# Multimedia Question Answering

**Richang Hong and Meng Wang**
*Hefei University of Technology, China*

**Guangda Li, Liqiang Nie, Zheng-Jun Zha, and Tat-Seng Chua**
*National University of Singapore*

**By surveying recent research in multimedia question answering, this article explores QA's evolution from text to multimedia and identifies the challenges in the field.**

Faced with the vast quantity of information returned by Web search engines such as Google, Bing, and Yahoo, users can easily become overwhelmed. Question-answering (QA) research attempts to tackle this information-overload problem. Instead of returning a ranked list of documents, as with current search engines, QA leverages advanced media content, linguistic analysis, and domain knowledge to return precise answers to users' natural-language questions.

However, to date, QA research has largely focused on text. Given that the vast amount of information on the Web is now in multimedia form, it is natural to extend text-based QA research to multimedia QA (MMQA). (We identify all types of answers except pure text as multimedia answers, including images, video, images and text, and so forth.) Further MMQA research must bear in mind several key points.[1] First, we must manage incomplete metadata and clean up noisy annotations. Second, appropriate multimedia answers are more intuitive for some questions. Third, multimedia answers are readily available for some types of questions given the popularity of video- and image-sharing sites. Thus, MMQA can complement text QA in a complete QA paradigm in which the best answers might be a combination of text and other mediums.

Thus far, few works have addressed MMQA services. Hui Yang and his colleagues presented an early system specifically designed to address video QA for news video.[2] Their work follows an architecture similar to text-based QA with video content analysis being performed at various stages of the QA pipeline. Following this work, several video QA systems were proposed, most of which relied on the use of textual transcripts derived from video optical character recognition (OCR) and automatic speech recognition (ASR) output.

Under the multimedia QA paradigm, Tom Yeh, John J. Lee, and Trevor Darrell were the first to present image-based QA.[3] They described a photo-based QA system for finding information about physical objects. Their approach consisted of three layers. The first layer performed template matching of a query photo to online images to extract structured data from multimedia databases in order to help answer questions about the photo. It used question text to filter images based on categories and keywords. The second layer performed search on the internal repository of resolved photo-based questions to retrieve relevant answers, and the third layer (human-computation QA layer) leveraged community experts to handle the most difficult cases.

Current technology is still far from enabling us to benefit from MMQA. Furthermore, none of these works fully exploit the rich content on Web 2.0. As we know, Web 2.0 facilitates interactive information sharing, interoperability, and collaboration on the Internet. Therefore, an emerging question is how to leverage user-contributed data such as tagging, comments, and ratings for MMQA. Such information is rapidly becoming more abundant with the popularity of social media sites. For example, YouTube serves 100 million distinct videos and 65,000 uploads daily, and the traffic of this site accounts for more than 20 percent of all Web traffic and 10 percent of the whole Internet, comprising 60 percent of the video watched online. The photo-sharing site Flickr contained more than 4 billion images as of October 2009, and more than 3 billion photos are being uploaded every month on Facebook.

This article briefly surveys the progress of MMQA research and details its future directions. Although search is certainly one of the key techniques in the QA paradigm,[4] here we focus on the problems introduced by MMQA.

## From Text to Multimedia

Research on text-based QA gained popularity following its introduction in the Text Retrieval



1070-986X/12/$31.00 © 2012 IEEE Published by the IEEE Computer Society

Conference (TREC) evaluations in the late 1990s. Depending on the type of questions and the expected answers, we can categorize QA results as factoid QA, list QA, definitional QA, and more recently, how-to, why, opinion, and analysis QA. A factoid QA returns as answers factual tidbits of information such as names, dates, locations, and quantities. In factoid and list QA, such as "What is the most populous country in Africa?" and "List the rice-producing countries," the system is expected to return one or more precise country names.[2] On the other hand, for definitional QA, such as "What is *X*?" or "Who is *X*?" the system should return a set of answer sentences that best describe the question topic.[5] Definitional QA is similar to query-oriented summarization.

Figure 1 illustrates a traditional QA framework, which consists of three main components: document retrieval, question analysis, and answer extraction (candidate answer retrieval, answer selection, and composition). Research in these three types of QA has achieved some success. For example, the commercial QA service Powerset is a factoid QA that focuses on returning factual answers by mining Wikipedia information.

