Application Potential of Multimedia Information Retrieval

Mohan S. Kankanhalli and Yong Rui, *Senior Member, IEEE*

***Abstract*—This paper will first briefly survey the existing impact of MIR in applications. It will then analyze the current trends of MIR research which can have an influence on future applications. It will then detail the future possibilities and bottlenecks in applying the MIR research results in the main target application areas, such as consumer (e.g. personal video recorders, web information retrieval), public safety (e.g. automated smart surveillance systems) and professional world (e.g. automated meeting capture and summarization). In particular, recommendations will be made to the research community regarding the challenges that need to be met to make the knowledge transfer towards the applications more efficient and effective. It will also attempt to study the trends in the applications which can inform the MIR community on directing intellectual resources towards MIR problems which can have a maximal real-world impact.**

***Index Terms*—Multimedia Systems, Information Retrieval, Applications, Consumer Electronics, Security.**

I. Introduction

Ever since Shannon formalized the notion of information, it has occupied a prominent position in the digital revolution. In fact, common parlance often equates information technology with computer science and engineering. The field of information retrieval has been an important focus area since the very early days [46]. While the original motivation stemmed from library science, many fundamental advances were made to expand the notion of retrieval beyond exact keyword matching. The area went through a relative lull in the 1980s and it suddenly was thrust into limelight in the 1990s after the emergence of world-wide web with the consequent search engines [4]. The seminal contribution of this area has been to establish the search paradigm at the center-stage of the information revolution.

Meanwhile, since the mid-eighties, another sub-theme emerged in the information retrieval research – content-based image retrieval [7]. This research area rapidly exploded in the 1990s by expanding into the field of multimedia information retrieval (MIR) [30]. While text information retrieval has had a radical impact on human society, it is natural to ponder about the influence of multimedia information retrieval. While most researchers will concede that the field is in relative infancy, it is worthwhile to analyze its potential capacity to have a significant impact in the future. In this vein, this paper aims to critically examine the long term application potential of MIR. We have deliberately chosen to concentrate on the application potential because the scientific challenges of the field have been well-understood even if they have not been adequately addressed. It is interesting to note that the growth of the MIR field has been organic in the sense many communities including database, computer vision, image processing, pattern recognition, and information retrieval have a slightly different and independent take on MIR. Our viewpoint in this paper will necessarily be integrative which is essential in order for real-world systems to be built.

In a spirit similar to that of [30], we construe MIR to be search for knowledge in all the digital media forms which can also include across multiple independent information attributes within a single data-stream. Thus web-image search, news video retrieval and music retrieval are all different manifestations of MIR.

We will first start with a brief survey of the existing works which will provide a quick overview of the scientific landscape of MIR. This will be followed by a brief review of the existing applications of MIR actually deployed in the real world. We will then analyze the trends in MIR research to speculate on the future application needs. We will also do the converse – cogitate on the application trends which will influence the research directions for MIR. We will attack this from broadly three application angles – consumer, security and public safety as well as professional world needs. We will then attempt to identify the technical challenges which need to be overcome in order to fructify the application potential. These challenges will then be prioritized in terms of the potential payoffs. And we will end the paper with some prescriptive recommendations.

II. EXISTING RESEARCH

While applications are the main motivation, a lot of research has been spurred by the genuine scientific challenge. In this

Manuscript received May 6, 2007.

Mohan S Kankanhalli is with the School of Computing, National University of Singapore (phone: +65 6516-6738; fax: +65 6779-4580; e-mail: mohan@comp.nus.edu.sg).
Yong Rui is with Microsoft China Research and Development Group, Beijing (e-mail: yongrui@microsoft.com).

section we will briefly survey state-of-the-art on both foundational works done in the field as well as on applications.

