# Content-Centric Computing in Visual Systems

Ramesh Jain

Center for Information Engineering
Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407

## 1 Introduction

The nature of computing has changed over the last four decades in almost all aspects. The progress in processing power and storage capacity is well known. Moore's law is now in the dictionary of a common man and internet years have replaced dog-years to represent compression of time. Computer was an esoteric device in even sixties and seventies; now it is an indispensable component of many household appliances.

Some aspects of computing have changed continuously, but relatively more silently compared to the impact of processing power and storage capacity. Computing has changed from alphanumeric computing to multimedia computing. It is now as common to find images on computer as text; and the amount of information in the form of images and video is increasing at a much faster rate than the textual information. This is resulting in many novel and interesting applications of computing. Another major change is that at one time computers were mainly used to compute and the data storage was the secondary function. Now data is becoming primary so much so that much of the computing is simply to find best ways to store and communicate the data to make it available in a right form, at a right place, at right time to information users. Finally, the early computers were self-sufficient; they did everything by themselves. Now we talk about network being the computer. One of the important debates among top computer entreprenuers is currently whether Network Computer (NC) is what people need or is it Networked PC. Important thing is that it is assumed that networking is essential; the issue is how to access it.

Obviously all these changes result in many interesting problems and opportunities. Computer vision and image processing community has a chance to influence the direction of computing significantly. In this paper we discuss two emerging concepts: Content-Centric Computing and Gestalt Vision. Content-centric computing is being developed to deal with the challenge of dealing with enormopus amount of multimedia data. Gestalt vision has been the dream that can now be realised using emerging computing and sensor technology.

## 2 Content Centric Computing

Traditional computing considers data and programs seperately. Though both are in computer memory, their role is very different. Programs work on data to produce desired operations. A program is basically a sequence of operations that specify what should be done with the data. In early days of computing, usually the operations were very compute-intensive. Many operations were performed on a given set of data. Even now, most of scientific computing involves performing lots of operation on given data. Databases bring different dimension to computing. Databases store a large volume of data. Most applications of databases are data intensive, rather than compute intensive. In databases, a large volume of data is structured in such a way that users can access this data efficiently. The basic function of a database is providing easy access to a large volume of data. The amount of computing done in a traditional database, let us say a payroll system, is usually insignificant compared to that in an engineering system to compute trajectory of a particle. In early applications of databases, this dichotomy of data-intensive and processing-intensive operations was very clear.

Databases evolved to provide access to data to a large number of users. The data must be organized to facilitate access by several users. A database system can be viewed at three differet levels: Physical Level, Conceptual Level, and User Level. At physical level, one worries about the issues related to storage of data on specific devices in specific form. Conceptual level deals with representation of entities and relationships among them. The user level provides different views and access to different subsets of information in the database to different users. Database designers try to keep these three levels independent of each other so that changes at one level do not necessarily result in significant changes at any other level.

Information retrieval systems have evolved and become another essential part of computing infrastructure in the last few years. These systems provide access to textual information either using keywords or using full text retrieval. These systems are the beginning of content-based retrieval in relatively unstructured information environment.

With advances in computing, particularly in storage technology, many applications emerged that require a large volume of data and lots of processing on this data. Engineering analysis systems and weather forecasting systems are a good example of this technology. These systems were used by a very few people, however. Also in most early applications, there used to be very little, if any, interactivity in these systems.

As the technology progressed, Computer Aided Design (CAD) systems started appearing. CAD systems provide interactivity to designers. These systems also provide access to objects at different levels of abstraction. In many engineering design systems, there is a large volume of data in these systems and hence these systems use increasingly sophisticated databases to manage access to large volume of data. In most applications, however, the database is private. The database is used only by one designer or a group of designers. These designers are usually very familiar with the system and hence can work with the system at a very sophisticated level.

In systems with a large volume of data, content-based interactivity is not only desirable, it is essential. As the amount of information grows, the human ability to remember correct information sources becomes overloaded and begins to fail. The success of databases is largely attributable to their ability to allow access to their content, based on the queries related to specific aspects of that content. On the World Wide Web also, search engines have played a major role in granting easier access to textual information. Currently, commercial tools to provide content-based access to visual information are in their early infancy.

