TUP-Springer Project

Edward Y. Chang

大规模多媒体信息管理与检索基础

模拟人类感知数学方法

Foundations of Large-Scale Multimedia Information Management and Retrieval

Mathematics of Perception

清华大学出版社

Springer

Edward Y. Chang

大规模多媒体信息管理与检索基础

模拟人类感知数学方法

Foundations of Large-Scale Multimedia Information Management and Retrieval

Mathematics of Perception

With 108 figures, 21 of them in color

TSINGHUA UNIVERSITY PRESS

Springer

内 容 简 介

大规模多媒体信息管理与检索面临着两大类艰巨的技术挑战。首先，这一工程问题的研究在本质上是多领域、跨学科的，涉及信号处理、计算机视觉、数据库、机器学习、神经科学和认知心理学；其次，一个有效的解决方案必须能解决高维数据和网络规模数据的可扩展性问题。本书第一部分（第1～8章）着重介绍如何采用多领域、跨学科算法来解决特征提取及选择、知识表示、语义分析、距离函数的制定等问题；第二部分（第9～12章）对解决高维数据和网络规模数据的扩展性问题提出了有效的处理方法。此外，本书的附录还给出了作者开发的开源软件的下载地址。

本书是作者在美国加州大学从事多年的教学科研及在 Google 公司工作多年的基础上编写的。本书适合多媒体、计算机视觉、机器学习、大规模数据处理等领域的研发人员阅读，也可作为高等院校计算机专业本科生及研究生的教材或教学参考书。

图书在版编目(CIP)数据

大规模多媒体信息管理与检索基础: 模拟人类感知数学方法: 英文 / 张智威著. -- 北京: 清华大学出版社, 2011.5
ISBN 978-7-302-24976-4

Ⅰ.①大… Ⅱ.①张… Ⅲ.①多媒体–信息管理–英文 ②多媒体检索系统–英文
Ⅳ.①TP37 ②G354.47

中国版本图书馆 CIP 数据核字(2011)第 041182 号

责任编辑：薛　慧
责任印制：李红英
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京雅昌彩色印刷有限公司
经　　销：全国新华书店
开　　本：153×235　**印　　张**：19.75
版　　次：2011 年 5 月第 1 版　　**印　　次**：2011 年 5 月第1次印刷
印　　数：1～2000
定　　价：59.00 元

产品编号：031394-01

To my family
Lihyuarn, Emily, Jocelyn, and Rosalind.

Foreword

The last few years have been transformative time in information and communication technology. Possibly this is one of the most exciting period after Gutenberg's moveable print revolutionized how people create, store, and share information. As is well known, Gutenberg's invention had tremendous impact on human societal development. We are again going through a similar transformation in how we create, store, and share information. I believe that we are witnessing a transformation that allows us to share our experiences in more natural and compelling form using audio-visual media rather than its subjective abstraction in the form of text. And this is huge.

It is nice to see a book on a very important aspect of organizing visual information by a researcher who has unique background in being a sound academic researcher as well as a contributor to the state of art practical systems being used by lots of people. Edward Chang has been a research leader while he was in academia, at University of California, Santa Barbara, and continues to apply his enormous energy and in depth knowledge now to practical problems in the largest information search company of our time. He is a person with a good perspective of the emerging field of multimedia information management and retrieval.

A good book describing current state of art and outlining important challenges has enormous impact on the field. Particularly, in a field like multimedia information management the problems for researchers and practitioners are really complex due to their multidisciplinary nature. Researchers in computer vision and image processing, databases, information retrieval, and multimedia have approached this problem from their own disciplinary perspective. The perspective based on just one discipline results in approaches that are narrow and do not really solve the problem that requires true multidisciplinary perspective. Considering the explosion in the volume of visual data in the last two decades, it is now essential that we solve the urgent problem of managing this volume effectively for easy access and utilization. By looking at the problem in multimedia information as a problem of managing information about the real world that is captured using different correlated media, it is possible to make significant progress. Unfortunately, most researchers do not have time and interest to look beyond their disciplinary boundaries to understand the real

problem and address it. This has been a serious hurdle in the progress in multimedia information management.

I am delighted to see and present this book on a very important and timely topic by an eminent researcher who has not only expertise and experience, but also energy and interest to put together an in depth treatment of this interdisciplinary topic. I am not aware of any other book that brings together concepts and techniques in this emerging field in a concise book. Moreover, Prof. Chang has shown his talent in pedagogy by organizing the book to consider needs of undergraduate students as well as graduate students and researchers. This is a book that will be equally useful for people interested in learning about the state of the art in multimedia information management and for people who want to address challenges in this transformative field.