A. A Survey on Existing Foundational Works

There are a variety of survey papers available which attempt to capture the scientific challenges and achievements in the various aspects area of multimedia information retrieval. The notion of using computational features to capture media semantics has been a consistent thread in MIR. Bridging this symbol to signal gap (semantic gap) is the core challenge in the field. Then efficiency issues stemming from computational complexity, indexing and usability are the second level of challenge. One of the early surveys [70] did a summarization from a database systems perspective. They not only presented the advances in characterizing content features such as color, shape, texture, they also considered temporal and spatial features for visual data. They also described the melodic features for audio such as pitch, amplitude and notes. Knowledge-based aspects to capture semantics as well as querying aspects were also considered and a taxonomy of existing systems was presented. One of the most complete surveys is presented in [50]. While their focus is on images, many of the scientific challenges of MIR like the sensory gap, semantic gap, content features, similarity functions, incorporation of domain knowledge, interaction, storage, indexing and evaluation have been precisely defined as well as surveyed. What is interesting is the early highlighting of the use of a learning approach (which was widely adopted subsequently) as well as the prescription for a holistic systems approach which has been embraced somewhat less enthusiastically by the community. [27] have presented a survey about the image retrieval from the world-wide web perspective. While the issues covered are similar to those of [50], there is a strong emphasis on the scalability issue due to the intrinsically global nature of the web. Many systems have been analyzed from this perspective. They also strongly emphasize on the need for understanding the users' perspective and also call for the incorporation of other media types in the spirit of MIR. [19] has brought up the issue of 3D models as an important media type to be considered. The detailed aspects of 3D model retrieval have been further elaborated by [15]. While the specifics differ, the main issues still are features, similarity and semantics. Both of the papers have their main focus on 3D shape characterization techniques. [14] presented an interesting survey which analyzed the publication trends in content-based image retrieval. Their survey focuses on features, annotation, concept-detection, relevance feedback and learning as well as interfaces. From their analysis, they found that work on application-oriented issues such as interfaces, visualization, scalability and evaluation has been under-emphasized with a propensity towards systems, feature extraction and relevance feedback. A recent survey of scientific challenges has been undertaken in [30]. This survey article presents the state of the art from various angles such as human-centered approach to retrieval, use of learning in semantics, feature extraction and similarity functions, browsing, summarization, indexing and evaluation. They also point out trends in new media such as 3D models, life-long archives, biological data as well as medical imaging. The paper ends with prescriptions for future work including having human in the retrieval loop, needs of collaborative environments, agent oriented architectures, novel learning approaches, use of folksonomies and crowd-sourcing. The paper intriguingly concludes that there are no completely solved scientific problems in MIR in a general setting. What it essentially means that considerable research efforts still need to be expended in MIR in order to solve some of the fundamental problems such as concept-based semantic search, multimodal analysis, novel interaction mechanisms and large-scale evaluation in realistic settings. A very recent survey on content-based image retrieval with a focus on high-level semantics is [29]. They consider five major approaches towards handling semantics – using ontologies, machine learning methods, relevance feedback, semantic templates and the use of web-based secondary information. They consider semantic templates as a promising way for reducing the semantic gap.

B. Existing Applications Survey

This section focuses on sampling the existing range of work in applications of MIR. One way to categorize MIR applications is to consider their sources, e.g., art and culture, medical, personal and the Web, from domain specific to generic. Today, what researchers can extract directly from images are low-level features, and there are several ways to utilize these features. 1). Users are interested directly in the low-level features, e.g., color and shape in an art object. 2). Users are interested in high-level concepts that have direct relationship with low-level features, e.g., a round dark area in a lung X-ray. 3). Users are interested in high-level concepts, but low-level features can not directly capture the high-level the concepts. Art and culture is an example in 1), medical imaging is an example in 2), and personal and Web are examples in 3). In the last case, we need to train some mid-level models, combine multiple modalities, or integrate human expertise, in order for the retrieval to be useful. This is true for generic MIR in personal content or on the Web.

Art and Culture: As many art objects, e.g., vases in the Getty Museum, have distinct color and texture patterns as well as well-defined shapes, this has been one of the obviously applicable domains for MIR. Some of the early research prototypes include MARS [44]. More recent work includes Zhejiang University's digital library project [73], where they support search for Chinese calligraphy. In general, MIR has been relatively successful in domains where low-level features are directly related with the user queries. .

Medical: While users in this domain are not interested in pure low-level features, they are interested in concepts that have direct relationship with low-level features. For example, a dark round area in a lung X-ray may mean a particular pathology that is of great interest to a medical professional. [1]. If MIR researchers and medical doctors join force, this domain

is likely to bear early fruits in MIR. A tutorial [33] summarizes existing approaches to medical image retrieval. There are promising attempts to link MIR techniques with genomics [54].

Personal: This is a very broad category which can include family albums [71], music recommendation (e.g., the Pandora system [37] and Last.fm [28]) and clothing search [31]. Note that while [71] uses the content-based approach, Pandora relies on human annotation and Last.fm utilizes usage data. Like.com [31] is Riya's offering [41] on clothing search, which currently focuses on handbags, shoes, watches, shirts, etc. It tries to address the "individual tastes" problem and uses a set of visual digital signatures to describe items' content. Like.com is one of the first commercial clothing search engines. More recent personal uses of MIR include personal video recorder, information access from mobile devices and location-based services. Personal video recorder allows a user to customarily store and construct entertainment content. An interesting topic in accessing from mobile device is how to best utilize its small screen. Another related topic is how to provide location-based service to meet user's need based on the context.

The Web: This category is the most diverse one. Given the data's diversity, it can potentially make the biggest impact and at the same time is the most challenging domain. Most of the today's MIR Web search engines, e.g., [20][34][68], are based on surrounding text and meta-data. However, true MIR search engines have begun to emerge, e.g., Blinkx [5] (uses both visual and speech features) and SearchVideo [47] (uses both text and visual features). A new trend in Web 2.0 is the long-tail effect. Compared to textual information on the Web, this is even more critical for the multimedia content. How to leverage meta-data on the Web and the semantic Web are probably key to successful Web-based MIR.

There are many other interesting and important domains, including industry (e.g., aircraft [60]), CAD/CAM [48], trademark [23], and in recent years security [36] and professional world [13].

While there have been a variety of applications attempted in several specialized as well as general domains, its widespread use is still yet to come. We strongly believe that this is an indicator of the incipient nature of the field. The text-based retrieval approach is relatively easier to exploit and hence more resources have been quite rightly concentrated in that area. However, while there are still substantial challenges in the text-only arena, we will analyze the application trends to argue that MIR will become necessary for several applications. And MIR will simultaneously and symbiotically coexist with text based information retrieval.

