How Far We've Come: Impact of 20 Years of Multimedia Information Retrieval

SHIH-FU CHANG, Columbia University

This article reviews the major research trends that emerged in the last two decades within the broad area of multimedia information retrieval, with a focus on the ACM Multimedia community. Trends are defined (nonscientifically) to be topics that appeared in ACM multimedia publications and have had a significant number of citations. The article also assesses the impacts of these trends on real-world applications. The views expressed are subjective and likely biased but hopefully useful for understanding the heritage of the community and stimulating new research direction.

Categories and Subject Descriptors: H.2.8 [**Database Applications**] : Image Database; H.3.3 [**Information Search and Retrieval**]; H.5.1 [**Multimedia Information Systems**]

General Terms: Algorithms

Additional Key Words and Phrases: Content-based image retrieval, video retrieval, music retrieval, multimedia information retrieval

ACM Reference Format:

Chang, S.-F. 2013. How far we've come: Impact of 20 years of multimedia information retrieval. ACM Trans. Multimedia Comput. Commun. Appl. 9, 1s, Article 42 (October 2013), 4 pages.
DOI: http://dx.doi.org/10.1145/2491844

1. INTRODUCTION

The ACM Multimedia research community has enjoyed two decades of vibrant success in tackling a large spectrum of challenges. Among them, multimedia analysis and retrieval constitutes a major area and has attracted substantial interest from researchers [Rui et al. 1999; Smeulders et al. 2000]. This article looks back and reviews trends that have occurred in the community and research outcomes that have propelled impact on real-world applications. Insight from such retrospective studies, at the risk of subjectivity and bias, may be useful for understanding the history (and heritage) of the community and may help stimulate new research direction in the future.

2. METHODOLOGY – A NONSCIENTIFIC APPROACH

Google Scholar was used to survey and find the most cited papers in the last 20 years published in ACM MM (Multimedia) and MIR (Multimedia Information Retrieval). Each of the top 600 papers from the Google Scholar search has more than 40 citations as of May 2013. By inspecting these papers and

Author's address: S.-F. Chang, 500 W. 120th St, S. W. Mudd Building Rm 1312, Columbia University, New York, NY 10027; email: shih.fu.chang@columbia.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2013 ACM 1551-6857/2013/10-ART42 $15.00
DOI: http://dx.doi.org/10.1145/2491844

Table I. Major Research Trends of Multimedia Analysis and Retrieval in the ACM MM Community

Year	Topics	Year	Topics	Year	Topics
1993	Content-based image search	2001 (cntd.)	3D model retrieval	2008 (cntd.)	Mobile user activity mining
1994	Motion-based video object search		Video browser behavior learning		Event detection in unconstrained videos
	Iconic video annotation	2002	MyLifeBits		Social signal processing
1995	Shot-based video parsing (e.g., *Virage*)		User attention model and affective interface		Mobile augmented reality (e.g., *Qualcomm Vuforia*)
	Music search by humming (e.g., *Shazam*)		Sports video event detection	2009	Image tag ranking and refinement
1997	Text OCR for video indexing	2003	Geo-location tagging of images (e.g., *Flickr geo-tagging*)		Image search via brain machine interface
	News video story segmentation		Cross modal analysis	2010	Image aesthetic
	Speech indexing of audio/video (e.g., *Blinkx*)		VideoQA		Affective gap
1998	Video summarization		Automatic consumer video editing	2011	Mobile visual search (e.g., *Google Goggles*)
	Relevance feedback in image retrieval	2004	Near-duplicate detection		Gesture recognition with depth sensor
	Hierarchical image classification		Manifold ranking		Visual memes
2000	Web image annotation		Large-scale concept detection	2012	Social event detection
	Sports video highlight (e.g., *Mitsubishi DVR*)	2008	Multimedia linking from papers (e.g., *Ricoh Hot Paper*)		Image emotion detection
2001	Active learning for image retrieval		Flickr distance for image retrieval		Multimodal beauty model
	Audio classification		Search tag suggestion		Video search intent

Note: Constructed based on inspection of 600 most cited papers (before 2009) and recent papers (starting 2009) published in ACM MM or MIR. Each topic is listed in the year when a highly cited paper of the topic was first published. Topics displayed in red indicate those with impacts on commercial products.

including only those related to multimedia analysis and retrieval, I tried to determine the earliest time when a new trend started to appear in the community. A separate query was used to include only recent papers published in these conferences starting 2009. The results are summarized in Table I. In addition, topics that are believed to have had successful impacts on real-world applications or products are highlighted in red.

