A Patterned Injury Digital Library for Collaborative Forensic Medicine

David Stotts, John Smith, Prasun Dewan, Kevin Jeffay, F. Donalson Smith, Dana Smith, Steven Weiss, James Coggins

Department of Computer Science ,CB 3175, Sitterson Hall, Chapel Hill, NC 27599-3175

William Oliver, MD

Armed Forces Institute of Pathology, Walter Reed Army Hospital, Washington, DC

Abstract

The UNC CS Collaboratory, in conjunction with the Armed Forces Institute of Pathology and the Office of the North Carolina Chief Medical Examiner, is constructing a digital library to augment the collaborative practice of forensic medicine. Called the Repository of Patterened Injury Data (RPID), the project is putting into digital form several important and heavily used collection of medical data, stored and presented to allow joint consultation by several pathologists. We are researching issues in data storage, retrieval, and delivery; in shared distributed user interfaces; and in collaborative work patterns for forensic medicine.

Keywords: Forensic medicine, pathology, patterned injury, hypermedia, collaboration, medical images, browsing, search, retrieval.

1. Project Overview

Prevention of violent crime continues to be an important national priority, and it is increasingly important in the daily lives of our citizenry. Those actively involved in investigating these crimes and in apprehending the people who commit them must be able to pool their knowledge and expertise to provide maximum effectiveness. The digital library testbed we are creating and the research we will do in relation to it will contribute to the more effective interaction of forensic pathologists. More specifically, it will enable forensic pathologists to consult a large testbed of data from forensic medicine, share new case data, apply enhancement and analyses algorithms to images and consult with one another. Our research will also apply broadly to other distributed groups of professionals who work intensively and over extended periods of time with multimedia data, including substantial volumes of image data. Examples include astronomers, art historians, and biochemical researchers.

We are creating a testbed digital library, called the Repository of Patterned Injury Data (RPID). It includes test, images, audio and video sequences, charts, diagrams, drawings, and numerical data. Information in the repository will support forensic pathologists collaborating on medical cases involving patterned injuries. Patterned injuries are those having distinctive characteristics as a result of the object causing the injury, such as a hammer or tire tread. Identifying the source of an injury provides valuable information not just for determining the cease of death but also in apprehending and prosecuting those responsible in cases where a violent crime is involved.

Individual data objects are linked in RPID into a hypermedia graph structure to support flexible browsing. Access is to be provided by query-based search both on attributes of images and on the text of accompanying material. The computer infrastructure provided by our project for working with the RPID will also include image enhancement, collaboration, and communication functions as integral parts of the environment.

The initial data for the repository will come from the extensive archives of AFIP; from the Office of the Armed Forces Medical Examiner's (OAFME) Lindenberg collection of microscope slides and pictorial data; and from the Milton Helpern Forensic Pathology Museum. Other specialized collections, such as images of tire tread marks, are maintained by individual pathologists and researchers; we intend to bring these data into the repository, once it has been established.

After the initial construction phase, we will proceed with connection to other consultation sites. Dr. Thomas Clarke in The Office of the Chief Medical Examiner for the State of North Carolina is enthusiastic about participating in this effort. Dr. Clarke's office performs 1300 autopsies a year, and funds 2500 additional ones; all these cases generate data for their collection. Their forensic archive goes back to 1970, with all data except photographs already in digital form. They are willing to incorporate this data into the RPID, and they will be the first intermediate site we connect.

We foresee approximately 2,000 images with associated case data being processed in the initial year,

In: *Proceedings of Digital Libraries '94, The First Annual Conference on the Theory and Practice of Digital Libraries*, College Station, TX, June 1994, pages 25-33.

with the addition of 5,000 images in year two, and growing to at least 20,000 total images with case data by project's end. With the inclusion of autopsy case data from the NC Chief Medical Examiner's Office, and with data being integrated from other experimental sites beginning in year 2, the amount of forensic material could grow to as much as 70,000 images with the associated case data.