III. TRENDS

In this section, we will analyze the trends in MIR research and its potential influence in several areas of interest. It will be a speculative attempt grounded in current trends. And the subject matter will be analyzed both ways -- how will MIR impact applications and also, it will outline the trend of applications which could impact what research problems should be studied by the MIR community. The research problems will be mentioned here in context and will be distilled as challenges in the next section.

A. Consumer World

Broadly speaking, the consumer multimedia content can be classified into two categories: the ones that possess specific structures and the ones that do not. The first category includes news videos and, to some extent, movies. The structure makes content analysis easier and has been exploited for automatic classification in [45][53]. A typical approach in these systems is a two stage scene classification scheme. First, the video stream is parsed and video shots are extracted. Second, each shot is then classified according to content classes such as *newscaster*, *report*, or *weather forecast*. The classification relies on the definition of one or more image/video templates for each content class. To classify a generic shot, a key frame is extracted and matched against the image template of every content class.

The second category is unstructured, e.g., home videos, or semi-structured, e.g., sports videos. Many researchers have studied the respective role of visual, audio and textual mode in sports video. For example, for the visual mode, Kawashima et al. [26] extracted bat-swing features based on the video signal. Xie et al. [66] segmented soccer videos into play and break segments using dominant color and motion information. For the audio mode, Rui et al. [43] detected the announcer's excited speech and ball-bat impact sound in baseball games using directional audio template matching. For the textual mode, Babaguchi et al. [3] look for time spans in which events are likely to take place through extraction of keywords from the closed captioning stream. An example approach using multimodal information fusion was reported by Snoek and Worring [51]. More recent work on sports video came from the Xu group [62][63][64] by using HMM analysis and integration of webcast text.

Some of the key trends in consumer multimedia content are:

- From structured media to unstructured media: This is the case from news [53] to sports [63], and is also the case within the sports video genre itself, i.e., from broadcast video [62] to non-broadcast video [63].
- From single media analysis [3][26] [43] to multimedia and multiple modality analysis [51].
- From analysis-alone [62][63] to integrated analysis and synthesis.

B. Public Safety

There are two main themes of research in the broad area of public safety. The first area is related to surveillance and monitoring while the second area is related to biometrics.

A significant amount of work has been done by the computer vision and multimedia researchers in the context of video surveillance, such as for face detection [65], moving object detection [25], object tracking [6], object classification [11], human behavior analysis [35], people counting [69], and

abandoned object detection [52]. A few works have also been reported for the surveillance using audio. The examples of various audio events detected in the past include glass breaks, explosions or door alarms [17], talking person, falling chair [12], impulsive gun shots [9] human's coughing in the office environment [21] and the working of an air-conditioner [32]. There are few works which have demonstrated the use of sensors other than video and audio. Pavlidis and Faltesek [38] used bio-chemical sensors and video camera to propose a security system against bio-chemical attacks. Foresti and Snidaro [18] used infrared cameras and color cameras to build a distributed sensor network for video surveillance for outdoor environments. They employed a linear fusion for combining the trajectory information about objects. Peralta and Peralta [39] presented a Perimeter Intruder Detection System (PIDS) for surveillance of risky environments such as swimming pools. They used infrared sensor-emitter and detectors units driven by the micro-controllers. Recently, Prati et al. [40] also presented a multisensor surveillance system consisting of video cameras and passive infra-red sensors (PIR). Their proposed used of multiple sensors helps in better object/person tracking. A comprehensive survey on visual surveillance of object motion and behaviors is presented in [22]. Valera and Velastin [57] have presented a survey on the state of the art of surveillance systems. They point to the trend of wide-area, multi-sensor systems with a high degree of automation.

Some of the key trends in surveillance and monitoring are:

- From rigid to flexible system design: Current surveillance systems are usually designed to handle only the specified tasks in rigid sensor settings. For example, if a surveillance system is designed to capture the faces of persons entering into a designated area, it is not used for any other task. The trend is to adopt a flexible approach and look at the surveillance systems in a "search paradigm" where an end-user can flexibly query the system, in a continuous or one-time manner, pretty much in the manner of search engines. Thus, it directly maps to the MIR problem [2].
- From camera only to multiple sensor types: Use of infrared, acoustic and chemical sensors in conjunction with video cameras is increasing. Visual sensors will continue to be dominant sensors but they will be opportunistically supplemented with other suitable sensors [10].
- From custom architectures to customizable architectures: Current systems tend to be built for a particular physical environment with particular sensor types and sensor placements. While this is efficient, it lacks portability and scalability necessary for widespread deployment. Given any physical surveillance environment, the system architecture should be able to register and identify the sensors and other sources that can be used to flexibly answer many expressive queries. In addition, addition and removal of sensors from the environment needs to be transparently handled [2].

Biometric applications represent a specialized application area of MIR related to public safety. Face recognition is an extremely well-studied area [72]. Given that many authentication and security applications need fast and accurate matching or searching of a given face from a large database, there have been significant efforts in this arena. This type of requirement also exists for other biometric signatures such as fingerprint, iris, palm and DNA recognition [16]. The main trend of the works here are robust feature extraction and improved matching. A comprehensive survey of issues in biometrics is discussed in [24].

