

MOBILE MEDIA SEARCH

Berna Erol¹, Jordan Cohen², Minoru Etoh³, Hsiao-Wuen Hon⁴, Jiebo Luo⁵, Johan Schalkwyk⁶

¹Ricoh Innovations, USA, ² SRI, USA, ³ NTT DoCoMo Research Laboratories, Japan,
⁴Microsoft Research Asia, China, ⁵Kodak, USA, ⁶Google, USA

ABSTRACT

This panel paper presents motivations for discussing mobile media search and contains statements from the panelists who are industry research leaders in this field.

Index Terms— mobile applications, media search, visual search, audio search, future of mobile

1. PANEL INTRODUCTION

Berna Erol, Ricoh Innovations, California Research Center, USA
berna_erol@rii.ricoh.com

Recently, many exciting media search applications have been introduced that take advantage of smart phones' audiovisual capture capabilities and their being always on and connected [1]-[13]. These applications address a real pain point for most mobile users and allow them to search with minimal text entry, if any. Is the mobile platform an ideal fit for media search? Are audio and visual signal processing technologies sufficiently accurate to support most mobile search applications? What are the killer applications of mobile media search? In this panel, we discuss the challenges and potentials of mobile media search, covering the following topics:

- Killer apps of mobile media search
- How media search on mobile devices compares to search on PCs
- Current and future technical challenges of mobile audio and visual search
- Using context information such as GPS for retrieval and presentation of results
- Future hardware and software capabilities of mobile devices for assisting search
- New interfaces for mobile search
- Network capabilities and speeds, and their role in media search
- Server based vs. distributed processing and retrieval
- Social aspects of mobile media search
- Economics of and business cases for mobile search

In the following sections we present introductory statements from each panelist.

2. PANEL MEMBERS AND STATEMENTS

Jordan Cohen, SRI, USA
jordan.cohen@sri.com

Availability of media search technologies and high speed networks enabled rapid development and deployment of various mobile media search applications for smart phones. One of the most interesting questions in this space is how the economics of mobile media search will evolve, and how this economic evolution will drive the technology.

Recently voice based directory search and other voice-enabled services of many types have been launched in the United States. Prominent among these is the Google voice search [2], in which voice is transcribed at a server and a Google search result is delivered to the phone, and Yahoo OneSearch [5] that allows web searching via submitting voice queries. Besides these, there are many other companies who support not only directory searching in the traditional sense, but retrieval of video and audio and other multimedia information such as maps and directions.

Emergence of killer applications in mobile media search area will depend -in a large part- on the economics and business models behind them. Currently the voice search applications are free. Google and Yahoo searches are supported by advertising and each API has provisions for delivering advertising to mobile phones. These services and their non-voice-driven brethren have already halved the revenue that carriers get for directory services over the past 5 years, and there is every indication that in the future the carrier revenues for these services will become negligible. This squeeze will ultimately require the carriers to find alternate revenue sources for delivery of this mobile information other than flat-rate digital services.

It is not obvious what the future holds, but it is interesting to contemplate. At the moment, web search companies have a clear advantage in launching and subsidizing voice search services. Alternatively, voice search applications can be

made available through the various smart phone application stores. For example, some directory services have launched with voice advertisements but obtained limited success. I encourage readers to think about not only the technology but the business models behind the mobile media search applications. The future killer applications should not only address the apparent user needs, but should be able to generate revenue for sustainability and further development.

Minoru Etoh, Research Laboratories, NTT DoCoMo, Japan
etoh@ieee.org

Many mobile phones are now being equipped with various I/O devices and sensors, such as accelerometers, GPS, microphones, and cameras. These allow applications to be context-aware and enable easy integration of user's real environment with applications.

Mobile device's peripheral and sensor enhancement will be an extremely fertile incubation environment for new and innovative killer applications in our daily life. That definitely differentiates mobile Internet from fixed-line Internet of PCs. That is already happening in far-east countries, where mobile devices are evolving not as miniaturized mobile PCs but as devices with unprecedented capabilities. Key enablers are the new-generation of sensors, and a 3G-and-beyond broadband cellular connection. The low-latency of networks and high network speed enables mobile applications perform many compute-intensive tasks, such as speech recognition, on the servers, helping to ease the computational burden as well as the battery power consumption. Nevertheless, I believe that as for mobile media search through the lens, we need more time for its generic use.

