

# Big Data Analytics for **Large-Scale Multimedia Search**

EDITED BY

Stefanos Vrochidis | Benoit Huet

Edward Y. Chang | Ioannis Kompatsiaris

with website



WILEY



## **Big Data Analytics for Large-Scale Multimedia Search**



# Big Data Analytics for Large-Scale Multimedia Search

*Edited by*

*Stefanos Vrochidis*

Information Technologies Institute, Centre for Research and Technology Hellas  
Thessaloniki, Greece

*Benoit Huet*

EURECOM  
Sophia-Antipolis  
France

*Edward Y. Chang*

HTC Research & Healthcare  
San Francisco, USA

*Ioannis Kompatsiaris*

Information Technologies Institute, Centre for Research and Technology Hellas  
Thessaloniki, Greece

**WILEY**

This edition first published 2019  
© 2019 John Wiley & Sons Ltd.

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, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Stefanos Vrochidis, Benoit Huet, Edward Y. Chang and Ioannis Kompatsiaris to be identified as the authors of the editorial material in this work asserted in accordance with law.

#### *Registered Offices*

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

#### *Editorial Office*

The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at [www.wiley.com](http://www.wiley.com).

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

#### *Limit of Liability/Disclaimer of Warranty*

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

#### *Library of Congress Cataloging-in-Publication Data*

Names: Vrochidis, Stefanos, 1975- editor. | Huet, Benoit, editor. | Chang, Edward Y., editor. | Kompatsiaris, Ioannis, editor.

Title: Big Data Analytics for Large-Scale Multimedia Search / Stefanos Vrochidis, Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece; Benoit Huet, EURECOM, Sophia-Antipolis, France; Edward Y. Chang, HTC Research & Healthcare, San Francisco, USA; Ioannis Kompatsiaris, Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece.

Description: Hoboken, NJ, USA : Wiley, [2018] | Includes bibliographical references and index. |

Identifiers: LCCN 2018035613 (print) | LCCN 2018037546 (ebook) | ISBN 9781119376989 (Adobe PDF) | ISBN 9781119377009 (ePub) | ISBN 9781119376972 (hardcover)

Subjects: LCSH: Multimedia data mining. | Big data.

Classification: LCC QA76.9.D343 (ebook) | LCC QA76.9.D343 V76 2018 (print) | DDC 005.7 – dc23

LC record available at <https://lcn.loc.gov/2018035613>

Cover design: Wiley

Cover image: © spainter\_vfx/iStock.com

Set in 10/12pt WarnockPro by SPi Global, Chennai, India

Printed and bound by CPI Group (UK) Ltd, Croydon, CR0 4YY

10 9 8 7 6 5 4 3 2 1

## Contents

<b>Introduction</b>	<i>xv</i>
<b>List of Contributors</b>	<i>xix</i>
<b>About the Companion Website</b>	<i>xxiii</i>

### Part I Feature Extraction from Big Multimedia Data 1

<b>1</b>	<b>Representation Learning on Large and Small Data</b>	<b>3</b>
	<i>Chun-Nan Chou, Chuen-Kai Shie, Fu-Chieh Chang, Jocelyn Chang and Edward Y. Chang</i>	
1.1	Introduction	3
1.2	Representative Deep CNNs	5
1.2.1	AlexNet	6
1.2.1.1	ReLU Nonlinearity	6
1.2.1.2	Data Augmentation	7
1.2.1.3	Dropout	8
1.2.2	Network in Network	8
1.2.2.1	MLP Convolutional Layer	9
1.2.2.2	Global Average Pooling	9
1.2.3	VGG	10
1.2.3.1	Very Small Convolutional Filters	10
1.2.3.2	Multi-scale Training	11
1.2.4	GoogLeNet	11
1.2.4.1	Inception Modules	11
1.2.4.2	Dimension Reduction	12
1.2.5	ResNet	13
1.2.5.1	Residual Learning	13
1.2.5.2	Identity Mapping by Shortcuts	14
1.2.6	Observations and Remarks	15
1.3	Transfer Representation Learning	15
1.3.1	Method Specifications	17
1.3.2	Experimental Results and Discussion	18
1.3.2.1	Results of Transfer Representation Learning for OM	19
1.3.2.2	Results of Transfer Representation Learning for Melanoma	20
1.3.2.3	Qualitative Evaluation: Visualization	21

1.3.3	Observations and Remarks	23
1.4	Conclusions	24
	References	25

## 2 Concept-Based and Event-Based Video Search in Large Video Collections 31

*Foteini Markatopoulou, Damianos Galanopoulos, Christos Tzelepis, Vasileios Mezaris and Ioannis Patras*