The trend here is to attempt to meet the grand challenge of simultaneously maximizing scale, usability and accuracy. Multimodal biometrics is then essential and will be increasingly adopted. This is in tandem with the advocacy of multi-factor authentication in the systems security field.

C. Professional World

Multimedia content not only exists in the consumer domain, they are also prevalent in the professional world. A good example is meeting content. Meetings are an important part of corporate life. Often, due to scheduling conflicts or travel constraints, people cannot attend all of their scheduled meetings. In addition, people are often only peripherally interested in a meeting such that they want to know what transpired there without actually attending. Being able to retrieve, browse and skim captured meeting content can therefore be very valuable [13]. For example, if a timeline can be generated showing when each person talked during the meeting, it can allow users to jump to interesting points, listen to a particular participant, and better understand the dynamics of the meeting.

To enable these functionalities, speaker tracking and clustering is the key. The core techniques include sound source localization, person tracking, and sensor fusion.

- *Sound Source Localization (SSL)*. This utilizes a microphone array to detect which meeting participant is talking [42].
- *Person Tracking* Although audio-based SSL can detect who is talking, its spatial resolution is not high enough to finely steer a virtual camera view. Occasionally it can also lose track due to room noise, reverberation, or multiple people speaking simultaneously. Vision-based person tracking is a natural complement to SSL. Though it does not know who is talking, it has higher spatial resolution and tracks multiple people at the same time.
- *Sensor Fusion.* The sensor fusion module gathers and analyzes reports from the SSL and person tracker to make an intelligent decision on who is talking. In [8], the authors use particle filters (PF) to obtain the speaker location. The proposal function of the PF is obtained from individual audio/video sensors, e.g., the person tracking module and SSL module. Particles are then drawn from the proposal function to be weighted and propagated through time [8].

While meetings are a widespread example, similar needs arise in other professional contexts such as archives of classroom

lectures, public speeches, workshops and conferences. There are several key trends in professional world MIR:

- From centralized to distributed. In the past, most of the professional content, e.g., meetings, notes and documents were stored centrally in a single location. Today, with the increase in project size and globalization, teams are becoming distributed and so is the professional content. How to effectively capture, store, share and collaborate on the content is therefore critical for project execution.
- From static to interactive. Traditional professional content is stored without much intelligence. When a content is viewed, the experience is static, one-way and passive – whatever is stored is played back to the user, no matter who that person is. Today, with more intelligence during the capture, e.g., who is talking, and post-processing, e.g., only show me the part where Tom speaks, the viewer experience becomes interactive, two-way and rich.

IV. TECHNICAL CHALLENGES

Given the existing and new applications, in this section, we will analyze what is needed in terms of technologies. This will be directly related to the research problems hinted at in the previous section. Some of these challenges have been identified earlier [67].

1) *Bridging the Semantic Gap*

What algorithms can automatic extract today are low-level features while what end users need are high-level concepts. This is called the semantic gap, and except in a few well-defined domains, e.g., medical and GIS, the gap is large. Researchers have tried both the pure manual labeling approach and the pure automatic content-based approach. Neither is completely successful. We argue that something in the middle, e.g., semi-automatic annotation approach, will bear fruit. TRECVID is making progress on that front [55]. It must be noted that some of the technical challenges outlined in this section are essentially alternative attempts at surmounting this gap (e.g. 2, 3, 4, and 7).

2) *How to Best Combine Human and Machine Intelligence?*

One advantage of MIR, compared with traditional pattern recognition, is that in most scenarios MIR systems are primarily designed with the human being as the user. Even the earliest compression algorithms recognized this fact and exploited the removal of perceptual redundancy in terms of the human visual system and the human psychoacoustic models. The work on the semantic and sensory gaps, which aims to link signals to symbols, is also facilitated by the recognition of the human in the loop. For instance, the work on relevance feedback for retrieval purposes utilizes the human's role as a consumer of multimedia information. There is a growing realization that fully automated systems are perhaps not always necessary where effective systems can be built in which tasks are apportioned based on the relative strengths of humans and machines. However, while the relevance feedback work has made an impact, it is still not sufficient. More advanced approaches need to be developed. One interesting development is the idea of human computation where large numbers of people are cleverly engaged in the task of tagging and region annotation via the ostensible mechanism of games [58][59]. This is a variation of crowd-sourcing unlike Flickr where the reliance is on uncoordinated individual voluntary efforts. Collaborative tagging using social networking is also an increasingly popular mechanism.

3) Active Multimedia Information Retrieval

Active feedback has been proposed for text-based information retrieval [49]. The basic idea is that the system should actively and collaboratively participate with the user in meeting the information need. It is different from relevance feedback in the sense that the system must decide which documents to present to the user in order to maximize the benefit of the user's judgment. That is, we can consider relevance feedback as "users provide the right answers" while active feedback as "the system asks the right questions". This paradigm can be particularly interesting in the MIR context since a combination of cross-modal information can be utilized to best learn the user intent.

