Socio-Affective Computing 6

Rajiv Shah
Roger Zimmermann

Multimodal Analysis of User-Generated Multimedia Content

Springer

Socio-Affective Computing

Volume 6

Series Editor
Amir Hussain, University of Stirling, Stirling, UK

Co-Editor
Erik Cambria, Nanyang Technological University, Singapore

This exciting Book Series aims to publish state-of-the-art research on socially intelligent, affective and multimodal human-machine interaction and systems. It will emphasize the role of affect in social interactions and the humanistic side of affective computing by promoting publications at the cross-roads between engineering and human sciences (including biological, social and cultural aspects of human life). Three broad domains of social and affective computing will be covered by the book series: (1) social computing, (2) affective computing, and (3) interplay of the first two domains (for example, augmenting social interaction through affective computing). Examples of the first domain will include but not limited to: all types of social interactions that contribute to the meaning, interest and richness of our daily life, for example, information produced by a group of people used to provide or enhance the functioning of a system. Examples of the second domain will include, but not limited to: computational and psychological models of emotions, bodily manifestations of affect (facial expressions, posture, behavior, physiology), and affective interfaces and applications (dialogue systems, games, learning etc.). This series will publish works of the highest quality that advance the understanding and practical application of social and affective computing techniques. Research monographs, introductory and advanced level textbooks, volume editions and proceedings will be considered.

More information about this series at http://www.springer.com/series/13199

Rajiv Shah • Roger Zimmermann

Multimodal Analysis of User-Generated Multimedia Content

Springer

Rajiv Shah
School of Computing
National University of Singapore
Singapore, Singapore

Roger Zimmermann
School of Computing
National University of Singapore
Singapore, Singapore

ISSN 2509-5706 ISSN 2509-5714 (electronic)
Socio-Affective Computing
ISBN 978-3-319-61806-7 ISBN 978-3-319-61807-4 (eBook)
DOI 10.1007/978-3-319-61807-4

Library of Congress Control Number: 2017947053

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

I dedicate this book to my late father,
Ram Dhani Gupta;
my mother, Girija Devi; and my other family
members for their
continuous support, motivation, and
unconditional love.
I love you all so dearly.

Foreword

We have stepped into an era where every user plays the role of both content provider and content consumer. With many smartphone apps seamlessly converting photographs and videos to social media postings, user-generated multimedia content now becomes the next big data waiting to be turned into useful insights and applications. The book *Multimodal Analysis of User-Generated Multimedia Content* by Rajiv and Roger very carefully selects a few important research topics in analysing big user-generated multimedia data in a multimodal approach pertinent to many novel applications such as content recommendation, content summarization and content uploading. What makes this book stand out among others is the unique focus on multimodal analysis which combines visual, textual and other contextual features of multimedia content to perform better sensemaking.

Rajiv and Roger have made the book a great resource for any reader interested in the above research topics and respective solutions. The literature review chapter gives a very detailed and comprehensive coverage of each topic and comparison of state-of-the-art methods including the ones proposed by the authors. Every chapter that follows is dedicated to a research topic covering the architecture framework of a proposed solution system and its function components. This is accompanied by a fine-grained description of the methods used in the function components. To aid understanding, the description comes with many relevant examples. Beyond describing the methods, the authors also present the performance evaluation of these methods using real-world datasets so as to assess their strengths and weaknesses appropriately.

Despite its deep technical content, the book is surprisingly easy to read. I believe the authors have paid extra attention to organizing the content for easy reading, careful proof editing and good use of figures and examples. The book is clearly written at the level suitable for reading by computer science students in graduate and senior years. It is also a good reference reading for multimedia content analytics researchers in both academia and industry. Whenever appropriate, the authors show their algorithms with clearly defined input, output and steps with

comments. This facilitates any further implementation as well as extensions of the methods. This is perhaps the part of the book which will attract the "programming-type" readers most.

I would like to congratulate both Rajiv and Roger for their pioneering work in multimodal analysis for user-generated multimedia content. I believe this book will become widely adopted and referenced in the multimedia community. It is a good guide for anyone who wishes to better understand the challenges and solutions of analysing multimedia data. I wish the authors all the best in their future research endeavour.

School of Information Systems,
Singapore Management University,
Singapore, Singapore
May 2017