2.1	Introduction	32
2.2	Video preprocessing and Machine Learning Essentials	33
2.2.1	Video Representation	33
2.2.2	Dimensionality Reduction	34
2.3	Methodology for Concept Detection and Concept-Based Video Search	35
2.3.1	Related Work	35
2.3.2	Cascades for Combining Different Video Representations	37
2.3.2.1	Problem Definition and Search Space	37
2.3.2.2	Problem Solution	38
2.3.3	Multi-Task Learning for Concept Detection and Concept-Based Video Search	40
2.3.4	Exploiting Label Relations	41
2.3.5	Experimental Study	42
2.3.5.1	Dataset and Experimental Setup	42
2.3.5.2	Experimental Results	43
2.3.5.3	Computational Complexity	47
2.4	Methods for Event Detection and Event-Based Video Search	48
2.4.1	Related Work	48
2.4.2	Learning from Positive Examples	49
2.4.3	Learning Solely from Textual Descriptors: Zero-Example Learning	50
2.4.4	Experimental Study	52
2.4.4.1	Dataset and Experimental Setup	52
2.4.4.2	Experimental Results: Learning from Positive Examples	53
2.4.4.3	Experimental Results: Zero-Example Learning	53
2.5	Conclusions	54
2.6	Acknowledgments	55
	References	55

## 3 Big Data Multimedia Mining: Feature Extraction Facing Volume, Velocity, and Variety 61

*Vedhas Pandit, Shahin Amiriparian, Maximilian Schmitt, Amr Mousa and Björn Schuller*

3.1	Introduction	61
3.2	Scalability through Parallelization	64
3.2.1	Process Parallelization	64
3.2.2	Data Parallelization	64
3.3	Scalability through Feature Engineering	65
3.3.1	Feature Reduction through Spatial Transformations	66
3.3.2	Laplacian Matrix Representation	66



3.3.3	Parallel latent Dirichlet allocation and bag of words	68
3.4	Deep Learning-Based Feature Learning	68
3.4.1	Adaptability that Conquers both Volume and Velocity	70
3.4.2	Convolutional Neural Networks	72
3.4.3	Recurrent Neural Networks	73
3.4.4	Modular Approach to Scalability	74
3.5	Benchmark Studies	76
3.5.1	Dataset	76
3.5.2	Spectrogram Creation	77
3.5.3	CNN-Based Feature Extraction	77
3.5.4	Structure of the CNNs	78
3.5.5	Process Parallelization	79
3.5.6	Results	80
3.6	Closing Remarks	81
3.7	Acknowledgements	82
	References	82

## Part II Learning Algorithms for Large-Scale Multimedia 89

<b>4</b>	<b>Large-Scale Video Understanding with Limited Training Labels</b>	<b>91</b>
	<i>Jingkuan Song, Xu Zhao, Lianli Gao and Liangliang Cao</i>	
4.1	Introduction	91
4.2	Video Retrieval with Hashing	91
4.2.1	Overview	91
4.2.2	Unsupervised Multiple Feature Hashing	93
4.2.2.1	Framework	93
4.2.2.2	The Objective Function of MFH	93
4.2.2.3	Solution of MFH	95
4.2.2.3.1	Complexity Analysis	96
4.2.3	Submodular Video Hashing	97
4.2.3.1	Framework	97
4.2.3.2	Video Pooling	97
4.2.3.3	Submodular Video Hashing	98
4.2.4	Experiments	99
4.2.4.1	Experiment Settings	99
4.2.4.1.1	Video Datasets	99
4.2.4.1.2	Visual Features	99
4.2.4.1.3	Algorithms for Comparison	100
4.2.4.2	Results	100
4.2.4.2.1	CC_WEB_VIDEO	100
4.2.4.2.2	Combined Dataset	100
4.2.4.3	Evaluation of SVH	101
4.2.4.3.1	Results	102
4.3	Graph-Based Model for Video Understanding	103
4.3.1	Overview	103
4.3.2	Optimized Graph Learning for Video Annotation	104

4.3.2.1	Framework	104
4.3.2.2	OGL	104
4.3.2.2.1	Terms and Notations	104
4.3.2.2.2	Optimal Graph-Based SSL	105
4.3.2.2.3	Iterative Optimization	106
4.3.3	Context Association Model for Action Recognition	107
4.3.3.1	Context Memory	108
4.3.4	Graph-based Event Video Summarization	109
4.3.4.1	Framework	109
4.3.4.2	Temporal Alignment	110
4.3.5	TGIF: A New Dataset and Benchmark on Animated GIF Description	111
4.3.5.1	Data Collection	111
4.3.5.2	Data Annotation	112
4.3.6	Experiments	114
4.3.6.1	Experimental Settings	114
4.3.6.1.1	Datasets	114
4.3.6.1.2	Features	114
4.3.6.1.3	Baseline Methods and Evaluation Metrics	114
4.3.6.2	Results	115
4.4	Conclusions and Future Work	116
	References	116
<b>5</b>	<b>Multimodal Fusion of Big Multimedia Data</b>	<b>121</b>
	<i>Ilias Gialampoukidis, Elisavet Chatzilari, Spiros Nikolopoulos, Stefanos Vrochidis and Ioannis Kompatsiaris</i>	
5.1	Multimodal Fusion in Multimedia Retrieval	122
5.1.1	Unsupervised Fusion in Multimedia Retrieval	123
5.1.1.1	Linear and Non-linear Similarity Fusion	123
5.1.1.2	Cross-modal Fusion of Similarities	124
5.1.1.3	Random Walks and Graph-based Fusion	124
5.1.1.4	A Unifying Graph-based Model	126
5.1.2	Partial Least Squares Regression	127
5.1.3	Experimental Comparison	128
5.1.3.1	Dataset Description	128
5.1.3.2	Settings	129
5.1.3.3	Results	129
5.1.4	Late Fusion of Multiple Multimedia Rankings	130
5.1.4.1	Score Fusion	131
5.1.4.2	Rank Fusion	132
5.1.4.2.1	Borda Count Fusion	132
5.1.4.2.2	Reciprocal Rank Fusion	132
5.1.4.2.3	Condorcet Fusion	132
5.2	Multimodal Fusion in Multimedia Classification	132
5.2.1	Related Literature	134
5.2.2	Problem Formulation	136
5.2.3	Probabilistic Fusion in Active Learning	137
5.2.3.1	If $P(S=0 V,T) \neq 0$ :	138

