Oriental Journal of Computer Science & Technology Vol. **3**(1), 125-137 (2010)

Video indexing and retrieval - Applications and challenges

M. RAVINDER, T. VENU GOPAL[2] and T. VENKAT NARAYANA RAO[3]

[1]Computer Science and Engineering, Geethanjali College of Engineering & Technology, Cheeryal, R.R.Dist, (India). [2]Computer Science and Engineering, JNTUCEH, Jagityal, Karimnagar (India). [3]Hyderabad Institute of Technology & Management (HITAM), R.R.District (India).

(Received: April 20, 2010; Accepted: May 21, 2010)

ABSTRACT

Video data indexing and retrieval which applies tags to large video databases, is useful as a complementary means for applications which are having multi media content and need faster search responses. Well-ordered and effective management of video documents depends on the availability of indexes. Manual indexing is not viable for large video collections in this modern era of information super highway. There is a want for a techniques and frameworks that can store, handle, search and retrieve the data from the huge media archive. This paper discusses about application areas, emerging challenges of video indexing & retrieval, and some future directions for video indexing, video retrieval management system .

Key words: Video indexing, Information retrieval, Annotations, Semantic gap, VDBMS.

INTRODUCTION

Video information plays a central role in multimedia database systems [1]. The amount of video information stored in archives worldwide is very huge. Conservative estimates state that there are more than 6 million hours of video already stored and this number grow at a rate of about 10 percent a year. Projections estimate that by the end of 2010, 50 percent of the total digital data stored worldwide will be video and rich media. Significant efforts have been spent in recent years to make the process of video archiving and retrieval faster, safer, more reliable and accessible to users anywhere in the world. Progress in video digitization and compression, together with advances in storage media, have made the task of storing and retrieving raw video data much easier. Evolution of computer networks and the growth with popularity of the Internet have made it possible to access these data from remote locations also.

Section II talk about video database management system, section III addresses about video indexing and retrieval, section IV focuses on applications, section V talks about technical challenges, section VI shows future directions, and section VII end up with conclusion.

Video database management system

The primary goal of a Video Database Management System (VDBMS) is to provide pseudo-random access to sequential video data. This goal is normally achieved by dividing a video clip into segments, indexing these segments, and representing the indexes in a way that allows easy browsing and retrieval. Therefore, it can be said that a VDBMS is basically a database of indexes (pointers) to a video recording [2], [8] process.

A. Components of VDBMS

The figure 2.1 presents a simplified block diagram of a typical VDBMS. Its main blocks constitute:

- Digitization and compression: Hardware and

software necessary to convert the video information into digital compressed format.
- Cataloguing: Process of extracting meaningful story units from the raw video data and building the corresponding indexes.
- Query / search engine: Responsible for searching the database according to the parameters provided by the user.
- Digital video archive: Repository of digitized, compressed video data.
- Visual summaries: Representation of video contents in a concise, typically hierarchical way.
- Indexes: Pointers to video segments or story units.
- User interface: Friendly, visually rich interface that allows the user to interactively query the database, browse the results, and view the selected video clips.

B. Video Content Organization

Because video is a structured medium in which actions and events in time and space convey stories, a video program must not be viewed as a non-structured sequence of frames, but instead it must be seen as a document.

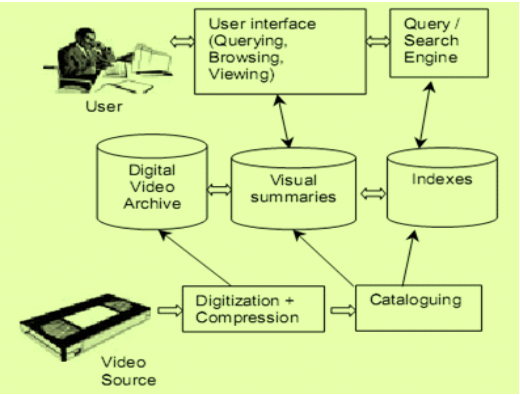

Fig. 2.1. Block diagram of a VDBMS

The sequential browsing method may not be suitable for long video sequences. In this case, methods using content summaries, such as those based on the Table of Contents (ToC), are very useful in providing quick access to structured video content. The process of converting raw video into structured units, used to build a visual table of contents (ToC) of a video program, referred to as video abstraction. The process of video content organization can be divided into three parts:

(1) Video modeling and representation.
(2) Video segmentation.
(3) Video summarization.