Recently, researchers have placed more emphasis on generating answers for questions such as "How to XX?" and "Why XX?" Multimedia answers are especially applicable to these types of questions. However, answering these questions requires analysis, synthesis, and aggregation of answer candidates from multiple sources at the semantic level. Web 2.0 has allowed social collaborative applications such as Wikipedia and YouTube to bloom on the Internet. An increasing number of Web-based community-based QA services (cQA) are bringing together a network of self-declared experts to answer posted questions. Launched on 21 2010, Quora has become an effective way for experts locate knowledge and disseminate information. Over time, a tremendous number of previous well-answered pairs have been stored in its repositories. Therefore, in most circumstances, users can find answers from searching the Quora archive rather than looking through a list of potentially relevant Web documents. Thus, instead of extracting answers from a certain document corpus, the retrieval task in cQA involves finding equivalent questions with readily available answers. The QA

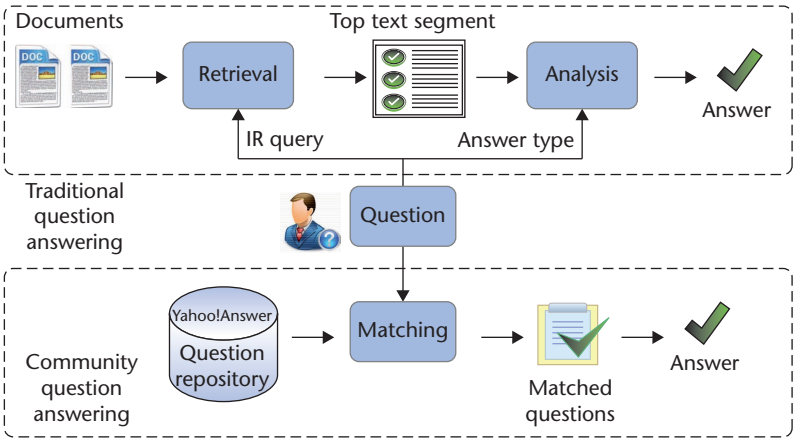


***Figure 1. A conceptual framework for traditional and community question answering (QA). The three main components are document retrieval, question analysis, and answer extraction.***

paradigm has shifted from composing answers to similar question matching.[6]

Given that the vast amount of Web content is nontextual media, it is natural to extend the text-based QA research to MMQA. MMQA is important for several reasons.

First, although most media contents are indexed with text metadata, most such metadata, such as those available on YouTube, is noisy and incomplete. As a result, much multimedia content will remain unretrievable without advanced media content analysis techniques.

Second, many questions are better explained with the help of a nontextual medium. For example, in providing textual answers to a definitional question such as "What is a thumb drive?" it would help to provide an image or video of a thumb drive with a textual description. Figure 2 illustrates how MMQA differs from other retrieval approaches.

Third, media content, especially videos, are now used to convey many types of information, as sites such as YouTube and other specialized video-/image-sharing sites and blogs have shown. Thus, many questions already have available answers in the form of video. This is especially true for the more challenging analysis and how-to questions. Answering such questions using traditional text-based framework is difficult because further analysis and composition is often necessary. It would be much clearer and more instructive to answer the question "How do I transfer photos from my camera to computer?" with a readily available how-to video than to direct people to a textual

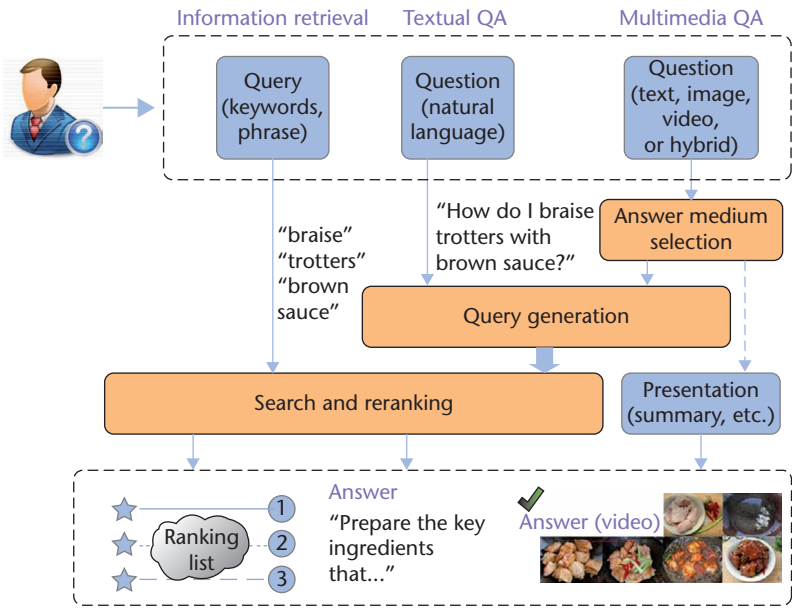


*Figure 2. Differences among information retrieval, textual QA, and MMQA. Search and reranking are common components among them, but MMQA differentiates itself with answer medium selection and answer presentation.*

instruction on how to do it in a step-by-step manner.