A video or a television event can be considered a vast stream of data representing intensity value at a point in an image. This intensity value represents some physical attributes in space for the scene captured by a camera. Viewers are interested in objects, their characteristics, relationships and temporal history. A video is interesting because it provides that information.

We can also view a physical event as the evolution of spatio-temporal characteristics at a certain location. Now, as the amount of data increases, human ability to specify the location decreases. Thus, a system that will provide facilities to specify objects and events and will return or retrieve data corresponding to those will be much more interesting and useful to humans.

In this paper, we discuss research in content-centric -computing at the Center for Information Engineering of University of California, San Diego. Our goal is provide an overview of research. Readers interested in the details will find them in the papers pointed out in specific research.

## 3 Image databases

Most research in image databases in our group is related to providing content-based addressability of large collection of images and video. Many new problems must be solved either by extending traditional approaches or developing some novel approaches. In this section we briefly discuss these research issues.

### 3.1 Data Model

A key problem that must be solved is how to represent image and related data in the system to allow an environment in which users from different backgrounds can retrieve information without much training. The database must strore image and video data and **features and metadata**.

Information in an image exists at several abstraction level and should be accessible at these levels. The datamodel used to store this information must allow existence of information at these multiple levels. Several data models have been proposed. Here we discuss one model that allows explicit representation of abstract levels in images. The VIMSYS data model uses a hierarchical representation of the data using various levels of semantic interpretation. At the image representation (IR) level, the actual image data is stored. Image objects (such as lines and regions) are extracted from the image and stored in the image object (IO) layer, with no domain interpretation. Each of these objects may be associated with a domain object in the DO layer. The semantic interpretation is incorporated in these objects. The domain event (DE) layer can then associate objects of the DO layer with each other, providing the semantic representation of spatial or temporal relationships. This hierarchy provides a mechanism for translating high-level semantic concepts into content-based queries using the corresponding image data. This allows queries based on object similarity to be generated, without requiring the user to specify the low-level image structure and attributes of the objects. Another very important aspect of this representation is that the first two levels, IR and IO, are domain-independent levels and the other two, DO and DS, are domain-dependent levels. We do not know any system yet where this goal of clearly organizing domain-dependent and domain-independent components can be cleanly partitioned and implemented. We believe, however, that is a worthwhile target. Most research in our group follows implicitly the VIMSYS model in our implementations.

## 3.2 Types of Features

Features must be extracted from input images and stored in the database. As is well known, different applications may require different features. Since the features must be stored at the time of data entry, one must carefully decide the features that will be used in a system. We consider that all features must be classified in one of the following classes:

**$F_u$:** This set contains the features which are commonly referred to as meta-features. Some of these features can be automatically acquired from the associated information on images. These features may include the size of the image, photographer, date taken, resolution and similar other information. This group also contains other features that can be called user- specified. Values are assigned to these features by the user at the time of insertion. Many of these features can be read by the system either from the header, filename, or other similar sources. These features can not be directly extracted from images.

**$F_d$:** This set contains the features which are derived directly from the image data at the time of insertion of the images in the database. Values are automatically calculated for these features using automatic or semiautomatic functions. These features are called derived features and include features that are commonly required in answering queries. These features are stored in the database.

**$F_c$:** This set contains the features whose values are not calculated until they are needed. Routines must be provided to calculate these values when they become necessary. These features may be computed from data at the query time. These features are called query-only features or computed features.

The first two types of features are actually stored in the database. Metadata can be frequently read from other sources or should be manually entered. Which feature should be in $F_d$ and which should be in $F_c$ is an engineering decision. One must study frequently asked queries and determine frequently required features. This determines the set to which a particular feature should belong.

The system interface encourages users to formulate his queries using metadata and derived features as much as possible. It reluctantly allows use of computed features. To access data, the system can purge the search space significantly using metadata and derived features and then apply computed features to only this reduced set of images. This strategy allows flexibility while maintaining a reasonable response time. The system may be able to predict wait time using number of images from which computed features must be extracted.