1) Video modeling and representation: Video modeling can be defined as the process of designing the representation of the video data based on its characteristics, the information content, and the applications it is intended for. Video modeling plays a key role in the design of VDBMSs, because all other functions are more or less dependent on it. Some of the requirements for a video data model are[2]:

- Support video data as one of its data types, just like textual or numeric data.
- Integrate content attributes of the video program with its semantic structure.
- Associate audio with visual information content.
- Express structural and temporal relationships between segments.
- Automatically extract low-level features (color, texture, shape, motion), and use them as attributes.

Video data models can usually be classified into the following categories [2] as shown in Figure 2.2.

- Models based on video segmentation: adopt a two-step approach, first segmenting the video stream into a set of temporally ordered basic units (shots), and then building domain-dependent models (either hierarchy or finite automata) upon the basic units.
- Models based on annotation layering (also known as stratification models): segment contextual information of the video and to approximate the movie editor's perspective on a movie, based on the assumption that if the annotation is performed at the finest grain (by a data camera), later coarser grain of information may be reconstructed easily if needed.
- Video object models: extend object-oriented data models to video. Their main advantages include the ability to represent and manage complex objects, handle object identities, encapsulate data and associated methods into objects, and inherit attribute structures

and methods based on class hierarchy.

- Algebraic video data models: define a video stream by recursively applying a set of algebraic operations on the raw video segment. Their fundamental entity is a presentation (multi-window, spatial, temporal, and content combination of video segments). Presentations are described by video expressions, constructed from raw segments using video algebraic operations.
- Statistical models: exploit knowledge of video structure as a means to enable the principled design of computational models for video semantics, and use machine learning techniques (e.g., Bayesian inference) to learn the semantics from collections of training, examples without having to rely on lower level attributes such as texture, color, or optical flow.

2) Video segmentation: Video segmentation (also referred to as video parsing) is the process of partitioning video sequences into smaller units. Video parsing techniques extract structural information from the video program by detecting temporal boundaries and identifying meaningful segments, usually called shots. The shot ("a continuous action on screen resulting from what appears to be a single run of the camera") is usually the smallest object of interest. Shots are detected automatically and typically represented by key-frames. Video segmentation can occur either at a shot level or at a scene level. The former is more often used and sometimes referred to as shot detection.

Shot detection can be defined as the process of detecting transitions between two consecutive shots, so that a sequence of frames

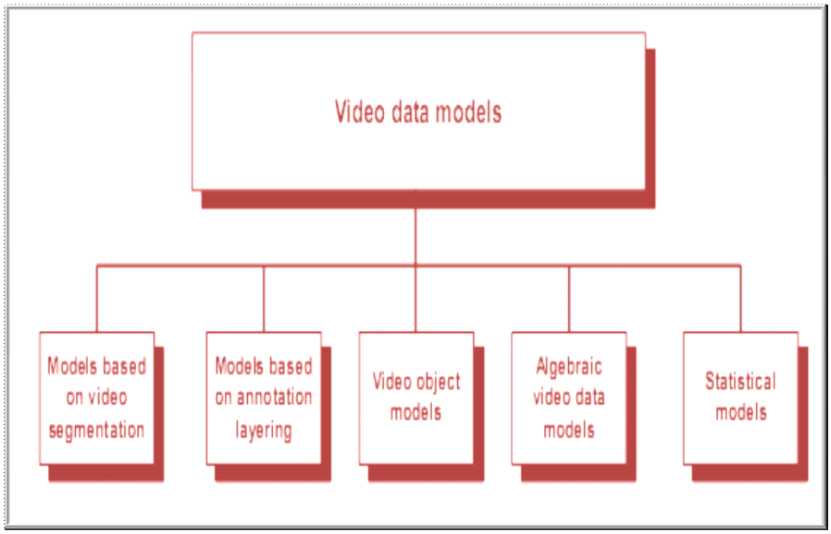

Fig. 2.2 Classification of video database models

belonging to a shot will be grouped together without ambiguity. There are two types of shot transitions: abrupt transitions (or cuts) and gradual transitions (e.g., fade-in, fade-out, dissolve). An alternative to shot detection, scene-based video segmentation consists of the automatic detection of semantic boundaries (as opposed to physical boundaries) within a video program. It is a much more challenging task, whose solution requires a higher level of content analysis.

