Rethinking Summarization and Storytelling for Modern Social Multimedia

Stevan Rudinac[1](✉), Tat-Seng Chua[2], Nicolas Diaz-Ferreyra[3], Gerald Friedland[4], Tatjana Gornostaja[5], Benoit Huet[6], Rianne Kaptein[7], Krister Lindén[8], Marie-Francine Moens[9], Jaakko Peltonen[10], Miriam Redi[11], Markus Schedl[12], David A. Shamma[13], Alan Smeaton[14], and Lexing Xie[15]

[1] University of Amsterdam, Amsterdam, The Netherlands
s.rudinac@uva.nl
[2] National University of Singapore, Singapore, Singapore
chuats@comp.nus.edu.sg
[3] Universität Duisburg-Essen, Duisburg, Germany
nicolas.diaz-ferreyra@uni-duisburg-essen.de
[4] University of California, Berkeley, USA
fractor@icsi.berkeley.edu
[5] Tilde, Riga, Latvia
gornostay@gmail.com
[6] EURECOM, Sophia Antipolis, France
Benoit.Huet@eurecom.fr
[7] Crunchr, Amsterdam, The Netherlands
amkaptein@hotmail.com
[8] University of Helsinki, Helsinki, Finland
krister.linden@helsinki.fi
[9] KU Leuven, Leuven, Belgium
sien.moens@cs.kuleuven.be
[10] Aalto University, Espoo, Finland
jaakko.peltonen@aalto.fi
[11] Nokia Bell Labs, Cambridge, UK
miriam.redi@gmail.org
[12] Johannes Kepler Universität Linz, Linz, Austria
markus.schedl@jku.at
[13] FX Palo Alto Laboratory, Inc., Palo Alto, USA
aymans@acm.org
[14] Dublin City University, Dublin, Ireland
alan.smeaton@dcu.ie
[15] Australian National University, Canberra, Australia
lexing.xie@anu.edu.au

Abstract. Traditional summarization initiatives have been focused on specific types of documents such as articles, reviews, videos, image feeds, or tweets, a practice which may result in pigeonholing the summarization

This position paper resulted from a working group discussion at the Schloss Dagstuhl seminar "User-Generated Content in Social Media", July 23–28, 2017, http://www.dagstuhl.de/17301.

K. Schoeffmann et al. (Eds.): MMM 2018, Part I, LNCS 10704, pp. 632–644, 2018.
https://doi.org/10.1007/978-3-319-73603-7_51

task in the context of modern, content-rich multimedia collections. Consequently, much of the research to date has revolved around mostly toy problems in narrow domains and working on single-source media types. We argue that summarization and story generation systems need to refocus the problem space in order to meet the information needs in the age of user-generated content in different formats and languages. Here we create a framework for flexible multimedia storytelling. Narratives, stories, and summaries carry a set of challenges in big data and dynamic multi-source media that give rise to new research in spatial-temporal representation, viewpoint generation, and explanation.

Keywords: Social multimedia · Summarization · Storytelling

1 Introduction

Social Multimedia [1] has been described as having three main components: content interaction between multimedia, social interaction around multimedia and social interaction captured in multimedia. Roughly speaking, this describes the interaction between traditional multimedia (photos and videos), mostly textual annotations on that media, and people interacting with that media. For almost a decade, fueled by the popularity of User-Generated Content (UGC), the bulk of research [2–8] has focused on meaningful extraction from any combination of these three points. With modern advancements in AI and computational resources [9,10], we now realize that multimedia summarization and story telling has worked in isolated silos, depending on the application and media (object detection, video summarization, Twitter sentiment, etc.); a broader viewpoint on the whole summarisation and reduction process is needed. Consequently, this realization gives rise to a second set of research challenges moving forward. In this paper, we revisit and propose to reshape the future challenges in multimedia summarization to identify a set of goals, prerequisites, and guidelines to address future UGC. Specifically, we address the problems associated with increasingly heterogeneous collections both in terms of multiple media and mixed content in different formats and languages, the necessity and complexities of dense knowledge extraction, and the requirements needed for sense making and storytelling.

