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

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.

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.

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://lccn.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

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

The website includes:

- Open source algorithms
- Data sets
- Tools materials for demostration 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

Big Data Analytics for Large-Scale Multimedia Search, First Edition.
Edited by Stefanos Vrochidis, Benoit Huet, Edward Y. Chang, and Ioannis Kompatsiaris.

Companion website: www.wiley.com/go/vrochidis/bigdata

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[1], and their AlexNet achieved breakthrough accuracy in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012 competition.[2] 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[3] 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.