Visual Information Retrieval: The Next Frontier in Search

Ramesh Jain

Abstract: *The first ten years of search techniques for WWW have been concerned with text documents. The nature of data on WWW and in computing is rapidly changing and is fast moving towards multimedia, including text. We believe that the next major frontier in search is the visual search based on visual information retrieval. This search will in fact take us to multimedia search. In this paper, we present some ideas related to the history of visual search and how it is evolving. We also present our perspective in this area.*

In the last ten years Search Systems have become one of the most important components of computing. The evolution and success of WWW is closely tied to the developments in search technology. Much of the content of WWW in its first decade of existence was in text oriented documents. To cope up with the increasing volume of information powerful search techniques evolved. These techniques were based on concepts and techniques in information retrieval. Interestingly the techniques that were developed to deal with contents in libraries, were refined, modified, and adopted for dealing with WWW.

The Web continues to evolve. Early Web was a web of text documents only. Tremendous progress made in the last decade in computing, storage, sensing, and communication technology has changed the landscape of data, computing, communication, and the Web. Images, Audio files (MP3), video, and sensor networks are rapidly becoming major part of the Web. This is transforming the nature and applications of the Web. Original search techniques developed to deal with primarily textual data need to evolve to cope with the transformation of the Web. In this paper we discuss the next frontier in search – the visual information retrieval or visual search.

A simple analysis of the nature of the data and expectations of users from current and next generation WWW shows some very important attributes of emerging issues and applications:

- Volume of data is growing exponentially every year.
- Multimedia and sensor data is becoming more and more common.
- Different data sources provide have to be combined to form a holistic picture.
- The spatiotemporal characteristics of the data must be taken into account
- Real-time data processing is becoming common
- Exploration, not querying, is the predominant mode of interaction, which makes context and state critical.
- The user is interested in experience and information, independent of the medium and the source.

In the following, we present our ideas to address some of the challenges being posed by the above changes and resulting requirements on the systems.

Brief History of Visual Search

In this section, we briefly review the state of art in research in image and video retrieval. Our goal is to present only issues that are relevant to the emerging approaches that will be useful in the next generation. We are presenting a perspective rather than exhaustive summary of approaches.

Images Search

Some early systems in visual information retrieval developed and utilized data models that looked at images and video at data and domain levels. VIMSYS model introduced in [1] considered data at 4 different levels: image representations, image objects, domain objects, and domain events. This exhaustive model has never been fully developed and implemented. We believe that this model is still relevant and could play a very important role in developing visual information retrieval systems. The face retrieval system presented in [2] was definitely influenced by VIMSYS model and addressed visual information retrieval problem that considered that the semantics in images is not only in the data, but also provided by the user. This notion of semantics emerging as an interaction of the user with the data was later more formally refined and developed as emergent semantics approach in [3].

Query-by-image-content [4] became a popular mechanism to search images in early days of image retrieval. In this approach, color histograms, texture features, and some structural features were used to characterize each image and then the distance to the example query image was calculated to rank all other images against it. To improve the results of these operations, relevance feedback was proposed as an alternative. These approaches received lots of attention from researchers and many variants of approaches based on color, texture, and shape were proposed to retrieve images based on pictorial content of an example image. Even now many research approaches are exploring refinement of approaches based on this concept.

Video Search

Much of the research in video search has really been addressing development of concepts and techniques to segment and represent video in segments that represent the structure of video. A major motivation behind this is to represent video similar to text in a hierarchical structure. Thus a video can be partitioned into stories or episodes; each story could be partitioned into scenes, scenes are composed of shots, and shots contain frames. Frames are considered an atomic unit of a video and using pictorial, audio, and semantic attributes one could group frames into shots; shots into scenes; scenes into episodes and finally episodes into a video.

Techniques for shot detection have matured reasonably well. Other techniques are being explored. News video has received significant attention from researcher. Soccer video has also been popular.

A serious self-impose limitation by researchers in this field has been use of only image data. Most video processing approaches did not even use audio in their analysis. This is very surprising because use of audio and other information sources simplifies the task of partitioning video into its constituents becomes easier.

Closed captions have been used to supplement information from images to detect shots. Some research has been done in detecting themes and stories based on closed captions also. In a few commercial systems, speech recognition technology has been used to convert speech in a video into text transcripts that could then be used by traditional information retrieval techniques. Two companies have shown success of this approach for News application. These companies are www.blinkx.tv and www.streamsage.com.