Ramesh Jain
Irvine, February 2011

Preface

The volume and accessibility of images and videos is increasing exponentially, thanks to the sea-change of imagery captured from film to digital form, to the availability of electronic networking, and to the ubiquity of high-speed network access. The tools for organizing and retrieving these multimedia data, however, are still quite primitive. One such evidence is the lack of effective tools to-date for organizing personal images or videos. Another clue is that all Internet search engines today still rely on the keyword search paradigm, which knowingly suffers from the semantic aliasing problem. Existing organization and retrieval tools are ineffective partly because they fail to properly model and combine "content" and "context" of multimedia data, and partly because they fail to effectively address the scalability issues. For instance, today, a typical content-based retrieval prototype extracts some signals from multimedia data instances to represent them, employs a poorly justified distance function to measure similarity between data instances, and relies on a costly sequential scan to find data instances similar to a query instance. From feature extraction, data representation, multimodal fusion, similarity measurement, feature-to-semantic mapping, to indexing, the design of each component has mostly not been built on solid scientific foundations. Furthermore, most prior art focuses on improving one single component, and demonstrates its effectiveness on small datasets. However, the problem of multimedia information management and retrieval is inherently an interdisciplinary one, and tackling the problem must involve collaboration between fields of machine learning, multimedia computing, cognitive science, and large-scale computing, in addition to signal processing, computer vision, and databases. This book presents an interdisciplinary approach to first establish scientific foundations for each component, and then address interactions between components in a scalable manner in terms of both data dimensionality and volume.

This book is organized into twelve chapters of two parts. The first part of the book depicts a multimedia system's key components, which together aims to comprehend semantics of multimedia data instances. The second part presents methods for scaling up these components for high-dimensional data and very large datasets. In part one we start with providing an overview of the research and engineering challenges in Chapter 1. Chapter 2 presents feature extraction, which obtains useful signals

from multimedia data instances. We discuss both model-based and data-driven, and then a hybrid approach. In Chapter 3, we deal with the problem of formulating users' query concepts, which can be complex and subjective. We show how active learning and kernel methods can be used to work effectively with both keywords and perceptual features to understand a user's query concept with minimal user feedback. We argue that only after a user's query concept can be thoroughly comprehended, it is then possible to retrieve matching objects. Chapters 4 and 5 address the problem of distance-function formulation, a core subroutine of information retrieval for measuring similarity between data instances. Chapter 4 presents Dynamic Partial function and its foundation in cognitive psychology. Chapter 5 shows how an effective function can also be learned from examples in a data-driven way. Chapters 6, 7 and 8 describe methods that fuse metadata of multiple modalities. Multimodal fusion is important to properly integrate perceptual features of various kinds (e.g., color, texture, shape; global, local; time-invariant, time-variant), and to properly combine metadata from multiple sources (e.g., from both content and context). We present three techniques: super-kernel fusion in Chapter 6, fusion with causal strengths in Chapter 7, and combinational collaborative filtering in Chapter 8.

Part two of the book tackles various scalability issues. Chapter 9 presents the problem of imbalanced data learning where the number of data instances in the target class is significantly out-numbered by the other classes. This challenge is typical in information retrieval, since the information relevant to our queries is always the minority in the dataset. The chapter describes algorithms to deal with the problem in vector and non-vector spaces, respectively. Chapters 10 and 11 address the scalability issues of kernel methods. Kernel methods are a core machine learning technique with strong theoretical foundations and excellent empirical successes. One major shortcoming of kernel methods is its cubic computation time required for training and linear for classification. We present parallel algorithms to speed up the training time, and fast indexing structures to speed up the classification time. Finally, in Chapter 12, we present our effort in speeding up Latent Dirichlet Allocation (LDA), a robust method for modeling texts and images. Using distributed computing primitives, together with data placement and pipeline techniques, we were able to speed up LDA 1,500 times when using 2,000 machines.

Although the target application of this book is multimedia information retrieval, the developed theories and algorithms are applicable to analyze data of other domains, such as text documents, biological data and motion patterns.

This book is designed for researchers and practitioners in the fields of multimedia, computer vision, machine learning, and large-scale data mining. We expect the reader to have some basic knowledge in Statistics and Algorithms. We recommend that the first part (Chapters 1 to 8) to be used in an upper-division undergraduate course, and the second part (Chapters 9 to 12) in a graduate-level course. Chapters 1 to 6 should be read sequentially. The reader can read Chapters 7 to 12 in selected order. Appendix lists our open source sites.