Ee-Peng Lim, PhD

Preface

The amount of user-generated multimedia content (UGC) has increased rapidly in recent years due to the ubiquitous availability of smartphones, digital cameras, and affordable network infrastructures. An interesting recent trend is that social media websites such as Flickr and YouTube create opportunities for users to generate multimedia content, instead of creating multimedia content by themselves. Thus, capturing UGC such as user-generated images (UGIs) and user-generated videos (UGVs) anytime and anywhere, and then instantly sharing them on social media platforms such as Instagram and Flickr, have become a very popular activity. Hence, user-generated multimedia content is now an intrinsic part of social media platforms. To benefit users and social media companies from an automatic semantics and sentics understanding of UGC, this book focuses on developing effective algorithms for several significant social media analytics problems. Sentics are common affective patterns associated with natural language concepts exploited for tasks such as emotion recognition from text/speech or sentiment analysis. Knowledge structures derived from the semantics and sentics understanding of user-generated multimedia content are beneficial in an efficient multimedia search, retrieval, and recommendation. However, real-world UGC is complex, and extracting the semantics and sentics from only multimedia content is very difficult because suitable concepts may present in different representations. Moreover, due to the increasing popularity of social media websites and advancements in technology, it is possible now to collect a significant amount of important contextual information (e.g., spatial, temporal, and preference information). Thus, it necessitates analyzing the information of UGC from multiple modalities for a better semantics and sentics understanding. Moreover, the multimodal information is very useful in a social network based news video reporting task (e.g., citizen journalism) which allows people to play active roles in the process of collecting news reports (e.g., CNN iReport). Specifically, we exploit both content and contextual information of UGIs and UGVs to facilitate different multimedia analytics problems.

Further advancements in technology enable mobile devices to collect a significant amount of contextual information in conjunction with captured multimedia content. Since the contextual information greatly helps in the semantics and sentics understanding of user-generated multimedia content, researchers exploit it in their research work related to multimedia analytics problems. Thus, the multimodal information (i.e., both content and contextual information) of UGC benefits several diverse social media analytics problems. For instance, knowledge structures extracted from multiple modalities are useful in an effective multimedia search, retrieval, and recommendation. Specifically, applications related to multimedia summarization, tag ranking and recommendation, preference-aware multimedia recommendation, multimedia-based e-learning, and news video reporting are built by exploiting the multimedia content (e.g., visual content) and associated contextual information (e.g., geo-, temporal, and other sensory data). However, it is very challenging to address these problems efficiently due to the following reasons: (i) difficulty in capturing the semantics of UGC, (ii) the existence of noisy metadata, (iii) difficulty in handling big datasets, (iv) difficulty in learning user preferences, (v) the insufficient accessibility and searchability of video content, and (vi) weak network infrastructures at some locations. Since different information knowledge structures are derived from different sources, it is useful to exploit multimodal information to overcome these challenges.

Exploiting information from multiple sources helps in addressing challenges mentioned above and facilitating different social media analytics applications. Therefore, in this book, we leverage information from multiple modalities and fuse the derived knowledge structures to provide effective solutions for several significant social media analytics problems. Our research focuses on the semantics and sentics understanding of UGC leveraging both content and contextual information. First, for a better understanding of an event from a large collection of UGIs, we present the EventBuilder system. It enables people to automatically generate a summary of the event in real-time by visualizing different social media such as Wikipedia and Flickr. In particular, we exploit Wikipedia as the event background knowledge to obtain more contextual information about the event. This information is very useful in an effective event detection. Next, we solve an optimization problem to produce text summaries for the event. Subsequently, we present the EventSensor system that aims to address sentics understanding and produces a multimedia summary for a given mood. It extracts concepts and mood tags from the visual content and textual metadata of UGCs and exploits them in supporting several significant multimedia analytics problems such as a musical multimedia summary. EventSensor supports sentics-based event summarization by leveraging EventBuilder as its semantics engine component. Moreover, we focus on computing tag relevance for UGIs. Specifically, we leverage personal and social contexts of UGIs and follow a neighbor voting scheme to predict and rank tags. Furthermore, we focus on semantics and sentics understanding from UGVs since they have a significant impact on different areas of a society (e.g., enjoyment, education, and journalism).

Since many outdoor UGVs lack a certain appeal because their soundtracks consist mostly of ambient background noise, we solve the problem of making UGVs more attractive by recommending a matching soundtrack for a UGV by exploiting content and contextual information. In particular, first, we predict scene moods from a real-world video dataset. Users collected this dataset from their daily outdoor activities. Second, we perform heuristic rankings to fuse the predicted confidence scores of multiple models, and, third, we customize the video soundtrack recommendation functionality to make it compatible with mobile devices. Furthermore, we address the problem of knowledge structure extraction from educational UGVs to facilitate e-learning. Specifically, we solve the problem of topic-wise segmentation for lecture videos. To extract the structural knowledge of a multi-topic lecture video and thus make it easily accessible, it is very desirable to divide each video into shorter clips by performing an automatic topic-wise video segmentation. However, the accessibility and searchability of most lecture video content are still insufficient due to the unscripted and spontaneous speech of speakers. We present the ATLAS and TRACE systems to perform the temporal segmentation of lecture videos automatically. In our studies, we construct models from visual, transcript, and Wikipedia features to perform such topic-wise segmentations of lecture videos. Moreover, we investigate the late fusion of video segmentation results derived from state-of-the-art methods by exploiting the multimodal information of lecture videos. Finally, we consider the area of journalism where UGVs have a significant impact on society.