Development work is taking place at two sites. The testbed collection of digital data is being constructed by the Armed Forces Institute of Pathology (AFIP), housed at the Walter Reed Army Hospital in Washington DC. Work on the computer and communications infrastructure and a related program of collaboration user studies is being done by the Collaboratory Project at the University of North Carolina at Chapel Hill (UNC). This infrastructure will be based on the Artifact-Based Collaboration (ABC) System developed by the UNC-CH Collaboratory project under the NSF CTCT program [12, 13, 25]. This will allow us to put a usable infrastructure in place quickly and also to carry out further computer systems research to develop new capabilities that can be tested and added incrementally to the installed system.

Users of the completed testbed will include three types of forensic professionals: consulting pathologists; forensic scientists specializing in patterned injury; and field processionals, including pathologists, criminologists, forensic scientists, and forensic odontologists. They will be situated at some twenty locations across the country, including AFIP; the UNC Medical School; the U.S. Department of Justice; the New Mexico State Medical Examiner's Office; the North Carolina Chief Medical Examiner's Office; and the Office of the Armed Forces Medical Examiner (OAFME). The initial repository will grow as a result of this use by forensic pathologists throughout the nation.

Use of the testbed will be closely monitored by project researchers, both by human observers and by automated tracking of users' actions, in order to document changes in the working practices of this community and to evaluate specific systems research components. Issues we will investigate include how to distribute multimedia data for effective access; how to organize the data for both hypermedia access and query-based search; how use of the digital library will affect the working practices of forensic pathologists; how to assist collaboration relative to the digital library; and how to build a collaborative infrastructure that will adapt to heterogeneous environments, both for mixed data forms and for mixed computer hardware and software platforms.

1.1 Relationship of RPID to efforts elsewhere

While there are several projects reported in the literature that provide hypermedia data in a medical context [7, 1] the RPID project has some unique aspects. The previous efforts have not been in the context of large or widely distributed data sets. They have allowed physicians to simulate in hypermedia their research notebooks, for example; or they have provided centralized facilities for small library subsets. A unique emphasis of the RPID is enabling of collaboration among users of the library. Another unique aspect of RPID is the research focus on building an infrastructure that will allow the easy integration of new user tools and source data as the library grows. Thus, we are designing for, and investigating the practicality of continual expansion of the data and increasing distribution among the library sites.

There are also numerous projects that provide image storage and manipulation facilities to medical personnel using images like radiographs. These systems tend to focus heavily on the graphics capabilities, and are little more than traditional databases otherwise.

2. Project Requirements and Scenarios

Our research on the Repository for Patterned Injury Data (RPID) will produce a large digital library of medical data (specifically, materials represented in various media, comprising patterned injury case records) for the collaborative use of forensic pathologists. RPID data will include test, images, video sequences, charts, diagrams, drawings, and numerical data. The project will make contributions to several areas of knowledge:

Organization and Structure of Information

- hypermedia data organization of a large collection of disparate materials
- strategies for moving back and forth between IR (searching) and hypermedia browsing
- strategies for constructing argumentative paths within a hypermedia data model

Computer and Communication Technology

- intelligent agents to traverse hypermedia structures and perform a variety of functions, including searching and linking
- tools and protocols for defining links into video data
- distributed storage and access to data using local- and wide-area networks
- adaptable caching and replication strategies for image-intensive data
- new architecture and implementation for supporting computer conferencing of arbitrary applications at the operating system level

User Behavior

- detailed task-based portraits of a user community prior to their use of the digital library, during their adaptation to it, and after they become familiar with it
- studies of the effects of digital library and collaboration technologies on communication patterns within the user community
- evaluation of search and browsing behaviors

- evaluation of the effect of specific computer and communication infrastructure features on task behavior
- examination of privacy and security issues within the user community and development of appropriate policies
- evaluation of effects of digital library technology on the conduct of science within a specific professional community

2.1 Motivation

Prevention of violent crime continues to be an important national priority, and it is increasingly important in the daily lives of our citizenry. Those actively involved in investigating these crimes and in apprehending the people who commit them must be able to pool their knowledge and expertise to provide maximum effectiveness. The digital library we will create, and our research with it, will enable more effective interaction of forensic pathologists. Specifically, it will enable forensic pathologists to consult a large testbed of patterned injury data, share new case data, apply enhancement and analysis algorithms to images, and consult remotely with one another. Our research results will apply broadly to other distributed groups of professionals who work intensively and over extended periods of time with multimedia data, including substantial volumes of image data.