3) Video Summarization: Video summarization is the process by which a pictorial summary of an underlying video sequence is presented in a more compact form, eliminating – or greatly reducing – redundancy. Video summarization focuses on finding a smaller set of images (key-frames) to represent the visual content, and presenting these key-frames to the user.

A still-image abstract, also known as a

static storyboard, is a collection of salient still images or key-frames generated from the underlying video. Most summarization research involves extracting key-frames and developing a browser-based interface that best represents the original video. The advantages of a still-image representation include:

- Still-image abstracts can be created much faster than moving image abstracts, since no manipulation of the audio or text information is necessary.
- The temporal order of the representative frames can be displayed so that users can grasp concepts more quickly.
- Extracted still images are available for printing, if desired.

An alternative to still-image representation is the use of video skims, which can be defined as short video clips consisting of a collection of image sequences and the corresponding audio, extracted from the original longer video sequence. Video skims represent a temporal multimedia abstraction that is played rather than viewed statically. They are comprised of the most relevant phrases, sentences, and image sequences and their goal is to present the original video sequence in an order of magnitude less time.

There are two basic types of video skimming:

- Summary sequences: Used to provide a user with an impression of the video sequence.
- Highlights: Contain only the most interesting parts of a video sequence.

Video Indexing and Retrieval

Video indexing is far more difficult and complex than its text-based counterpart. While on traditional DBMS, data are usually selected based on one or more unique attributes (key fields), it is neither clear nor easy to determine what to index on a video data. Therefore, unlike textual data, generating video data indexes automatically is much harder and tedious. The process of building indexes for video programs can be divided into three main steps: (1) Parsing: temporal segmentation of the video contents into smaller units. (2) Abstraction: extracting or building a representative subset of video data from the original video. (3) Content analysis: extracting visual features from representative video frames.

Video indexing can be classified into three categories:

A. Annotation-Based Indexing

Annotation is usually a manual process performed by an experienced user, and subject to problems, such as: time, cost, specificity, ambiguity, and bias, among several others. A commonly used technique consists of assigning keyword(s) to video segments (shots). Annotation-based indexing techniques are primarily concerned with the selection of keywords, data structures, and interfaces, to facilitate the user's effort[2,9,10].

But even with additional help, keyword-based annotation is inherently poor, because keywords:

- Do not express spatial and temporal relationships.
- Cannot fully represent semantic information and do not support inheritance, similarity, or inference between descriptors.
- Do not describe relations between descriptions.

B. Feature-Based Indexing

Feature-based indexing techniques have been extensively researched over the past decade. Their goal is to enable fully automated indexing of a video program based on its contents. They usually rely on image processing techniques to extract key visual features (color, texture, object motion, etc.) from the video data and use these features to build indexes [8], [9]. The main open problem with these techniques is the semantic gap between the extracted features and the human interpretation of the visual scene.

C. Domain-Specific Indexing

Techniques that use logical (high-level) video structure models (a priori knowledge) to further process the results of the low-level video feature extraction and analysis [2]. Some of the most prominent examples of using this type of indexing technique have been found in the area of summarization of sports events.

A typical scheme of video-content analysis and indexing, as proposed by many researchers, involves four primary processes: feature extraction, structure analysis, abstraction, and indexing as shown in Fig. 3.1

D. Retrieval By Querying

The video data retrieval process consists of four main steps: (1) User specifies a query using the GUI resources, (2) Query is processed and evaluated, (3) The value or feature obtained is used to match and retrieve the video data stored in the Video Data Base, (4) The resulting video data is displayed on the user's screen for browsing, viewing, and (optionally) query refining (relevance feedback).

Queries to a Video Data Base Management System can be classified in a number of ways, according to their content type, matching type, granularity, behavior, and specification [2,7] as illustrated in Fig. 3.2. The semantic information query is the most difficult type of query, because it requires understanding of the semantic content of the video data. The Meta information query relies on metadata that has been produced as a result of the annotation process, and therefore, is similar to conventional database queries. The audiovisual query is based on the low-level properties of the video program and can be further subdivided into: spatial, temporal, and spatio-temporal. In case of deterministic query, the user has a clear idea of what he expects as a result, whereas in the case of browsing query, the user may be vague about his retrieval needs or unfamiliar with the structures and types of information available in the video database[2].

