Zhu Liu and Michiel Bacchiani

TechWare: Mobile Media Search Resources

Please send suggestions for Web resources of interest to our readers, proposals for columns, as well as general feedback, by e-mail to Dong Yu ("Best of the Web" associate editor) at dongyu@microsoft.com.

In this issue, "Best of the Web" focuses on mobile media search, which has enjoyed rapid growth in consumer demand and hence receives a lot of attention from content providers and device manufacturers. The combination of a large increase in computational power and an always-on broadband connection available everywhere at all times makes for an ideal device for all our social, business, and entertainment needs. In addition, such devices provide a wide array of multimedia sensors such as a high-quality camera, a touch screen, audio-capturing abilities, and a global positioning system (GPS). This makes the devices not only suited for multimedia consumption but also a powerful platform for multimedia content generation. The relatively small form factor of the mobile devices in comparison to desktop machines also brings challenges as it is much more difficult to provide textual input to the device. On the other hand, the wide availability of the multimedia sensors makes additional input modalities more natural and many applications make good use of the added opportunities this provides. As a result, the importance of multimedia search has risen considerably given the explosion of devices attempting to locate content and the explosion of content generation itself. In addition, the spectrum of the problem of multimedia search has widened given the additional input modalities now commonly available through the sensor rich clients.

This column first reviews the recent progress in the field of mobile media search and then discusses the enabling media indexing and query technologies. Some useful online links and resources will be listed for the interested readers for further research. Unless otherwise noted, the resources are free.

MOBILE MEDIA SEARCH

Not being mined, a diamond is simply an ordinary stone buried in the dirt. That is also true for multimedia content. With the growing availability of low-cost multimedia capture devices (video, image, and sound) and the availability of easy-to-use video editing software, shooting a video clip and publishing it online is no longer a complicated task. For example, in the last year, video material uploaded to the YouTube video sharing service has increased from about 20 h/min to about 35 h/min. This explosion of content makes the problem of multimedia search both more pressing as well as challenging. As a result, the area of multimedia search has received a lot of attention as it moved from a text-only retrieval problem to a much richer and multifaceted problem. It brought in several other research fields like speech recognition, image/video processing, natural language understanding, computer vision, and music identification to better understand the content being indexed. In addition, the group-based content generation and consumption has added user signals that can be used in retrieval (e.g., a video might "go viral"). And the many sensor-rich clients make the use of multimodal input a more realistic usage setting. Users might use speech to input a text query, or they might use their cameras to input an image-based query. Touch screens make multimodal consumption of complex results an obvious presentation method. The swift change in the problem setting, the size and nature of the content indexed, and the proliferation of devices used to engage in the content have created a new and exciting field of multimedia search.

The momentum of this technology is not only evident in the many products produced by industry in this space, it is also evident in the academic community. Recently, the "Mobile Media Search: Has Media Search Finally Found Its Perfect Platform?" panel series at the ACM International Multimedia 2009 Conference and ICASSP 2009 addressed this field. In these two panels, leading experts in the field from both industry and academia shared their opinions on mobile media search.

In the following two sections, we introduce two dimensions of mobile media search applications and systems: multimodal query interfaces and multimedia retrieved results.

MULTIMODAL QUERY IN MOBILE MEDIA SEARCH

The dominant search interface both on desktop as well as mobile devices is based on text-query input. Such queries might be keywords, key phrases, or natural language questions. Alternatively, some search engines provide a hierarchical categorization of the available content and provide the users an interface to navigate that hierarchy. Contrary to desktops, input of a text query can be quite challenging on a mobile device due to its small form factor. On the other hand, unlike desktops, mobile devices

Digital Object Identifier 10.1109/MSP.2011.941095
Date of publication: 15 June 2011

1053-5888/11/$26.00©2011IEEE

are equipped with a large array of sensors making alternate input modalities a widely available option. This section focuses on the alternate interfaces mobile devices provide. Specifically, we highlight voice-based, music-based, image-based, and multimodal-based query input methods.

