

# **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. Chong***

HTC Research & Healthcare  
San Francisco, USA

***Ioannis Kompatsiaris***

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

**WILEY**

# Contents

Introduction **xv**

List of Contributors **xix**

About the Companion Website **xxiii**

## Part I Feature Extraction from Big Multimedia Data 1

1	Representation Learning on Large and Small Data 3
	<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
	<i>Foteini Markatopoulou, Damianos Galanopoulos, Christos Tzelepis, Vassilios Mezaris and Ioannis Patras</i>	
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
	<i>Vedhas Pandit, Shahin Amiriparian, Maximilian Schmitt, Amr Mousa and Bjorn Schuller</i>	
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	<b>68</b>
3.4	Deep Learning-Based Feature Learning	<b>68</b>
3.4.1	Adaptability that Conquers both Volume and Velocity	<b>70</b>
3.4.2	Convolutional Neural Networks	<b>72</b>
3.4.3	Recurrent Neural Networks	<b>73</b>
3.4.4	Modular Approach to Scalability	<b>74</b>
3.5	Benchmark Studies	<b>76</b>
3.5.1	Dataset	<b>76</b>
3.5.2	Spectrogram Creation	<b>77</b>
3.5.3	CNN-Based Feature Extraction	<b>77</b>
3.5.4	Structure of the CNNs	<b>78</b>
3.5.5	Process Parallelization	<b>79</b>
3.5.6	Results	<b>80</b>
3.6	Closing Remarks	<b>81</b>
3.7	Acknowledgements	<b>82</b>
	References	<b>82</b>

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

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

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
5	Multimodal Fusion of Big Multimedia Data	121
	<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 $P(S=0 V,T)^0$ :	138

5.2.3.2	IfP(S=0 V,T)^0:	138
5.2.3.3	Incorporating Informativeness in the Selection (P^1^)	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
6	Large-Scale Social Multimedia Analysis	157
	<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

7	Privacy and Audiovisual Content: Protecting Users as Big Multimedia Data Grows Bigger 183 <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
7.1.1	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
8	Data Storage and Management for Big Multimedia 211 <i>Bjorn PorJonsson, Gylfi Por Gudmundsson, 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>	Perceptual Hashing for Large-Scale Multimedia Search	239
	<i>Li Weng, I-HongJhuo 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	AT-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

10	Image Tagging with Deep Learning: Fine-Grained Visual Analysis	269
	<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
11	Visually Exploring Millions of Images using Image Maps and Graphs	289
	<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
12	Medical Decision Support Using Increasingly Large Multimodal Data Sets	317
	<i>Henning Müller and Devrim Unay</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