The Next Frontier

When considering important changes in data sources and the operations required, one very interesting fact emerges. Let us consider data sources to be broadly of two types. In many situations, we know the data source precisely, though in cyberspace it may be distributed. In other case we only know that what we need may be available somewhere. Our needs may also vary. In some cases, we need precise information and want that answer. In other situations, we are trying to gain some insights in the behavior of a system, event, or concept and we want to explore and understand what is going on. In Table 1, we show this situation. The first column is for known sources, and the second is for unknown sources. The bottom row is for getting precise information, while the top row is for gaining insights. As we all know, for getting precise information from a precise source, current databases are excellent. For gaining insights based on a precise source, visualization environments and tools are emerging. For imprecise sources, like the web, lot of research is being done in the area of search engines. Most of the applications that we discussed in the above section, however, fall in the top quadrant where one is trying to gain insight from imprecise sources. This quadrant leads to what we call experiential environment that will be increasingly common in most data intensive applications.

Insight	Visualization	Experiential Environments
Information	Current Databases	Search Engines
	Precise Source	Imprecise Source

Table 1: Data sources and access goals

Experiential Environment

Current database and search environments are information-centric in which it is expected that the user knows what information he needs and will articulate his needs in terms of a query. To provide efficiency and scalability, database systems had to be stateless. A stateless system can interact with multiple users in multiple contexts. In early days of computing, people considered computer's time lot more valuable and wanted to adapt their behavior to meet requirements of computing environment. Things have changed significantly. Now people want systems that can be personalized and can provide information in a specific context that they are exploring. The system should remember how a user got to a particular state and should answer questions in that context. The responsibility of articulating a precise query is being transferred from users to systems.

A very important but often ignored fact in design of computing environment is that humans are very efficient in conceptual and perceptual analysis and relatively weak in mathematical and logical analysis; computers are exactly opposite. Computers can perform mathematical and logical operations millions of times faster than any human can. On the other hand, the perceptual capabilities of a computer, even after all the progress made in this area in the last 40 years, remain very primitive. Presenting sequential and logical information to humans does not allow them to utilize their strength. Similarly, expecting computers to detect complex patterns in data does not utilize their strength. By using computers and users synergistically as a system, a very different type of environment, called experiential environment, can be developed. **Experiential environments allow a user to directly utilize his senses to observe data and information of interest related to an event and to interact with the data based on his interests in the context of that event.**

This is the next natural step and takes computing and communications to the next level of applicability and usability. In future one is likely to see an increasing trend in this direction. Considering the applications that we discussed above, we identified some important aspects of experiential environments in [5], which are: natural application environments, same query and presentation spaces, maintaining and using user state and context, and multimedia immersion.

Assimilating Data into Unified Information and Knowledge

Data is collected in many forms using different sensors, including humans. Data could be audio, video, text, alphanumeric, infrared, and many other forms depending on the sensor employed. In the next generation search and retrieval systems, it will be important to consider all data sources in a unified format. While searching for visual information, it is important to combine visual information with other sources and present to a user a unified picture.

Current indexing techniques for different data types depend upon metadata for that type. Metadata plays a key role in introducing semantics in the data and is important in utilization of the data. Relational approaches were developed to introduce semantics in databases. Database schemas provide semantics in relational tables. Recently XML has become very popular for introducing semantics in text. I find it amusing that initially researchers worked hard to develop automatic approached to deduce semantics from the data and many researchers still pursue that goal. That is a worthwhile goal, but it has become clear that it is significantly more complex than initially thought to be. And this led to developing a mark-up approach to semantics. Languages were also designed using this approach XML allows introduction of semantics in strongly human mediated environments. For sensory data like audio and video, people are developing feature based techniques. These techniques are in their early stage. Not much attention was given to indexing sensory data and even now, not enough attention is being given to sensory data.

For images commonly used features are color histograms and simple measures of texture and structure. Interestingly, most measures used are for global images not for objects in images. In most applications, people are interested in objects. It is a problem to get to those objects, and remain an ignored problem. It is true for most other signals also. Signals are usually indexed using features that capture global or semi-global features.