5.2.3.2	If $P(S=0 V,T) \neq 0$ :	138
5.2.3.3	Incorporating Informativeness in the Selection ( $P(S V)$ )	139
5.2.3.4	Measuring Oracle's Confidence ( $P(S T)$ )	139
5.2.3.5	Re-training	140
5.2.4	Experimental Comparison	141
5.2.4.1	Datasets	141
5.2.4.2	Settings	142
5.2.4.3	Results	143
5.2.4.3.1	Expanding with Positive, Negative or Both	143
5.2.4.3.2	Comparing with Sample Selection Approaches	145
5.2.4.3.3	Comparing with Fusion Approaches	147
5.2.4.3.4	Parameter Sensitivity Investigation	147
5.2.4.3.5	Comparing with Existing Methods	148
5.3	Conclusions	151
	References	152
<b>6</b>	<b>Large-Scale Social Multimedia Analysis</b>	<b>157</b>
	<i>Benjamin Bischke, Damian Borth and Andreas Dengel</i>	
6.1	Social Multimedia in Social Media Streams	157
6.1.1	Social Multimedia	157
6.1.2	Social Multimedia Streams	158
6.1.3	Analysis of the Twitter Firehose	160
6.1.3.1	Dataset: Overview	160
6.1.3.2	Linked Resource Analysis	160
6.1.3.3	Image Content Analysis	162
6.1.3.4	Geographic Analysis	164
6.1.3.5	Textual Analysis	166
6.2	Large-Scale Analysis of Social Multimedia	167
6.2.1	Large-Scale Processing of Social Multimedia Analysis	167
6.2.1.1	Batch-Processing Frameworks	167
6.2.1.2	Stream-Processing Frameworks	168
6.2.1.3	Distributed Processing Frameworks	168
6.2.2	Analysis of Social Multimedia	169
6.2.2.1	Analysis of Visual Content	169
6.2.2.2	Analysis of Textual Content	169
6.2.2.3	Analysis of Geographical Content	170
6.2.2.4	Analysis of User Content	170
6.3	Large-Scale Multimedia Opinion Mining System	170
6.3.1	System Overview	171
6.3.2	Implementation Details	171
6.3.2.1	Social Media Data Crawler	171
6.3.2.2	Social Multimedia Analysis	173
6.3.2.3	Analysis of Visual Content	174
6.3.3	Evaluations: Analysis of Visual Content	175
6.3.3.1	Filtering of Synthetic Images	175
6.3.3.2	Near-Duplicate Detection	177
6.4	Conclusion	178
	References	179

<b>7</b>	<b>Privacy and Audiovisual Content: Protecting Users as Big Multimedia Data Grows Bigger</b>	<b>183</b>
	<i>Martha Larson, Jaeyoung Choi, Manel Slokom, Zekeriya Erkin, Gerald Friedland and Arjen P. de Vries</i>	
7.1	Introduction	183
7.1.1	The Dark Side of Big Multimedia Data	184
7.1.2	Defining Multimedia Privacy	184
7.2	Protecting User Privacy	188
7.2.1	What to Protect	188
7.2.2	How to Protect	189
7.2.3	Threat Models	191
7.3	Multimedia Privacy	192
7.3.1	Privacy and Multimedia Big Data	192
7.3.2	Privacy Threats of Multimedia Data	194
7.3.2.1	Audio Data	194
7.3.2.2	Visual Data	195
7.3.2.3	Multimodal Threats	195
7.4	Privacy-Related Multimedia Analysis Research	196
7.4.1	Multimedia Analysis Filters	196
7.4.2	Multimedia Content Masking	198
7.5	The Larger Research Picture	199
7.5.1	Multimedia Security and Trust	199
7.5.2	Data Privacy	200
7.6	Outlook on Multimedia Privacy Challenges	202
7.6.1	Research Challenges	202
7.6.1.1	Multimedia Analysis	202
7.6.1.2	Data	202
7.6.1.3	Users	203
7.6.2	Research Reorientation	204
7.6.2.1	Professional Paranoia	204
7.6.2.2	Privacy as a Priority	204
7.6.2.3	Privacy in Parallel	205
	References	205