2 Related Work

Summarization problems. Content summarisation problems arise in different application domains and are a long-standing interest of the natural language processing, computer vision, and multimedia research communities. Summarising long segments of text from a single or multiple documents is often done with extractive techniques, on which extensive surveys exist [11]. The problem of summarising image collections arises when there are e.g. large amounts of images from many users in a geographic area [12,13], or about a particular social event [14], or when it is necessary to generate a summarizing description (caption) [8]. Similarly, it is often needed to shorten or visualize long video sequences. Early solutions for video-to-image summarisation include automatic story-boards [15] and video summaries in forms of manga comic-book layout [16]. Audiovisual video summaries involve the processing of both audio and visual channels through e.g. joint optimisation of cross-modal coherence [17], or matching of audio segments [18]. In recent years, researchers have explored the summarization of ego-centric or surveillance videos by detecting important objects and actions [19] or constructing a map of key people in a known environment [20]. In the last decade, research on video summaries for real-world events increasingingly focused on large-scale social events reported online [18,21]. This position paper examines the summarization problem more broadly, taking a step back from one particular media format to be summarized, and targeting a large range of applications.

Relevance criteria for summarization. Early approaches to information retrieval (IR) and summarization focused on relevance of the content presented to the user. However, by the end of 90s the community realized that users prefer diversified search results and summaries instead of results lists produced based on relevance criterion only [15]. While the application domains varied, since then most summarization approaches focused on finding a balance between relevance, representativeness and diversity. The Informedia project is one of the best known early examples following such paradigm in addressing, amongst others, the problem of video summarization [22]. However, as users may be more sensitive to irrelevant than (near) duplicate items, enforcing diversity without hurting relevance is very challenging. This is witnessed by a large body of research on e.g. image search diversification [23–27]. Social multimedia summarization has further found its way in diverse applications ranging from personalized tweet summarization [28] to visual summarization of geographic areas and tourist routes [12,13,23,29] for POI recommendation and exploration. With the increased availability of affordable wearables, in recent years lifelogging has started gaining popularity, where the goal is to generate a diary or a record of the day's activities and happenings by creating a summary or a story from the video/image data gathered [30,31]. Progress has been made in summarizing heterogeneous user-generated content with regards to relevance, representativeness, and diversity [15]. However, relevance criteria and their interplay may be much more complex than commonly assumed [12] and, in case of visual content, include additional factors such as content popularity, aesthetic appeal and sentiment. Thus we call for rethinking the foundations of summarization and storytelling.

Benchmarks and formalization efforts. For almost two decades, common datasets, tasks, and international benchmarks fueld research on summarization and storytelling [32–34]. A typical task involved automatically generating a shorter (e.g. 100-word) summary of a set of news articles. TRECVID BBC Rushes summarization was probably the first systematic effort in the multimedia and computer vision communities focusing on video summarization [35]. The task involved reducing a raw and unstructured video captured during the recording of a TV series to a short segment of just a couple of minutes. Another well-known example is the ImageCLEF 2009 Photo Task, which revolved around image diversification [25]. The participants were expected to produce image search results covering multiple aspects of a news story, such as the images of "Hillary Clinton", "Obama Clinton" and "Bill Clinton" for a query "Clinton". Image search diversification has also been a topic of an ongoing MediaEval Diverse Social Images Task, run annually since 2013 [27].