Edward Y. Chang
Palo Alto, February 2011

Acknowledgements

I would like to thank contributions of my Ph.D students and research colleagues (in roughly chronological order): Beitao Li, Simon Tong, Kingshy Goh, Yi Wu, Navneet Panda, Gang Wu, John R. Smith, Bell Tseng, Kevin Chang, Arun Qamra, Wei-Cheng Lai, Kaihua Zhu, Hongjie Bai, Hao Wang, Jian Li, Zhihuan Qiu, Wen-Yen Chen, Dong Zhang, Xiance Si, Hongji Bao, Zhiyuan Liu, Maosong Sun, Dingyin Xia, Zhiyu Wang, and Shiqiang Yang. I would also like to thank the funding supported by three NSF grants: NSF Career IIS-0133802, NSF ITR IIS-0219885, and NSF IIS-0535085.

Contents

Chapter 1
Introduction — Key Subroutines of Multimedia Data Management

Abstract This chapter presents technical challenges that multimedia information management faces. We enumerate five key subroutines required to work together effectively so as to enable robust and scalable solutions. We provide pointers to the rest of the book, where in-depth treatments are presented.
Keywords: Mathematics of perception, multimedia data management, multimedia information retrieval.

1.1 Overview

The tasks of multimedia information management such as clustering, indexing, and retrieval, come up against technical challenges in at least three areas: data representation, similarity measurement, and scalability. First, data representation builds layers of abstraction upon raw multimedia data. Next, a distance function must be chosen to properly account for similarity between any pair of multimedia instances. Finally, from extracting features, measuring similarity, to organizing and retrieving data, all computation tasks must be performed in a scalable fashion with respect to both data dimensionality and data volume. This chapter outlines design issues of five essential subroutines, and they are:

1. Feature extraction,
2. Similarity (distance function formulation),
3. Learning (supervised and unsupervised),
4. Multimodal fusion, and
5. Indexing.

1.2 Feature Extraction

Feature extraction is fundamental to all multimedia computing tasks. Features can be classified into two categories, *content* and *context*. Content refers directly to raw imagery, video, and mucic data such as pixels, motions, and tones, respectively, and their representations. Context refers to metadata collected or associated with content when a piece of data is acquired or published. For instance, EXIF camera parameters and GPS location are contextual information that some digital cameras can collect. Other widely used contextual information includes surrounding texts of an image/photo on a Web page, and social interactions on a piece of multimedia data instance. Context and content ought to be fused synergistically when analyzing multimedia data [1].

Content analysis is a subject studied for more than a couple of decades by researchers in disciplines of computer vision, signal processing, machine learning, databases, psychology, cognitive science, and neural science. Limited progress has been made in each of these disciplines. Many researchers now are convinced that interdisciplinary research is essential to make ground breaking advancements. In Chapter 2 of this book, we introduce a model-based and data-driven hybrid approach for extracting features. A promising model-based approach was pioneered by neural scientist Hubel [2], who proposed a feature learning pipeline based on human visual system. The principal reason behind this approach is that human visual system can function so well in some challenging conditions where computer vision solutions fail miserably. Recent neural-based models proposed by Lee [3] and Serre [4] show that such model can effectively deal with viewing of different positions, scales, and resolutions. Our empirical study confirmed that such model-based approach can recognize objects of rigid shapes, such as watches and cars. However, for objects that do not have invariant features such as pizzas of different toppings, and cups of different colors and shapes, the model-based approach loses its advantages. For recognizing these objects, the data-driven approach can depict an object by collecting a representative pool of training instances. When combining model-based and data-driven, the hybrid approach enjoys at least three advantages:

1. *Balancing feature invariance and selectivity*. To achieve feature selectivity, the hybrid approach conducts multi-band, multi-scale, and multi-orientation convolutions. To achieve invariance, it keeps signals of sufficient strengths via pooling operations.
2. *Properly using unsupervised learning to regularize supervised learning*. The hybrid approach introduces unsupervised learning to reduce features so as to prevent the subsequent supervised layer from learning trivial solutions.
3. *Augmenting feature specificity with diversity*. A model-based only approach cannot effectively recognize irregular objects or objects with diversified patterns; and therefore, we must combine such with a data-driven pipeline.

Chapter 2 presents the detailed design of such a hybrid model involving disciplines of neural science, machine learning, and computer vision.