Past video retrieval system has focus on primitive features. These systems were fully automatic, but the retrieval is not effective and accurate. In order to achieve accuracy and effectiveness[8,10], the research community move from syntactic content based retrieval to semantic content based retrieval. Semantic content extraction is more complex because it does not only based on low level feature (color, texture, shape, object etc) i.e. visual similarity but also required domain knowledge, user interaction and semantic extraction. Retrieval of multimedia on the basis of semantic features is a solution to the drawback of the syntactic content based information retrieval. Semantic features involves different level of semantics in a multimedia data, it permits the queries like find the video the clips that contains car show or find the video clips of cycling race. Translating the user requirements only to the low level feature seen by the computer is the primary cause of the gap. So the solution is to Understand the meaning behind the query[6].

Query processing usually involves four steps: (1) Query parsing: where the query condition or assertion is usually decomposed into the basic unit and then evaluated. (2) Query evaluation: uses pre-extracted (low level) visual features of the video data. (3) Database index search. (4) Returning of results: the video data is retrieved if the assertion or the similarity measurement is satisfied.

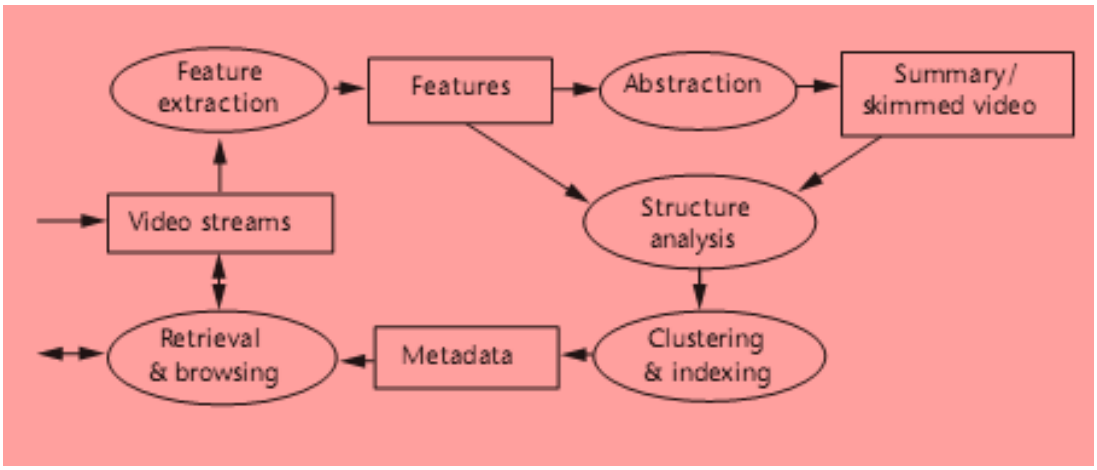

Fig. 3.1. Process diagram for video content analysis and indexing

Fig. 3.2 Classifications of queries to a Video Data Base Management System

Applications

While applications are the main motivation, a lot of research has been prompted by the genuine scientific challenge. This section focuses on applications of video Indexing and Retrieval (VIR).

A. Existing Applications

1) 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 VIR. In general, VIR has been relatively successful in domains where low-level features are directly related with the user queries.

2) 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 VIR researchers and medical doctors join force, this domain is likely to bear early fruits in VIR.

3) Personal: This is a very broad category which can include family albums [1], music recommendation and clothing search. Like.com is one of the first commercial clothing search engines. More recent personal uses of VIR 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 utilize its small screen best. Another related topic is how to provide location-based service to meet user's need based on the context.

4) The Web: This category is the most diverse field. 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 VIR Web search engines are based on surrounding text and meta-data. However, true VIR search engines have begun to emerge (uses both visual and speech features) and Search Video (uses both text and visual features). 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 is the probable key to successful Web-based VIR.

B. Upcoming Applications

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. It is strongly accepted as true 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.

1) 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[1]. 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. Some of the key trends in consumer multimedia content are:

- From structured media to unstructured media: This is the case from news to sports, and is also the case within the sports video genre itself, i.e., from broadcast video to non-broadcast video.
- From single media analysis to multimedia and multiple modality analysis[1].
- From analysis-alone to integrated analysis and synthesis.

2) 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, moving object detection, object tracking, object classification, human behavior analysis, people counting, and abandoned object detection. A few works have also been reported for the surveillance using audio.

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 VIR problem.
- 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.
- 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[4].

C. Professional World

Professional activities that involve generating or using large volumes of video and multimedia data are prime candidates for taking advantage of video-content analysis techniques. Here we discuss several such applications.

