Inf Retrieval (2010) 13:398–402
DOI 10.1007/s10791-010-9128-6

BOOK REVIEW

David Gibbon, Zhu Liu: Introduction to video search engines

Springer, 2008, 276 pp, price: 64.95 euros, ISBN: 978-3-540-79336-6

Colum Foley

Published online: 16 February 2010

Video search is a complex and wide ranging research domain combining researchers across many disciplines. The amount of video content on the web has exploded in recent times with more and more broadcasters making their content available online and end users generating their own content and sharing it on websites such as YouTube.

Research into video processing and retrieval has been ongoing for many years and whilst full semantic understanding of the content of multimedia artifacts is still some way off, much progress has been made into developing automated processing solutions to attach some meaning to the content of video. Many groups have proposed content processing and delivery systems, indeed a look at the roll-call of groups involved in the annual TRECVid (Smeaton et al. 2006) benchmarking workshop over previous years provides an appreciation for the efforts in this community. Despite this progress, however, when searching for video on the internet our interactions are still mainly through a basic keyword search.

As such this book by David C. Gibbon and Zhu Liu from AT&T is a timely analysis of the current state of the art in video search; it provides the reader with an understanding and grounding in the major processes involved in delivering a rich multimedia experience to the end user. As the authors state, this book is intended for senior undergraduates or first-year graduate students. For the beginner it provides a good introduction to the area of video search. The reader is given an understanding of the types of video content available on the web today, how this content is described through metadata, and the content processing components in video, audio, and text which enable features to be extracted for higher level processing. For practitioners operating in the area the book can provide an appreciation of aspects of video search that they might not have direct experience with.

The book is divided into three main sections. Section 1 deals with the background and fundamentals of video search engines including video types, sources and associated metadata, and the specific issues involved with delivering internet video. Section 2 deals specifically with media processing, which is taking a video, audio, or a text signal, and extracting information in the form of features. Section 3 deals with case studies showing practical systems in use today, the idea being to showcase some of the underlying

C. Foley (✉)
CLARITY: Centre for Sensor Web Technologies, Dublin City University, Dublin 9, Dublin, Ireland
e-mail: colum.foley@computing.dcu.ie; columfoley@gmail.com

technologies discussed in previous chapters as they operate in state of the art video search systems.

Chapter 1 outlines the different types of video web sites and content providers that exist on the internet, including traditional broadcasters, aggregators, sharing sites, and domain specific sites. Using a cost/value categorization, the chapter discusses the different sources of video found on today's internet working its way from low cost web-cams to high cost production quality TV series and feature films. The chapter discusses the challenges of video search by comparing the issues faced by video search engines to their textual counterparts, including multiple file formats, data transport, and browsing of content. The chapter concludes by outlining some advantages video data has over textual data.

Chapter 2 provides a thorough description of the metadata available for different types of video. The authors provide good motivation for the importance of metadata for content retrieval as it "captures the subjective essence of the media". The chapter discusses the major types of metadata standards in use today including Dublin Core, MPEG-7, and MPEG-21. The authors then describe the types of metadata found across digital video beginning firstly with the essential video metadata common across all media such as file size, title, type, bit rates etc. Next is a discussion on the types of metadata specific to personal media collections, broadcast television, video on demand, and timed text formats. The chapter also discusses RSS (Really Simple Syndication), which the authors describe as the de-facto metadata standard for Web media.

Chapter 3: Internet Video provides the reader with an appreciation of the complexities involved in delivering video over the internet. The authors discuss the features of digital video including aspect ratio, frame rates, and compression rates and the standards that exist. Next the authors discuss the types of internet media systems and the issues involved in the delivery of IP video including transport, rights management, multi-rate encoding, the delivery of a television like experience via IPTV, and the emergence of flash video as a means for greater interoperability.

Chapter 4 concludes the first section and background of the book. It begins by discussing three typical architectural components common to video search engines: content acquisition, content processing, and content retrieval. The authors describe how a user typically interacts with a video search engine by looking at the interactive cycle of video search and the granularity of search results. The chapter provides a comprehensive discussion on the factors affecting scalability in each of the three main stages of a video search system previously described. The chapter concludes with an outline of the various interfaces which enable developers to access video search services and a summary of typical system features or services users can expect from a video search engine.

Chapter 5 is an introductory chapter for the second section of the book and provides an overview of media processing. The chapter introduces the elements of media processing common to video, audio, and text processing, discussing in turn the processes of feature extraction and feature selection, media segmentation, and clustering and structure generation.

Chapter 6 introduces and discusses components for video content analysis namely shot boundary detection, representative image selection, face detection, face recognition, optical character recognition, concept detection, and video browsing. Shot boundary detection, the segmentation of video into its constituent shots, is discussed by using as an example the shot boundary detection system developed by AT&T and evaluated at the annual TRECVid benchmarking workshop in 2006. Next the authors discuss the process of selecting representative images from the shot, from the simple method of choosing the first, last, or middle frame of the video, to more complex methods based on heuristics and taking

into account camera motion and object motion. Face detection and recognition is next and the authors highlight the difficulty of detecting faces due to both environmental changes and the orientation of faces. The authors discuss the process of video optical character recognition (OCR), highlighting the challenges of resolution and text size in video OCR as compared with OCR operating in still images before describing some methods in use today. Concept detection is used to derive high level semantic concepts from video shots based on extracting low-level features and learning patterns of occurrence of these features using machine learning techniques. The final section of the chapter is devoted to exploring different types of video browsing interfaces from state of the art video retrieval systems.