Data Silos

We all get information about objects and events from different sources in different data types. So what I know about the war against the terrorism is based on what I saw on TV, what I read in newspapers and magazines, what I heard on radio and discussed with my friends. My perception of this situation is based on assimilation of information assimilated from all these data sources. In our head, somehow we assimilate all the information and represent this information in a form that is independent of the media.

In information systems we create data silos. The metadata is defined and introduced for a data of a particular type. And all these metadata collections, or indices, form silos. It is not possible for a video collection to interact with text collection and prepare a unified collection. Currently all these silos are very strong and have negligible, if any, interactions among them. The situation is shown in Fig 1 below.

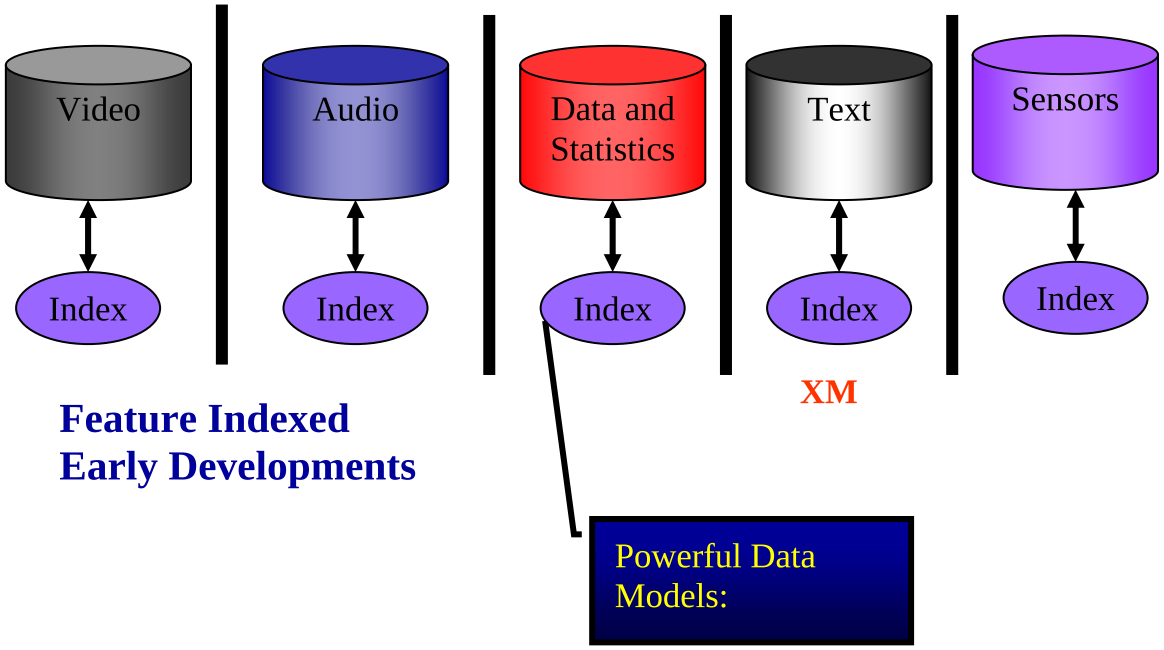

Figure 1: Different data sources have different indexing mechanisms, but these sources live in their own silos.

Information Assimilation

As discussed above, in most applications, information about an object or event may be drawn from different data sources that may be of different data types. This data and information from different sources must be combined to provide complete information. A serious problem in existing information systems is that most of the data sources are designed to behave like independent silos. It is assumed that after they are analyzed and their metadata has been extracted, somehow the metadata can be combined to provide correct results. Many research efforts are underway to develop approaches for information integration by looking at the ways to use metadata from each silo.

A commonly used approach in engineering systems is to use a strong domain based estimation of parameters of the system based on many disjoint and disparate sources of information. In all these approaches, a mathematical model of the system is successively developed by observing data from disparate sources. Each data source is just a data source and contributes to the refinement of the model. The goal is to get as precise a model as possible. In this process, it is possible to completely ignore data from a specific source at some stage. Thus, a data source is just that a data source and model represents the current knowledge of the system based on evidence from all the data sources.

Conceptually this stage is very different from the current information integration systems where a particular data source is analyzed and then its results are combined with other data sources. In computer science also, a concept similar to the above is used for translation of different languages into a neutral representation. In these cases, however, the representation is suppose to facilitate the translation process; in knowledge system, the model represents the goal. A very important result of the assimilation approach is that the system can efficiently deal with real time data by keeping only what is important for the goal of the system. This approach also allows a very smooth and effective introduction of semantics in the process.

