

Multimodal Learning toward Micro-Video Understanding

Synthesis Lectures on Image, Video, and Multimedia Processing

Editor

Alan C. Bovik, *University of Texas, Austin*

The Lectures on Image, Video and Multimedia Processing are intended to provide a unique and groundbreaking forum for the world's experts in the field to express their knowledge in unique and effective ways. It is our intention that the Series will contain Lectures of basic, intermediate, and advanced material depending on the topical matter and the authors' level of discourse. It is also intended that these Lectures depart from the usual dry textbook format and instead give the author the opportunity to speak more directly to the reader, and to unfold the subject matter from a more personal point of view. The success of this candid approach to technical writing will rest on our selection of exceptionally distinguished authors, who have been chosen for their noteworthy leadership in developing new ideas in image, video, and multimedia processing research, development, and education.

In terms of the subject matter for the series, there are few limitations that we will impose other than the Lectures be related to aspects of the imaging sciences that are relevant to furthering our understanding of the processes by which images, videos, and multimedia signals are formed, processed for various tasks, and perceived by human viewers. These categories are naturally quite broad, for two reasons: First, measuring, processing, and understanding perceptual signals involves broad categories of scientific inquiry, including optics, surface physics, visual psychophysics and neurophysiology, information theory, computer graphics, display and printing technology, artificial intelligence, neural networks, harmonic analysis, and so on. Secondly, the domain of application of these methods is limited only by the number of branches of science, engineering, and industry that utilize audio, visual, and other perceptual signals to convey information. We anticipate that the Lectures in this series will dramatically influence future thought on these subjects as the Twenty-First Century unfolds.

Multimodal Learning toward Micro-Video Understanding

Liqiang Nie, Meng Liu, and Xuemeng Song

2019

Virtual Reality and Virtual Environments in 10 Lectures

Stanislav Stanković

2015

Dictionary Learning in Visual Computing

Qiang Zhang and Baoxin Li

2015

Combating Bad Weather Part II: Fog Removal from Image and Video

Sudipta Mukhopadhyay and Abhishek Kumar Tripathi

2015

Combating Bad Weather Part I: Rain Removal from Video

Sudipta Mukhopadhyay and Abhishek Kumar Tripathi

2014

Image Understanding Using Sparse Representations

Jayaraman J. Thiagarajan, Karthikeyan Natesan Ramamurthy, Pavan Turaga, and Andreas Spanias

2014

Contextual Analysis of Videos

Myo Thida, How-lung Eng, Dorothy Monekosso, and Paolo Remagnino

2013

Wavelet Image Compression

William A. Pearlman

2013

Remote Sensing Image Processing

Gustavo Camps-Valls, Devis Tuia, Luis Gómez-Chova, Sandra Jiménez, and Jesús Malo

2011

The Structure and Properties of Color Spaces and the Representation of Color Images

Eric Dubois

2009

Biomedical Image Analysis: Segmentation

Scott T. Acton and Nilanjan Ray

2009

Joint Source-Channel Video Transmission

Fan Zhai and Aggelos Katsaggelos

2007

Super Resolution of Images and Video

Aggelos K. Katsaggelos, Rafael Molina, and Javier Mateos

2007

Tensor Voting: A Perceptual Organization Approach to Computer Vision and Machine Learning

Philippos Mordohai and Gérard Medioni

2006

Light Field Sampling

Cha Zhang and Tsuhan Chen

2006

Real-Time Image and Video Processing: From Research to Reality

Nasser Kehtarnavaz and Mark Gamadia

2006

MPEG-4 Beyond Conventional Video Coding: Object Coding, Resilience, and Scalability

Mihaela van der Schaar, Deepak S Turaga, and Thomas Stockhammer

2006

Modern Image Quality Assessment

Zhou Wang and Alan C. Bovik

2006

Biomedical Image Analysis: Tracking

Scott T. Acton and Nilanjan Ray

2006

Recognition of Humans and Their Activities Using Video

Rama Chellappa, Amit K. Roy-Chowdhury, and S. Kevin Zhou

2005

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2019

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Multimodal Learning toward Micro-Video Understanding
Liqiang Nie, Meng Liu, and Xuemeng Song