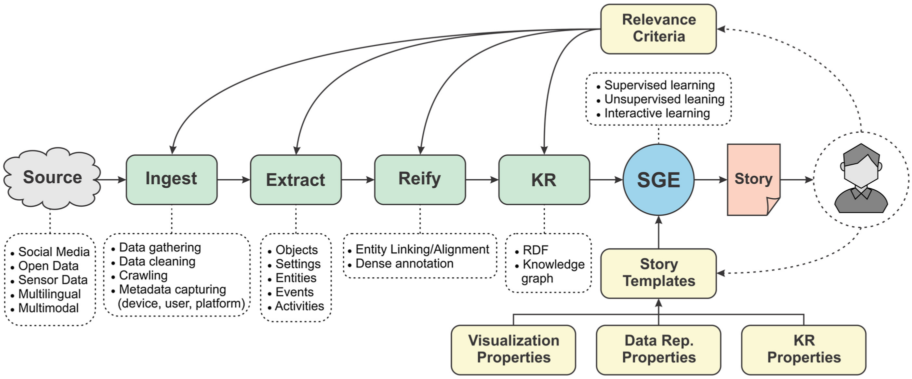

Fig. 1. Pipeline of our proposed framework for generating narratives, stories and summaries from heterogeneous collections of user generated content and beyond.

Although many people intuitively understand the concept of summarization, the complexity of the problem is best illustrated by the difficulties in even unequivocally defining a summary [36]. So, instead of focusing on strict definitions, most benchmarks took a pragmatic approach by conducting either intrinsic or extrinsic evaluation [32]. In intrinsic evaluation an automatically generated summary is compared directly against a "standard", such as summaries created by the humans. On the other hand, extrinsic evaluation measures the effectiveness of a summary in fulfilling a particular task as compared with the original set of documents (e.g. text, images or videos). Over the years many interesting metrics for evaluating (text) summaries were proposed, such as recall-oriented understudy for gisting evaluation (ROUGE) [37], bilingual evaluation understudy (BLEU) [38] and Pyramid Score [39]. Some of these were later on successfully adapted to the visual domain [12,40]. These initiatives had an impact on the progress in the field of summarization. However, their almost exclusive focus on a single modality (e.g. text or visual) or language and the traditional tasks (e.g. text document and video summarization or search diversification) does not reflect the richness of social multimedia and the complex use cases it brings.

3 Proposed Framework

First, we take a step back and look at a media-agnostic birds-eye view of the problem. We therefore imagine a generic framework that follows the requirements as driven by the user, instead of the technology. Figure 1 shows an overview of the concept, which follows the standard pattern of a media pipeline along the "ingest," "extract," "reify" paradigm. The goal of the framework is to create a story for the user, who is querying for information using a set of relevance criteria. Before doing that, we assume the user has configured the framework somehow, e.g. to choose some visualization template and define basic properties of the story. We then assume a tool that would query a set of sources from the Internet or elsewhere, download ("ingest") the data, "extract" relevant information and then "reify" it in a way that it can be added into some standardized Knowledge Representation (KR). The knowledge representation would then, in connection with the initial configuration, be used to create the final story. We will next discuss technical and other challenges to be addressed by the community in order to put flesh onto our bare bones framework.

3.1 Challenges and Example Application Domains

A framework for holistic storytelling brings a new set of research challenges and also reshapes some of the more traditional challenges in UGC. We identify these as *storytelling challenges* which include handling of time/temporality/history, dynamic labeling of noise, focused story generation, tailoring to impartiality or a viewpoint, quality assessment and explainability as well as *UGC challenges* which include ethical use, multi source fusion, multilinguality and multimodality, information extraction, knowledge update and addition of new knowledge, staying agnostic to specific application, supporting various types of open data, portability and finding a balance between depth and breadth. We now describe a set of application domains that illustrate some of the aforementioned challenges.

Smart urban spaces. Increased availability of open data and social multimedia has resulted in the birth of urban computing [41] and created new possibilities for better understanding cities. Although spontaneously captured, social multimedia may provide valuable insights about geographic city regions and their inhabitants. For example, user-generated content has been used to create summaries of geographic areas and tourist routes in location recommendation systems [12,13]. Sentiment extracted from social multimedia, in combination with neighborhood statistics was also proven invaluable for a more timely estimation of city livability and its causes [42]. Similarly, when looking for signs of issues such as neighborhood decay or suboptimal infrastructure, city administrators are increasingly monitoring diverse UGC streams, ranging from social media and designated neighborhood apps to data collected by mobile towers and wearables. Efficient approaches to summarization and storytelling are needed to facilitate exploration in such large and heterogeneous collections.