### **Part III Scalability in Multimedia Access 209**

<b>8</b>	<b>Data Storage and Management for Big Multimedia</b>	<b>211</b>
	<i>Björn Þór Jónsson, Gylfi Þór Guðmundsson, Laurent Amsaleg and Philippe Bonnet</i>	
8.1	Introduction	211
8.1.1	Multimedia Applications and Scale	212
8.1.2	Big Data Management	213
8.1.3	System Architecture Outline	213
8.1.4	Metadata Storage Architecture	214
8.1.4.1	Lambda Architecture	214
8.1.4.2	Storage Layer	215
8.1.4.3	Processing Layer	216

8.1.4.4	Serving Layer	216
8.1.4.5	Dynamic Data	216
8.1.5	Summary and Chapter Outline	217
8.2	Media Storage	217
8.2.1	Storage Hierarchy	217
8.2.1.1	Secondary Storage	218
8.2.1.2	The Five-Minute Rule	218
8.2.1.3	Emerging Trends for Local Storage	219
8.2.2	Distributed Storage	220
8.2.2.1	Distributed Hash Tables	221
8.2.2.2	The CAP Theorem and the PACELC Formulation	221
8.2.2.3	The Hadoop Distributed File System	221
8.2.2.4	Ceph	222
8.2.3	Discussion	222
8.3	Processing Media	222
8.3.1	Metadata Extraction	223
8.3.2	Batch Processing	223
8.3.2.1	Map-Reduce and Hadoop	224
8.3.2.2	Spark	225
8.3.2.3	Comparison	226
8.3.3	Stream Processing	226
8.4	Multimedia Delivery	226
8.4.1	Distributed In-Memory Buffering	227
8.4.1.1	Memcached and Redis	227
8.4.1.2	Alluxio	227
8.4.1.3	Content Distribution Networks	228
8.4.2	Metadata Retrieval and NoSQL Systems	228
8.4.2.1	Key-Value Stores	229
8.4.2.2	Document Stores	229
8.4.2.3	Wide Column Stores	229
8.4.2.4	Graph Stores	229
8.4.3	Discussion	229
8.5	Case Studies: Facebook	230
8.5.1	Data Popularity: Hot, Warm or Cold	230
8.5.2	Mentions Live	231
8.6	Conclusions and Future Work	231
8.6.1	Acknowledgments	232
	References	232
<b>9</b>	<b>Perceptual Hashing for Large-Scale Multimedia Search</b>	<b>239</b>
	<i>Li Weng, I-Hong Jhuo and Wen-Huang Cheng</i>	
9.1	Introduction	240
9.1.1	Related work	240
9.1.2	Definitions and Properties of Perceptual Hashing	241
9.1.3	Multimedia Search using Perceptual Hashing	243
9.1.4	Applications of Perceptual Hashing	243
9.1.5	Evaluating Perceptual Hash Algorithms	244

9.2	Unsupervised Perceptual Hash Algorithms	245
9.2.1	Spectral Hashing	245
9.2.2	Iterative Quantization	246
9.2.3	$K$ -Means Hashing	247
9.2.4	Kernelized Locality Sensitive Hashing	249
9.3	Supervised Perceptual Hash Algorithms	250
9.3.1	Semi-Supervised Hashing	250
9.3.2	Kernel-Based Supervised Hashing	252
9.3.3	Restricted Boltzmann Machine-Based Hashing	253
9.3.4	Supervised Semantic-Preserving Deep Hashing	255
9.4	Constructing Perceptual Hash Algorithms	257
9.4.1	Two-Step Hashing	257
9.4.2	Hash Bit Selection	258
9.5	Conclusion and Discussion	260
	References	261

## Part IV Applications of Large-Scale Multimedia Search 267

<b>10</b>	<b>Image Tagging with Deep Learning: Fine-Grained Visual Analysis</b>	<b>269</b>
	<i>Jianlong Fu and Tao Mei</i>	
10.1	Introduction	269
10.2	Basic Deep Learning Models	270
10.3	Deep Image Tagging for Fine-Grained Image Recognition	272
10.3.1	Attention Proposal Network	274
10.3.2	Classification and Ranking	275
10.3.3	Multi-Scale Joint Representation	276
10.3.4	Implementation Details	276
10.3.5	Experiments on CUB-200-2011	277
10.3.6	Experiments on Stanford Dogs	280
10.4	Deep Image Tagging for Fine-Grained Sentiment Analysis	281
10.4.1	Learning Deep Sentiment Representation	282
10.4.2	Sentiment Analysis	283
10.4.3	Experiments on SentiBank	283
10.5	Conclusion	284
	References	285
<b>11</b>	<b>Visually Exploring Millions of Images using Image Maps and Graphs</b>	<b>289</b>
	<i>Kai Uwe Barthel and Nico Hezel</i>	
11.1	Introduction and Related Work	290
11.2	Algorithms for Image Sorting	293
11.2.1	Self-Organizing Maps	293
11.2.2	Self-Sorting Maps	294
11.2.3	Evolutionary Algorithms	295
11.3	Improving SOMs for Image Sorting	295