Thus, MMQA can complement text QA in the whole QA paradigm. Image, video, and audio QA aim to return precise images, video clips, or audio fragments as answers to users' questions. In fact, the TREC Video Retrieval Evaluation (TRECVID) has partially addressed the factoid QA problem of finding precise video content at the shot (or keyframe) level. This is done in the form of automated and interactive (shot) video retrieval, where the aim is to find a ranked list of shots that contain the desired query target, such as finding shots of George Bush. Hang Cui, Min-Yen Kan, and Tat-Seng Chua designed an early system to address the multimedia factoid QA that follows a similar architecture as text-based QA, with video content analysis being performed at various stages of the QA pipeline to obtain precise video answers.[2] Their work also includes a simple video summarization process to provide the contextual aspects of the answers. Other than factoid QA, we are unaware of other research that addresses the equivalence of multimedia definition and how-to QA.

Web 2.0 has helped QA exploit Web knowledge and human computation. However, little research has explored using information from social sharing sites for MMQA.[7–9] Given an event query, previous research presented a solution for summarizing returned YouTube videos by identifying key shots and composing them in a concise summary.[7] That approach preprocessed community contributed tags and propagated them among key shots to increase the readability of the produced storyboard. For how-to videos that provide information on completing a specific task, another work exploited the use of diverse Web 2.0 knowledge on YouTube and Yahoo!Answers to locate the most appropriate video answers.[9] This work mainly discusses how to combine textual analysis and visual content analysis to rerank potential video answers. Richang Hong and his colleagues developed a prototype system to automatically construct a multimedia encyclopedia for photo-based factoid QA. That system leverages high-resolution color photos from Flickr.

Nevertheless, MMQA is still far from industrial application and faces many obstacles. Howcast (www.howcast.com) is a new Web-based video sharing hub for how-to QA. However, it has a limited number of videos, so it can only retrieve answers for a few questions. Liqiang Nie and his colleagues proposed a method that enriches text answers with image and video information.[10] Given a question, their system searches for the best text answers in cQA and then automatically collects media information from the Web to enrich the text answer.

## Peripheral Vision in MMQA

Peripheral vision in MMQA focuses on the facets that should be emphasized when designing a MMQA system. So far, only a few systems can present a multimedia answer, mainly because algorithms can automatically extract low-level features but users need high-level concepts, creating a semantic gap. Although researchers have attempted surmount this gap, it remains a challenge.

IBM's DeepQA project has built a computer system that can perform open-domain QA using a range of knowledge. The ultimate goal is for computers to understand complex information requirements and deliver precise, meaningful responses, even synthesizing, integrating, and rapidly reasoning over the breadth of human knowledge. As a complement to text QA, MMQA should be included in the DeepQA project.

Figure 3 illustrates the facets of the MMQA research field. In this study, we identify the challenges in achieving MMQA from the aspects of user intent, data scope, question processing, and answer presentation.