We propose algorithms for news video (UGV) reporting to support journalists. An interesting recent trend, enabled by the ubiquitous availability of mobile devices, is that regular citizens report events which news providers then disseminate, e.g., CNN iReport. Often such news are captured in places with very weak network infrastructure, and it is imperative that a citizen journalist can quickly and reliably upload videos in the face of slow, unstable, and intermittent Internet access. We envision that some middleboxes are deployed to collect these videos over energy-efficient short-range wireless networks. In this study we introduce an adaptive middlebox design, called NEWSMAN, to support citizen journalists. Specifically, the NEWSMAN system jointly considers two aspects under varying network conditions: (i) choosing the optimal transcoding parameters and (ii) determining the uploading schedule for news videos. Finally, since the advances in deep neural network (DNN) technologies enabled significant performance boost in many multimedia analytics problems (e.g., image and video semantic classification, object detection, face matching and retrieval, text detection and recognition in natural scenes, and image and video captioning), we discuss their roles to solve several multimedia analytics problems as part of future directions to readers.

Singapore, Singapore — Rajiv Ratn Shah
Singapore, Singapore — Roger Zimmermann
May 2017

Acknowledgements

Completing this book has been a truly life-changing experience for me, and it would not have been possible to do without the blessing of God. I praise and thank God almighty for giving me strength and wisdom throughout my research work to complete this book. I am grateful to numerous people who have contributed toward shaping this book.

First and foremost, I would like to thank my Ph.D. supervisor Prof. Roger Zimmermann for his great guidance and support throughout my Ph.D. study. I would like to express my deepest gratitude to him for encouraging my research and empowering me to grow as a research scientist. I could not have completed this book without his invaluable motivation and advice. I would like to express my appreciation to the following professors at the National University of Singapore (NUS) for their extremely useful comments: Prof. Mohan S. Kankanhalli, Prof. Wei Tsang Ooi, and Prof. Teck Khim Ng. Furthermore, I would like to thank Prof. Yi Yu, Prof. Suhua Tang, Prof. Shin'ichi Satoh, and Prof. Cheng-Hsin Hsu who have supervised me during my internships at National Tsing Hua University, Taiwan, and National Institute of Informatics, Japan. I am also very grateful to Prof. Ee-Peng Lim and Prof. Jing Jiang for their wonderful guidance and support during my research work in the Living Analytics Research Centre (LARC) at Singapore Management University, Singapore. A special thanks goes to Prof. Ee-Peng Lim for writing the foreword for this book.

I am very much thankful to all my friends who have contributed immensely to my personal and professional time in different universities, cities, and countries during my stay there. Specifically, I would like to thank Yifang Yin, Soujanya Poria, Deepak Lingwal, Vishal Choudhary, Satyendra Yadav, Abhinav Dwivedi, Brahmraj Rawat, Anwar Dilawar Shaikh, Akshay Verma, Anupam Samanta, Deepak Gupta, Jay Prakash Singh, Om Prakash Kaiwartya, Lalit Tulsyan, Manisha Goel, and others. I would also like to acknowledge my debt to my friends and relatives for encouraging throughout my research work. Specifically, I would like to

thank Dr. Madhuri Rani, Rajesh Gupta, Priyanka Agrawal, Avinash Singh, Priyavrat Gupta, Santosh Gupta, and others for their unconditional support.

Last but not the least, I would like to express my deepest gratitude to my family. A special love goes to my mother, Girija Devi, who has been a great mentor in my life and had constantly encouraged me to be a better person, and my late father, Ram Dhani Gupta, who has been a great supporter and torchbearer in my life. The struggle and sacrifice of my parents always motivate me to work hard in my research work. The decision to leave my job as a software engineer and pursue higher studies was not easy for me, but I am grateful to my brothers Anoop Ratn and Vikas Ratn for supporting me in the time of need. Without love from my sister Pratiksha Ratn, my sisters-in-law Poonam Gupta and Swati Gupta, my lovely nephews Aahan Ratn and Parin Ratn, and my best friend Rushali Gupta, this book would not have been completed.

Singapore, Singapore Rajiv Ratn Shah
May 2017

Contents

About the Authors

Rajiv Ratn Shah received his B.Sc. with honors in mathematics from Banaras Hindu University (BHU), India, in 2005. He received his M.Tech. in computer technology and applications from Delhi Technological University (DTU), India, in 2010. Prior to joining Indraprastha Institute of Information Technology Delhi (IIIT Delhi), India, as an assistant professor, Dr. Shah has received his Ph.D. in computer science from the National University of Singapore (NUS), Singapore. Currently, he is also working as a research fellow in Living Analytics Research Centre (LARC) at the Singapore Management University (SMU), Singapore. His research interests include the multimodal analysis of user-generated multimedia content in the support of social media applications, multimodal event detection and recommendation, and multimedia analysis, search, and retrieval. Dr. Shah is the recipient of several awards, including the runner-up in the Grand Challenge competition of ACM International Conference on Multimedia 2015. He is involved in reviewing of many top-tier international conferences and journals. He has published several research works in top-tier conferences and journals such as Springer MultiMedia Modeling, ACM International Conference on Multimedia, IEEE International Symposium on Multimedia, and Elsevier *Knowledge-Based Systems*.