ISBN: 978-3-031-01127-6 paperback
ISBN: 978-3-031-02255-5 ebook
ISBN: 978-3-031-00216-8 hardcover

DOI 10.1007/978-3-031-02255-5

A Publication in the Springer series
SYNTHESIS LECTURES ON IMAGE, VIDEO, AND MULTIMEDIA PROCESSING

Lecture #20
Series Editor: Alan C. Bovik, *University of Texas, Austin*
Series ISSN
Print 1559-8136 Electronic 1559-8144

Multimodal Learning toward Micro-Video Understanding

Liqiang Nie, Meng Liu, and Xuemeng Song
Shandong University, Jinan, China

*SYNTHESIS LECTURES ON IMAGE, VIDEO, AND MULTIMEDIA
PROCESSING #20*

ABSTRACT

Micro-videos, a new form of user-generated contents, have been spreading widely across various social platforms, such as Vine, Kuaishou, and TikTok. Different from traditional long videos, micro-videos are usually recorded by smart mobile devices at any place within a few seconds. Due to its brevity and low bandwidth cost, micro-videos are gaining increasing user enthusiasm. The blossoming of micro-videos opens the door to the possibility of many promising applications, ranging from network content caching to online advertising. Thus, it is highly desirable to develop an effective scheme for the high-order micro-video understanding.

Micro-video understanding is, however, non-trivial due to the following challenges: (1) how to represent micro-videos that only convey one or few high-level themes or concepts; (2) how to utilize the hierarchical structure of the venue categories to guide the micro-video analysis; (3) how to alleviate the influence of low-quality caused by complex surrounding environments and the camera shake; (4) how to model the multimodal sequential data, i.e., textual, acoustic, visual, and social modalities, to enhance the micro-video understanding; and (5) how to construct large-scale benchmark datasets for the analysis? These challenges have been largely unexplored to date.

In this book, we focus on addressing the challenges presented above by proposing some state-of-the-art multimodal learning theories. To demonstrate the effectiveness of these models, we apply them to three practical tasks of micro-video understanding: popularity prediction, venue category estimation, and micro-video routing. Particularly, we first build three large-scale real-world micro-video datasets for these practical tasks. We then present a multimodal transductive learning framework for micro-video popularity prediction. Furthermore, we introduce several multimodal cooperative learning approaches and a multimodal transfer learning scheme for micro-video venue category estimation. Meanwhile, we develop a multimodal sequential learning approach for micro-video recommendation. Finally, we conclude the book and figure out the future research directions in multimodal learning toward micro-video understanding.

KEYWORDS

micro-video understanding, multimodal transductive learning, multimodal cooperative learning, multimodal transfer learning, multimodal sequential learning, popularity prediction, venue category estimation, micro-video recommendation

Contents

	Preface	xiii
	Acknowledgments	xv
1	Introduction	1
1.1	Micro-Video Proliferation	1
1.2	Practical Tasks	3
1.2.1	Micro-Video Popularity Prediction	3
1.2.2	Micro-Video Venue Categorization	3
1.2.3	Micro-Video Routing	4
1.3	Research Challenges	5
1.4	Our Solutions	7
1.5	Book Structure	9
2	Data Collection	11
2.1	Dataset I for Popularity Prediction	11
2.2	Dataset II for Venue Category Estimation	12
2.3	Dataset III for Micro-Video Routing	16
2.4	Summary	16
3	Multimodal Transductive Learning for Micro-Video Popularity Prediction ..	19
3.1	Background	19
3.2	Research Problems	19
3.3	Feature Extraction	20
3.3.1	Observations	20
3.3.2	Social Modality	21
3.3.3	Visual Modality	21
3.3.4	Acoustic Modality	23
3.3.5	Textual Modality	23
3.4	Related Work	24
3.4.1	Popularity Prediction	24
3.4.2	Multi-View Learning	25