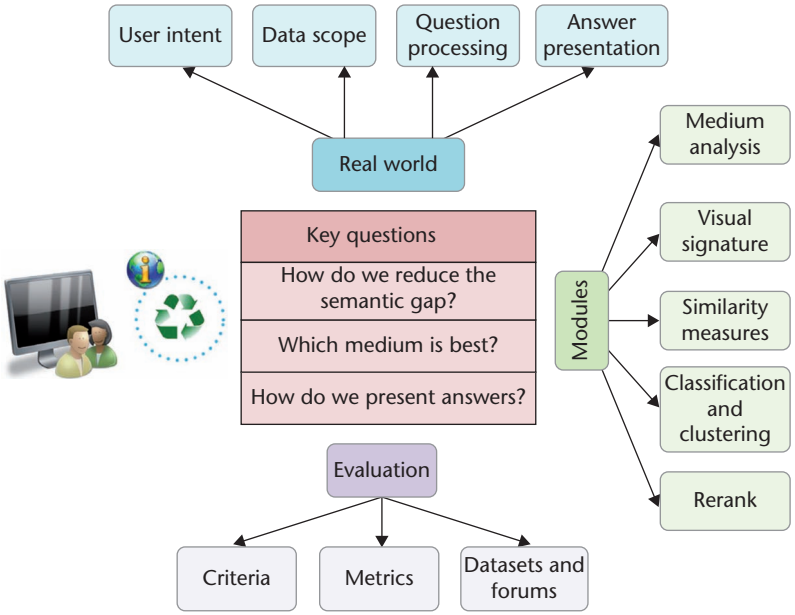


*Figure 3. Aspects of the multimedia question answering (MMQA) research field. This article focuses on user intent, data scope, question processing, and answer presentation.*

### Determining User intent

When users search for pictures, they might not have a clear idea about what they are looking for. For example, Ritendra Datta and his colleagues broadly characterized users as browsers, surfers, and searchers based on the clarity of their intent.[4] However, with QA, users can more specifically express what they want. Still, this field requires meaningful work toward more precisely capturing user intent and improving answer quality by better understanding user's intent, for example, using question suggestion.

Alexander Kotov and ChengXiang Zhai proposed a framework for question-guided search.[11] Their system automatically generates potentially interesting questions to users. In case of imprecise or ambiguous queries, it naturally engages users with feedback circles to refine their information need and guide them toward their search goal. In this way, the system can also guide users toward useful answers because the answers to the suggested questions are already known to exist.

### Choosing the Proper Medium

A QA system's design should be influenced by the nature and scope of the data. In QA, *data scope* refers to data sources and mediums—that is, determining the best sources and mediums (image, audio, video, or a hybrid) to answer a query. The scheme from Nie and his colleagues supplies the best answers with multimedia content using cQA.[10] Their process covers three components. First, given a question and its community-contributed answer in the cQA corpus, it determines which type of medium we should add to enrich the textual answers. By deeply exploring the linguistic cues from QA contexts and potential media answers, each QA pair is categorized into one of the four predefined classes: text only; text and image; text and video, or text, image, and video. The proposed scheme then automatically extracts and selects the more informative query for multimedia search. Following that, with the generated query, the system vertically collects and selects relevant image and video data with visual content analysis.

### Extracting Queries from Questions

The querying modalities supported by search systems include keywords, free text, image, graphics, and composites. Their processing methods have been broadly characterized from a system perspective as text-based, content-based, composite, interactive-simple, and interactive-composite.[4] However, the QA paradigm mainly utilizes the free-text modality because users enter natural-language questions. Samuel Huston and W. Bruce Croft examined query processing techniques that can be applied to verbose queries prior to submission to a search engine to improve the search engine's results.[12] Because search is key to QA, their proposed verbose query processing would be meaningful in a QA strategy design using Web knowledge.

Image modality has solicited much research interest, especially with the widespread use of mobile devices.[3] This modality can benefit from content-based image retrieval (CBIR) techniques. Giridhar Kumaran and Vitor R. Carvalho presented techniques to shorten long queries by removing extraneous terms.[13] They transformed the problem into learning to rank all subsets of the original subqueries

based on their predicted quality and selecting the top subquery.

### Presenting Answers

Traditional search presents results using a sorted list of descending relevancy. QA attempts to return a precise answer. Traditional QA employs a search technique for retrieving potential documents, further locates the paragraph that likely contains the answer, and finally analyzes the text segment to compose an answer, in essence providing a summarization. In contrast, MMQA can use multimodality summarization or semantic summarization to present an answer by summarizing the retrieved potential answers from various sources (text, image, audio, video, or a hybrid) at the semantic level.

Richang Hong and his colleagues attempted to summarize videos from social sharing sites for event-based questions, such as ''What was the Hong Kong handover ceremony?''[7] In these cases, two types of summaries (dynamic skimming and storyboards) are produced for the answer presentation.

## Evaluation Schemes

For any QA system, evaluation schemes must involve determining the three aspects:

- *Dataset evaluation*. The dataset should be large enough for the evaluation to be statistically significant.
- *Ground truth for relevance to the problem at hand*. Although ground truth is subjective, it is indispensable in evaluation.
- *Metrics for evaluating competing approaches*. The evaluation criteria in metrics should try to model human requirements from a population perspective.

Although some evaluation datasets and benchmarks exist for evaluating the QA performance, only a few datasets have been designed specifically for MMQA, including some subcollections of TRECVID and the NUS-WEBV, which was released by a team from National University of Singapore.