Roger Zimmermann is associate professor of computer science in the *School of Computing* at the *National University of Singapore* (NUS). He is also deputy director with the *Interactive and Digital Media Institute* (IDMI) at NUS and codirector of the *Centre of Social Media Innovations for Communities* (COSMIC), a research institute funded by the National Research Foundation (NRF) of Singapore. Prior to joining NUS, he held the position of research area director with the Integrated Media Systems Center (IMSC) at the University of Southern California (USC). He received his M.S. and Ph.D. degree from the University of Southern California in 1994 and 1998, respectively. Among his research interests are mobile video management, streaming media architectures, distributed and peer-to-peer systems, spatiotemporal data management, and location-based services. Dr. Zimmermann is a senior member of *IEEE* and a member of *ACM*. He has

coauthored a book, seven patents, and more than 220 conference publications, journal articles, and book chapters in the areas of multimedia, GIS, and information management. He has received funding from NSF (USA), A*STAR (Singapore), NUS Research Institute (NUSRI), NRF (Singapore), and NSFC (China) as well as several industries such as Fuji Xerox, HP, Intel, and Pratt & Whitney. Dr. Zimmermann is on the editorial boards of the IEEE Multimedia Communications Technical Committee (MMTC) R-Letter and the Springer International Journal of *Multimedia Tools and Applications* (MTAP). He is also an associate editor for the *ACM Transactions on Multimedia Computing, Communications, and Applications* journal (ACM TOMM), and he has been elected to serve as secretary of ACM SIGSPATIAL for the term 1 July 2014 to 30 June 2017. He has served on the conference program committees of many leading conferences and as reviewer of many journals. Recently, he was the general chair of the ACM *Multimedia Systems 2014* and the *IEEE ISM 2015* conferences and TPC cochair of the *ACM TVX 2017* conference.

Abbreviations

UGC	User-generated content
UGI	User-generated image
UGT	User-generated text
UGV	User-generated video
HMM	Hidden Markov model
EventBuilder	Real-time multimedia event summarization by visualizing social media
EventSensor	Leveraging multimodal information for event summarization and concept-level sentiment analysis
DNN	Deep neural network
UTB	User tagging behavior
PD	Photo description
NV	Neighbor voting-based tag ranking system
NVGC	NV corresponding to geo concepts
NVVC	NV corresponding to visual concepts
NVSC	NV corresponding to semantics concepts
NVGVC	NV corresponding to the fusion of geo and visual concepts
NVGSC	NV corresponding to the fusion of geo and semantics concepts
NVVSC	NV corresponding to the fusion of visual and semantics concepts
NVGVSC	NV corresponding to the fusion of geo, visual, and semantics concepts
EF	Early fusion-based tag ranking system
LFE	NV based on late fusion of different modalities with equal weights
LFR	NV based on late fusion with weights determined by the recall of different modalities
DCG	Discounted cumulative gain
NDCG	Normalized discounted cumulative gain
PROMPT	A personalized user tag recommendation for social media photos leveraging personal and social contexts

CRAFT	Concept-level multimodal ranking of Flickr photo tags via recall-based weighting
ADVISOR	A personalized video soundtrack recommendation system
EAT	Emotion annotation tasks
AI	Artificial intelligence
NLP	Natural language processing
E-learning	Electronic learning
ATLAS	Automatic temporal segmentation and annotation of lecture videos based on modeling transition time
TRACE	Linguistic-based approach for automatic lecture video segmentation leveraging Wikipedia texts
NPTEL	National Programme on Technology Enhanced Learning
MIT	Massachusetts Institute of Technology
NUS	National University of Singapore
CNN	Cable News Network
NEWSMAN	Uploading videos over adaptive middleboxes to news servers
PSNR	Peak signal-to-noise ratio
R–D	Rate–distortion
TI	Temporal perceptual information
SI	Spatial perceptual information
Amazon EC2	Amazon Elastic Compute Cloud
EDF	Earlier deadline first
FIFO	First in, first out
SNG	Satellite news gathering
SMS	Short message service
MMS	Multimedia messaging service
FAQ	Frequently asked questions
MKL	Multiple kernel learning

Chapter 1
Introduction

Abstract The amount of user-generated multimedia content (UGC) has increased rapidly in recent years due to the ubiquitous availability of smartphones, digital cameras, and affordable network infrastructures. However, real-world UGC is complex, and extracting the semantics and sentics from only multimedia content is very difficult because suitable concepts may be exhibited in different representations. Since it is possible now to collect a significant amount of relevant contextual information due to advancements in technology, we analyze the information of UGC from multiple modalities to facilitate different social media applications in this book. Specifically, we present our solutions for applications related to multimedia summarization, tag ranking and recommendation, preference-aware multimedia recommendation, multimedia-based e-learning, and news videos uploading by exploiting the multimedia content (*e.g.*, visual content) and associated contextual information (*e.g.*, geo-, temporal, and other sensory data). Moreover, we presented a detailed literature survey and future directions for research on user-generated multimedia content.