11.3.1	Reducing SOM Sorting Complexity	295
11.3.2	Improving SOM Projection Quality	297
11.3.3	Combining SOMs and SSMs	297
11.4	Quality Evaluation of Image Sorting Algorithms	298
11.4.1	Analysis of SOMs	298
11.4.2	Normalized Cross-Correlation	299
11.4.3	A New Image Sorting Quality Evaluation Scheme	299
11.5	2D Sorting Results	301
11.5.1	Image Test Sets	301
11.5.2	Experiments	302
11.6	Demo System for Navigating 2D Image Maps	304
11.7	Graph-Based Image Browsing	306
11.7.1	Generating Semantic Image Features	306
11.7.2	Building the Image Graph	307
11.7.3	Visualizing and Navigating the Graph	310
11.7.4	Prototype for Image Graph Navigation	312
11.8	Conclusion and Future Work	313
	References	313
<b>12</b>	<b>Medical Decision Support Using Increasingly Large Multimodal Data Sets</b>	<b>317</b>
	<i>Henning Müller and Devrim Ünay</i>	
12.1	Introduction	317
12.2	Methodology for Reviewing the Literature in this chapter	320
12.3	Data, Ground Truth, and Scientific Challenges	321
12.3.1	Data Annotation and Ground Truthing	321
12.3.2	Scientific Challenges and Evaluation as a Service	321
12.3.3	Other Medical Data Resources Available	322
12.4	Techniques used for Multimodal Medical Decision Support	323
12.4.1	Visual and Non-Visual Features Describing the Image Content	323
12.4.2	General Machine Learning and Deep Learning	323
12.5	Application Types of Image-Based Decision Support	326
12.5.1	Localization	326
12.5.2	Segmentation	326
12.5.3	Classification	327
12.5.4	Prediction	327
12.5.5	Retrieval	327
12.5.6	Automatic Image Annotation	328
12.5.7	Other Application Types	328
12.6	Discussion on Multimodal Medical Decision Support	328
12.7	Outlook or the Next Steps of Multimodal Medical Decision Support	329
	References	330

## Conclusions and Future Trends 337

## Index 339





## Introduction

In recent years, the rapid development of digital technologies, including the low cost of recording, processing, and storing media, and the growth of high-speed communication networks enabling large-scale content sharing, has led to a rapid increase in the availability of multimedia content worldwide. The availability of such content, as well as the increasing user need of analysing and searching into large multimedia collections, increases the demand for the development of advanced search and analytics techniques for big multimedia data. Although multimedia is defined as a combination of different media (e.g., audio, text, video, images etc.) this book mainly focuses on textual, visual, and audiovisual content, which are considered the most characteristic types of multimedia.

In this context, the big multimedia data era brings a plethora of challenges to the fields of multimedia mining, analysis, searching, and presentation. These are best described by the Vs of big data: volume, variety, velocity, veracity, variability, value, and visualization. A modern multimedia search and analytics algorithm and/or system has to be able to handle large databases with varying formats at extreme speed, while having to cope with unreliable “ground truth” information and “noisy” conditions. In addition, multimedia analysis and content understanding algorithms based on machine learning and artificial intelligence have to be employed. Further, the interpretation of the content over time may change, leading to a “drifting target” with multimedia content being perceived differently in different times with often low value of data points. Finally, the assessed information needs to be presented in comprehensive and transparent ways to human users.

The main challenges for big multimedia data analytics and search are identified in the areas of:

- multimedia representation by extracting low- and high-level conceptual features
- application of machine learning and artificial intelligence for large-scale multimedia
- scalability in multimedia access and retrieval.

Feature extraction is an essential step in any computer vision and multimedia data analysis task. Though progress has been made in past decades, it is still quite difficult for computers to accurately recognize an object or comprehend the semantics of an image or a video. Thus, feature extraction is expected to remain an active research area in advancing computer vision and multimedia data analysis for the foreseeable

future. The traditional approach of feature extraction is model-based in that researchers engineer useful features based on heuristics, and then conduct validations via empirical studies. A major shortcoming of the model-based approach is that exceptional circumstances such as different lighting conditions and unexpected environmental factors can render the engineered features ineffective. The data-driven approach complements the model-based approach. Instead of human-engineered features, the data-driven approach learns representation from data. In principle, the greater the quantity and diversity of data, the better the representation can be learned.

An additional layer of analysis and automatic annotation of big multimedia data involves the extraction of high-level concepts and events. Concept-based multimedia data indexing refers to the automatic annotation of multimedia fragments with specific simple labels, e.g., “car”, “sky”, “running” etc., from large-scale collections. In this book we mainly deal with video as a characteristic multimedia example for concept-based indexing. To deal with this task, concept detection methods have been developed that automatically annotate images and videos with semantic labels referred to as concepts. A recent trend in video concept detection is to learn features directly from the raw keyframe pixels using deep convolutional neural networks (DCNNs). On the other hand, event-based video indexing aims to represent video fragments with high-level events in a given set of videos. Typically, events are more complex than concepts, i.e., they may include complex activities, occurring at specific places and times, and involving people interacting with other people and/or object(s), such as “opening a door”, “making a cake”, etc. The event detection problem in images and videos can be addressed either with a typical video event detection framework, including feature extraction and classification, and/or by effectively combining textual and visual analysis techniques.