The most popular search evaluation metrics can be used in QA evaluation: precision $p$ (referring to the percentage of the retrieved entries that are relevant to the query) and recall $r$ (pertaining to the percentage of all the relevant items in the search database that are retrieved).[14]

Two other popular metrics are the mean reciprocal rank and $F$-measure. The traditional $F$-measure ($F_1$ score) is the harmonic mean of precision and recall: $F_1 = 2 \times p \times r/(p + r)$. The general formula for a nonnegative real $\beta$ is

$$F_\beta = \frac{(1+\beta^2)(p+r)}{\beta^2 \times p} + r$$

Therefore, $F_\beta$ measures the effectiveness of the retrieval with respect to a user who attaches $\beta$ times as much importance to recall as to precision.

Mean reciprocal rank helps us evaluate any process that produces a list of possible responses to a query ordered by probability of correctness. A query response's reciprocal rank is the multiplicative inverse of the rank of the first correct answer. The mean reciprocal rank is the average of the reciprocal ranks of the results for a sample of queries.

An ad hoc MMQA system might also employ the adapted metrics from search for evaluation, and a user-based evaluation (a subjective test) is a feasible way to determine QA system performance in most cases.

Evaluation dataset design is always a delicate issue. TRECVID is one of the most significant datasets, and it consists of highly edited videos as well as social media Web videos.[15] Other Web video datasets are being constructed to evaluate the performance of various systems designed to exploit user-contributed content from social media. We categorize these as either data- or application-driven datasets, where the former aims to analyze the attributes of Web videos and characterize large volume of videos for modeling data distribution and the latter targets some specific application (such as Web video categorization or MMQA).

Here, we briefly introduce NUS-WEBV, the first dataset specifically designed for MMQA.[16] We believe that a larger-scale dataset construction for MMQA will benefit from the design strategy used to develop NUS-WEBV.

When developing NUS-WEBV, we limited it to news- and event-related Web videos—those retrieved by news event queries—utilizing Wikipedia's hierarchical topic relationships. We initially selected 98 raw event queries from Wikipedia and remove some of them are because their hit numbers on YouTube were

lower than eight. These queries covered the natural, air show, political, entertainment, social, and sports categories, and the number of queries for each category varied from four to 13. For each query, we downloaded either all videos or the top-$N$, where $N$ was decided based on query hits. We then crawled the associated contextual information such as tags, titles, and descriptions as well as links to related videos. The number of videos downloaded per query ranged from eight to 300, with an average of 168.85 videos per query.

Using this method, we eventually formed a data collection of 10,130 videos with a total duration of 751 hours, for an average of 12.52 hours per query. About 10 volunteers were involved in the creation of the ground truth by annotating each video with respect to relevance, category, and quality. Relevance was determined at the shot level, while the other two tasks were at video level.

### Conclusion

Multimedia QA is an emerging topic, and so far, the research and the achievements in this area are preliminary. Several follow-on research directions can be identified. First, an urgent need exists to set up large corpora to promote multimedia QA research, especially for definition and how-to QA. Second, the field must develop better techniques for concept detection and multimedia event detection. Concept annotation is important to uncovering additional visual contents such as images and video clips. Google identified (by posting on a Google Research Blog) that Web video shot annotations as one area of meaningful research in the near future. To ensure scalability of such techniques to Web-scale problems, we need to exploit the various online visual databases with comprehensive visual concept coverage and visual examples. Finally, we need to extend the existing approaches to general domains. MMQA can certainly benefit from achievements in this direction. **MM**

## Acknowledgment

This article was supported by NUS-Tsinghua Extreme Search project under the grant number R-252-300-001-490, in part by the Natural Science Foundation of China (NSFC) under grant 61172164, in part by the Open Project Program of National Laboratory of Pattern Recognition, and in part by the NSFC under grant 61272214.