Business intelligence. User generated content is a valuable source of information for companies and institutions. Business information can be obtained by analyzing what the public is saying about a company, its products, marketing campaigns and competitors. Traditionally business intelligence relied on facts and figures collected from within the organisation, or provided by third-party reports and surveys. Instead of surveys, direct feedback can be obtained by listening to what people are saying on social media, either directed at their own social circle, or directly at the company in the case of web care conversations. Content can consist of textual messages or videos, for example product reviews. Besides the volume of messages, the sentiment of messages is important to analyze into positive and negative aspects. The amount of user generated content can easily add up to thousands of messages on a single topic, so summarization techniques are needed to efficiently process the wealth of information available [43].

Health and Wellness. There is a wealth of data about our health and wellness which is generated digitally on an individual basis. This includes genomic information from companies like 23andme[1] which uses tissue samples from individuals to generate information about our ancestry as well as about our possible susceptibility to a range of inherited diseases. We also have information about our lifestyles which can be gathered from our social media profiles and information about our physical activity levels and sports participation from any fitness trackers that we might wear or use. When we have health tests or screening we can have indications of biomarkers from our clinical tests for such things as cholesterol levels, glucose levels, etc. We have occasional once-off readings of our physiological status via heart and respiration rates and increasingly we can use wearable sensors to continuously monitor glucose, heart rate etc. to see how these change over time. From all of this personal sensor data there is a need to generate the *"story of me"*, telling my health professional and me how well or healthy I am now, whether my health and wellness is improving or is on the slide, and if there's anything derived from those trends that I should know.

Lifelogs. In this use case a large amount of first-person ethnographic video or images taken from a wearable camera over an extended period of days, weeks, months or even years, has been generated. Such a collection may be augmented and aligned with sensor data such as GPS location or biometric data like heart rate, activity levels from wearable accelerometers or stress levels from galvanic skin response sensors. There is a need to summarize each day's or week's activities to allow reviewing or perhaps to support search or browsing through the lifelog. Summaries should be visual, basically selecting a subset of images of videos, and applications could be in memory support where a summary of a day can be used to trigger memory recall [44]. In this case the visual summary should incorporate events, objects or activities which are unusual or rare throughout the lifelog in preference to those which are mundane or routine like mealtimes, watching TV or reading a newspaper which might be done every day [45].

Field study/survey. The relevance of consumer-produced multimedia often transcends the reason for creating and sharing it. As a side effect this information could be used for field studies of other kinds, if it can be retrieved in a timely fashion. The framework we propose could enable empirical scientists of many disciplines to leverage this data for field studies based on extracting required information from huge datasets. This currently constitutes a gap between the elements

[1] http://www.23andme.com.

of what multimedia researchers have shown is possible to do with consumer-produced big data and the follow-through of creating a comprehensive field study framework supporting scientists across other disciplines. To bravely bridge this gap, we must meet several challenges. For example, the framework must handle unlabeled and noisily labeled data to produce an altered dataset for a scientist who naturally wants it to be both as large and as clean as possible. We must also design an interface that will be intuitive and yet enable complex search queries that rely on feature and statistics generation at a large scale.

Entertainment. Multimedia summarization and storytelling can also serve to fulfill a pure entertainment need. Respective approaches could, for instance, support event-based creation of videos from pictures and video clips recorded on smart phones. To this end, they would automatically organize and structure such user-generated multimedia content, possibly in low quality, and subsequently determine the most interesting and suited parts in order to tell the story of a particular event, e.g., a wedding. The multimedia material considered by such an event-based storytelling approach is not necessarily restricted to a single user, but could automatically determine and select the best images/scenes from the whole audience at the event, or at least those choosing to share material.

3.2 Use Cases

When rethinking the requirements, we primarily analyzed two types of use cases: summarization and storyboarding.