### 3.3 Indexing

Image retrieval is accomplished using many features. Indexing techniques for spatial data have been developed by many researchers. These techniques are very limited when it comes to addressing the problem of similarity indexing. Techniques like TV-trees are a good step in the right direction but lack several important features. Performance of the most indexing techniques degrades significantly with increase in dimension. Another complicating problem is that in image databases, it is usually desired to develop indexing techniques that will allow ranking of data, rather than filtering used in conventional databases. Image databases use a similarity function to rank all images with respect to a prototype. Most similarity functions use weights that may be adjusted at run time. These requirements suggest that a fresh look at the indexing approaches is required.

In our work on indexing, we articulated this problem as ``similarity indexing'' and developed the fundamental types of ``similarity queries'' that we believe should be supported. We also proposed a new dynamic structure for similarity indexing called the similarity search tree or SS-tree. In nearly every test we performed on high dimensional data, we found that this structure performed better than the R*-tree. Our tests also showed that the SS-tree is much better suited for approximate queries than the R*-tree.

One of the major difficulties in solving the indexing problem in image databases is the high dimension (6-100) of the feature vectors that are used to represent objects. We studied different indexing structures by applying them to a set of high-dimensional data and later developed a variant of the optimized k-d tree, that we call the VAM k-d tree, and an optimized R-tree we call the VAMSplit R-tree. We found that the VAMSplit R-tree provides better overall performance than all competing structures we tested for main memory and secondary memory applications. We observed large improvements in performance relative to the R*-tree and SS-tree in secondary memory applications, and modest improvements relative to optimized k-d tree variants. Extensive empirical tests on synthetic and real datasets show that our optimized structures, the VAMSplit k-d tree and VAMSplit R-tree, are superior to the R*-tree and SS-tree in terms of query performance, time to create an index, and space utilization.

## 3.4 Interfaces

Informations systems are used by users with disparate backgrounds. The interfaces to these systems should be such that any novice can use very intuitive methods. The operations used in these interactions must require almost no knowledge of the organization of the data and information. Many of these operations can not be conveniently performed using traditional interfaces. Here we discuss some general issues in designing interfaces for image databases. Due to the nature of the data and several abstraction levels, it is expected that users will require multimodal interface mechanisms.

**General Search:** In general, there will be two modes of navigation: locating and browsing. In the location mode, a user knows what he or she wants and his queries will be to get precisely that information. In the location mode, many queries may be symbolic because what is required can be articulated using meta data. Some location queries may require visual data. It is expected that search queries will deal mostly with meta data. For these queries some query language, possibly a variant of SQL, may be used.

**Query by Pictorial Example (QPE):** A very powerful expression of a query is to point to a picture and expect that the system will show all pictures similar to the example. This approach is easy to use, but very complex to implement. The system must use certain features and some similarity measures to evaluate other pictures that are similar to the example. Effectively, the system must rank all data with respect to the example and then display pictures that are closest to the example. Interestingly, this approach has been a very popular approach in the image databases that are being designed.

In QPE, features and similarity measures must be clearly defined for use in retrieving images. Similarity judgement has been a difficult problem and continues to attract attention of several researchers. The most interesting fact about similarity measures is that they are domain dependent and very subjective. Assuming that we have identified a measure that is acceptable to a user for his or her domain, we face some interesting problems in QPE. All images are compared to the example to evaluate their similarity. This is possible in those cases where the size of the database is such that computations can be done in reasonable time. When the size of the database grows such that it is not possible to accomodate all data in main memory and such computations become impractical, one must resort to indexing techniques.

**Query Canvas:** Queries may be formulated by starting with an existing picture, scanning a new picture and modifying these by using visual and graphical tools available in common picture editing programs, such as Adobe Photoshop. One may cut-and-paste from several images to articulate a query in the form of an image. It is also possible to start from a clean image and then draw an image using different tools. The basic idea in this approach is to provide a tool to define a picture that may be used in a QPE. This approach allows a user to define a picture that they are looking for.

**Containment Queries:** In many cases, a user may point to an object, or circle an area in an image and request all images that contain similar regions. These queries appear very easy, and will be very easy, if complete segmentation of images is performed and then all region properties are stored. Most image database systems store only global characteristics of image. In these cases, one is looking for all images that are a superset of the region attributes. Once all such images are retrieved, some other filtering techniques could be developed to solve this problem.