The analysis of patterned injury in forensic pathology presents a challenge that draws uniquely upon both the medical and forensic expertise of the investigator. As a problem in wounding, it is a medical challenge. As a problem in image enhancement and pattern analysis, it is a forensic challenge. This skill benefits heavily from experience. The investigator much have both experience in wound analysis and a knowledge of the universe of discourse, wherein lies the object which caused the injury.

For instance, discerning the general properties of an object (that it is rounded, sharp-edged, etc.) is well within the general knowledge base of most forensic pathologists. Recognizing that a specific injury is likely to have been caused by, say, a socket wrench requires the examiner not only to know about pathology but also to have some knowledge of automotive tools. This need for an encyclopedic recall of objects and object properties makes much patterned injury analysis difficult, especially when injuries occur in areas which abound with specialized tools, implements, and objects (*e.g.,* construction sites, factories, hair-dressing salons, etc.). Moreover, recognizing the types of marks made by partial or oblique blows is a further challenge in analyzing the geometry of impressions.

No forensic pathologist or odontologist can be expert in all areas of hardware manufacture and utilization. Instead, we must rely on the experience of our colleagues. Unfortunately, when an expert in the field retires, a wealth of specialized experience is lost, usually along with a career's worth of valuable patterned injury data and analysis results. The RPID project is an attempt to save such knowledge and to make such experience permanently available to pathologists and investigators across the country. We will do this by establishing an electronic registry of solved patterned injuries, and by developing a computing and communications infrastructure wherein pathologists and investigators can electronically access the data, search for cases germane to their current problems, and conveniently consult with one another.

2.2 User community requirements

Primary characteristics of the users of the library for forensic medicine include:

- Highly skilled and experienced group of expert users.
- Users widely distributed throughout the nation.
- Work intensively over sustained periods of time and require access for both casual browsing and quick retrieval of specific information.
- Work with multiple forms of data, including photographic images, computer enhanced images, text, video, audio, and statistical data.
- Consult and communicate with one another in direct reference to the data.
- Produce new information (*e.g.,* sequences of images, both photographic and generated) that is integrated with archival data.

A computer and communications infrastructure that can support these users in their accustomed patterns of work as well as enable them to work in new and more productive ways in the future must have the following features:

- Augment the skills and knowledge of users, not try to replace them or reduce them to algorithm.
- Provide support for multimedia, initially including image, text, video, and statistical data, but extensible to other forms.
- Provide mechanisms to link data in *ad hoc* as well as systematic ways.
- Support extensive, free-ranging browsing.
- Support automated search according to coded attributes (*e.g.,* specific characteristics of wounds, such as length, depth, shape, etc.).
- Support easy integration of new material into the RPID.
- Provide an open architecture that can accommodate arbitrary applications, such as image analysis and enhancement tools, and integrate their output into the RPID for future use.
- Support collaborative viewing and interaction with RPID data and associated applications.
- Support conversations and discussions among small groups of users.
- Support exchange of case-specific collections/structures of information, including commentary, for asynchronous consultation/collaboration.

2.3 System Usage

The RPID will use the collaborative computing and communications infrastructure developed by the Collaboratory at UNC-CH for dissemination of, exploration of, and interactive consultation with the medical case data from AFIP. There are several distinct forms of interaction required by this digital library:

• *exploration*

The RPID will be organized along feature vectors incorporating both image-based geometric information and "expert" knowledge. Associated with each wound image will be the image information regarding the object which caused the wound, a text file containing pertinent case data and notes, a text file containing links for further searching, and other media as appropriate (*e.g.,* video clips, scene images, site diagrams, etc.). For each wound in the wound image, a link will be established to adjoining images in the feature vector space.

To explore this space, then, an investigator will choose a starting place, and then move from image to image along change vectors. For instance, assume an investigator has a case and has found an image which is similar to, but still significantly different from the unknown mark in class characteristics. Let's assume that the unknown mark has the same general shape but is in some sense "longer," and that "length" is one of the feature metrics upon which the graph is built. The investigator can then move to the next "longer" wound mark which is similar in the other search dimensions. At that image, he or she can then continue along that dimension, or choose another dimension. A similar hypertext linking for the text portion of the nodes is also appropriate.