Summarization has traditionally involved document summarization, i.e. reducing pieces of text into a shorter version, and video summarization, where multiple or long videos are reduced to a shorter version. As data is increasingly available in many modalities and languages, it is possible to generate a multilingual and multimodal summary according to the user's information request. Large events such as elections or important sports competitions are covered by many channels, including traditional media and different social media. New directions for summarization include interactive summaries of UGC opinions or sentiment-based data visualization, and forecasting including prediction of electoral results, product sales, stock market movements and influenza incidence [46]. Getting an overview of a certain music genre or style requires algorithms capable of identifying the most representative and important music [47–49], which should ideally also take cultural aspects into account when analysing meaning of a genre [50].

A **storyboard** is a summary that conveys a change over time. This may include a recount of the given input in order to tell an unbiased story of an event, e.g. the Fall of the Berlin wall or the Kennedy murder. It may also aim to present or select facts to persuade a user to perform a particular action or change opinion, e.g. pointing out the likely murderer in the Kennedy case. If the input is open-ended, the summary may be structured by background information, e.g. a composite clip giving a visual summarization of an event (such as a concert, a sports match, etc.) where the summarized input is provided by those attending the event but the story is structured according to a timeline given by background information.

4 Prerequisites

Once user generated content has been gathered, extracted, and reified, it should be expressed in a KR. This is a step prior to the generation of stories and summaries which aims to describe the information of interest following a representation formalism. Some of the knowledge representation formalisms widely adopted in the multimedia community are Resource Description Framework (RDF) and Knowledge Graph. The selection of one approach over the others is tightly connected with the purpose of the summary/story and the technique used for its construction. This means that knowledge must be represented using a language with which the Story Generation Engine can reason in order to satisfy complex relevance criteria and visualization requirements (templates) specified by the users. These relevance criteria and visualization requirements imply a set of desired properties on the data and KR, as well as the end result presented to the user, which are fundamental for summarization and storytelling.

4.1 Data Representation Properties

Complex user information needs and the relevance criteria stemming from them require novel (multimodal and multilingual) data representations. In Table 1 we list some critical prerequisites they should fulfill.

Table 1. Properties data representation should have for facilitating effective summarization and storytelling.

Data representation properties		
Location	Time	Observed
Single ⇌ Distributed	Scheduled ⇌ Unplanned	Entity-driven ⇌ Latent
Physical ⇌ Virtual	Short ⇌ Long	
Personal ⇌ Public/Shared	Recurrent ⇌ One-off	
Independent ⇌ Cascaded		

Time: The "events" described by a story could have very different properties. For example, an event could be *scheduled* (e.g. Olympic Games) or *unplanned* (e.g. a terrorist attack). In the former case relevance criteria and the visualization templates could be easier to foresee, but an effective data representation should accommodate the latter use case as well. Similarly, the events could have a *longer* (e.g., studies abroad) or *shorter* (e.g., birthday) duration. Finally, data representation should ideally accommodate both *recurrent* and *once off* events.

Location: Although multimedia analysis has found its way in modeling different aspects of geographic locations [13,27,51], most related work addressed specific use cases and little effort has been made in identifying general "spatial" criteria underlying data representations should satisfy. In this regard, the representation should account for the events occurring at a *single* (e.g. rock concert) or *distributed* location (e.g. Olympic Games). In both cases those locations can be further *physical* or *virtual*. On the other hand, the events of interest can be *personal* or *public/shared*. While in the former case the content interpretation and relevance criteria may have a meaning for a particular individual only, the later is usually easier to analyze due to a higher "inter-user agreement". Finally, data representation should be designed with the awareness that the aforementioned types of events could additionally be *independent* or *cascaded*.

Observed: In many analytic scenarios the summaries and stories presented to the user contain well-defined named *entities*, i.e. topics, people, places and organizations. An example would be a well-structured news article covering a political event. Yet the topics of interest may be *latent*, which is particularly common in social media discussions. For example, a public servant sifting through millions of social media posts in an attempt to verify an outbreak of a new virus may be interested in various unforeseen and seemingly unrelated conversations, which together provide conclusive evidence. Therefore, a good data representation should ideally provide support for both.