VOICE-BASED SEARCH

A major shortcoming of mobile devices is the effort needed to input a textual query to the device. Several companies are addressing this issue by providing a speech recognition-based solution. This in itself is a hard speech recognition problem. Queries have a very large vocabulary. They tend to be short, meaning that a language model provides a flat prior to the search problem. The user population of this technology is very large. And the most common use of the technology is on-the-go, likely in a noisy environment like on a city street or in an airport. The application also provides modeling opportunities as the query context might be available from the text input location (the text input field and surrounding Web page) and the profile of the user is likely consistent as most mobile devices are owned and operated by a single user.

- ***Vlingo*** [www.vlingo.com (mobile app)]: As a voice recognition app, Vlingo takes voice commands to send e-mail and text messages, dial the phone, update Twitter and Facebook statuses, and search for information on the Web.
- ***Google*** [www.google.com/mobile/voice-search (integral to Android and as an app)]: Voice search allows users to input search queries by voice. In addition, the app allows dictation of a text message and voice-controlled navigation and calling.
- ***Tell me/Microsoft*** [www.microsoft.com/en-us/Tellme (mobile app)]: Users are able to use this app to make a call, open an application, and search the Web.
- ***Dragon*** [www.dragonmobileapps.com (mobile app)]: The Dragon dictation app allows users to speak and instantly see the message, e-mail, or social network updates. The Dragon search app searches a variety of top Web sites using spoken queries.

MUSIC-BASED SEARCH

Often while on the go, we hear a catchy tune or a song that we recognize but can't quite place. This presents a search problem in a modality that isn't covered well by classical text-based information retrieval. One could query based on part of the lyrics but no such option exists for instrumental music, and for many songs the lyrics that one might remember are not very descriptive. Given the availability of a microphone on all mobile devices, capturing a snippet of the song becomes a possibility and several companies provide a music-based query search. In fact, improvements in the robustness of the technology allow the users in cases to simply hum the tune as the input query to the search and successfully retrieve the tune in question.

- ***Shazam*** [www.shazam.com (mobile app)]: Shazam identifies recorded music by listening to the song for about 5–7 s. After recognizing the song, it returns the song name, artist, album, link to buy the song, and YouTube video if available.
- ***SoundHound*** [www.soundhound.com (free and paid mobile apps)]: Similar to Shazam, SoundHound also allows users to search music by singing/humming the song or speaking a title or band name. Lyrics, artist, and other relevant information are retrieved.
- ***MusicID*** [musicid2.com (paid mobile service)]: MusicID identifies music by hearing it. Additional features include viewing lyrics and artist biographies and downloading songs.

IMAGE-BASED SEARCH

Similar to music-based search, some queries are hard to express in text but easy to express in the form of an image. Artwork such as paintings or album and book covers are easily captured with a camera but much harder to accurately describe in text. In other cases, a visual can directly encode information such as universal product code (UPC) or quick response (QR) barcodes or written text that can be automatically analyzed and transformed to a query. A barcode read might lead to a URL or a product search. A text snippet might be a query in itself or be the input to machine translation. In other cases, the image-based query can be a complex scene such as a photo of a product or a characteristic building. Several companies provide image-based query processing as apps on mobile devices since they universally are equipped with high-quality cameras, making the image capture an easily accessible modality to feed such searches.

- ***Google*** [www.google.com/mobile/goggles (mobile app)]: The Goggles app uses camera input to search different kinds of objects and places, including text, landmarks, books, contact info, artwork, wine, and logos.
- ***SnapTell*** [www.snaptell.com (mobile app)]: The SnapTell app finds information about any book, DVD, CD, or video game using a picture of its cover or the UPC/European article number (EAN) barcode image.
- ***Kooaba*** [www.kooaba.com (mobile app)]: The kooaba visual search app recognizes media covers, including books, CDs, DVDs, games, newspapers, and magazines, and provides relevant product information.

MULTIMODAL-BASED SEARCH

Finally, some queries are most naturally expressed by a combination of modalities. Location is easily expressed by a GPS or gesturing on a map whereas a named location is best described by a text query. Furthermore, in cases where multiple locations are the results of a query, gesture to pinpoint a particular result among the set will be a natural choice. Various companies have provided map-based search applications that allow a combination of text- and gesture-based query inputs to location-related search results.

- ***AT&T*** (www2.research.att.com/~johnston): This link introduces the concept of multimodal interfaces and describes some early work of Michael Johnston at AT&T in this field.