Most current image retrieval systems use holistic comparisons that require a global match between images or presegmented object in images. However, often the user of an image database system is interested in a local match between images. For example, ``Find images from the database with something like this anywhere in the image," or ``Find images with something like this in some region of any image in the database," or ``Find images with this spatial configuration of regions like this." White developed a new framework that should help to allow these types of queries to be answered efficiently. In order to illustrate the usefulness of this framework, called ImageGREP, a complete image retrieval system based on local color information was developed. Our system features fully automatic insertion and very efficient query execution, rivaling the efficiency of systems that can only handle global image comparisons. The query execution engine, called the ImageGREP Engine, can process queries at a speed of approximately 3000 images per second (or better) on a standard workstation when the index can be stored in main memory. In the future, we believe our framework should be used in other domains and applications, to handle queries based on texture or other material properties and perhaps domain specific image properties.

**Semantic Queries:** All the above queries are based on image attributes. In most applications, an image database is likely to be prepared for a specific domain-dependent application, such as human faces, icefloe images, or retinal images. It is important that users can then interact using domain-dependent terms. It is common that people may describe a person using terms like *big eyes, wide mouth, small ears,* rather than the corresponding image objects.

Our research recognizes that image databases are systems for doing retrieval of images, as opposed to recognition. There are essential differences between retrieval and recognition. Retrieval systems require much more flexibility than most recognition systems and, on the other hand, can make use of more sophisticated interfaces to be guided by the user. The flexibility requirement prevents use of traditional symbolic objects semantics, based on region segmentation and object models. We believe that better results can be achieved using simple perceptual clues which guarantee the required domain-independence and flexibility.

The main belief behind this research is that, if perceptual clues are used in connection with the right similarity measure, there is a significant correlation between them and the object semantics of the user. In particular, we are developing a similarity model, based on three concepts: multi-resolution decomposition, a careful analysis of the geometry of the color space, and the definition of a general nonlinear metric for image distance. This model allows us to obtain perceptually and semantically sensible results and, at the same time, gives us enough flexibility to incorporate some kind of domain specific information when this is necessary.

**Object Related Queries:** These queries are semantic queries that ask for presence of an object. These queries may deal with three-dimensional objects. Since three-dimensional objects are difficult to recognize using automated techniques, these queries may become very complex. Three-dimensional object recognition is a very active research area in machine vision. Queries based on recognizing objects in a query image may be, therefore, very difficult to execute.

**Spatio-temporal Queries:** In video sequences, and in many other applications where pictures are obtained over a long period, a user may want to get answers to some spatio-temporal events and concepts. Answers to such questions may require complete analysis of all video sequences and storing some important features from there. Considering the fact that methods to represent temporal events are not well developed yet, this area requires much research before one can design a system to deal with spatio- temporal queries at the natural language level.

**Video Databases: TV News on Demand** Video is rapidly becoming the preferred mode of receiving information. Video is most certainly the most vivid medium for conveying information. Video has gained tremendous popularity since it appeared on the scene. As is well known, television has been one of the most influential inventions of this century. The last decade has seen growth in the use of camcorders in all aspects of human activities.

Video is the most impressive medium for communicating and recording events in our life. Its use is limited, however, by its basically sequential nature. To access a particular segment of interest on a tape, one must spend significant time is searching for the segment. Video databases have potential to change the way we access and use video. By storing each individual shot in the database, one can then access any individual frame based on the content of the shot. Each shot can be analyzed to find what is contained in each shot. Frames in each shot can be analyzed to find events in it. By segmenting videos into shots and analyzing those shots, one can extract information that can be put into a database. This database can then be searched to find sequences of interest.

Video databases can be useful in many applications. One application is news on demand. Suppose that each sequence is analyzed and the information in it is stored in a database with pointers to the relevant frames. This database then can be used to view the news of choice to the depth desired by a user, in the sequence desired.

## 4 Gestalt Vision

At any given time, we can only see the world, or the environment, from one perspective. To acquire other perspectives, we must move our eyes. To explore the environment from other viewpoints, we have to physically move. When we view the environment from one perspective, we are limited to what one may call tunnel vision or more precisely, considering the nature of image formation process, "funnel vision." Remember the famous fable about the six blind men and an elephant? Cameras have similar limitations. Thus, when a scene is captured using only one camera, the perspective is limited One could obtain more information about the environment by panning and tilting the camera so that one could see a complete view from one position. QuickTime VR has attracted attention by providing a mechanism to record these scenes from one position and then allowing a user to view the scene in any direction, but from this viewpoint. Similar efforts are being made in many research groups by taking multiple images of a scene and then using software to merge these images to provide a larger picture than is possible from any single camera view.