3.4.3	Low-Rank Subspace Learning	26
3.5	Notations and Preliminaries	28
3.6	Multimodal Transductive Learning	28
3.6.1	Objective Formulation	29
3.6.2	Optimization	31
3.6.3	Experiments and Results	33
3.7	Multi-Modal Transductive Low-Rank Learning	39
3.7.1	Objective Formulation	39
3.7.2	Optimization	44
3.7.3	Experiments and Results	47
3.8	Summary	56
4	Multimodal Cooperative Learning for Micro-Video Venue Categorization	59
4.1	Background	59
4.2	Research Problems	59
4.3	Related Work	60
4.3.1	Multimedia Venue Estimation	60
4.3.2	Multi-Modal Multi-Task Learning	61
4.3.3	Dictionary Learning	62
4.4	Multimodal Consistent Learning	62
4.4.1	Optimization	65
4.4.2	Task Relatedness Estimation	66
4.4.3	Complexity Analysis	67
4.4.4	Experiments	68
4.5	Multimodal Complementary Learning	73
4.5.1	Multi-Modal Dictionary Learning	74
4.5.2	Tree-Guided Multi-Modal Dictionary Learning	75
4.5.3	Optimization	77
4.5.4	Online Learning	78
4.5.5	Experiments	81
4.6	Multimodal Cooperative Learning	90
4.6.1	Multimodal Early Fusion	92
4.6.2	Cooperative Networks	93
4.6.3	Attention Networks	96
4.6.4	Experiments	98
4.7	Summary	107

5	Multimodal Transfer Learning in Micro-Video Analysis	109
5.1	Background	109
5.2	Research Problems	109
5.3	Related Work	110
5.4	External Sound Dataset	111
5.5	Deep Multi-Modal Transfer Learning	112
5.5.1	Sound Knowledge Transfer	113
5.5.2	Multi-Modal Fusion	114
5.5.3	Deep Network for Venue Estimation	115
5.5.4	Training	116
5.6	Experiments	117
5.6.1	Experimental Settings	117
5.6.2	Acoustic Representation (RQ1)	119
5.6.3	Performance Comparison (RQ2)	119
5.6.4	External Knowledge Effect (RQ3)	120
5.6.5	Visualization	121
5.6.6	Study of DARE Model (RQ4)	121
5.7	Summary	124
6	Multimodal Sequential Learning for Micro-Video Recommendation	125
6.1	Background	125
6.2	Research Problems	125
6.3	Related Work	126
6.4	Multimodal Sequential Learning	127
6.4.1	The Temporal Graph-Based LSTM Layer	129
6.4.2	The Multi-Level Interest Modeling Layer	131
6.4.3	The Prediction Layer	131
6.5	Experiments	133
6.5.1	Experimental Settings	133
6.5.2	Baselines	133
6.5.3	Overall Comparison	134
6.5.4	Component-Wise Evaluation of ALPINE	136
6.5.5	Justification of the Temporal Graph	138
6.5.6	Attention Visualization	139
6.6	Summary	140

7	Research Frontiers	141
7.1	Micro-Video Annotation	141
7.2	Micro-Video Captioning	143
7.3	Micro-Video Thumbnail Selection	145
7.4	Semantic Ontology Construction	146
7.5	Pornographic Content Identification	147
	Bibliography	149
	Authors' Biographies	169

Preface

The unprecedented growth of portable devices contributes to the success of micro-video sharing platforms such as Vine, Kuaishou, and TikTok. These devices enable users to record and share their daily life within a few seconds in the form of micro-videos at any time and any place. As a new media type, micro-videos have garnered great enthusiasm due to brevity, authenticity, communicability, and low-cost. The proliferation of micro-videos confirms the old saying that good things come in small packages.

Like traditional long videos, micro-videos are a combination of textual, acoustic, and visual modalities. These modalities are correlated rather than independent, and they essentially characterize the same micro-videos from distinct angles. Effectively fusing heterogeneous modalities toward video understanding indeed has been well-studied in the past decade. Yet, micro-videos have their unique characteristics and corresponding research challenges, including but not limited to the following.