## References

1. T.–S. Chua et al., ''From Text Question-Answering to Multimedia QA on Web-Scale Media Resources,'' *Proc. ACM Multimedia Workshop Large-Scale Multimedia Retrieval and Mining* (LS-MMRM), ACM Press, 2009, pp. 51–58.
2. H. Yang et al., ''Structured Use of External Knowledge for Event-based Open-Domain Question-Answering,'' *Proc. 26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* ACM Press, 2003, pp. 33–40.
3. T. Yeh, J.J. Lee, and T. Darrell, ''Photo-Based Question Answering,'' *Proc. 16th ACM Int'l Conf. Multimedia,* ACM Press, 2008, pp. 389–398.
4. R. Datta et al., ''Image Retrieval: Ideas, Influence, and Trends of the New Age,'' *ACM Computing Surveys,* vol. 40, no. 2, 2008, article no. 5.
5. H. Cui, M.–Y. Kan, and T.–S. Chua, ''Soft Pattern Matching Models for Definitional Question Answering,'' *ACM Trans. Information Systems,* vol. 25, no. 2, 2007, article no. 8.
6. K. Wang et al., ''Segmentation of Multi-Sentence Questions: Towards Effective Question Retrieval in cQA Services,'' *Proc. 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* AMC Press, 2010, pp. 387–394.
7. R. Hong et al., ''Beyond Search: Event Driven Summarization for Web Videos,'' *ACM Trans. Multimedia Computing, Comm., and Applications,* vol. 7, no. 4, 2011, article no. 35.
8. R. Hong et al., ''Mediapedia: Mining Web Knowledge to Construct Multimedia Encyclopedia,'' *Advances in Multimedia Modeling,* LNCS 5916, Springer, 2010, pp. 556–566.
9. G. Li et al., ''Video Reference: Question Answering on YouTube,'' *Proc. 17th Int'l ACM Conf. Multimedia,* ACM Press, 2009, pp. 773–776.
10. L. Nie et al., ''Multimedia Answering: Enriching Text QA with Media Information,'' *Proc. 34th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* ACM Press, 2011, pp. 695–704.
11. A. Kotov and C. Zhai, ''Towards Natural Question Guided Search,'' *Proc. 19th Int'l Conf. World Wide Web* (WWW), ACM Press, 2010, pp. 541–550.
12. S. Huston and W.B. Croft, ''Evaluating Verbose Query Processing Techniques,'' *Proc. 33rd Int'l ACM SIGIR Conf. Research and Development*

*in Information Retrieval,* ACM Press. 2010, pp. 291–298.
13. G. Kumaran and V.R. Carvalho, ''Reducing Long Queries Using Query Quality Predictors,'' *Proc. 32nd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* ACM Press, 2009, pp. 564–571.
14. M. Wang et al., ''Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation,'' *IEEE Trans. Multimedia,* vol. 11, no. 3, 2009, pp. 465–476.
15. M. Wang et al., ''Unified Video Annotation via Multigraph Learning,'' *IEEE Trans. Circuits and Systems for Video Technology,* vol. 19, no. 5, 2009, pp. 733–746.
16. R. Hong et al., ''Exploring Large Scale Data for Multimedia QA: An Initial Study,'' *Proc. ACM Int'l Conf. Image and Video Retrieval* (CIVR), ACM Press, 2010, pp. 74–81.

**Richang Hong** is a professor at the Hefei University of Technology (HFUT). His research interests include multimedia question answering (MMQA), content-based image retrieval, and video content analysis. Hong has a PhD from the University of Science and Technology of China (USTC), Hefei, China. Contact him at hongrc.hfut@gmail.com.

**Meng Wang** is a professor at the Hefei University of Technology (HFUT). His current research interests include multimedia content analysis, computer vision, and pattern recognition. Wang has a PhD from the Department of Electronics Engineering and Information Science at the University of Science and Technology of China (USTC), China. Contact him at eric.mengwang@gmail.com.

**Guangda Li** is a postdoctoral research fellow in the School of Computing at the National University of Singapore. His research interests include object retrieval and video content analysis. Li has a PhD from the National University of Singapore (NUS). Contact him at Guangda10@gmail.com.

**Liqiang Nie** is a doctoral student in the School of Computing at the National University of Singapore. His research interests include multimedia content analysis, search, large-scale computing, and multimedia applications such as MMQA, image reranking, and recommendation. Contact him at nieliqiang@gmail.com.

**Zheng-Jun Zha** is a postdoctoral research fellow in School of Computing at the National University of Singapore. His research interests include content-based image retrieval and pattern recognition. Zha has a PhD from the Department of Automation at the University of Science and Technology of China (USTC), China. Contact him at zhazj@comp.nus.edu.sg.

**Tat-Seng Chua** is a professor in the School of Computing at the National University of Singapore. His research interests include multimedia information processing, especially text and video information extraction, retrieval, and question answering (QA). Chua has a PhD from the University of Leeds, UK. Contact him at chuats@comp.nus.edu.sg.



*Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.*