Keywords Semantics analysis • Sentics analysis • Multimodal analysis • User-generated multimedia content • Multimedia fusion • Multimedia analysis • Multimedia recommendation • Multimedia uploading

1.1 Background and Motivation

User-generated multimedia content (UGC) has become more prevalent and asynchronous in recent years with the advent of ubiquitous smartphones, digital cameras, affordable network infrastructures, and auto-uploaders. A survey [6] conducted by Ipsos MediaCT, Crowdtap, and the Social Media Advertising Consortium on 839 millennial persons (18–36 years old) indicates that (i) every day, millennials spend a significant amount of time with different types of media, (ii) they spend 30% of the total time with UGC, (iii) millennials prefer social media above all other media types, (iv) they trust information received through UGC 50% more than information from other media sources such as newspapers, magazines, and television advertisement, and (v) UGC is 20% more influential in purchasing decisions of Millennials than other media types. Thus, UGC such as

R. Shah, R. Zimmermann, *Multimodal Analysis of User-Generated Multimedia Content*, Socio-Affective Computing 6, DOI 10.1007/978-3-319-61807-4_1

user-generated texts (UGTs), user-generated images (UGIs), and user-generated videos (UGVs) play a pivotal role in e–commerce, specifically in social commerce. Moreover, instantly sharing UGC anytime and anywhere on social media platforms such as Twitter, Flickr, and NPTEL [3] has become a very popular activity. For instance, in a very popular photo sharing website Instagram, over 1 billion UGIs have been uploaded so far and it has more than 500 million monthly active users [11]. Similarly, over 10 billion UGIs have been uploaded so far in another famous photo sharing website Flickr[1] which has over 112 million users, and an average of 1 million UGIs has uploaded daily [10].

Thus, it is required to extract knowledge structures from UGC on such social media platforms to provide various multimedia-related services and solve several significant multimedia analytics problems. The extracted knowledge structures are very useful in the semantics and sentics understanding of UGC and facilitate several significant social media applications. Sentics are common affective patterns associated with natural language concepts exploited for tasks such as emotion recognition from text/speech or sentiment analysis [19]. Sentics computing is a multidisciplinary approach to natural language processing and understanding at the crossroads between affective computing, information extraction, and commonsense reasoning, which exploits both computer and human sciences to interpret better and process social information on the web [18]. Sentics is also the study of waveforms of touch, emotion, and music, and named by Austrian neuroscientist Manfred Clynes. However, it is a very challenging task to extract such knowledge structures because real-world UGIs and UGVs are complex and noisy, and extracting semantics and sentics from the multimedia content alone is a very difficult problem. Hence, it is desirable to analyze UGC from multiple modalities for a better semantics and sentics understanding. Different modalities uncover different aspects that are useful in determining useful knowledge structures. Such knowledge structures are exploited in solving different multimedia analytics problems.

In this book, we investigate the usage of multimodal information and the fusion of user-generated multimedia content in facilitating different multimedia analytics problems [242, 243]. First, we focus on the semantics and sentics understanding of UGIs to address the multimedia summarization problem. Such summaries are very useful in providing overviews of different events automatically without looking into the vast amount of multimedia content. We particularly address problems related to recommendation and ranking of user tags, summarization of events, and sentics-based multimedia summarization. These problems are very important in providing different significant services to users. For instance, recommendation and ranking of user tags are very beneficial in an effective multimedia search and retrieval. Moreover, multimedia summarization is very useful in providing an overview of a given event. Subsequently, we also focus on the semantics and sentics understanding of UGVs. Similar to the processing of UGIs, we exploit the multimodal

[1] www.flickr.com

information in the semantics and sentics understanding of UGVs, and address several significant multimedia analytics problems such as soundtrack recommendation for UGVs, lecture videos segmentation, and news videos uploading. All such UGVs have a significant impact on a society. For instance, soundtrack recommendation enhances the viewing experience of a UGV, lecture videos segmentation assist in e–learning, and news videos uploading supports citizen journalists.