So far most commonly used camera-based mobile applications remain at the level of 2D barcode and color code recognition. Media search technologies beyond barcodes, however, are now emerging in the market, as distributed speech recognition and image fingerprinting recognition are being commercialized [6]. I anticipate future killer applications exist in tagging real visual objects such as magazine covers or signboards with various Internet contents. It's a market timing issue; technologies have to meet the needs of general population in the future. Moreover, several technology challenges exist, such as feature extraction in noisy environments and compact but rich feature description for communication.

Hsiao-Wuen Hon, Microsoft Research Asia, China
hon@microsoft.com

Combination of location-aware and always-connected mobile devices and servers with abundant computing power for handling rich media enables many potential killer apps, including the ones in the area of media search. Finding and

accessing information is one of the basic user needs. In addition to traditional text-based keyword queries, powerful mobile devices can support richer and hybrid queries, such as images, audio, video, and their combinations.

In Microsoft Research Asia, we are developing mobile search systems that support image queries and audio queries [7] and mobile visual-search system for product image categorization [8]. Figure 1 shows a typical design for mobile media search systems that is designed in a client-server scheme [7]. The architecture is mainly composed of four parts. The **Mobile Client** with embedded microphone, camera and other sensors, can generate multimodal queries through audio recording and photo taking, and to receive search results as well. The **Carrier Server Interface** for the mobile client to interact with multimodal query and obtain search results, may include IVR (interactive voice response) server, WAP (wireless application protocol) server or MMS (multimedia message service) server. The **Search Server** is the core logic of matching algorithm runs here for matching multimodal queries against the indexed database in a distributed computing framework. The Carrier Server Interface can then access the **Content Database** for the desired media content based on the matched results sent back by the Search Server.

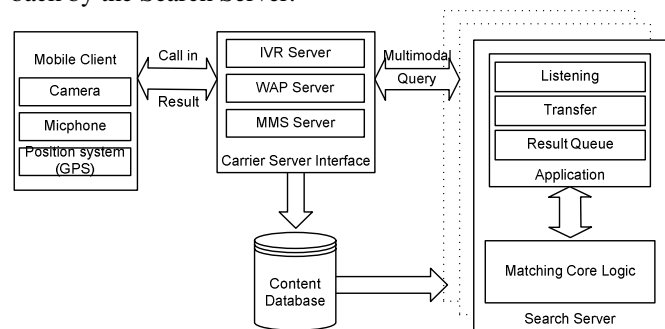


Figure 1. Mobile media search system design

Building robust and efficient mobile search systems is still a challenging task, with regards to relevance, speed, integration of other sensors (e.g. GPS), context, and user experience. Substantial research in these areas is essential to make mobile media search a powerful way to access information in everyday life.

Jiebo Luo, Kodak
jiebo.luo@kodak.com

I think that mobile platform is the best platform for media search today and in the near future. And the key reason is the availability of rich and valuable context information – right here, right now, and sharing between the right people. Indeed, time, location, and people are the three most important aspects of context for media. The goals of media search given the Internet connectivity are to harness the knowledge of numerous connected users, and to socialize on large scales and with enormous efficiency.

Our research group at Kodak has been focused on one particular area of research in recent years — integration of content and context for media management. Content refers to the information contained in the pixels and frames, while context refers to the information surrounding the pixels and frames. Integration of content and context is crucial to human-human communication and human understanding of multimedia. Likewise, such integration is likely to make media search more effective, especially given the well known semantic gaps. A series of our papers show that camera metadata and GPS can be used in conjunction with visual content analysis to boost media understanding and media search [9][10].

On a mobile device such as the iPhone and Nokia N95, not only time and location is known, we also know who is doing the search and whom the information is shared with. This is powerful information to have. More importantly, we have access to the wealth of information on the web and in the other devices and people connected by the communication network and social network. The Zonetag system by Yahoo [11] is a good example of successful applications developed for mobile devices. It suggests tags for your photos based on geo tagging and existing community tags for the same location, making it easy to tag from your phone. Snaptell [12] is another useful app for iPhone where a user can take a picture of a product and use it to find useful information such as price, seller rating, and product review. One can also ask a friend by sharing the picture and tapping into the social networks. Another context can be provided by a mobile device user to help media search is audio input (description, sound). It is also possible to search for people related information in the same fashion with a face image. Clearly, this type of real-time, on-the spot, context-enabled media search would not be possible with stationary PCs. Digital cameras are also catching up with mobile phones by having built-in GPS and Wi-Fi capabilities.