(1) Information sparseness. Micro-videos are very short, lasting for 6–15 s, and they hence usually convey only a few concepts. In light of this, we need to learn their sparse and conceptual representations for better discrimination. (2) Hierarchical structure. Micro-videos are implicitly organized into a four-layer hierarchical tree structure with respect to their recording venues. We should leverage such a structure to guide the organization of micro-videos by categorizing them into the leaf nodes of this tree. (3) Low-quality. Most portable devices have nothing to offer for video stabilization. Some recorded videos can thus be visually shaky or bumpy, which greatly hinders the visual expression. Furthermore, the audio track that comes along with the video can differ in terms of distortion and noise, such as buzzing, hums, hisses, and whistling, which is probably caused by the poor microphones or complex surrounding environments. We thus have to harness the external visual or sound knowledge to compensate the shortest boards. (4) Multimodal sequential data. Beyond textual, acoustic, and visual modalities, micro-videos also have social modality. In such a context, a user is enabled to interact with micro-videos and other users via social actions, such as click, like, and follow. As time goes on, multiple sequential data in different forms emerge and reflect users' historical preferences. To strengthen micro-video understanding, we have to characterize and model the sequential patterns. (5) The last challenge we are facing is the lack of benchmark datasets to justify our ideas.

In this book, to tackle the aforementioned research challenges, we present some state-of-the-art multimodal learning theories and verify them over three practical tasks of micro-video understanding: popularity prediction, venue category estimation, and micro-video routing. In particular, we first construct three large-scale real-world micro-video datasets corresponding to the three practical tasks. We then propose a multimodal transductive learning framework

to learn the micro-video representations in an optimal latent space via unifying and preserving information from different modalities. In this transductive framework, we integrate the low-rank constraints to somehow alleviate the information sparseness and low-quality problems. This framework is verified on the popularity prediction task. We next present a series of multimodal cooperative learning approaches, which explicitly model the consistent and complementary modality correlations. In the multimodal cooperative learning approaches, we make full use of the hierarchical structure by the tree-guided group lasso, and further solve the information sparseness via dictionary learning. Following that, we work toward compensating the low-quality acoustic modalities via harnessing the external sound knowledge. This is accomplished by a deep multimodal transfer learning scheme. The multimodal cooperative learning approaches and the multimodal transfer learning scheme are both justified over the task of venue category estimation. Thereafter, we develop a multimodal sequential learning approach, relying on temporal graph-based long short-term memory networks, to intelligently route micro-videos to the target users in a personalized manner. We ultimately summarize the book and figure out the future research directions in multimodal learning toward micro-video understanding.

This book represents a preliminary research on learning from multiple correlated modalities of given micro-videos, and we anticipate that the lectures in this series will dramatically influence future thought on these subjects. If in this book we have been able to dream further than others have, it is because we are standing on the shoulders of giants.

Liqiang Nie, Meng Liu, and Xuemeng Song
July 2019

Acknowledgments

It is a pleasure to acknowledge many colleagues who have made this time-consuming book project possible and enjoyable. In particular, many members of the iLearn Center in Shandong University and the LMS Lab in National University of Singapore have co-worked on various aspects of multimodal learning and its applications in micro-video understanding. Their efforts have supplied ingredients for insightful discussions related to the writing of this book, and hence we are greatly appreciative.

Our first thanks undoubtedly goes to Dr. Peiguang Jing at Tianjin University, Dr. Jingyuan Chen at Alibaba Group, Dr. Jianglong Zhang at Information & Telecommunication Company SGCC, as well as Mr. Yongqi Li and Mr. Yinwei Wei at Shandong University. We consulted with them on some specific technical chapters of the book and they are also the major contributors of some chapters. Their constructive feedback and comments at various stages have been significantly helpful in shaping the book. We also take this opportunity to thank Prof. Tat-Seng Chua at National University of Singapore who never hesitated to offer his advice and share his valuable experience whenever the authors needed him. Particular thanks go to Miss Qian Liu and Miss Xiaoli Li, who read the earlier drafts of the manuscript and provided helpful comments to improve the readability.

We are very grateful to the anonymous reviewers. Despite their busy schedules, they read the book very carefully and gave us many insightful suggestions, which were the key to making this book as sound as possible.

We are grateful to Morgan & Claypool and particularly the Vice President & Publisher Mr. Joel Claypool for his help and patience throughout the writing of this book. He has managed to get everything done on time and provided us with many pieces of valuable advice. This book would not have been completed, or at least not be what it looks like now, without the support, direction, and help of him and his team.

Last, but certainly no least, our thanks go to our beloved families for their unwavering support during this fun book project, as well as for their understanding and tolerance of many weekends and long nights spent on the book by the authors. We dedicate this book to them, with love.

Liqliang Nie, Meng Liu, and Xuemeng Song
July 2019