When it comes to multimedia analysis, machine learning is considered to be one of the most popular techniques that can be applied. These include CNN for representation learning such as imagery and acoustic data, as well as recurrent neural networks for series data, e.g., speech and video. The challenge of video understanding lies in the gap between large-scale video data and the limited resource we can afford in both label collection and online computing stages.

An additional step in the analysis and retrieval of large-scale multimedia is the fusion of heterogeneous content. Due to the diverse modalities that form a multimedia item (e.g., visual, textual modality), multiple features are available to represent each modality. The fusion of multiple modalities may take place at the feature level (early fusion) or the decision level (late fusion). Early fusion techniques usually rely on the linear (weighted) combination of multimodal features, while lately non-linear fusion approaches have prevailed. Another fusion strategy relies on graph-based techniques, allowing the construction of random walks, generalized diffusion processes, and cross-media transitions on the formulated graph of multimedia items. In the case of late fusion, the fusion takes place at the decision level and can be based on (i) linear/non-linear combinations of the decisions from each modality, (ii) voting schemes, and (iii) rank diffusion processes. Scalability issues in multimedia processing systems typically occur for two reasons: (i) the lack of labelled data, which limits the scalability with respect to the number of supported concepts, and (ii) the high computational overload in terms of both processing time and memory complexity. For the first problem, methods that learn primarily on weakly labelled data (weakly supervised learning, semi-supervised learning)

have been proposed. For the second problem, methodologies typically rely on reducing the data space they work on by using smartly-selected subsets of the data so that the computational requirements of the systems are optimized.

Another important aspect of multimedia nowadays is the social dimension and the user interaction that is associated with the data. The internet is abundant with opinions, sentiments, and reflections of the society about products, brands, and institutions hidden under large amounts of heterogeneous and unstructured data. Such analysis includes the contextual augmentation of events in social media streams in order to fully leverage the knowledge present in social media, taking into account temporal, visual, textual, geographical, and user-specific dimensions. In addition, the social dimension includes an important privacy aspect. As big multimedia data continues to grow, it is essential to understand the risks for users during online multimedia sharing and multimedia privacy. Specifically, as multimedia data gets bigger, automatic privacy attacks can become increasingly dangerous. Two classes of algorithms for privacy protection in a large-scale online multimedia sharing environment are involved. The first class is based on multimedia analysis, and includes classification approaches that are used as filters, while the second class is based on obfuscation techniques.

The challenge of data storage is also very important for big multimedia data. At this scale, data storage, management, and processing become very challenging. At the same time, there has been a proliferation of big data management techniques and tools, which have been developed mostly in the context of much simpler business and logging data. These tools and techniques include a variety of noSQL and newSQL data management systems, as well as automatically distributed computing frameworks (e.g., Hadoop and Spark). The question is which of these big data techniques apply to today's big multimedia collections. The answer is not trivial since the big data repository has to store a variety of multimedia data, including raw data (images, video or audio), meta-data (including social interaction data) associated with the multimedia items, derived data, such as low-level concepts and semantic features extracted from the raw data, and supplementary data structures, such as high-dimensional indices or inverted indices. In addition, the big data repository must serve a variety of parallel requests with different workloads, ranging from simple queries to detailed data-mining processes, and with a variety of performance requirements, ranging from response-time driven online applications to throughput-driven offline services. Although several different techniques have been developed there is no single technology that can cover all the requirements of big multimedia applications.

Finally, the book discusses the two main challenges of large-scale multimedia search: accuracy and scalability. Conventional techniques typically focus on the former. However, recently attention has mainly been paid to the latter, since the amount of multimedia data is rapidly increasing. Due to the curse of dimensionality, conventional feature representations of high dimensionality are not in favour of fast search. The big data era requires new solutions for multimedia indexing and retrieval based on efficient hashing. One of the robust solutions is perceptual hash algorithms, which are used for generating hash values from multimedia objects in big data collections, such as images, audio, and video. A content-based multimedia search can be achieved by comparing hash values. The main advantages of using hash values instead of other content representations is that hash values are compact and facilitate fast in-memory indexing and search, which is very important for large-scale multimedia search.

Given the aforementioned challenges, the book is organized in the following chapters. Chapters 1, 2, and 3 deal with feature extraction from big multimedia data, while Chapters 4, 5, 6, and 7 discuss techniques relevant to machine learning for multimedia analysis and fusion. Chapters 8, and 9 deal with scalability in multimedia access and retrieval, while Chapters 10, 11, and 12 present applications of large-scale multimedia retrieval. Finally, we conclude the book by summarizing and presenting future trends and challenges.