One of the biggest technical challenges of mobile platform is the scarce computational resources. Therefore, many computationally intensive visual computing algorithms cannot be deployed on the mobile platform yet. Nevertheless computation power of the mobile devices is increasing and some algorithms, such as face detection, are becoming available in many mobile devices such as digital cameras. More importantly, with advances in high speed networks and cloud computing, it is possible to perform heavy computing on servers. This enables many new media-rich applications on mobile devices.

Johan Schalkwyk, Google, USA
johans@google.com

Voice is an important input modality for mobile devices. Typing on small devices can be slow and difficult. Moreover, many typical mobile usage scenarios (e.g., driving, walking) are best accommodated by voice input.

Speech recognition integrated with other input modes (typing, pointing, etc.) can facilitate the use of search and other mobile services in many instances that would otherwise be impractical. This is one of the motivations for our Google Search by Voice project [2].

One of the challenges of mobile media search is to decide on what media to return or what actions to take after processing a query. For example an audio search can result in one of placing a phone call, playing a music clip, or performing an Internet search. The type of returned results can also depend on the query and the context of the query. When the user asks for "pictures of the Golden Gate Bridge", images can be returned. If the user does a location search, a map may be the most appropriate to return. But if the user is driving at the time, returning the spoken directions may be the best.

Another challenge of mobile media search is about having efficient interfaces to consume the found media. Audio and video clips on the Internet are mostly described by the meta-data only (for example the title and keywords). Automatic transcription of these clip potentially provide a much richer search experience. Not only does it allow discovery of media which would not be found by meta data alone, but it also facilitates the ability to browse the contents more effectively. The Google Elections Gadget is a prime example of this [13].

Combination of an efficient multi modal input interface and appropriate rendering of search results may help reach a tipping point where mobile devices are an indispensable tool for accessing media on the Internet.

BIOGRAPHY

Berna Erol received her PhD. Degree in Electrical and Computer Engineering at the University of British Columbia. Since 2001 she has been a senior research scientist at Ricoh California Research Center, USA. Berna Erol has authored or co-authored more than 40 journal and conference papers, three book chapters, and more than 50 patent applications in the area of multimedia signal processing. Her main contributions to research and development consist of content-based video and image retrieval, image and video coding and representation, E-meeting systems, text retrieval, mobile applications, and multimedia processing for mobile devices. She had been an active participant in the video coding standardization activities such as ITU-T H.263 and MPEG-7. She has served in the program and organizing committees of ACM, SPIE and IEEE conferences and she is an associated editor of the IEEE Signal Processing Magazine.

Web page: <http://www.bernaerol.com>

Jordan Cohen is a Senior Scientist at SRI International, specializing in language based applications. He is the

Principal Investigator for the DARPA GALE program for SRI. He has worked in the Department of Defense, IBM, and at the Institute for Defense Analyses in Princeton, NJ. Jordan was previously the CTO of Voice Signal Technologies, a company which produces multimodal speech-centric interfaces for mobile devices.

Web page: www.sri.com

Minoru Etoh received M.S.E.E. from Hiroshima University and Ph.D. degree from Osaka University. Dr. Etoh is Deputy Managing Director of Research Laboratories at NTT Docomo, Japan. Before joining to Docomo he lead an image communication research team in Panasonic and participated in the MPEG-4 standardization activities. He joined Multimedia Laboratories of NTT Docomo in 2000, where he contributed to launching Docomo's 3G mobile multimedia services including video phones and audio-visual content download applications. He became Managing Director of Docomo USA Labs in 2002 and Multimedia Labs in 2005 respectively. Recently he drafted Docomo corporate R&D strategic vision, and developed distributed speech recognition (DSR) applications in 2006, and recently has engaged in data mining and mobile search engine personalization research so as to foster new killer application environments.