3. TREND ANALYSIS

The earliest trend that attracted a massive following of research is content-based image retrieval, which aimed to measure visual content similarity based on matching low-level features. Early papers starting in 1993 utilized simple visual features, such as color, texture, shape, region, and motion. This concept of low-level features was then extended to the audio domain by using acoustic features for music retrieval (1995). Additional features for video indexing, such as shot boundaries, automatic speech recognition (ASR), text optical character recognition (OCR), news story segmentation, and

sports domain-specific event detection were introduced in the following years (1995–2002). User interaction was incorporated in the form of relevance feedback or active learning during 1998–2001. With the goal of overcoming the semantic gap, research on image classification, audio classification, and Web image annotation emerged in 1998–2001, and then started exploration in cross-media correlation in 2003. This finally culminated with efforts in large-scale concept detection in benchmarking efforts like TRECVID in 2004 and event detection in unconstrained videos in 2008. Along the way, researchers started to draw on and develop more sophisticated statistical and machine learning techniques. In addition, instead of solely relying on content features, interesting ideas combining context and geographic location information were published around 2003–2004.

The availability of Web-scale data spurred many new ideas, including using Flickr to define concept distance for image search (2008), disambiguating search tag suggestions (2008), and refining Web image tags that are rich but often noisy (2009). The pervasive adoption of mobile devices as a platform for information consumption and communication has shaped new research trends in mobile geo-tagging (2003), mobile user activity pattern mining (2008), and mobile product search (2011).

In recent years, the community started to explore research beyond matching content similarity and recognizing semantics, such as those just discussed. Along this line, two new trends are attracting broad interest: first, rich social interactions between users and content in highly connected social media platforms, and second, higher-level interactions between users and content involving new concepts related to affect, aesthetics, emotion, and intent.

Besides algorithmic advances, many interesting and potentially useful applications enabled by intelligent multimedia content analysis have also been reported, such as music search by humming (1995), sports video highlighting (2000), MyLifeBits (2002), multimedia presentation linking (2003), and recently, mobile augmented reality (2008) and mobile visual search (2011). Example commercial products are highlighted in Table I.

4. CONCLUSION AND OPEN ISSUES

We have witnessed exciting impacts in areas such as mobile audio search (e.g., Shazam), visual search (e.g., Ricoh Hot Paper and Google Goggles), and emerging products with augmented reality capabilities (e.g., Qualcomm Vuforia and Google Glass). I would accredit the success of these stories more to the availability of large-scale data storage and computing than to true advances in solving fundamental challenges of media content recognition. Several applications previously listed apply feature matching techniques without requiring complex recognition models. The feasibility of storing large datasets and matching them in real time makes these novel applications possible.

Predicting trends and impact of future research is difficult, but in drawing lessons from past trends and advances in the related fields, like information retrieval and recommendation systems, I believe near-term opportunities continue to lie in areas external to content-based analysis. Much could be gained by exploring the vast amounts of data now available about how users create, share, and interact with content, and how multimedia content is used throughout explosively growing social media. Multimedia retrieval could be made much more successful if we have more knowledge about the real-world events at which the multimedia content is captured, the intention and social context when a piece of content is shared, and how the audience responds (by analyzing the popularity, viewer comments, and response sentiments). In today's highly connected world, what is being written, favored, or shared would add a lot to determining how the content could be indexed and searched. Information like these is useful for helping researchers working on content analysis and machine learning to define and formulate the right problems that can be solved in the near term and have promise of impact on practical applications of interest to users. In some sense, these may be considered timely lower-hanging fruit, the harvesting of which still requires innovative ideas and ingenious system design.

Another opportunity comes from the emerging availability of big data for multimedia analysis, which is actually a double-edged sword. On one hand, it provides an extremely valuable resource that has been shown instrumental in driving progress in many research areas, such as automatic speech recognition; on the other hand, it poses great challenges in developing new problem formulations, efficient large-scale machine learning algorithms, and scalable system architectures that allow researchers to explore new ideas and repeat results easily. Collaborative community efforts are critically needed to address this issue.

Facing the challenges and opportunities previously mentioned, the multimedia community is well positioned to continue its very successful track record accomplished in the past two decades. So far, the community has shown excellent agility in adapting to new challenges and fusing new knowledge from other disciplines. Going forward, such abilities will be very important for solving new problems, especially those involving high-level information and human interaction, such as affect, social sentiment, and user emotion. Theories and models from other fields like psychology, media, and cognitive sciences will again prove very useful for developing computational approaches to solving these new problems.

REFERENCES

RUI, Y., HUANG, T. S., AND CHANG, S.-F. 1999. Image retrieval: Current techniques, promising directions, and open issues. *J. Visual Commun. Image Rep. 10*, 1, 39–62.

SMEULDERS, A. W. M., WORRING, M., SANTINI, S., GUPTA, A., AND JAIN, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Machine Intell. 22,* 12, 1349–1380.

Received May 2013; accepted June 2013