Capturing UGVs has also become a very popular activity in recent years due to advancements in the manufacturing of mobile devices (*e.g.*, smartphones and tablets) and network engineering (*e.g.*, wireless communications). People now can easily capture UGVs anywhere, anytime, and instantly share their real-life experiences via social websites such as Flickr and YouTube. Enjoying videos has become a very popular entertainment as compared to traditional ways due to its easy access. Thus, besides traditional videos provided by professionals such as movies, music videos, and advertisements, UGVs are also getting higher popularity. UGVs are instantly shareable on social websites. For instance, video hosting services such as YouTube,[2] Vimeo,[3] Dailymotion,[4] and Veoh[5] allow individuals to upload their UGVs and share with others through their mobile devices. In the most popular video sharing website YouTube which has more than 1 billion users, everyday people watch hundreds of millions of hours of UGVs and generate billions of views [21]. Moreover, users upload 300 h of videos every minute on YouTube [21]. Almost 50% of the global viewing time comes from mobile devices, and this is expected to increase rapidly shortly because prices of mobile devices and wireless communications are getting much cheaper. Music videos enhance video watching experience because they do provide not only visual information but also involve music which matches with scenes and locations. However, many outdoor UGVs lack a certain appeal because their soundtracks consist mostly of ambient background noise (*e.g.*, environmental sounds such as cars passing by, *etc.*). Since sound is a very important aspect that contributes greatly to the appeal of a video when it is being viewed, a UGV with a matching soundtrack has more appeal for sharing on social media websites than a normal video without interesting sound.

Considering that a UGV with a matching soundtrack has more appeal for sharing on social media websites (*e.g.*, Flickr, Facebook, and YouTube) and with today's mobile devices that allow immediate sharing of UGC on such social media websites, it is desirable to easily and instantly generate an interesting soundtrack for the UGV before sharing. However, generating soundtracks for UGVs is not easy in the mobile environment due to the following reasons. Firstly, traditionally it is tedious and time-consuming for a user to add a custom soundtrack to a UGV. Secondly, an important aspect is that a good soundtrack should match and enhance the overall mood of the UGV and meet the user's preferences. Lastly, automatically

[2] www.youtube.com

[3] www.vimeo.com

[4] www.dailymotion.com

[5] www.veoh.com

generating a matching soundtrack for the UGV with less user intervention is a challenging task. Thus, it is necessary to construct a music video generation system that enhances the experience of viewing a UGV by adding a soundtrack that matches with both scenes of the UGV and the preferences of a user. In this book, we exploit both multimedia content such as visual features and contextual information such as spatial metadata of UGVs to determine sentics and generate music videos for UGVs. Our study confirms that multimodal information facilitates the understanding of user-generated multimedia content in the support of social media applications. Furthermore, we also consider two more areas where UGVs have a significant impact on a society: (1) education, and (2) journalism.

The number of digital lecture videos has increased dramatically in recent years due to the ubiquitous availability of digital cameras and affordable network infrastructures. Thus, multimedia-based e–learning systems which use electronic educational technologies as a platform for teaching and learning activities have become an important learning environment. It makes distance learning possible by enabling students to learn remotely without being in class. For instance, MIT OpenCourseWare [16] provides open access of virtually all MIT course content using a web-based publication. Now, it is possible to learn from experts in any area through e–learning (*e.g.*, MIT OpenCourseWare [16], and Coursera [12]), without any barriers such as time and distance. Many institutions such as National University of Singapore (NUS) have already started e–learning components in the practice of instructions to prepare themselves for continuing classes even if it is not possible for students to visit the campus due to certain calamities. Thus, e–learning helps in lowering cost, effective learning, faster delivery, and lowering environmental impact in educational learning systems. A long lecture video recording often discusses a specific topic of interest in only a few minutes within the video. Therefore, the requested information may bury within a long video that is stored along with thou-sands of others. It is often relatively easy to find the relevant lecture video in an archive, but the main challenge is to find the proper position within that video. Several websites such as VideoLectures.NET [20] which host lecture videos enable students to access different topics within videos using the annotation of segment boundaries derived from crowd-sourcing. However, the manual annotation of segment boundaries is very time-consuming, subjective, error-prone, and a costly process. Thus, it requires the implementation of a lecture video segmentation system which can automatically segment videos as accurately as possible even if qualities of lecture videos are not sufficiently high. Automatic lecture video segmentation will be very useful in e–learning when it combines with automatic topic modeling, indexing, and recommendation [31]. Subsequently, to facilitate journalists in the area with weak network infrastructures, we propose methods for efficient uploading of news videos.

Citizen journalism allows regular citizens to capture (news) UGVs and report events. Courtney C. Radsch defines citizen journalism as "*an alternative and activist form of newsgathering and reporting that functions outside mainstream media in-stitutions, often as a response to shortcomings in the professional journalistic field, that uses similar journalistic practices but is driven by different*

objectives and ideals and relies on alternative sources of legitimacy than traditional or mainstream journalism" [163]. Citizens often can report a breaking news more quickly than traditional news reporters due to the advancement in technology. For instance, on April 4, 2015, Feidin Santana, an American citizen recorded a video that showed a former South Carolina policeman shooting and killing the unarmed Michael Scott [7]. This video has gone viral on social media before it was taken up by any mainstream news channels. This video has helped in revealing the truth about this incident. Thus, the ubiquitous availability of smartphones and cameras has increased the popularity of citizen journalism. However, there is also some incident when any false news is reported by some citizen reporter that causes loss to some organization or person. For instance, Apple suffered a temporary drop in its stock due to a false report which is generated by CNN iReport about Steve Jobs' health in 2008 [1]. CNN allows citizens to report news using modern smartphones, tablets, and websites through its CNN iReport service. This service has more than 1 million citizen journalist users [5], who report news from places where traditional news reporters may not have access. Every month, it garners an average of 15,000 news reports and its content nets 2.6 million views [4]. It is, however, quite challenging for reporters to timely upload news videos, especially from developing countries, where Internet access is slow or even intermittent. Thus, it entails to enable regular citizens to report events quickly and reliably, despite weak network infrastructure at their places.