Event Graphs for Unified Indexing

We have developed an information assimilation approach to build a unified indexing system that introduces a layer on top of metadata layer of disparate data sources. This layer uses event-based domain model and the metadata to construct a new index that is independent of the data types in different data sources. Based on analysis of many applications and theories in human memory organization we decided to use events as the basic organization entity for unified indexing. An event is defined as a significant occurrence or happening located at a single point in space-time. An application domain can be modeled in terms of events and objects. Events are hierarchical and have all desirable characteristics that have made objects so popular in software development. In fact, events could be considered objects that have time and space as primary attributes. Details about events modeling and powerful features of events are beyond the scope of this paper and will be discussed elsewhere. This approach is shown in Fig. 2.

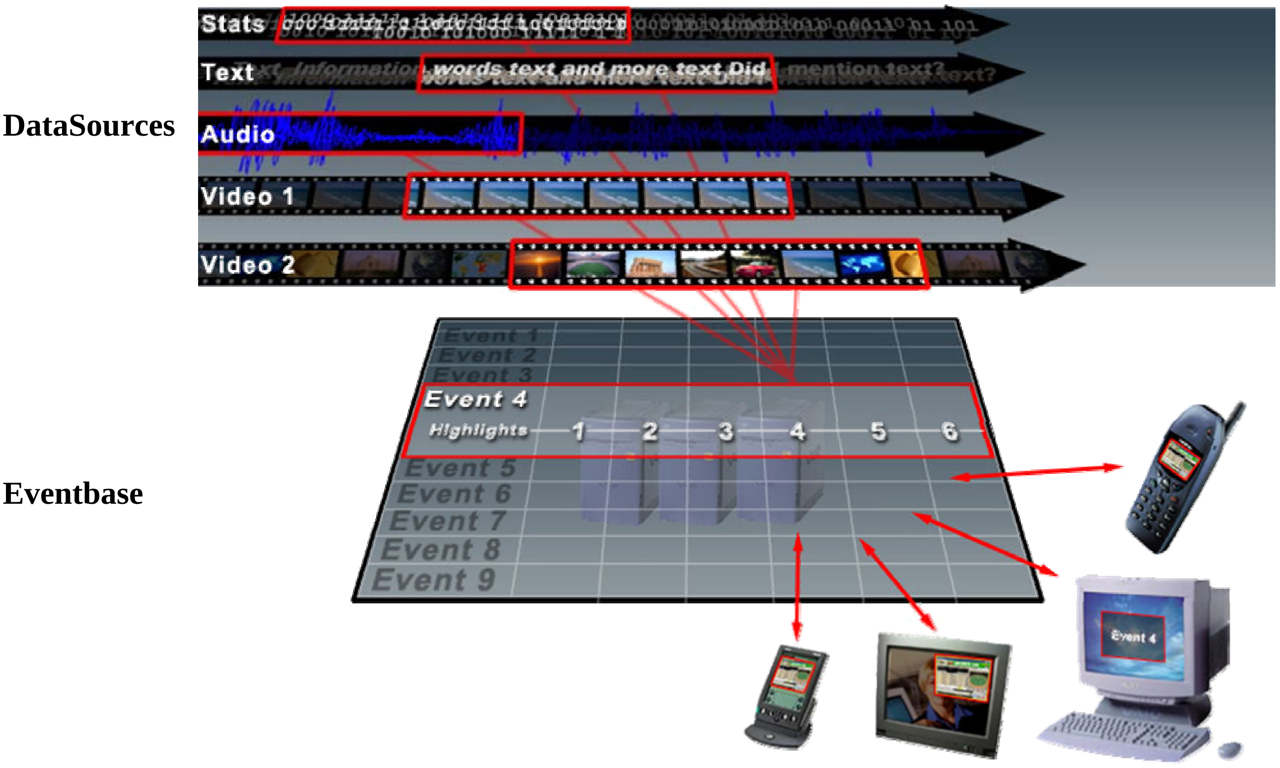

Figure 2: Event Graphs are used to unify different data sources by providing a semantic indexing and linking approach.

