic cost function for learning jointly a sparsifying dictionary and a dimensionality reduction transformation. The proposed cost function covers a wide range of classic low dimensional representation methods, such as Principal Component Analysis, Local Linear Embedding, Laplacian Eigenmap, Linear Discriminant Analysis (LDA), Semi-supervised LDA, and more. Experimental evaluations on image classification, clustering, 2/3 D visualization, and object categorization demonstrate the strong competitive performance in comparison with state-of-the-art algorithms.

Contents