1) Automated authoring of Web content: Media organizations and TV broadcasting companies have

shown considerable interest in presenting their information on the Web. A survey conducted in 1998 by the Pew Research Center for the People and the Press indicated that the number of Americans who obtained their news on the Internet was growing at an astonishing rate. This survey indicated that 36 million people got their news online at least once a week. This number had more than tripled in a two-year period (The full survey results are available at http://people-press.org/reports/).

The process of generating Web-accessible content usually involves using one of several existing Web-authoring tools to manually compose documents consisting of text, images, and possibly audio and video clips. This process usually consumes considerable amounts of time. Pictorial Transcripts uses video and text analysis techniques to convert closed-captioned video programs to Hypertext Markup Language (HTML) presentations with still frames containing the visual information accompanied by text derived from the closed captions. A content-based sampling method14 performs the task of reducing the video frames into a small set of images that represent the visual contents of each scene in a compact way. This sampling process is based on detecting cuts and gradual transitions, as well as a quantitative analysis of the camera operations. Linguistic analysis of closed-caption text refines the text, generates textual indices, and creates links to supplementary information.

2) Searching and browsing large video archives: Another professional application of automated media content analysis is in organizing and indexing large volumes of video data to facilitate efficient and effective use of these resources for internal use. Major news agencies and TV broadcasters own large archives of video that have been accumulated over many years. Besides the producers, others outside the organization use the footage from these archives to meet various needs. These large archives usually exist on numerous different storage media, ranging from black-and-white ûlm to magnetic-tape formats.

Traditionally, the indexing information used to organize these large archives has been limited to titles, dates, and human-generated synopses. We generally use this information to select video programs possibly relevant to the application at hand. We ultimately distinguish the material's relevance by viewing the candidate programs linearly or nonlinearly. Converting these large archives into digital form is a ûrst step in facilitating the search process. This in itself is a major improvement over the old methods. These large video libraries create a unique opportunity for using intelligent media analysis techniques to create advanced searching and browsing techniques to ûnd relevant information quickly and inexpensively.

3) Easy access to educational material*:* The availability of large multimedia libraries that we can efficiently search has a strong impact on education. Students and educators can expand their access to educational material. The Telecommunications Act of 1996 has acknowledged the signiûcance of this. It has special provisions for providing Internet access to schools and public libraries. This holds the promise of turning small libraries that contain a small number of books and multimedia sources into ones with immediate access to every book, audio program, video program, and other multimedia educational material. It also gives students access to large data resources without even leaving the classroom.

4) Indexing and archiving multimedia presentations: Intelligently indexing multimedia presentations is another area where content-based analysis can play a major role. Existing video compression and transmission standards have made it possible to transmit presentations to remote sites. We can then store these presentations for on-demand replay. Different media components of the presentation can be processed to characterize and index it. Such processing could include analyzing the speaker's gestures, slide transition detection, extracting textual information by performing optical character recognition (OCR) on the slides, speech recognition, speaker identification, discrimination and audio event detection. The information extracted by this processing generates powerful indexing capabilities that would enable content-based retrieval of different segments of a presentation. Users can search an archive of presentation to locate information about desired topic.

5) Indexing and archiving multimedia collaborative sessions: Multimedia collaborative systems can also beneût from effective multimedia understanding and indexing techniques. Communication networks give people the ability to work together despite geographic distances. The multimedia collaborative sessions involve real-time exchange of visual, textual, and auditory information. The information retained is often limited to the collaboration's end result and doesn't include the steps that were taken or the discussions that took place. We can set up archiving systems to store all the information together with relevant synchronization information. Content-based analysis and indexing of these archives based on multiple information streams enable the retrieval of segments of the collaborative process. Such a process lets users not only access the end result but also the process that led to those results. When the communication links used for the collaborative session are established by a conferencing bridge, we can use the available data in the indexing process, thereby reducing the processing required to identify each stream's source.

Technical Challenges

This section explores on the subject of what is needed in terms of technologies, and here eight challenges have been discussed [1].

A. Bridging the Semantic Gap

What algorithms can automatically 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.

B. How to Best Combine Human and Machine Intelligence?

One advantage of VIR, compared with traditional pattern recognition, is that in most scenarios VIR 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.

C. Active Multimedia Information Retrieval

Active feedback has been proposed for text-based information retrieval [5]. 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 VIR context since a combination of cross-modal information can be utilized to best learn the user intent.

D. Judicious Use of Secondary Sources of Information

The text retrieval problems have immensely benefited from the use of Word internet 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.