• *searching*

To find a starting point for this search, it will be necessary to find a first approximate match. For this, once the appropriate feature vector space is determined, an appropriate norm, perhaps weighted will be chosen, and a distance measure from the unknown image will be calculated. This distance function will then be used to find an appropriate starting point.

• *consultation*

While an investigator should be able quickly and intuitively to traverse the graph which organizes the registry without assistance, the power of this proposal is greatly enhanced by the ability to allow interactive consultation on the images during exploration. As stated in the introduction, patterned injury analysis is a profoundly experience-determined skill - a picture may be worth a thousand words, but it is only ten minutes of good experience. The RPID will allow an investigator in a local office to collaborate with specialists at the AFIP to receive timely aid in the evaluation of his or her images and in the exploration of the existing library. By allowing real-time interaction in image consultation, this will increase productivity of forensic pathologists many fold.

By allowing interactive consultation to involve application software as well as image display, the RPID will also enable pathologists to demonstrate their image processing skills during a consultation. This can be extremely important for explaining the forensic use of image processing methods to criminal investigators and lawyers, who often are not technologically sophisticated.

2.4 Digital library structure

The computer and communications infrastructure will be built by the UNC-CH Collaboratory Project in close consultation with the American Registry of Pathologists (ARP) and AFIP. The system will integrate several basic technologies: distributed data storage, hypermedia, information search and retrieval, generalized computer conferencing, intelligent agents, and tools for representing and browsing large databases.

The core of the library will be distributed hypermedia repository, based on ABC/DGS, a distributed storage subsystem and applications-support layer developed at UNC [12, 13, 25]. Users will interact with the repository and with each other through a network, such as the Internet. The hypermedia repository will support forms of information, including individual images, blocks of text, statistical data, and audio/video sequences, as individual content objects. Individual objects will be organized as a large graph structure that can be both browsed and searched.

To support browsing, objects will be linked with one another along a variety of dimensions. For example, all of the data associated with a given case could be linked into a tree that included as one branch the images, as another branch the various reports associated with the case, and, as a third branch, video clips of the crime scene. However, images in one case could also be linked to similar or related images in other cases according to specific features, such as length of wound, depth, or shape. Thus, users will be able to browse within the material from a given case but also across the primary structure of the library to data associated with other cases. The system will record a trace of users' browsing paths to facilitate subsequent analysis, as a learning aid, and as a means of capturing one form of expert knowledge.

To support search, data will be characterized in terms of specific features and parameters on those features when they are entered in the repository. Features will be used to generate links in the hypermedia graph structure. They will also be stored in a conventional information retrieval system. Consequently, users will be able to submit queries

to the retrieval system, obtain a set of objects, which they may then view or otherwise access through the hypermedia system. Once a user has arrived at a given object, he or she will be able to branch out using hypermedia browsing facilities to other objects linked to it. Thus, the system will combine capabilities of both hypermedia browsing and information retrieval.

The system will be based on an open architecture so that arbitrary applications can be included in the environment. This will enable users to invoke specialized tools - such as image enhancement programs, other software they are accustomed to using, and new tools as they are developed - on data included in the repository. Once such applications have been run, their output can also be stored in the repository for future use. Thus, the system will provide a flexible, easy to use environment for exploratory data analysis. A conferencing feature will permit two or more users, possibly located at remote sites, to share the same view of the data, or of an application being used with the data, through a computer network. It will also permit users to move smoothly from individual to collaborative work, and back.

While a major use of the system will be through browsing or searches controlled directly the expert user, the system will also include intelligent agents that can traverse the hypermedia store and apply algorithms to content objects. Thus, for example, users will be able to provide an agent with an image of a wound and with an image comparison algorithm and have the agent (operating asynchronously and concurrently with the user) explore the graph structure of the repository searching for similar images. When if finds candidate images, it could construct a link between the original and the candidate for subsequent consideration by the user.

The system must also include provision for audio and, ideally, video conversations/discussions. Initially, this will be limited to channels supplied by the telephone system. As digital technologies mature, we hope to include video and to support both forms through the computer network. This will enable the system to offer communication through the workstation as opposed to requiring the user to go outside the computing environment.