Web page: <http://micketoh.web.fc2.com>

Hsiao-Wuen Hon is the Managing Director of Microsoft Research Asia (MSRA). An IEEE fellow, Dr. Hon is an internationally recognized expert in speech technology. He serves on the editorial board of the international journal of the *Communication of the ACM*. Dr. Hon has published more than 100 technical papers in international journals and at conferences. He co-authored a book, *Spoken Language Processing*, which is a graduate-level textbook in speech technology. Dr. Hon holds more than three dozens of patents in several technical areas. Prior to MSRA, Dr. Hon was architect with the Natural Interaction Service Division at Microsoft Corporation. He was responsible for architectural and other technical aspects of the award-winning Microsoft Speech Server product. Dr. Hon joined Microsoft Research as a senior researcher in 1995. He previously worked at Apple Computer, where he led research and development for Apple's Chinese Dictation Kit. Dr. Hon received a B.S. in EE from National Taiwan University and Ph.D in Computer Science from Carnegie Mellon University.

Web page: <http://research.microsoft.com/en-us/people/hon/>

Jiebo Luo (S'93–M'96–SM'99–F'09) received his B.S. and M.S. degrees from the University of Science and Technology of China (USTC) in 1989 and 1992, respectively, and a Ph.D. degree from the University of Rochester in 1995, all in Electrical Engineering. He is currently a Senior Principal Scientist with Kodak Research Laboratories, Rochester, NY. His research interests include image processing, computer vision, multimedia data mining,

computational photography, and biomedical informatics. He is the author of over 130 technical papers and 50 granted U.S. patents. He has served on the editorial boards of the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Multimedia, Pattern Recognition, and Journal of Electronic Imaging. He is also a Fellow of SPIE.

Web page: <http://www.linkedin.com/in/jieboluo>

Johan Schalkwyk is Senior Staff Engineer at Google. Johan has been working in the Speech industry for over 15 years primarily working on technology to bring Speech to the masses. While at SpeechWorks Johan worked on the Open Speech Recognizer which is deployed in thousands of applications around the world, including many Fortune 500 companies. Johan joined Google in 2005 where among other projects he worked on the OpenFst as a joint collaboration between NYU and Google. In 2008 Johan lead the Google Search by voice project which is currently available on iPhone and Android G1 phones.

11. REFERENCES

- [1] Berna Erol, Emilio Antúnez, Jonathan J. Hull, "HOTPAPER: Multimedia Interaction with Paper using Mobile Phones", ACM Multimedia Conference, pp. 399-408, 2008.
- [2] Google Mobile App with Voice Search <http://www.google.com/mobile/apple/app.html>
- [3] Shazam Audio Search, <http://www.shazam.com/>
- [4] Kooaba Image Recognizer, <http://www.kooaba.com/>
- [5] Yahoo OneSearch with Voice, <http://mobile.yahoo.com/onesearch/voice>
- [6] Minoru Etoh, "Cellular Phones as Information Hubs," Proc ACM SIGIR Workshop on Mobile Information Retrieval, Singapore, July, 2008.
- [7] Xing Xie, Lie Lu, Menglei Jia, Hua Li, Frank Seide, Wei-Ying Ma, Mobile Search with Multimodal Queries, Proceedings of the IEEE, Vol. 96, No. 4, Apr. 2008.
- [8] Lie Lu, Frank Seide. "Mobile Ringtone Search through query by Humming," Proceedings of ICASSP, 2008.
- [9] J. Luo, M. Boutell, C. Brown, "Exploiting context for semantic scene content understanding," IEEE Signal Processing Magazine, pp. 101-114, 2006.
- [10] J. Luo, J. Yu, D. Joshi, W. Hao, "Event Recognition with a Third Eye," ACM Multimedia Conference, 2008.
- [11] S. Ahern, M. Davis, et al., "Zonetag: Designing context-aware mobile media capture to increase participation", Workshop on Pervasive Image Capture and Sharing, 2006.
- [12] SnapTell, <http://www.Snaptell.com>
- [13] C. Alberti, M. Bacchiani, A. Bezman et al., "An Audio Indexing System for Election Video Material", Proceedings of ICASSP, 2009.