Using a powerful information system to mediate between viewers and multiple cameras, it is possible to provide Gestalt Vision, which is more than is possible using any individual camera. Gestalt Vision provides a holistic view by combining localized views. A viewer then can see the scene from any position and may walk through a dynamic scene without disturbing the events in the scene.

### 4.1 Multiple Perspective Interactive Video

Content-based interactivity and Gestalt Vision can be implemented by combining the tools and techniques currently being developed in different disciplines of computer science. In many applications, such as sports broadcasting, traffic monitoring and visual surveillance, multiple cameras are placed at strategically selected points to provide an operator a global view of events. In all these applications, different camera shots are fed to one location and there all these camera views are displayed. In a broadcast application, one of these views is selected by the editor or producer of the program to be broadcast to consumers. Clearly, this is intelligent multiplexing where the operator plays the role of the intelligent multiplexor.

Using evolving information systems and the delivery mechanisms created by the network infrastructure commonly available now, it is possible to develop systems that allow content-based interactivity and Gestalt Vision by strategically placing multiple cameras in an environment of interest. The image stream from each camera is processed to extract task-dependent information and is fed to an information system, called an Environmental Model. The Environmental Model assimilates information received from all cameras into one information system, which represents that information at multiple levels of abstraction.

This information system offers two major facilities. A user can interact with the information system at many different levels of information abstraction, and select the visualization mode in which to view the desired information. Also, a user can view any information of interest from any viewpoint of interest. Thus, the human multiplexor is removed and the user becomes the producer of information. Another major advantage is that the Environmental Model can be used by several users to view different information at the same time. Since the Environmental Model is an information system, it can be designed to reside at one or multiple locations and satisfy the information or entertainment needs of a diverse group of users at the same time.

Multiple Perspective Interactive video provides a framework for the management of and interactive access to multiple streams of data capturing different perspectives of an event. It has strong database and hypermedia components that allow a user to interact with live events and browse the underlying database for similar or related events or to construct interesting queries.

## 5 Role of Computer Vision

Computer vision and image analysis techniques are basic tools required for content-based operations in visual information systems. Many traditional approaches need to be reexamined, however. If traditional approaches are applied without considering the needs of visual information management systems, only some simple problems may be solved. To implement content-based approaches for interaction and retrieval, it is essential to analyze and represent the data at multiple levels of abstractions. These levels of abstractions should be determined by what is needed by an application. Computer vision research has developed tools for abstraction, reasoning across these levels for recognition and many other tasks. Most of those tasks were autonomous. In information systems, humans are in the loop. The presence of humans in the loop results in a interesting and significant difference in the role of computer vision. Now computer vision has to be a mediator between a human and data sources. On one hand, this results in less reasoning and decision making related tasks, on the other now the system must use abstractions that make sense to humans. We discovered that this results in many interesting new challenges.

## 6 Reference Sources

Here we list several papers from our group that are addressing issues related to the topics discussed above. This list is intentionally limited to our own research and is intended to provide more information on topics presented in the paper. Details about related research can be found from these papers. All these papers and pointers to several research papers is provided on our web page "http://vision.ucsd.edu".

1. Arun Hampapur and Ramesh Jain and Terry Weymouth, ``Production Model based Digital Video Segmentation'' *Journal of Multimedia Tools and Applications*, Vol. 1, No. 1, pp. 9-46, 1995.

2. Ramesh Jain and Arun Hampapur, ``Metadata in Video Databases'', *Sigmod Record: Special Issue On Metadata For Digital Media*, Dec. 1994.

3. Ramesh Jain. Telepresence in education: Building the universal university. Educom Review, 32(3):49-55, May/June 1997.

4. Arun Katkere, Saied Moezzi, Don Y. Kuramura, Patrick Kelly, and Ramesh Jain. Toward video-based immersive environments. Multimedia Systems, 5(2):69-85, 1997.

5. David A. White and Ramesh Jain. ImageGREP: Fast Visual Pattern Matching in Image Databases. In Proceedings of the SPIE: Storage and Retrieval for Image and Video Databases V, San Jose, CA, USA, volume 3022, pages 96-107, February 1997.