E. Centralized Services versus Distributed Services

In a lot of VIR 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 system. In the long run, there will be perhaps a mixed ecology of systems and architectures over which VIR systems will be built. Choosing the right architecture for a particular VIR system is an open problem. However, it is not always strictly a technical issue; economic pricing models often contribute to the prevalence of one service architecture over the other.

F. Handling of Live Multimedia Data

Current VIR 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 [6]. Moreover, information needs arise anywhere due to increased mobility. With mobile phones becoming ubiquitous, universal VIR is a possibility. How does one effectively retrieve multimedia data in such a scenario?. The search paradigm will evolve to 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.

G. 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 VIR?. 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.

H. Expanding the Search Paradigm into Newer Areas

Many new and old problems are being recast as a VIR 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 VIR 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 VIR problems. The search paradigm will be critical in order to handle the data interdependence complexity.

Future Directions

Despite the considerable progress of academic research in multimedia information retrieval, there has been relatively little impact of VIR research into commercial applications with some niche exceptions such as video segmentation. One example of an attempt to merge academic and commercial interests would be Riya (www.riya.com). Their goal is to have a commercial product that uses the academic research in face detection and recognition and allows the users to search through their own photo collection or through the Internet for particular persons. Another example is the Magic Video Browser (www.magicbot.com) which transfers VIR research in video summarization to household desktop computers and has a plug-in architecture intended for easily adding new promising summarization methods as

they appear in the research community. An interesting long-term initiative is the launching of Yahoo! Research Berkeley (research.yahoo.com/Berkeley), a new research partnership between Yahoo! Inc. and UC Berkeley with the declared scope to explore and invent social media and mobile media technology and applications that will enable people to create, describe, find, share, and remix media on the web. Nevenvision (www.nevenvision.com) is developing technology for mobile phones that utilizes visual recognition algorithms for bringing in ambient finding technology. However, these efforts are just in their formative days and there is a need for avoiding a future where the VIR community is isolated from real world interests. We believe that the VIR community has a golden opportunity to the growth of the multimedia search field that is commonly considered to be the next major frontier of search [Battelle 2005].

The potential landscape of multimedia information retrieval is quite wide and diverse.Below is some potential areas for additional VIR research challenges:

A. Human Centered Methods

We should focus as much as possible on the user, who may want to explore intensively on of search media. It has been noted that decision makers need to explore an area to acquire valuable insight, thus experiential systems which stress the exploration aspect are strongly encouraged. Studies on the needs of the user are also highly encouraged toward giving us understanding of their patterns and desires. New interactive devices have largely been overlooked and should be tested to provide new possibilities, such as human emotional state detection and tracking.

B. Multimedia Collaboration

Discovering more effective means of human- computer-mediated interaction is increasingly important as our world becomes more wired or wirelessly connected. In a multimodal collaboration environment many questions remain: How do people find one another? How does an individual discover meetings/collaborations? What are the most effective multimedia interfaces in these environments for different purposes, individuals, and groups? Multimodal processing has many potential roles ranging from transcribing and summarizing meetings to correlating voices, names, and faces, to tracking individual (or group) attention and intention across media. Careful and clever instrumentation and evaluation of collaboration environments will be a key to learning more about just how people collaborate.

Very important here is the query model which should benefit from the collaboration environment. One solution would be to use an event-based query approach that can provide the users a more feasible way to access the related media content with the domain knowledge provided by the environment model. This approach would be extremely important when dealing with live multimedia where the multimedia information is captured in a real-life setting by different sensors and streamed to a central processor.

C. Interactive Search and Agent Interfaces

Emergent semantics and its special case of relevance feedback methods are quite popular because they potentially allow the system to learn the goals of the user in an interactive way. Another perspective is that relevance feedback is serving as a special type of smart agent interface. Agents are present in [3] learning environments, games, and customer service applications. They can mitigate complex tasks, bring expertise to the user, and provide more natural interaction. For example, they might be able to adapt sessions to a user, deal with dialog interruptions or follow-up questions, and help to manage focus of attention. Agents raise important technical and social questions but equally provide opportunities for research in representing, reasoning about, realizing agent belief and attitudes (including emotions). Creating natural behaviors and supporting speaking and gesturing agent displays are important user interface requirements. Research issues include what the agents can and should do, how and when they should do it (e.g., implicit versus explicit tasking, activity, and reporting), and by what means should they carry out communications (e.g., text, audio, video). Other important issues include how do we instruct agents to change their future behavior and who is responsible when things go wrong.