## List of Contributors

***Laurent Amsaleg***

Univ Rennes, Inria, CNRS  
IRISA  
France

***Shahin Amiriparian***

ZD.B Chair of Embedded Intelligence for  
Health Care and Wellbeing  
University of Augsburg  
Germany

***Kai Uwe Barthel***

Visual Computing Group  
HTW Berlin  
University of Applied Sciences  
Berlin  
Germany

***Benjamin Bischke***

German Research Center for Artificial  
Intelligence and TU Kaiserslautern  
Germany

***Philippe Bonnet***

IT University of Copenhagen  
Copenhagen  
Denmark

***Damian Borth***

University of St. Gallen  
Switzerland

***Edward Y. Chang***

HTC Research & Healthcare  
San Francisco, USA

***Elisavet Chatzilari***

Information Technologies Institute  
Centre for Research and Technology  
Hellas  
Thessaloniki  
Greece

***Liangliang Cao***

College of Information and Computer  
Sciences  
University of Massachusetts Amherst  
USA

***Chun-Nan Chou***

HTC Research & Healthcare  
San Francisco, USA

***Jaeyoung Choi***

Delft University of Technology  
Netherlands  
and  
International Computer Science Institute  
USA

***Fu-Chieh Chang***

HTC Research & Healthcare  
San Francisco, USA

***Jocelyn Chang***

Johns Hopkins University  
Baltimore  
USA

**Wen-Huang Cheng**

Department of Electronics Engineering  
and Institute of Electronics  
National Chiao Tung University  
Taiwan

**Andreas Dengel**

German Research Center for Artificial  
Intelligence and TU Kaiserslautern  
Germany

**Arjen P. de Vries**

Radboud University  
Nijmegen  
The Netherlands

**Zekeriya Erkin**

Delft University of Technology and  
Radboud University  
The Netherlands

**Gerald Friedland**

University of California  
Berkeley  
USA

**Jianlong Fu**

Multimedia Search and Mining Group  
Microsoft Research Asia  
Beijing  
China

**Damianos Galanopoulos**

Information Technologies Institute  
Centre for Research and Technology  
Hellas  
Thessaloniki  
Greece

**Lianli Gao**

School of Computer Science and Center  
for Future Media  
University of Electronic Science and  
Technology of China  
Sichuan  
China

**Ilias Gialampoukidis**

Information Technologies Institute  
Centre for Research and Technology  
Hellas  
Thessaloniki  
Greece

**Gylfi Þór Guðmundsson**

Reykjavik University  
Iceland

**Nico Hezel**

Visual Computing Group  
HTW Berlin  
University of Applied Sciences  
Berlin  
Germany

**I-Hong Jhuo**

Center for Open-Source Data & AI  
Technologies  
San Francisco  
California

**Björn Þór Jónsson**

IT University of Copenhagen  
Denmark

and

Reykjavik University  
Iceland

**Ioannis Kompatsiaris**

Information Technologies Institute  
Centre for Research and Technology  
Hellas  
Thessaloniki  
Greece

**Martha Larson**

Radboud University and  
Delft University of Technology  
The Netherlands

**Amr Mousa**

Chair of Complex and Intelligent Systems  
University of Passau  
Germany

**Foteini Markatopoulou**

Information Technologies Institute  
Centre for Research and Technology  
Hellas  
Thessaloniki  
Greece

and

School of Electronic Engineering and  
Computer Science  
Queen Mary University of London  
United Kingdom

**Henning Müller**

University of Applied Sciences Western  
Switzerland (HES-SO)  
Sierre  
Switzerland

**Tao Mei**

JD AI Research  
China

**Vasileios Mezaris**

Information Technologies Institute  
Centre for Research and Technology  
Hellas  
Thessaloniki  
Greece

**Spiros Nikolopoulos**

Information Technologies Institute  
Centre for Research and Technology  
Hellas  
Thessaloniki  
Greece

**Ioannis Patras**

School of Electronic Engineering and  
Computer Science  
Queen Mary University of London  
United Kingdom

**Vedhas Pandit**

ZD.B Chair of Embedded Intelligence for  
Health Care and Wellbeing  
University of Augsburg  
Germany

**Maximilian Schmitt**

ZD.B Chair of Embedded Intelligence for  
Health Care and Wellbeing  
University of Augsburg  
Germany

**Björn Schuller**

ZD.B Chair of Embedded Intelligence for  
Health Care and Wellbeing  
University of Augsburg  
Germany

and

GLAM - Group on Language, Audio and  
Music  
Imperial College London  
United Kingdom

**Chuen-Kai Shie**

HTC Research & Healthcare  
San Francisco, USA

**Manel Slokom**

Delft University of Technology  
The Netherlands

**Jingkuan Song**

School of Computer Science and Center  
for Future Media  
University of Electronic Science and  
Technology of China  
Sichuan  
China

***Christos Tzelepis***

Information Technologies Institute  
Centre for Research and Technology  
Hellas  
Thessaloniki  
Greece  
and

School of Electronic Engineering and  
Computer Science  
QMUL, UK

***Devrim Ünay***

Department of Biomedical Engineering  
Izmir University of Economics  
Izmir  
Turkey

***Stefanos Vrochidis***

Information Technologies Institute  
Centre for Research and Technology  
Hellas  
Thessaloniki  
Greece

***Li Weng***

Hangzhou Dianzi University  
China  
and  
French Mapping Agency (IGN)  
Saint-Mande  
France

***Xu Zhao***

Department of Automation  
Shanghai Jiao Tong University  
China



## About the Companion Website

This book is accompanied by a companion website:

**[www.wiley.com/go/vrochidis/bigdata](http://www.wiley.com/go/vrochidis/bigdata)**



The website includes:

- Open source algorithms
- Data sets
- Tools materials for demonstration purpose

Scan this QR code to visit the companion website.