4.2 Knowledge Representation Properties

Building on best practices from the semantic web community, the results of ingestion, extraction and reification (cf. Fig. 1) should be further organized in a knowledge representation. Example candidates include RDFs and knowledge graph. The KR should be flexible enough to allow for *temporal*, *spatial* and *observed* properties of the events discussed in Sect. 4.1. It should further support both *implicit* and *explicit* relations between the items, as well as their modification "on the fly" (cf. Table 2). The events and their building blocks could further be *independent* and *correlated/causal*. To facilitate a wide range of possible relevance criteria as well as their complex interplay, the KR should also include notions of importance, representativeness and frequency. Finally, the content interpretation and user information needs can be specified at different semantic levels, which in case of multimedia range from e.g. term or pixel statistics, semantic concepts, events and actions to the level of semantic theme and complex human interpretations involving aesthetic appeal and sentiment. Supporting a wide range of relevance is therefore a necessary condition for facilitating creation of effective and engaging summaries and stories.

Table 2. Properties a knowledge representation should have.

Knowledge representation properties	
Implicit ⇌ Explicit	Independent ⇌ Correlated/Causal
Uniqueness/Representativeness	Support for different semantic levels

4.3 Story Properties

Given the content, data and KRs and the user information needs, the output of the pipeline depicted in Fig. 1 is the story (or summary) presented to the user. Below we enumerate a number of criteria an ideal set of "story templates" should satisfy (see Table 3). A story should satisfy both *functional* (e.g. fulfilling a purpose) and *quality* (e.g. metrical) requirements [32]. The importance of a particular requirement should ideally be learned from user interactions. The system should further support *self-contained/interpretable* and *stepping-stone/connector* type of summaries. While the former by itself provides an insight into a larger multimedia item or a collection, the later serves a goal further on the horizon, such as faster collection exploration. Additionally, the design should accommodate both *succinct* and *narrative*, as well as *abstractive* and *generative* stories. With regard to the input and output modalities and languages, support should be provided for *modality-preserving* and *cross-modal* and/or *cross-lingual* use-cases. In many scenarios, user information needs can be satisfied with a *static* story. However, the size and heterogeneity of a UGC collection as well as the complexity of user information needs make *interactive* summarization and storytelling increasingly popular. Depending on the information needs, a *factual* or *stylistic* summary may be desirable, which is why the system should support both flavors and perhaps allow for interactive learning of their balance. Finally, while a *generic* story may be sufficient for some, *personalization* should also be supported.

Table 3. Story properties that should be facilitated by the story generator engine.

Story properties	
Functional ⇌ Quality	Modality-preserving ⇌ Cross-modal
Self-contained ⇌ Stepping-stone	Static ⇌ Dynamic/Interactive
Succinct ⇌ Narrative	Factual ⇌ Stylistic
Abstractive ⇌ Generative	Generic ⇌ Personalized

5 Conclusion

Motivated by an observation about discrepancies between state of the art research on the one hand and the increasing richness of user generated content and the accompanying complex user information needs on the other, we revisit the requirements for multimedia summarization and storytelling. We reiterate the importance of summarization and storytelling for facilitating efficient and appealing access to large collections of social multimedia and interaction with them. Our proposed framework identifies a set of challenges and prerequisites related to data and KR as well as the process of their creation, i.e. ingestion, extraction and reification. We further make an inventory of the desirable properties a story should have for addressing a wide range of user information needs. Finally, we showcase a number of application domains and use cases that could serve as the catalyst for future research on the topic.