Mobile search. Cartoon by Tayfun Akgul (tayfun.akgul@ieee.org).

- ***Google Maps for Mobile*** (www.google.com/mobile/maps): Provides detailed road and satellite maps, possibly with driving directions and location of business listings on the map as markers. Search input is in the form of text queries or speech. The output is consumed mainly by gesture to move the map or get details for a marked business. Driving directions can also produce speech output.
- ***AT&T/YP*** [www.yellowpages.com/products/ypmobile (mobile app)]: The YP app provides voice- and map-based search for the users to find closer results from millions of listings nationwide. Search results can be shared via text, e-mail, Facebook, and Twitter.

MULTIMEDIA CONTENT IN MOBILE MEDIA SEARCH

Multimedia content consumption by mobile users is growing exponentially. According to the CISCO Global Mobile Data Traffic Forecast (http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf), mobile video traffic will exceed 50% of mobile network traffic for the first time in 2011. In light of this trend, major Web search engines such as Google, Bing, Yahoo, and Baidu, have significantly boosted their support for multimedia content search on mobile devices. Support is provided either through customized Web portals or dedicated mobile applications. Content creators like *The New York Times*, NBC, ABC, and others have likewise reached out to mobile consumers by supplying native mobile applications for easy access to their content. Within these applications, users can find news articles as a mix of text, photos, and audio/video. Alternatively, for larger pieces of content like audio books, podcasts, or feature-length movies, desktop media management software might be used to acquire the content and synchronize with the mobile device and then makes the media available for mobile consumption. Examples of such applications are iTunes for iPhone/iPod and Zune for Windows-based devices.

The primary types of multimedia that mobile users are looking for are audio/music and image/video. In this section, we review a few mobile applications that are specialized in these categories.

- ***Pandora*** [www.pandora.com (free registration required)]: Based on a seed of the user preferences on some artists/songs, the application uses collaborative filtering built on the user population to create a user-tailored inventory of music.
- ***Cross Forward*** [www.crossforward.com (mobile app)]: The audiobooks app offers more than 3,500 classics for free. The paid version has a growing collection of premium audiobooks. Many of the recordings are taken from the LibriVox project (librivox.org), where volunteers record chapters of books and make those available in the public domain as shared audio files.
- ***MIT*** [ocw.mit.edu/index.htm (mobile app)]: The MIT LectureHall app provides a dynamic environment for accessing MIT OpenCourseWare video lectures on the go.
- ***Cooliris***: [www.cooliris.com (mobile app)]: Cooliris presents images and videos on an infinite three-dimensional (3-D) wall that lets the user enjoy the content without clicking page to page. It allows the user easily search across YouTube, Google, Flickr, Picasa, deviantART, and Yahoo. The user is able to control the interface by simple gestures and tilting the mobile device.
- ***Facebook*** [www.facebook.com (needs free subscription)]: The tag suggestions function in Facebook is a useful tool that makes face tagging no longer a chore. When the users upload a set of photos that feature the same friends, Facebook groups together faces and suggests the name based on existing tags using face recognition software.
- ***YouTube*** (m.youtube.com): YouTube provides an application that allows search and playback of video content on mobile devices as well as user interaction in the form of posting comments. In addition, YouTube allows users to upload videos they capture with their mobile device directly from their mobile device.
- ***AT&T*** [www.att.com/u-verse (for U-verse service subscribers)]: The U-verse mobile app from AT&T lets U-verse customers search and download popular shows from the mobile library to watch on a mobile device.
- ***Netflix*** (netflix.com): Netflix offers an Internet subscription service for enjoying movies and full episodes of TV shows.
- ***Blinkx*** (m.blinkx.com): With an index of over 35 million h of searchable video, blinkx's video search engine is one of the largest on the Web. At blinkx, the users can create personal video playlists and build a customized video wall for their blog page.

- *TRUVEO* (www.truveo.com): TRUVEO indexes more than 300 million videos from thousands of sources across the Web. Users can search videos, optionally within categories such as news, TV shows, movies, music, and celebrities.