## **Part I**

### **Feature Extraction from Big Multimedia Data**



## 1

## Representation Learning on Large and Small Data

*Chun-Nan Chou, Chuen-Kai Shie, Fu-Chieh Chang, Jocelyn Chang and Edward Y. Chang*

### 1.1 Introduction

Extracting useful features from a scene is an essential step in any computer vision and multimedia data analysis task. Though progress has been made in past decades, it is still quite difficult for computers to comprehensively and accurately recognize an object or pinpoint the more complicated semantics of an image or a video. Thus, feature extraction is expected to remain an active research area in advancing computer vision and multimedia data analysis for the foreseeable future.

The approaches in feature extraction can be divided into two categories: *model-centric* and *data-driven*. The model-centric approach relies on human heuristics to develop a computer model (or algorithm) to extract features from an image. (We use imagery data as our example throughout this chapter.) Some widely used models are Gabor filter, wavelets, and scale-invariant feature transform (SIFT) [1]. These models were engineered by scientists and then validated via empirical studies. A major shortcoming of the model-centric approach is that unusual circumstances that a model does not take into consideration during its design, such as different lighting conditions and unexpected environmental factors, can render the engineered features less effective. In contrast to the model-centric approach, which dictates representations independent of data, the data-driven approach learns representations from data [2]. Examples of data-driven algorithms are multilayer perceptron (MLP) and convolutional neural networks (CNNs), which belong to the general category of neural networks and deep learning [3, 4].

Both model-centric and data-driven approaches employ a model (algorithm or machine). The differences between model-centric and data-driven can be described in two related aspects:

- Can data affect model parameters? With model-centric, training data does not affect the model. With data-driven, such as MLP or CNN, their internal parameters are changed/learned based on the discovered structure in large data sets [5].
- Can more data help improve representations? Whereas more data can help a data-driven approach to improve representations, more data cannot change the

features extracted by a model-centric approach. For example, the features of an image can be affected by the other images in the CNN (because the structure parameters modified through back-propagation are affected by all training images), but the feature set of an image is invariant of the other images in a model-centric pipeline such as SIFT.

The greater the quantity and diversity of data, the better the representations can be learned by a data-driven pipeline. In other words, if a learning algorithm has seen enough training instances of an object under various conditions, e.g., in different postures, and has been partially occluded, then the features learned from the training data will be more comprehensive.

The focus of this chapter is on how *neural networks*, specifically CNNs, achieve effective representation learning. Neural networks, a kind of neuroscience-motivated models, were based on Hubel and Wiesel's research on cats' visual cortex [6], and subsequently formulated into computation models by scientists in the early 1980s. Pioneer neural network models include Neocognitron [7] and the shift-invariant neural network [8]. Widely cited enhanced models include LeNet-5 [9] and Boltzmann machines [10]. However, the popularity of neural networks surged only in 2012 after large training data sets became available. In 2012, Krizhevsky [11] applied deep convolutional networks on the ImageNet dataset<sup>1</sup>, and their AlexNet achieved breakthrough accuracy in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 competition.<sup>2</sup> This work convinced the research community and related industries that representation learning with big data is promising. Subsequently, several efforts have aimed to further improve the learning capability of neural networks. Today, the top-5 error rate<sup>3</sup> for the ILSVRC competition has dropped to 3.57%, a remarkable achievement considering the error rate was 26.2% before AlexNet [11] was proposed.

We divide the remainder of this chapter into two parts before suggesting related reading in the concluding remarks. The first part reviews representative CNN models proposed since 2012. These key representatives are discussed in terms of three aspects addressed in He's tutorial presentation [14] at ICML 2016: (i) representation ability, (ii) optimization ability, and (iii) generalization ability. The representation ability is the ability of a CNN to learn/capture representations from training data assuming the optimum could be found. Here, the optimum refers to attaining the best solution of the underlying learning algorithm, modeled as an optimization problem. This leads to the second aspect that He's tutorial addresses: the optimization ability. The optimization ability is the feasibility of finding an optimum. Specifically on CNNs, the optimization problem is to find the optimal solution of the stochastic gradient descent. Finally, the generalization ability is the quality of the test performance once model parameters have been learned from training data.

The second part of this chapter deals with the small data problem. We present how features learned from one source domain with big data can be transferred to a different target domain with small data. This transfer representation learning approach is critical

1 ImageNet is a dataset of over 15 million labeled images belonging to roughly 22,000 categories [12].

2 The ILSVRC [13] evaluates algorithms for object detection and image classification on a subset of ImageNet, 1.2 million images over 1000 categories. Throughout this chapter, we focus on discussing image classification challenges.

3 The top-5 error used to evaluate the performance of image classification is the proportion of images such that the ground-truth category is outside the top-5 predicted categories.