4) Judicious Use of Secondary Sources of Information

The text retrieval problems have immensely benefited from the use of WordNet which essentially acts like a secondary source of information. Preliminary attempts to mimic this in the content-based image retrieval arena have shown that the use of secondary information in the form of web-data can substantially improve precision and recall. It would be interesting to formalize this notion of secondary sources of information in a rigorous framework. The challenge will be in suitably utilizing the noisy and variable quality information from diverse sources (such as emails, calendars, blogs, spreadsheets, voicemail etc) to amplify the information gain. Cross-modal retrieval will be crucial in helping achieve this.

5) Centralized Services versus Distributed Services

In a lot of MIR applications, the traditional flip-flop between centralized and distributed architectures will arise. Should one store one's personal photos on the home PC or a corporation's server like Flickr? This will also be true for videos, songs, emails, spreadsheets, 3D models and all kinds of multimedia information. While a personal collection provides greater control, it also demands greater peripheral responsibilities. Reliability, fault tolerance, ease of access, handling of legacy formats and trust will be critical issues. These issues can be directly mapped to technical features related to the underlying hardware architecture and database deployed for the MIR system. However, it is not always strictly a technical issue -- economic pricing models often contribute to the prevalence of one service architecture over the other. For example, advertisement supported services tend to favor a centralized

system. In the long run, there will be perhaps a mixed ecology of systems and architectures over which MIR systems will be built. Choosing the right architecture for a particular MIR system is an open problem.

6) *Handling of Live Multimedia Data*

Current MIR systems such as those for web search tacitly assume the relatively static crawl-able and indexable nature of data. But if there are many live sensory feeds, these assumptions are no longer valid. Crawling over several days will be useless for live data and massive-scale real-time indexing will be infeasible. Moreover, information needs arise anywhere due to increased mobility. With mobile phones becoming ubiquitous, universal MIR is a possibility. How does one effectively retrieve multimedia data in such a scenario? The search paradigm will evolve to an *information-on-demand* paradigm – the minimal amount of information needed by the user to accomplish the task at hand is to be delivered in the right mode in a timely fashion at the right place. This trend may call for revolutionary advances in system architectures.

7) *Impact of Novel Sensors*

With the increasing variety and decreasing cost of various types of sensors, there will be an increase in the use of radically different media such as infrared, motion sensor information, text in assorted formats, optical sensor data, telemetric data of various sorts (biological and satellite), transducers data, financial data, location data captured by GPS devices, spatial data, haptic sensor data, graphics and animation data. Some other developments are moving cameras on vehicles such as public buses (which is essentially the issue of mobile sensors). Humans are also mobile sensors recording information in various media such as blogs. It would be useful to speculate on which type of new sensors could help cross difficult hurdles of MIR? For example, if every interesting object/building/place in the world had RFID tags with some indexed information, then cameras equipped with RFID readers would greatly simplify the annotation problem. Therefore, it may be worthwhile to think of opportunistically enhancing the environment with suitable sensors to overcome the sensory and semantic gaps. This approach may yield rich pay-offs.

8) *Expanding the Search Paradigm into Newer Areas*

Many new and old problems are being recast as a MIR problem. Desktop search is one example. Continuously archived data like in myLifeBits is another. Multimedia surveillance is also witnessing this transformation. Corporate databases and national archives also naturally lend themselves to be recognized as MIR systems. Data mining needs will force a re-look of massive, spatially distributed, temporally dynamic data such as in finance, customer relationship management and transport arenas to be considered as MIR problems. The search paradigm will be critical in order to handle the data interdependence complexity.

V. DISCUSSION

This section will essentially prioritize the identified challenges in research problems and applications in terms of the potential impact and also will discuss about obstacles in achieving them.

The eight technical challenges in the previous section can be classified into 3 categories. The first 4 concern the semantic gap, where challenges 2-4 are using different techniques to bridge the gap. Challenges 5 and 6 concern the architecture, system and performance of the system. Finally, challenges 7 and 8 look into future sensors, paradigms and applications.

From the application point of view, the most immediate challenge lies in the overcoming of the sensory and semantic gaps. Many challenges outlined in the previous section either directly or indirectly try to overcome this obstacle. For MIR, there are three paradigms on the research spectrum that ranges from the least automatic to the most automatic. On the far left end, there is the pure manual labeling paradigm that labels multimedia content, e.g., images and video clips, manually with text labels and then use text search to search multimedia content indirectly. On the far right end, there is the content-based search paradigm that can be fully automatic by using low-level features from multimedia analysis. Over the past decade, there has been a gradual realization that neither of the above two extremes provides good search/retrieval results. Instead, something in the middle strikes a better balance, and that is the automated annotation paradigm. This paradigm is not purely automatic, as people need to label/annotate some content. But it is not purely manual either - once a concept detector is trained based on the labeled data, the detector can automatically annotate the same concept for other video clips. Furthermore, within the annotation paradigm, the relationship-based annotation approach outperforms other existing annotation approaches, because individual concepts are considered jointly instead of independently. This can be suitably enhanced using novel sensors and secondary sources of information.

In the medium term, more domain specific applications on personal systems can proliferate provided significant advances are made in incremental learning, active retrieval techniques and appropriate query paradigms.