MEDIA SEARCH TECHNOLOGIES

The most commonly used features for making multimedia content discoverable is the tag and metadata associated with that content when it is made available on the Internet. For example, pictures might have captions, videos might have metadata text describing the content, or the content might be annotated with category tags. State-of-the-art search engines increasingly also rely on content-based media processing techniques to provide a richer representation of the indexed content. For example, speech recognition can provide a text signal from the spoken content of videos and image analysis might provide richer content tagging. In this section, we briefly cover the enabling technology in four areas: speech/audio processing, image processing, video processing, and machine learning.

SPEECH/AUDIO PROCESSING

- *SOX* (sox.sourceforge.net): SOX allows file format conversion among commonly used containers. In addition, it allows signal manipulations such as sample rate conversions, channel multiplexing, and filtering.
- *Wavesurfer* (www.speech.kth.se/wavesurfer): Visual analysis of audio recording allows navigation, playback, segmentation, labeling and fundamental pitch, and frequency (spectral) analysis.
- *Transcriber* (trans.sourceforge.net): Transcriber is a complex labeling tool intended for speech transcription to label training data for automatic speech recognition system development.
- *MaART* (maart.sourceforge.net): Tools for music feature extraction for music-based search and retrieval.
- *HTK* (htk.eng.cam.ac.uk): Full-featured hidden Markov model toolkit allowing the design and training from data of a speech recognition system.
- *OpenFst* (www.openfst.org): Full-featured library of finite state transducers manipulation algorithms that can be used to implement large vocabulary speech recognition systems as well as natural language processing systems.
- *SRILM* [www-speech.sri.com/projects/srilm (free for academic use only)]: Toolkit for estimation of statistical language models with applications to automatic speech recognition and natural language processing.

IMAGE PROCESSING

- *OpenCV* (opencv.willowgarage.com/wiki): Library optimized for real-time computer vision application such as image processing, image/video file IO, image segmentation, object detection and tracking, image pyramids, geometric transforms, camera calibration, stereo image processing, and machine learning (C/C++ and Python).
- *MATLAB* [www.mathworks.com/products/image (requires license)]: Tools for image processing and visualization, including image enhancement, image deblurring, feature detection, noise reduction, image segmentation, spatial transformations, and image registration.
- *ImageMagick* (www.imagemagick.org): Tools for creating, editing, and converting images in various image formats.
- *PittPatt* [www.pittpatt.com (Software development kit (SDK) is not free)]: SDK implements face-finding, tracking, and recognition algorithms.

VIDEO PROCESSING

- *TRECVID* (trecvid.nist.gov): Developed in light of the video retrieval task sponsored by NIST. Provides publicly available tools for shot boundary detection, event detection, and copy detection. The data sets used in the evaluations are only available to registered participants.
- *CCExtractor* (ccextractor.sourceforge.net): A closed-caption extracting tool for MPEG files. Supports H.264 streams and CEA/EIA 708/608 standards. Produces closed captions in several formats like SubRip, SAMI, and others.
- *ComSkip* (www.comskip.org): Analyzes MPEG files to detect commercial breaks based on silence, black frame, logo, aspect ratio change, and closed-caption information features.
- *ffmpeg* (www.ffmpeg.org): Tool for transcoding and editing videos. In addition it can be used as a front-end video parser in a video analysis system.

MACHINE LEARNING

Machine learning and pattern recognition are the key technologies that media content analysis systems rely on. The following are a few general-purpose machine learning tools that can be used as the building blocks for content analysis systems:

- *Weka* (www.cs.waikato.ac.nz/ml/weka): A collection of machine learning algorithms in Java for data mining tasks. It contains data preprocessing, classification, regression, clustering, association rules, and visualization.
- *LIBSVM* (www.csie.ntu.edu.tw/~cjlin/libsvm): A library for support vector machines implementing support vector classification, regression, and distribution estimation, as well as multiclass classifiers.
- *Torch* (www.torch.ch): Torch is a machine learning library written in C++, and distributed under a BSD license.

AUTHORS

Zhu Liu (zliu@research.att.com) is a principal member of technical staff in the Video and Multimedia Technologies and Services Research Department of AT&T Labs–Research, USA.

Michiel Bacchiani (michiel@google.com) is a senior staff research scientist leading the speech indexing efforts at Google, USA.

[SP]