Chapter 7 discusses the fundamentals of audio analysis, content-based audio processing and indexing techniques, and audio query and browsing methods. The chapter begins by describing the fundamentals of audio as it is represented in both the time and frequency domains. Next the authors discuss the two levels of audio features, frame-level and clip level. Audio segmentation, finding abrupt changes along the audio stream, is discussed next and the authors describe approaches for both speaker segmentation and audio scene segmentation. Audio content categorization attempts to classify each audio segment into predefined categories based on its low-level features. Here the authors describe three different audio classification situations in detail: speaker recognition, speech/non-speech classification, and music genre classification. Automatic speech recognition (ASR) is a process to convert spoken speech into a sequence of words. The authors describe the commonly used statistical n-gram language model approach to ASR. The chapter concludes with a review of state of the art audio querying and browsing techniques.

Chapter 8 discusses text processing techniques. Text associated with video data can provide crucial information as to the content of the video, for example the subtitles distributed with DVDs and many TV programs provide a full transcript of the spoken text. In this chapter the authors describe some common text processing techniques including story segmentation, named entity extraction, part of speech tagging, capitalization, information retrieval, and text summarization. Story segmentation can be used to divide a program into cohesive stories and the authors discuss different approaches from the literature. Next the authors describe named-entity extraction, the extraction of elements with their associated categories from sentences (e.g. “Bill” is a person’s name). Part-of-speech tagging is used to associate words in text to particular parts of speech such as nouns, verbs, and adjectives. The authors describe capitalization which they say is an important factor in determining the quality of speech transcripts and for identifying parts of speech. Information retrieval is described next, in particular the authors describe the processes of stemming, term weighting, and ranking of documents. The chapter concludes with a description of the process of text summarization.

Chapter 9 concludes the second section of the book on media processing by discussing multimodal processing. The authors describe how a multimedia document’s semantics are embedded in multiple forms that are usually complimentary to each other. Using case-studies, and calling on the fundamentals of media processing described in the previous three chapters, the authors describe three multimodal processing modules through case studies: closed caption alignment using both audio and text, multimodal news story segmentation using audio, video, and text, and major cast detection using face detection and speech.

The final section of the book begins with Chapter 10 which provides a brief look at a handful of the well known research systems and some commercial systems in the area of video search. The chapter also looks at some early internet deployments, focusing on systems which automate media analysis for retrieval. Next the authors talk about some of

the resources available for evaluating the quality of video retrieval systems. The chapter concludes with a case study of the AT&T MIRACLE (Multimedia Information Retrieval by Content) system, and provides details of the different components of the system including data acquisition, content processing, and querying.

In the concluding chapter of the book, Chapter 11, the authors observe certain trends in video search giving the reader a general direction of the ongoing research in its constituent technologies. In particular the authors discuss trends in video production, distribution, the web and user interaction, television technology, and new media devices.

Discussion

The opening section of the book provides a good introduction to the challenges of video search and sets the scene nicely for the remainder of the book. The final chapter of the section, Chapter 4, provides a nice overview of the main architectural components of a video search system. Here, I found the discussion on content acquisition very informative as the authors covered topics such as metadata normalization, user contributed content, syndicated contribution, and broadcast acquisition. For content processing the authors provide a high level overview of the stages involved in processing video and managing the video asset, and while the discussion lacked depth here this is understandable given that the topic occupies the entire second section of the book. The section on retrieval also lacked sufficient depth however, unlike content processing, I feel that the authors never fully return to retrieval and search later in the book in sufficient detail.

The second section of the book covers media processing. This section of the book provides good background on the fundamentals of processing in each of the three main modalities of video, audio, and text. Personally, I found it a good revision for text and video processing and for audio processing it provided me with an appreciation of the aspects of the process that I was less familiar with. While the description of the shot boundary detection system in Chapter 6 was very thorough, the first section on feature extraction has a slightly rushed feel to it. The reader is introduced to many new concepts quickly. As feature extraction plays such a major role in the video processing process I feel that the underlying concepts such as color histograms, motion intensity etc. should have been teased out in greater detail here so as not to lose the reader. In Chapter 8: Text Processing, the discussion on ranking felt a little lightweight, the authors only briefly mention term weighted retrieval, and no mention is made of the common probabilistic or vector space models for retrieval.

The final section of the book describes case-studies of practical video search systems in use today. I found this section interesting as it both informs the reader of how the fundamentals described in previous chapters are put to use in real world systems and provides a sense of the direction of the trends in the technology. One area I was hoping to see more detail in was in the evaluation of video search. Whilst the authors devote a section in Chapter 10 to the topic, the section reads more of an overview of resources available to researchers and practitioners rather than an in depth description of the process of evaluation which I was hoping for. The authors do not tackle issues of evaluation metrics or performance measures here, rather pointers are given to multimedia retrieval datasets available and ongoing evaluation initiatives.

Overall I very much enjoyed reading this book and would recommend it to colleagues. Coming at it as an information retrieval researcher with some experience in video search I found this book provided me with an appreciation for some aspects of video search I was less familiar with including metadata standards, the types and sources of video content

available, and the processing of video. I feel that this book will serve as a useful reference for me in future work. I would also recommend this book to others working in information retrieval who want to understand video content and the processing of this multimedia data source. As a complete text on video search however, I feel that this book is lacking the required depth in the search and retrieval aspect of the process. In particular I would have liked to have seen more emphasis given to search processes including weighting, retrieval, multimodal fusion, and video search evaluation.

Reference

Smeaton, A. F., Over, P., & Kraaij, W. (2006). Evaluation campaigns and TRECVid. In Proceedings of the 8th ACM international workshop on multimedia information retrieval (Santa Barbara, California, USA, 26–27, October, 2006). MIR '06. ACM Press, New York, NY, pp. 321–330. doi:http://doi.acm.org/10.1145/1178677.1178722.