In the long run, handling of cross-modal information especially in the mixed setting of live and archived data aided by optimal contextual processing will be the key to the spread of the MIR systems. However, instead of recognizable standalone MIR systems, perhaps seamlessly embedded MIR applications selectively cooperating with complementary systems may turn out to be the ultimate manifestation of the information on demand paradigm.

VI. Conclusion

We have first presented a brief survey on the foundational works and existing applications in MIR. A study of several surveys reveals prolific activity in this area which has been matched by the sheer diversity of applications developed. However, widespread commercial use is still at a nascent stage. We then studied the application trends to identify eight technical challenges. They fit into three broad categories related to the semantic gap, system issues and future sensor-based emerging paradigms. We speculate that the optimal delivery of apposite information utilizing parsimonious resources for satisfying the information needs of the globally mobile end-user will be the long term driver for a profusion of diverse MIR services.

Acknowledgment

We thank Pradeep Atrey, Min-Yen Kan, Tat-Seng Chua and Ramesh Jain for the insightful discussions.

References

[1] J. Amores and P. Radeva, Medical Image Retrieval Based on Plaque Appearance, Chapter in "Plaque Imaging Book", IFMBE press, ed. J. Suri et.al.

[2] P.K. Atrey, M.S. Kankanhalli and R. Jain, Information Assimilation Framework for Event Detection in Multimedia Surveillance Systems, ACM Multimedia Systems Journal, Vol. 12, No. 3, pp. 239-253, December 2006.

[3] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event-based indexing of broadcasted sports video by intermodal collaboration," IEEE Transactions on Multimedia, vol. 4, no. 1, pp. 68-75, March 2002.

[4] J. Battelle, "*The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*", Penguin, New York, 2005.

[5] Blinkx.com, http://www.blinkx.com

[6] E. Chang and Y. F.Wang. Multi-camera spatio-temporal fusion and biased sequence-data learning for security surveillance. *Proc. ACM International Workshop on Video Surveillance*, Berkeley, CA, USA, November 2003.

[7] S.K. Chang, "Image Information Systems", Special Issue on Visual Communications Systems, *Proceedings of the IEEE*, Vol. 73, No. 4, pp. 754-764, April 1985.

[8] Y. Chen and Y. Rui Real-time Speaker Tracking Using Particle Filter Sensor Fusion, Proceedings of the IEEE , vol. 92, no. 3, pp. 485-494, Mar. 2004

[9] C. Clavel, T. Ehrette, and G. Richard. Event detection for an audio-based surveillance system. *Proc. IEEE International Conference on Multimedia and Expo*, Amsterdam, July 2005.

[10] S. Calderara, R. Cucchiara**,** A. Prati, "Multimedia Surveillance: Content-based Retrieval with Multicamera People Tracking" in *Proc. ACM International Workshop on Video Surveillance and Sensor Networks (VSSN 2006)*, Santa Barbara, USA, Oct. 2006.

[11] R. T. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of IEEE*, Vol. 89 , No. 10, pp. 1456–1477, 2001.

[12] M. Cristani, M. Bicego, and V. Murino. Online adaptive background modeling for audio surveillance. *Proc. IEEE International Conference on Pattern Recognition*, pp. 399–402, Cambridge, UK, August 2004.

[13] R. Cutler, Y. Rui, et. al. Distributed meetings: a meeting capture and broadcasting system, Proc. of ACM Multimedia 2002

[14] R. Datta, J. Li and J. Wang, Content-Based Image Retrieval – Approaches and Trends of the New Age, *Proc. ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR 2005)*, pp. 253-262, Singapore,, November 2005.

[15] A. Del Bimbo and P. Pala, Content-Based Retrieval of 3D Models, *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 2, No. 1, pp. 20-43, February 2006.

[16] K. Delac and M. Grgic, A Survey of Biometric Recognition Methods, *Proc. International Symposium Electronics in Marine (ELMAR 2004)*, 184-193, June 2004.

[17] A. Dufaux, L. Bezacier, M. Ansorge, and F. Pellandini. Automatic sound detection and recognition for noisy environment. *Proc. European Signal Processing Conference*, pp. 1033–1036, Finland, September 2000.

[18] G. L. Foresti and L. Snidaro. A distributed sensor network for video surveillance of outdoor environments. *Proc. IEEE International Conference on Image Processing,* Rochester, New York, USA, September 2002.

[19] T. Funkhouser, M. Kazhdan, P. Min and P. Shilane, Shape-Based Retrieval and Analysis of 3D Models, *Communications of the ACM*, Vol. 48, No. 6, pp. 58-64, June 2005.

[20] Google Video, http://video.google.com/

[21] A. Harma, M. F. McKinney, and J. Skowronek. Automatic surveillance of the acoustic activity in our living environment. *Proc. IEEE International Conference on Multimedia and Expo*, Amsterdam, July 2005.

[22] W. Hu, T. Tan, L. Wang and S. Maybank, A Survey on Visual Surveillance of Object Motion and Behaviors, *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, Vol. 34, No. 3, pp. 334-352, August 2004.

[23] A. K. Jain and A. Vailaya, Shape-based retrieval: a case study with trademark image databases, Pattern Recognition, Vol. 31, No. 9, Sept 1998, pp. 1369-1390.