D. Neuroscience and New Learning Models

Observations of child learning and

neuroscience suggest that exploiting information from multiple modalities (i.e., audio, imagery, haptic) reduces processing complexity. For example, researchers have begun to explore early word acquisition from natural acoustic descriptions and visual images (e.g., shape, color) of everyday objects in which mutual information appears to enabling vast reduction in computational complexity. This work, which exploits results from speech processing, computer vision, and machine learning, is being validated by observing mothers in play with their pre-linguistic infants performing the same task. Neuroscientists and cognitive psychologists are only beginning to discover and, in some cases, validate abstract functional architectures of the human mind. However, even the relatively abstract models available from today's measurement techniques (e.g., low fidelity measures of gross neuroanatomy via indirect measurement of neural activity such as cortical blood flow) promise to provide us with new insight and inspire innovative processing architectures and machine learning strategies.

E. No Solved Problems

From the most recent panel discussions at the major VIR scientific conferences including ACM VIR and CIVR, it is generally agreed that there are no "solved" problems. In some cases a general problem is reduced to a smaller niche problem where high accuracy and precision can be quantitatively demonstrated, but the general problem remains largely unsolved. In summary, all of the general problems need significant further research.

CONCLUSSION

We have first offered a tiny assessment on the initial works and existing applications in VIR. Significant work had been done from the past but yet a number of problems still want the research interest. The major meet up head-on is the semantic gap; key in trouble is how to calculate semantic features from primal features. This study showcases that the majority of the problems in existing approaches are due to the lack of semantic extraction and user behavior concern. There is a need for a system that can understand the user query according to their necessities and make searching and retrieval efficient and ideal. Current systems are trying to travel around the requisite of the rising industries and trends , on the other hand certain capabilities are still not up to the mark, thus this domain of research can bring about exemplary revolution in the near future.

REFERENCES

1. Mohan S. Kankanhalli and Yong Rui, Senior Member, IEEE "Application Potential of Multimedia Information Retrieval" 0092-SIP-2007-PIEEE.R1.
2. Oge Marques and Borko Furht, Department of Computer Science and Engineering ,Florida Atlantic University ,Boca Raton, FL, USA "INTRODUCTION TO VIDEO DATABASES".
3. MICHAEL S. LEW Leiden University, The Netherlands and NICU SEBE University of Amsterdam, The Netherlands and CHABANE DJERABA LIFL, France and RAMESH JAIN University of California at Irvine, USA "Content-based Multimedia Information Retrieval: State of the Art and Challenges".
4. Nevenka Dimitrova, Philips Research, Hong-Jiang Zhang, Microsoft Research, Behzad Shahraray, AT&T Labs Research Ibrahim Sezan Sharp Laboratories of America, Thomas Huang University of Illinois at Urbana–Champaign Avideh Zakhor, University of California at Berkeley "Applications of Video-Content Analysis and Retrieval" Feature Article 1070-986X/02/ © 2002 IEEE.
5. Shih-Fu Chang, Qian Huang, Thomas Huang, Atul Puri, and Behzad Shahraray "Multimedia Search and Retrieval" Published as a chapter in Advances in Multimedia: Systems, Standards, and Networks, A. Puri

and T. Chen (eds.). New York: Marcel Dekker, 1999.

6. Nida Aslam, Irfanullah, Kok-Keong Loo, and Roohullah "Limitation and Challenges: Image/Video Search & Retrieval" International Journal of Digital Content Technology and its Applications Volume 3, Number 1, March 2009.
7. Sharon McDonald and John Tait, "Search Strategies in Content-Based Image Retrieval", 2003 ACM 1-58113-646-3/03.
8. Chirag Shah W. Bruce Croft ,"Evaluating High Accuracy Retrieval Techniques" *SIGIR'04,* July t 2004 ,ACM 1-58113-881
9. James Z. Wang1 (Chair), Nozha Boujemaa2, Alberto Del Bimbo3, Donald Geman4, "Panel Diversity in Multimedia Information Retrieval Research", MIR'06,2006, ACM 1-59593-495-2
10. Bill Kules, Jack Kustanowitz and Ben Shneiderman,"Categorizing Web Search Results into Meaningful and Stable Categories Using Fast-Feature Techniques" 2006 ACM 1-59593-354-9/06.