However, since consultation is so important and since it is so hard to get people together at the same time, particularly if their skills are in high demand, the system will include facilities to enable a user to organize bodies of material, including his or her recorded statements, that can later be viewed and responded to by another, consulting user. Thus, collaborators will be able to carry on extended, asynchronous "conversations" with regard to specific data without both having to be available at the same time. A similar facility for collecting, recording, and playback will also be useful to forensic pathologists when presenting evidence in the courtroom.

3 Research Issues

We have already mentioned that user studies are an integral component of the RPID project, but we will not further discuss this aspect here. The other major dimension of our research is in technical support for the collaborating users. We have identified hypotheses in three broad technical areas: user functions, distributed hypermedia data storage and access, and support for collaboration between heterogeneous hardware and software platforms. All three areas present fundamental problems that must be solved if the information highway and the digital library concept are to be broadly useful for technical and scientific work, but we only have space in this report to adequately outline our investigations in the user functions area.

The problem of different data types and formats being non-uniformly mixable in current hypermedia systems is being addressed by developing a set of orthogonal linking operations. These allow full interconnection of heterogeneous data in a uniform manner. We are concentrating especially on video, seeking to make it a first class hypermedia data form capable of being manipulated both by sequence *and content* (as is text in current hypermedia systems). We are also comparing various hypermedia models, classifying them according to their performance and suitability for the different data forms and uses in a large digital library.

3.1 Image analysis services within hypermedia

In support of our system-required techniques for linking uniformity, and for general support of forensic investigation, the RPID provides integrated image analysis services. Of all the various data types that will appear in the RPID, photographic images (and by extension, frame-based video) require the most expertise to address by content. In addition, image analysis is used by forensic pathologists to enhance difficult-to-see information from images, or to enhance the appearance of images to highlight information. The RPID must offer an efficient and flexible image processing component. Previous image analysis research has shown that:

- that domain-specific libraries or packages fare better in terms of user satisfaction and power than general image analysis packages.
- that when elaborate user interfaces are included, their care and feeding tends to overwhelm the image analysis content of the package, and
- that the field of image analysis is fragmented such that every laboratory has its own private image library structure, none of which are ideal for PIDL.

We are selecting specific image analysis tools for inclusion in the RPID infrastructure. Users may bring

other tools they select into the environment as well, using the results of our interoperability research.

3.2 Orthogonal data types: linking uniformity

We are creating algorithms that will allow users of video data to identify and hypertextually annotate moving objects in a video sequence. These methods will work in real-time, and will apply both to stored video (*e.g..,* clips from a crime scene[1]) and to real-time video (*e.g.,* a teleconference). We will require algorithms to find object edges in an image and outline them; algorithms for tracking these outlines from image to image in the frame sequence; and integration of this information with a hypertext interface for collecting and organizing information to be linked to the video objects.

We are basing our image manipulation algorithms on work done at the University of Florida on face tracking in an image sequence [3, 4]. This project demonstrated that tracking a face moving in frame-based video sequence could be done at a rate of 16 frames per second using an Intel 386 processor. Initial acquisition of a face, though, in the first frame took about 0.5 minute. We expect that considerable improvement can be obtained with workstations and new algorithms. Our goal is to be able to acquire and track objects in video in real-time; this would allow hyperlinks to be anchored into video sequences as they are generated (teleconferencing) in addition to after they are stored (archival clips).

The results of this work will allow the video (and audio) sequences to be treated uniformly as a component of a hyperlinked information network, alongside other forms of information such as text and still graphics. Frame-based video technology is not specific to this environment and the work will immediately apply to hypermedia systems in general. In current hypermedia systems, video clips are added to the basic information structure mainly as "view-only" components. A reader may play the video, reverse it, freeze it and so on, as with a VCR. However, no interesting interactions are possible with the *content* of the video information, as is possible for the content of text and static graphics. This research will allow video information to be more fully integrated into hypermedia systems, increasing the orthogonality of the set of operations in such systems by allowing video information to be interacted with and manipulated in the same ways that static information is now. The result for hypermedia will be systems that are less "modal" and more seamless.