[24] A. K. Jain, A. Ross and S. Pankanti, " Biometrics: A Tool for Information Security", *IEEE Transactions on Information Forensics and Security* Vol. 1, No. 2, pp. 125-143, June 2006.

[25] O. Javed, Z. Rasheed, O. Alatas, and M. Shah. M-KNIGHT: A real time surveillance system for multiple overlapping and non-overlapping cameras. *Proc. IEEE International Conference on Multimedia and Expo*, pp. I:649–652, Baltimore, MD, USA, July 2003.

[26] T. Kawashima, K. Tateyama, T. Iijima, and Y. Aoki, "Indexing of baseball telecast for content-based video retrieval," in Proceedings of the International Conference on Image Processing, 1998, pp. 871-874.

[27] M. Kherfi, D. Ziou and A. Bernardi, Image Retrievaal from the World Wide Web: Issues, Techniques, and Systems, *ACM Computing Surveys*, Vol. 36, No. 1, pp. 35-67, March 2004.

[28] Last.fm, http://www.last.fm

[29] L. Liu, D. Zhang, G. Lu and W.Y. Ma, A Survey of Content-based Image Retrieval with High-level Semantics, Pattern Recognition, Vol. 40, No. 1, pp. 262-282, 2007.

[30] M.S. Lew, N. Sebe, C. Djeraba and R. Jain, "Content-Based Multimedia Information Retrieval: State of the Art and Challenges", *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 2, No. 1, pp. 1-19, February 2006.

[31] Like.com, http://www.like.com

[32] M. McHugh and A. F. Smeaton. Towards event detection in an audio-based sensor network. *Proc. The ACM International Workshop on Video Surveillance and Sensor Networks*, pp. 87–94, Singapore, November 2005.

[33] Medical, http://www-i6.informatik.rwth-aachen.de/~deselaers/mie2005-tutorial-dvd/

[34] MSN Soapbox, http://soapbox.msn.com/

[35] W. Niu, L. Jiao, D. Han, and Y. F.Wang. Human activity detection and recognition for video surveillance. *Proc. IEEE conference on Multimedia and Expo*, Taiwan, June 2004.

[36] ObjectVideo, http://objectvideo.com/

[37] Pandora.com, http://www.pandora.com/

[38] I. Pavlidis and T. Faltesek. A video-based surveillance solution for protecting the air-intakes of buildings from bio-chem attacks. *Proc. IEEE International Conference on Image Processing*, Rochester, New York, USA, September 2002.

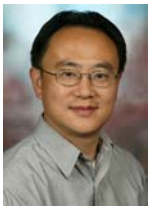
[39] J. O. Peralta and M. T. C. Peralta. Security PIDS with physical sensors, real-time pattern recognition, and continuous patrol. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, Vol. 32, No. 4, pp. 340–346, November 2002.

[40] A. Prati, R. Vezzani, L. Benini, E. Farella, and P. Zappi. An integrated multi-modal sensor network for video surveillance. *Proc. The ACM International Workshop on Video Surveillance and Sensor Networks*, pp. 95–102, Singapore, November 2005.