The presence of contextual information in conjunction with multimedia content has opened up interesting research avenues within the multimedia domain. Thus, the multimodal analysis of UGC is very helpful for an effective information access. It assists in an efficient multimedia analysis, retrieval, and services because UGC is often unstructured and difficult to access in a meaningful way. Moreover, it is difficult to extract relevant content from only one modality because suitable concepts may exhibit in different representations. Furthermore, multimodal information augments knowledge bases by inferring semantics from unstructured multimedia content and contextual information. Therefore, we leverage information from multiple modalities in our solutions to the problems mentioned above. Specifically, we exploit the knowledge structures derived from the fusion of heterogeneous media content to solve different multimedia analytics problems.

1.2 Overview

As illustrated in Fig. 1.1, this book concentrates on the multimodal analysis of user-generated multimedia content (UGC) in the support of social media applications. We determine semantics and sentics knowledge structures from UGC and leverage them in addressing several significant social media problems. Specifically, we present our solutions for five multimedia analytics problems that benefit by leveraging multimodal information such as multimedia content and contextual information (*e.g.*, temporal, geo-, crowdsourced, and other sensory data). First, we solve the

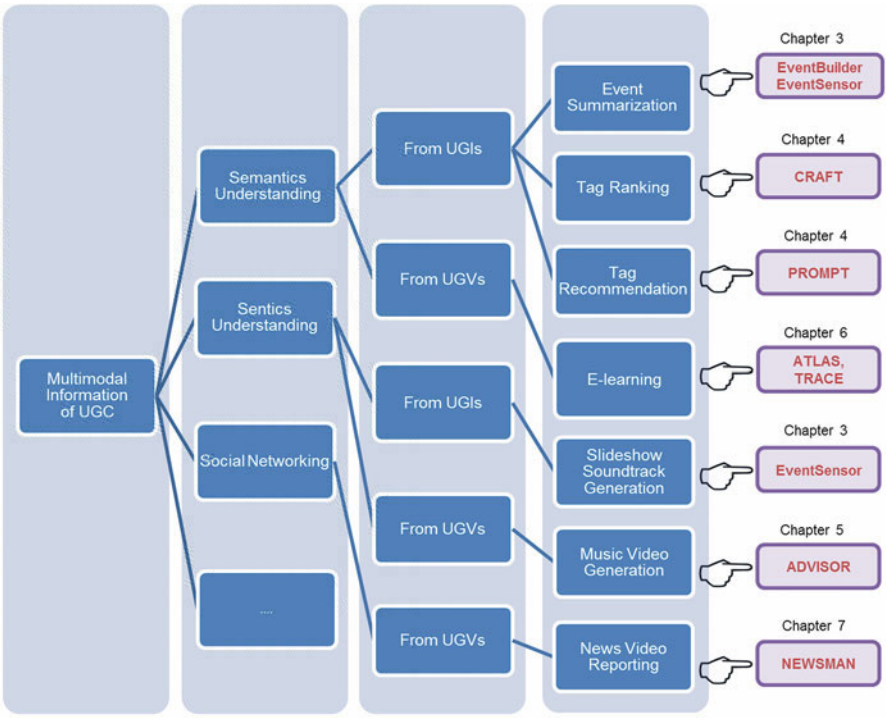

Fig. 1.1 Multimedia applications that benefit from multimodal information

problem of event understanding based on semantics and sentics analysis of UGIs on social media platforms such as Flickr [182, 186]. Subsequently, we address the problem of computing tag relevance for UGIs [181, 185]. Tag relevance scores determine tag recommendation and ranking of UGIs which are subsequently very useful in the searching and retrieval of relevant multimedia content. Next, we answer the problem of soundtrack recommendation for UGVs [187, 188]. A UGV with matching soundtrack enhance video viewing experience. Furthermore, we address research problems in two very important areas (journalism and education) where UGVs have a significant impact on society. Specifically, in the education area, we work out the problem of automatic lecture video segmentation [183, 184]. Finally, in the journalism area, we resolve the problem of user-generated news videos uploading over adaptive middleboxes to news servers in weak network infrastructures [180]. Experimental results have shown that our proposed approaches perform well. Contributions of each work are listed below:

1.2.1 Event Understanding

To efficiently browse multimedia content and obtain a summary of an event from a large collection of UGIs aggregated in social media sharing platforms such as Flickr and Instagram, we present the EventBuilder system. EventBuilder deals with semantics understanding and automatically generates a multimedia summary of a given event in real-time by leveraging different social media such as Wikipedia and Flickr. EventBuilder has two novel characteristics: (i) leveraging Wikipedia as event background knowledge to obtain additional contextual information about an input event, and (ii) visualizing an interesting event in real-time with a diverse set of social media activities. Subsequently, we enable users to obtain a sentics-based multimedia summary from the large collection of UGIs through our proposed sentics engine called, EventSensor. The EventSensor system addresses the sentics understanding from UGIs and produces a multimedia summary for a given mood. It supports sentics-based event summarization by leveraging EventBuilder as its semantics engine component. EventSensor extracts concepts and mood tags from visual content and textual metadata of UGC and exploits them in supporting several significant multimedia-related services such as a musical multimedia summary. Experimental results confirm that both EventBuilder and EventSensor outperform their baselines and effectively summarize knowledge structures on the YFCC100M dataset [201]. The YFCC100M dataset is a collection of 100 million photos and videos from Flickr.

1.2.2 Tag Recommendation and Ranking

Social media platforms such as Flickr allow users to annotate UGIs with descriptive keywords, called, tags which significantly facilitate the effective semantics understanding, search, and retrieval of UGIs. However, manual annotation is very time-consuming and cumbersome for most users, making it difficult to find relevant UGIs. Though there exist some deep neural networks based tag recommendation systems, tags predicted by such systems are limited because most of the available deep neural networks are trained with a few visual concepts. For instance, Yahoo's deep neural network can identify 1756 visual concepts from its publicly available dataset of 100 million UGIs and UGVs. However, the number of concepts that deep neural network can identify is rapidly increasing. For instance, the Google Cloud Vision API [14] can quickly classify photos into thousands of categories such as a *sailboat*, *lion*, and *Eiffel Tower*. Furthermore, Microsoft organized a challenge to recognize faces of 1 million celebrities [65]. Facebook claims to be working on identifying 100,000 objects. However, merely tagging a UGI with the identified objects may not describe the objective aspects of the UGI since often users tag UGIs with some user-defined concepts (*e.g.*, associate objects with some actions, attributes, and locations). Thus, it is very important to learn the tagging behavior of

users for tag recommendation. Moreover, recommended tags for a UGI are not necessarily relevant to users' interests. Furthermore, often annotated or predicted tags of a UGI are in a random order and even irrelevant to the visual content. Thus, it necessitates for automatic tag recommendation and ranking systems that consider users' interests and describe objective aspects of the UGI such as visual content and activities. To this end, this book presents a tag recommendation system, called, PROMPT, and a tag ranking system, called, CRAFT. Both systems leverage the multimodal information of a UGI to compute tag relevance. Specifically, for tag recommendation, first, we determine a group of users who have similar interests (tagging behavior) as the user of the UGI. Next, we find candidate tags from visual content and textual metadata leveraging tagging behaviors of users determined in the first step. Particularly, we determine candidate tags from the textual metadata and compute their confidence scores using asymmetric tag co-occurrence scores. Next, we determine candidate user tags from semantically similar neighboring UGIs and compute their scores based on voting counts. Finally, we fuse confidence scores of all candidate tags using a sum method and recommend top five tags to the given UGI. Similar to the neighbor voting based tag recommendation, we propose a tag ranking scheme based on a voting from the UGI neighbors derived from multimodal information. Specifically, we determine the UGI neighbors leveraging geo, visual, and semantics concepts derived from spatial information, visual content, and textual metadata, respectively. Experimental results on a test set from the YFCC100M dataset confirm that the proposed algorithm performs well. In the future, we can exploit our tag recommendation and ranking techniques in SMS/MMS bases FAQ retrieval [189, 190].

1.2.3 Soundtrack Recommendation for UGVs

Most of the outdoor UGVs are captured without much interesting background sounds (*i.e.*, environmental sounds such as cars passing by, *etc.*). Aimed at making outdoor UGVs more attractive, we introduce *ADVISOR*, a personalized video soundtrack recommendation system. We propose a fast and effective heuristic ranking approach based on heterogeneous late fusion by jointly considering three aspects: venue categories, visual scene, and the listening history of a user. Specifically, we combine confidence scores produced by SVM^{hmm} [2, 27, 75] models constructed from geographic, visual, and audio features, to obtain different types of video characteristics. Our contributions are threefold. First, we predict scene moods from a real-world video dataset that was collected from users' daily outdoor activities. Second, we perform heuristic rankings to fuse the predicted confidence scores of multiple models, and third, we customize the video soundtrack recommendation functionality to make it compatible with mobile devices. A series of extensive experiments confirm that our approach performs well and recommends appealing soundtracks for UGVs to enhance the viewing experience.