The unifying index is organized using event model for the application domain and is stored in a database, called Event Base. Event base contains all information about events that is assimilated from different sources and also links to original data sources. The links to data sources are particularly important to present appropriate media in the context of events. A user directly interacts with the event base; the event base uses original data sources as and when required. This has several advantages including pre-processing important information related to events and objects based on domain knowledge; presenting information using domain based visualization, and providing a unified access to all information related to an event independent of the time the data became available or was entered in an associated database. As we see in the following, the above help in providing an experiential environment for a user to access the information.

Event Graphs are first used to parse the data as it is coming and for assimilating data to build an environment model reflecting knowledge based on the information that has been collected so far. Thus, Event graphs are used to parse data and create a list of spatio-temporal events as they take place. This becomes the database that describes domain semantics using individual data streams as simple data streams only to visit and study details as may be required. Of course one may visit this stream with the idea to re-experience the event.

Event graphs capture the entities and their roles in the event, its location, and time. It also captures event transition information. Causality is captured in event transition mechanism.

Eventbase

Event graph partitions the data and generates events. These events are stored in a database that stores the name of the event, nature of the event, and all other relevant information. The relevant information may not be available at the time of the event creation; it may become available later and should be appropriately attached to the event.

Eventbase is an organic database that keeps growing as a result of many different processes running and is different from the database in this respect.

Next Generation Search

The nature of search is following both evolutionary as will as disruptive paths. It is natural that in the next few years many of the current search paradigms will be refined to produce better search experience. I believe, however, that in a few years there will be revolutionary changes in this area due to changing nature of data and expectations of new applications. The sequence of evolution and its timing are subject to so many variables that it is almost impossible to predict. The changes that I mention below are in no particular order of importance and time. I do believe, however, that all these technological challenges are being addressed at different places in different forms and will come together in not so distant future.

- Current search techniques will be modified by post-processing the results. Several sites are already using clustering and grouping into folders. These are the first steps. There will be significant amount of activity in this area.
- Metadata is already starting to get used by search engines. This will increase. Text mining and ontological techniques will start embedding meta-data in text documents and use this in search. This will improve the quality of results significantly.
- Powerful visualization techniques for presenting the results to users will emerge. One already sees different approaches tried by some sites. None of them has been compelling so far. This direction has strong potential. A good indication of the success of these techniques is how OLAP and visualization environments evolved for large databases and data warehouses. Holistic visualization of search results will be common.

- Multimedia search will slowly become really multimedia. Initially it will still be extension of current approaches. Research in this area will mature to start defining metadata specific to different media and semi-automatic and automatic techniques to extract this metadata. A major change will be that unlike current search engines that deal with images and other media by assigning textual tags, future search engines will consider text as one of the media and deal with using more powerful and generalized techniques.
- Current search environments evolved from traditional information-centric systems. Future systems will rely on emergent approaches implemented in experiential environments. Thus, sentient exploration of data will be common. Current query and keyword oriented environments will be subsumed by exploratory environments.
- Novel techniques will evolve to detect events in live data and automatically communicate these events to servers that will build powerful event bases. Search and exploration techniques will be developed to provide users immediate access to all information and data related to events of interest with minimal delay.

References

[1] A. Gupta, T. Weymouth, and R. Jain, "Semantic Queries with Pictures, The VIMSYS Model," *Proceedings of VLDB'91, 17th International Conference* on *Very Large Data Bases*, Barcelona, Spain. Sept. 3-6, 1991

[2] J. Bach, S. Paul, and R. Jain, "An Interactive Image Management System for Face Information Retrieval," IEEE Transactions on Knowledge and Data Engineering, Special Section on Multimedia Information Systems. Publication. 1993.

[3] Simone Santini, Amarnath Gupta, and Ramesh Jain "Emergent semantics through interaction in Image Databases" IEEE Transactions on Knowledge and Data Engineering, summer 2001.

[4] Y. Rui, T.S. Huang, S.F. Change, Image retrieval: current techniques, promising directions, and open issues, J. Visual Commun. Image Representation 10 (1) (1999) 39–62.

[5] Ramesh Jain, "Experiential Computing", in Comm. Of ACM, July 2003.

ABOUT THE AUTHOR

Ramesh Jain, Donald Bren Professor of Information & Computer Sciences, Department of Computer Science, Donald Bren School of Computer Sciences,University of California,Irvine CA 92687, jain@ics.uci.edu