References

1. Tian, Y., Srivastava, J., Huang, T., Contractor, N.: Social multimedia computing. Computer **43**(8), 27–36 (2010)
2. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: the good the bad and the OMG!, pp. 538–541. AAAI Press (2011)
3. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: ACM IMC 2007, pp. 1–14 (2007)
4. Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweet the debates: understanding community annotation of uncollected sources. In: ACM WSM 2009, pp. 3–10 (2009)
5. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: ACM WWW 2010, pp. 591–600 (2010)
6. Bian, J., Yang, Y., Zhang, H., Chua, T.S.: Multimedia summarization for social events in microblog stream. IEEE Trans. Multimed. **17**(2), 216–228 (2015)
7. Hong, R., Tang, J., Tan, H.K., Ngo, C.W., Yan, S., Chua, T.S.: Beyond search: event-driven summarization for web videos. ACM Trans. Multimed. Comput. Commun. Appl. **7**(4), 35:1–35:18 (2011)
8. Gornostay (Gornostaja), T., Aker, A.: Development and implementation of multilingual object type toponym-referenced text corpora for optimizing automatic image description generation. In: Dialogue 2009 (2009)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS 2013, pp. 3111–3119. CAI (2013)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS 2012, pp. 1097–1105. CAI (2012)
11. Nenkova, A., McKeown, K.: A survey of text summarization techniques. In: Aggarwal, C., Zhai, C. (eds.) Mining Text Data, pp. 43–76. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-3223-4_3
12. Rudinac, S., Larson, M., Hanjalic, A.: Learning crowdsourced user preferences for visual summarization of image collections. IEEE Trans. Multimed. **15**(6), 1231–1243 (2013)
13. Rudinac, S., Hanjalic, A., Larson, M.: Generating visual summaries of geographic areas using community-contributed images. IEEE Trans. Multimed. **15**(4), 921–932 (2013)
14. Xie, L., Sundaram, H., Campbell, M.: Event mining in multimedia streams. Proc. IEEE **96**(4), 623–647 (2008)
15. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: ACM SIGIR 1998, pp. 335–336. ACM, New York (1998)
16. Uchihashi, S., Foote, J., Girgensohn, A., Boreczky, J.: Video manga: generating semantically meaningful video summaries. In: ACM MM 1999, pp. 383–392. ACM (1999)

17. Sundaram, H., Xie, L., Chang, S.F.: A utility framework for the automatic generation of audio-visual skims. In: ACM MM 2002, pp. 189–198. ACM (2002)
18. Kennedy, L., Naaman, M.: Less talk, more rock: automated organization of community-contributed collections of concert videos. In: WWW 2009, pp. 311–320. ACM (2009)
19. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: IEEE CVPR 2013, pp. 2714–2721 (2013)
20. Yu, S.I., Yang, Y., Hauptmann, A.: Harry potter's marauder's map: localizing and tracking multiple persons-of-interest by nonnegative discretization. In: IEEE CVPR 2013, pp. 3714–3720 (2013)
21. Xie, L., Natsev, A., Kender, J.R., Hill, M., Smith, J.R.: Visual memes in social media: tracking real-world news in YouTube videos. In: ACM MM 2011, pp. 53–62. ACM (2011)
22. Wactlar, H.D., Kanade, T., Smith, M.A., Stevens, S.M.: Intelligent access to digital video: informedia project. Computer **29**(5), 46–52 (1996)
23. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. In: ACM WWW 2008, pp. 297–306 (2008)
24. van Leuken, R.H., Garcia, L., Olivares, X., van Zwol, R.: Visual diversification of image search results. In: ACM WWW 2009, pp. 341–350 (2009)
25. Lestari Paramita, M., Sanderson, M., Clough, P.: Diversity in photo retrieval: overview of the ImageCLEFPhoto task 2009. In: Peters, C., Caputo, B., Gonzalo, J., Jones, G.J.F., Kalpathy-Cramer, J., Müller, H., Tsikrika, T. (eds.) CLEF 2009. LNCS, vol. 6242, pp. 45–59. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15751-6_6
26. Sanderson, M., Tang, J., Arni, T., Clough, P.: What else is there? Search diversity examined. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 562–569. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00958-7_51
27. Ionescu, B., Popescu, A., Radu, A.L., Müller, H.: Result diversification in social image retrieval: a benchmarking framework. Multimed. Tools Appl. **75**(2), 1301–1331 (2016)
28. Ren, Z., Liang, S., Meij, E., de Rijke, M.: Personalized time-aware tweets summarization. In: ACM SIGIR 2013, pp. 513–522 (2013)
29. Hao, Q., Cai, R., Wang, X.J., Yang, J.M., Pang, Y., Zhang, L.: Generating location overviews with images and tags by mining user-generated travelogues. In: ACM MM 2009, pp. 801–804. ACM, New York (2009)
30. Gurrin, C., Smeaton, A.F., Doherty, A.R.: Lifelogging: personal big data. Found. Trends Inf. Retr. **8**(1), 1–125 (2014)
31. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: IEEE CVPR 2012, pp. 1346–1353, June 2012
32. Harman, D., Over, P.: The DUC summarization evaluations. In: HLT 2002, San Francisco, CA, USA, pp. 44–51. Morgan Kaufmann Publishers Inc. (2002)
33. Dang, H.T.: Overview of DUC 2006. In: DUC 2006 (2006)
34. Owczarzak, K., Dang, H.T.: Overview of the TAC 2011 summarization track: guided task and AESOP task. In: TAC 2011 (2011)
35. Over, P., Smeaton, A.F., Awad, G.: The TRECVid 2008 BBC rushes summarization evaluation. In: ACM TVS 2008, pp. 1–20 (2008)
36. Radev, D.R., Hovy, E., McKeown, K.: Introduction to the special issue on summarization. Comput. Linguist. **28**(4), 399–408 (2002)
37. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: ACL 2004 Workshop, pp. 74–81 (2004)

38. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL 2002, pp. 311–318 (2002)
39. Nenkova, A., Passonneau, R.J.: Evaluating content selection in summarization: the pyramid method. In: HLT-NAACL, pp. 145–152 (2004)
40. Li, Y., Merialdo, B.: VERT: automatic evaluation of video summaries. In: ACM MM 2010, pp. 851–854 (2010)
41. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. **5**(3), 38:1–38:55 (2014)
42. Boonzajer Flaes, J., Rudinac, S., Worring, M.: What multimedia sentiment analysis says about city liveability. In: Ferro, N., Crestani, F., Moens, M.-F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 824–829. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1_74
43. Dey, L., Haque, S.M., Khurdiya, A., Shroff, G.: Acquiring competitive intelligence from social media. In: MOCR AND 2011, p. 3. ACM (2011)
44. Doherty, A.R., Hodges, S.E., King, A.C., Smeaton, A.F., Berry, E., Moulin, C.J., Lindley, S., Kelly, P., Foster, C.: Wearable cameras in health. Am. J. Prev. Med. **44**, 320–323 (2013)
45. Lee, H., Smeaton, A.F., O'Connor, N.E., Jones, G., Blighe, M., Byrne, D., Doherty, A., Gurrin, C.: Constructing a SenseCam visual diary as a media process. Multimed. Syst. **14**(6), 341–349 (2008)
46. Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., Gloor, P.: The power of prediction with social media. Internet Res. **23**(5), 528–543 (2013)
47. Tian, M., Sandler, M.B.: Towards music structural segmentation across genres: features, structural hypotheses, and annotation principles. ACM Trans. Intell. Syst. Technol. **8**(2), 23:1–23:19 (2016)
48. Goto, M.: A chorus section detection method for musical audio signals and its application to a music listening station. IEEE Trans. Audio Speech Lang. Process. **14**(5), 1783–1794 (2006)
49. Chai, W.: Semantic segmentation and summarization of music. IEEE Sig. Process. Mag. **23**(2), 124–132 (2006)
50. Schedl, M., Flexer, A., Urbano, J.: The neglected user in music information retrieval research. J. Intell. Inf. Syst. **41**, 523–539 (2013)
51. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: YFCC100M: the new data in multimedia research. Commun. ACM **59**(2), 64–73 (2016)