6. Ramesh Jain, Arun Katkere, and Jennifer Schlenzig. MPI Video: Content-Centric Interactivity and Gestalt Video. In Imagina '97, Monaco, January 1997.

7. Saied Moezzi, Li-Cheng Tai, and Philippe Gerard. Virtual View Generation for 3D Digital Video. IEEE Multimedia, 4(1):18-26, January 1997.

8. Saied Moezzi, Arun Katkere, Don Y. Kuramura, and Ramesh Jain. Reality Modeling and Visualization from Multiple Video Sequences. IEEE Computer Graphics and Applications, 16(6):58-63, November 1996.

9. Michael H. Goldbaum, Saied Moezzi, Adam Taylor, Shankar Chatterjee, Edward Hunter, and Ramesh Jain. Automated diagnosis and image understanding with object extraction, object classification, and inferencing in retinal images. In IEEE International Conference on Image Processing, November 1996.

10. Amarnath Gupta, Saied Moezzi, Adam Taylor, Shankar Chatterjee, Ramesh Jain, Michael H. Goldbaum, and S. Burgess. Content-based retrieval of ophthalmological images. In IEEE International Conference on Image Processing, November 1996.

11. David A. White and Ramesh Jain. Algorithms and Strategies for Similarity Retrieval. Technical Report VCL-96-101, Visual Computing Laboratory, University of California, San Diego, 9500 Gilman Drive, Mail Code 0407, La Jolla, CA 92093-0407, July 1996.

12. Ramesh Jain and Amarnath Gupta. Computer Vision and Visual Information Retrieval. In Festschrift for Prof. Azriel Rosenfeld. IEEE Computer Soc., 1996.

13. Arun Katkere, Jennifer Schlenzig, Amarnath Gupta, and Ramesh Jain. Interactive Video on WWW: Beyond VCR-like Interfaces. In Fifth International World Wide Web Conference (WWW5), Computer Networks and ISDN Systems, volume 28, pages 1559-1572, Paris, France, May 6-10 1996.

14. Simone Santini and Ramesh Jain. Gabor Space and the development of preattentive similarity. In Proceedints of ICPR 96, International Conference on Pattern Recognition, volume 1, pages 40-44, August 1996.

16. Simone Santini and Ramesh Jain. Similarity Queries in Image Databases. In IEEE International Conference on Computer Vision and Pattern Recognition, pages 646-651, San Francisco, CA, USA, June 1996.

17. Patrick H. Kelly, Amarnath Gupta, and Ramesh Jain. Visual Computing Meets Data Modeling: Defining Objects in Multi-Camera Video Databases. In Proceedings of the SPIE: Storage and Retrieval for Image and Video Databases IV, San Jose, CA, USA, volume 2670, pages 120-131, February 1996.

18. David A. White and Ramesh Jain. Similarity Indexing: Algorithms and Performance. In Proceedings of the SPIE: Storage and Retrieval for Image and Video Databases IV, San Jose, CA, USA, volume 70, pages 62-75, February 1996.

19. David A. White and Ramesh Jain. Similarity Indexing with the SS-tree. In Proc. 12th IEEE International Conference on Data Engineering, pages 516-523, New Orleans, Louisiana, USA, February 1996.

20. Arun Katkere, Saied Moezzi, Don Kuramura, Patrick Kelly, and Ramesh Jain. Towards Video-Based Immersive Environments. ACM-Springer Multimedia Systems Journal: Special Issue on Multimedia and Multisensory Virtual Worlds, Spring 1996. A version available as Technical Report VCL-95-105.

21. J.R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphreys, R.C. Jain, and C. Shu, ``Virage Image Search Engine: An open framework for image management" Proceedings of SPIE Conf. on Storage and Retrieval for Still Image and Video Databases, Feb. 1996.

22. Ramesh Jain and Koji Wakimoto, ``Multiple perspective interactive video" In IEEE Multimedia Computing Systems, pages 202--211, May 1995.

23. Jill Kliger, Deborah Swanberg, and Ramesh Jain, ``Concept Clustering in a Query Interface to an Image Database", the Proceedings of UIST, Atlanta GA, Nov. 1993.