Specific research issues: The questions we will seek to answer in our experiments are these:

- Can a good orthogonal set of linking operations be designed and implemented for the varied data forms in RPID?
- What data abstractions are needed to support uniform treatment of linking among heterogeneous forms?
- What linking operations can be devised for audio sequences?
- Can object location and tracking algorithms be developed for content-based video linking?
- Can the video content-linking algorithms be made efficient enough to operate in real-time (*e.g.,* for linking into teleconferencing sessions)?

3.3 Authoring and browsing in different data models

Several major hypermedia data models have been proposed in the past decade, but no comprehensive study has been done to determine the relative effectiveness of each, or to compare them against one another for a classification of utility in various domains. We propose to perform such a comparative study using the digital forensic medicine library. Before describing the goals and structure of our experiments, we outline the major models we will implement and work with:

- *Dexter (structured hierarchical graphs) [9].* Dexter was an effort, after a decade of applied hypermedia research, to formalize the features of successful system developments. The model owes much to earlier work on the HAM [2]. The basic Dexter model contains a structured graph, and an execution rule for presenting graph elements and traversing its links.

- *ABC (hierarchical directed graphs) [12].* ABC uses hierarchical directed graphs as the basic structure for hypermedia information, much like Dexter, but has some differences in link types and access methods. It is also the native model atop the current DGS system, and will be the primary implementation vehicle for the initial RPID.

- *WWW (unstructured flat graphs) [10].* The basic data model for the World Wide Web (WWW) hypermedia facility is a graph with no central definition, and with nodes distributed very widely geographically. The graph is flat, in that no node itself contains a graph. Each node is a "document component" containing names of links to other content elements (Uniform resource Locators, or URLs). WWW graphs tend to be exceptionally dynamic, with a non-trivial possibility of finding no nodes at the ends of some links.

- *Hypersets (sets) [16].* The basic model underlying the Hypersets hypermedia system is defined in terms of mathematical set theory. At any point in browsing, a reader is presented with a set of nodes, all related by

[1] Consider the video we all saw recently on television, of mobs draggin the bodies of U.S. soldiers through the streets of Mogadishu, Somalia.

sharing some characteristic (*e.g.,* all concerning injuries from dog bites). The reader may select any member of the set for viewing, or may "browse" by requesting a list of all other sets of which a particular element is a member. Thus, the notion of "link" found in other models is served in Hypersets by set intersection. Set-based hyperdocuments have been shown to be especially useful for taxonomic organization.

- *Trellis (hierarchical parallel automata) [19].* Trellis is distinguished from other models by its use of Petri nets, a class of parallel automation, to define both the static and the dynamic structures of a hyperdocument. Information elements are mapped to the places in a Petri net, and links are mapped to the transitions in the automation. Browsing proceeds as allowed by valid execution sequences of the net. Browsing may create and synchronize parallel paths of activity. The Trellis model has been shown to be effective for defining collaboration protocols within hyperdocuments [20].

These models have been extensively explored and reported on individually, but very little comparative research has been done among these models.

The ABC hypermedia system is built on top of the DGS distributed graph server [25]. The DGS is a general distributed storage system with directed graphs as its main storage abstraction (comparable to files in traditional storage systems). The interpretation of graphs required by the ABC system is supplied collectively in two places: in the DGS engine itself, and in each application interface. We can construct all these hypermedia models on top the DGS with an appropriate semantics for interpretation of the stored graphs.

Each hypermedia data model is either an interpreted graph (*e.g.,* a Petri net, a semantic net) or can be easily and efficiently represented as a graph (*e.g.,* sets). To implement this variety of models in the DGS we will first create a *semantics layer* for the DGS. This addition will allow rules to be specified telling how the components of a graph will be used at execution time. For example, the Trellis model can be represented by specifying that a graph has two node types (place, transitions), that the graph is bipartite, and that "execution" takes place by moving markers among the place nodes in a certain way. For initial experimentation with the semantics layer concept we expect to specify these rules in a Prolog-like notation: after determining the best way to encode structural and behavioral characteristics, we will move into a compiled format such as an object library.