[41] Riya.com, http://www.riya.com

[42] Y. Rui and D. Florencio, Time Delay Estimation in the Presence of Correlated Noise and Reverberation, *Proc. of IEEE ICASSP 2004* , Montreal, Quebec, Canada, May 17-21.
[43] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in Proceedings of the Eighth ACM International Conference on Multimedia, 2000, pp. 105-115.
[44] Y. Rui, T. Huang, and S. Mehrotra, Content-based image retrieval with relevance feedback in MARS, Proc. of IEEE ICIP, 1997
[45] Y. Rui, T. S. Huang, and S. Mehrotra, "Exploring video structures beyond the shots," in Proc. of IEEE Conf. Multimedia Computing and Systems, pp. 237-240, 1998.
[46] G. Salton, "Automatic *Information Organization and Retrieval*", McGraw-Hill, New York, 1968.
[47] SearchVideo.com, http://www.searchvideo.com
[48] C. Shahabi and S. Maytham, An experimental study of alternative shape-based image retrieval techniques, Multimedia Tools and Applications, 32: 29-48, 2006.
[49] X. Shen and C. Zhai. Active feedback in ad hoc information retrieval. Proc. International ACM Conference on Research and Development in Information Retrieval (SIGIR 2005), pages 59 -- 66, 2005.
[50] A. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 22, No. 12, pp. 1349-1380, December 2000.
[51] C. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," Tech. Rep., Intelligent Sensory Information Systems Group, University of Amsterdam, Technical Report 2001-20, 2001.
[52] M. Spengler and B. Schiele. Automatic detection and tracking of abandoned objects. *Proc. Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Nice, France, October 2003.
[53] D. Swanberg, C. Shu, and R. Jain, "Knowledge guided parsing in video databases," SPIE Processings, vol. 1908, pp. 13-24, 1993.
[54] T. Syeda-Mahmood, Content-based Retrieval in Gene Expression Databases, *Proc. ACM International Conference on Multimedia*, pp. 78-787, 2004.
[55] TRECVID. *TREC video retrieval evaluation.* http://www-nlpir.nist.gov/projects/trecvid/
[56] University of Amsterdam, http://www.mediamill.nl
[57] M. Valera and S. A. Velastin. Intelligent Distributed Surveillance Systems: A Review. *IEE Proceedings on Visual Image Signal Processing*, Vol. 152, No. 2, pp. 192–204, April 2005.
[58] L. von Ahn and L. Dabbish. Labeling Images with a Computer Game. In *ACM Conference on Human Factors in Computing Systems, CHI 2004*. Pages 319-326, 2004.
[59] L. von Ahn**,** R. Liu and M. Blum. Peekaboom: A Game for Locating Objects in Images. In *ACM Conference on Human Factors in Computing Systems, CHI 2006*. Pages 55-64, 2006.
[60] J. Wang, and G. Wiederhold, Simplicity: semantics-sensitive integrated matching for picture libraries. IEEE Trans. PAMI, 23, 9, 947-963
[61] Z. Wang, M. Hoffman, P. Cook and K. Li, VFerret: Content-Based Similarity Search Tool for Continuous Archived Video, *Proc. ACM Workshop on. Capture, Archival and Retrieval of Personal Experiences (CARPE 2006)*, pp. 19-25, Santa Barbara, October 2006.
[62] J. Wang, C. Xu, E. S. Chng, K. Wah and Q. Tian, Automatic Replay Generation for Soccer Video Broadcasting", *Proc. of ACM MultiMedia'04*, pp 32-39, New York, USA, 2004
[63] J. Wang, C. Xu, E. S. Chng, X. Yu and Q. Tian, Event Detection Based on Non-Broadcast Sports Video", *Proc. of IEEE ICIP'04*, pp. 1637-1640, Singapore, 2004
[64] J.Wang, C. Xu, E. S. Chng, L. Duan, K.Wan and Q. Tian, "Automatic Generation of Personalized Music Sports Video", *Proc. of ACM MultiMedia'05*, pp 735-744 Singapore, 2005.
[65] B.Wu, H. Ai, C. Huang, and S. Lao. Rotation invariant neural network based face detection. *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, pp. 79– 84, Seoul, Korea, May 2004.
[66] L. Xie, S. F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," in Proceedings of the International Conference on Acoustic, Speech, and Signal Processing, vol. 4, May 2002, pp. 4096-4099.
[67] Z. Xiong, X.S. Zhou, Q. Tian, Y. Rui and T.S. Huang, Semantic Retrieval of Video, IEEE Signal Processing Magazine, Vol. 23, No. 2, pp. 18-27, 2006.
[68] Yahoo! Video, http://video.yahoo.com
[69] D. B. Yang and H. H. Gonzalez-Banos. Counting people in crowds with a real-time network of simple image sensors. *Proc. IEEE International Conference on Computer Vision*, Nice, France, October 2003.
[70] A. Yoshitaka and T. Ichikawa, A Survey on Content-Based Retrieval for Multimedia Databases, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 11, No. 1, pp. 81-93, 1999.
[71] L. Zhang, L. Chen, M. Li, and H. Zhang, Automated annotation of human faces in family albums, Proc. of ACM multimedia, 2003
[72] W. Zhao, R. Chellappa, P. Phillips and A. Rosenfeld, Face Recognition: A Literature Survey, *ACM Computing Surveys*, Vol. 35, No. 4, pp. 399-458, December 2003.
[73] ZheJiang University, http://www.cadal.zju.edu.cn/IM/IM.htm

Mohan S. Kankanhalli is a Professor at the Department of Computer Science of the School of Computing at the National University of Singapore. He obtained his BTech (Electrical Engineering) from the Indian Institute of Technology, Kharagpur, and his MS and PhD (Computer and Systems Engineering) from the Rensselaer Polytechnic Institute. He has worked at the Institute of Systems Science in Singapore and at the Department of Electrical Engineering of the Indian Institute of Science, Bangalore. His current research interests are in Multimedia Signal Processing (sensing, content analysis, retrieval) and Multimedia Security (surveillance, digital rights management and forensics). He is on the editorial board of several journals including the IEEE Transactions on Multimedia and the IEEE Transactions on Information Forensics and Security.



Yong Rui serves as Director of Strategy of Microsoft China R&D (CRD) Group. Before this role, Dr. Rui spent seven years and managed the Multimedia Collaboration team at Microsoft Research Redmond. Dr. Rui is a Senior Member of both ACM and IEEE. He is an Associate Editor of ACM Transactions on Multimedia Computing, Communication and Applications (TOMCCAP), IEEE Transactions on Multimedia, and IEEE Tran on Circuits and Systems for Video Technologies.. He received his PhD from University of Illinois at Urbana-Champaign (UIUC). Dr. Rui contributes significantly to the research communities in computer vision, signal processing, machine learning, and their applications in communication, collaboration, and multimedia systems. His contribution to relevance feedback in image search created a new research area in multimedia. He has published twelve books and book chapters, and over seventy referred journal and conference papers. Dr. Rui holds 30 issued and pending US patents. He is a General Chair of Int. Conf. Image and Video Retrieval (CIVR) 2006, a Program Chair of ACM Multimedia 2006, and a Program Chair of Pacific-Rim Conference on Multimedia (PCM) 2006.