Specific research issues: The questions we will seek to answer in our experiments are these:

- is one model more appropriate than the others for organizing very large data sets? relatedly, is there one model that is more capable than the others of being scaled up?
- is there one model that is more appropriate for organizing heterogeneous data sets?
- is one model more appropriate than the others for organizing information systems that are meant to support collaborative work?
- Can a hierarchy of functionality, or expressive power, be placed on these models? Which models can readily be described in terms of the others? Is there one model that can be said to he the most general?

3.4 Search and retrieval within hypermedia

Hypermedia provides an excellent browsing capability once a starting point has been found. Bur to find one or more data objects from which to initiate hypermedia searching requires a different form or retrieval. Earlier discussed a strategy for retrieving data objects based on the value of various attributes found in those data objects (for example, wound length, or the particular caliber of the weapon used). The specific attributes to be used and the specific values for these attributes are hand coded by trained specialists from the various forms of information found in the data object. To hypermedia browsing and retrieval based on attributes we add a third strategy: access to data objects using classifications and information retrieval techniques applied to the text portion of the data object.

In addition to photos, drawings, videos, etc., each data object will have a text component. This, for example, could be the report written by the police, coroner, or attending physician, or could be the pathologists report written after analysis. We propose to use text-based information retrieval techniques on these reports as a means of entry into the hypermedia database.

This is a particularly attractive application for text-based information retrieval. The textual material tends to be brief and to the point with very little that is irrelevant to the case. The context of the written material is limited and, especially with material written by physicians, the vocabulary is precise and restricted.

We propose to experiment with several strategies for automatic classification and retrieval with this portion of the database. Initially, we propose to classify textual material using a standard medical thesaurus with additional terms unique to forensic pathology added where necessary [15]. This will create for each text object a vector of keywords representing the content of that piece of text. Experimental results will determine whether any of the various keyword weighting schemes are useful or whether simple boolean weighting performs satisfactory [26, 24].

Users can access the textual database using three basic strategies. First, the user can enter a free form query which is then subjected to the same classification process as was the data. The query and data descriptor vectors are then compared [26]; those data objects scoring highest are presented to the user ordered by their score. Second, the user can be presented with the thesaurus and instructed to formulate his or her query from this controlled vocabulary. The user will then be presented with preliminary search results (*e.g.,* "your query matched 50 data objects") and will be able to accept, broaden, or narrow the query using the thesaurus hierarchy. And third, we can combine these techniques with relevance feedback [17,5, 8] allowing the user to formulate a free form query and then, after an initial search, have the system revise the query based on explicit or implicit user interaction and on the data objects that the user has judged relevant.

Additionally, we propose to subject the text portion of the database to automatic classification and thesaurus construction strategies [24, 6, 14, 11] to determine whether purely automatic techniques provide satisfactory results. We have an advantage over traditional information retrieval applications in that this portion of the system does not need to provide high recall and only moderately high precision [26, 24] in order for it to be judged successful. In our application, retrieval from the text portion serves mainly to provide an entry point into the hypermedia which is then browsed extensively. A search is successful if it provides one or more data objects from which a complete (*i.e.,* high recall) search of the hypermedia can be made.

In the initial version of the RPID, the user will be presented with data objects that match the query, but the user will not be given more detailed information as to exactly what in the data object text matched the query. As an enhancement to the basic system, we propose to embed our existing full text retrieval system [18, 22, 21, 23] into the hypermedia system. This will direct the user to those portions of the text where matches occur. This will make it much easier for the user to scan the text and judge relevance, and will also be the basis for text-based hyperlinking from one data object text to another.

References

[1] A. M. Burger, B. D. Meyer, C. P. Jung, and K. B. Long. The virtual notebook system. In *Proceedings of ACM Hypertext '91,* pages 395-401. ACM, December 1991.

[2] Brad Campbell and Joseph M. Goodman. HAM: A general purpose hypertext abstract machine. *Communications of the ACM,,* 31(7):856-861, July 1988.

[3] K. Deng and J. N. Wilson. An approximation-based video tracking system. In *Proc. of SPIE: image Algebra and Morphological Image Processing II,* volume 1568, Jule 1991.

[4] K. Deng and J. N. Wilson. Contour estimation using global shape constraints and local forces. In *Proc. of SPIE: Geometric Methods in Computer Vision,* volume 1570, July 1991.

[5] S. R. Friedman, J. A. Maceyak, and S. F. Weiss. A relevance feedback system based on document transformations. In F. Salton, editor, *The SMART Retrieval System,* Prentice-Hall, 1971.

[6] E. Fox, T. Nutter, T. Ahlswede, M. Evens, and J. Markowitz. Building a large thesaurus for information retrieval. In *Proc. of the Second Conference on Applied Natural Language Processing.* Association for Computational Linguistics, February 1988.

[7] Mark E. Frisse. Searching for information in a hypertext medical handbook. *Communications of the ACM,* 31(7):880-886, July 1988.

[8] D. Harman. Relevance feedback revisited. In *Proc. of the fifteenth annual international ACM SIGIR Conference on Research and Development in Information Retrieval.* June 1992.

[9] Frank Halasz and Mayer Schwartz. The Dexter hypertext reference model. In Judi Moline, Dan Benigni, and Jean Baronas, editors, *Proceedings of the Hypertext Standardization Workshop,* pages 95-133. National Institute of Standards and Technology, February 1990. NIST Special Publication 500-178. Workshop held January 16-18, 1990.

[10] K. Hughes. Entering the world-wide web. Technical report. Honolulu Community College, September 1993.

]11] K. Sparck Jones and D. M. Jackson. The use of automatically obtained keyword classification for information retrieval. *Information Storage and Retrieval,* 5, 1970.

[12] K. Jeffay, J. K. Lin, J. Menges, F.D. Smith, and J. B. Smith. Architecture of the artifact-based collaboration system matrix. In *Proceedings of CSCW '92 (Toronto),* pages 195-202. ACM Press, 1992.

[13] K. Jeffay, J. Menges, J. K. Lin, F. D. Smith , and J. B. Smith. Architecture of the artifact-based collaboration system matrix. In *CSCW '92:*

Proc. of the Conf. on Computer-Supported Cooperative Work. ACM Press, November 1992.

[14] K. Sparck Jones. Automatic indexing. *Journal of Documentation,* 30, 1974.

[15] Mesh (medical subject headings 1988). 29(1, part 2), January 1988.

[16] H. Van Dyke Parunak. Don't link me in: Set-based hypermedia for taxonomic reasoning. In *Proceedings of Hypertext 91,* pages 233-242. ACM, December 1991.

[17] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System.* Prentice Hall, 1971.

[18] G. Salton. Another look at automatic text-retrieval systems. *Communications of the ACM,* 29(7), July 1986.

[19] P. David Stotts and Richard Furuta, Petri-net-based hypertext: Document structure with browsing semantics. *ACM Transactions on Information Systems,* 7(1):3-29, January 1989.

[20] P. D. Stotts and R. Furuta. Modeling and prototyping collaborative software processes. In *Proceedings of the NATO Advanced Research Workshop on Integration of Information and Collaboration Models,* June 1993. To appear. Also published as Technical Report TR93-020, Computer Science Collaboratory, Univ. of North Carolina at Chapel Hill, 1993; and as Tech Report TAMU-HRL 93-006, Hypermedia Research Laboratory, Texas A&M University, July 1993.

[21] J. B. Smith, G. J. Ferguson, J. D. Bolter, M. Lansman, D. V. Beard, and S. F. Weiss. We: A writing environment for professionals. In *Proc. of the 1987 National Computer Conference,* 1987.

[22] J. B. Smith, G. F. Ferguson, and S. F. Weiss. A hypertext writing environment and its cognitive basis. In *Proc. of the Hypertext '87 Workshop,* November 1987.

[23] J. B. Smith, G. J. Ferguson, and S. F. Weiss. Microarras: An advanced full-text retrieval and analysis system. In *1987 Intl. Conf. on Research and Development in Information Retrieval,* 1987.

[24] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw Hill, 1983.

[25] D. E. Shackelford, J. B. Smith, and F. D. Smith. The architecture and implementation of a distributed hypermedia storage system. In *Proceedings of ACM Hypertext '93,* pages 1-13. ACM, November 1993.

[26] C. J. van Rijsbergen. Butterworths, 1979.