# Conceptual Indexing for Video Retrieval

**Andrew S. Gordon and Eric A. Domeshek**
The Institute for the Learning Sciences
Northwestern University
1890 Maple Avenue
Evanston, IL 60201
gordon@ils.nwu.edu, domeshek@ils.nwu.edu

**Abstract**

There is a growing demand for technology that will allow computers to better manage and manipulate video and other media sources. We are building a system to help create and search an archive of stock video clips in order to lower the cost of new video and multimedia productions. The goal is to provide on-line access to stored video clips with flexible, effective, and efficient retrieval. Our approach to this retrieval problem is to construct a rich conceptual indexing system, a simple retrieval algorithm, and an easy-to-use browsing interface. Because the current state of the art in video analysis does not support automated extraction of many interesting conceptual features from video, we focus our efforts on elaborating a vocabulary of what those features should be and on building a system in which it is practical to hand-code indexes. We argue that the conceptual indexing approaches of existing case-based reasoning systems in AI can be adapted to meet the representational requirements of intelligent multimedia retrieval systems. We present six categories of conceptual indexes that are applicable to stock video clips, argue for our approach to the development of effective end-user indexing and retrieval systems, and describe our current implementations.

## 1. Introduction

Advances in computer technologies are realizing the dreams of a growing community of people that use multimedia information for work, education, and entertainment. It is now possible to create database systems that provide instant access to text, graphics, sound, and video, enabling a wide range of previously impossible applications. Unfortunately, the multimedia database dream is followed by the indexing and retrieval nightmare: How can we analyze and organize the contents of a multimedia database to support the various retrieval tasks presented by its users? Rich databases of multimedia information are rendered useless without the ability to access the right piece of information at the right time.

Consider the problems presented to us by Andersen Telemedia, a medium-sized video production facility eager to adopt new multimedia database technology. Andersen Telemedia, part of Arthur Andersen & Co., maintains a library of the video that it produces and acquires. This includes, among other forms, stock video, i.e. video clips general enough to be reused in later productions. Until recently, they managed their stock video by storing the raw footage tape containing potential stock video clips in their stock footage library and creating a database record including a short text description of the tape's contents. Dedicated librarians were used to handle assistant producers' requests for stock video, which ranged from the very specific, e.g. "close up of water ripples", to the very abstract, e.g. "symbolic images of people designing the future".

These librarians, who were often very familiar with the breadth of the library, would service a stock video request by searching the database for key words or phrases that would direct them to raw footage tapes. They would then view those tapes to determine if, in fact, they contained clips that would satisfy the request.

Tempted by the multimedia database dream, Andersen Telemedia has decided to digitize each stock footage clip in its library into a on-line video server and make its library directly accessible to producers and assistant producers, as well as people outside of the video production facility. The new system will likely greatly decrease the amount of time it takes a stock video librarian to service stock video requests; making video clips instantly accessible on-line eliminates the need to locate, load, and view a physical raw footage tape, which was the time bottleneck of their previous system.

But what about users of the database who are not already familiar with its contents? As Andersen Telemedia has come to realize, much of the success of their previous retrieval method must be attributed to the knowledge of the video librarian rather than the database search mechanisms that were provided. Removing the video librarian from the retrieval process highlights the inadequacies of the database storage and retrieval methods they are using. If Andersen Telemedia's new system is to be successful for a large population of users, they must determine how to replace the intelligence that video librarians used to make the retrieval process effective.

The need for intelligent multimedia information retrieval systems presents a new opportunity to apply artificial intelligence research to large-scale real-world problems. Developing intelligent retrieval systems is not a new problem in AI; many areas of AI research have recognized that flexible and efficient manipulation of stored information is a central component of intelligence. In particular, indexing and retrieval has been identified as the central issue in all case-based reasoning systems (Schank, 1982; Kolodner, 1993), including those designed for planning and natural-language understanding (Riesbeck & Schank, 1989) as well as generating explanations (Schank, Kass, & Riesbeck, 1994).

There are some obvious parallels between indexing and retrieval in case-based reasoning systems and multimedia information storage and retrieval. Video clips, sound bites, pictures, and texts can be viewed as cases in a multimedia database, which can itself be viewed as a case memory system. Entering multimedia cases into a database and retrieving them later can be viewed as types of case indexing and reminding, respectively. Given these obvious parallels, we would be poor problem-solvers, indeed, if we did not consider the indexing and retrieval solutions offered in existing case-base reasoning systems, and determine if this work can be adapted to address the real-world problems presented by users of multimedia databases.

This paper investigates the challenge of intelligent multimedia information retrieval as a case indexing and retrieval problem. Specifically, we shall be exploring solutions to the problem of stock video indexing and retrieval described above. While we recognize and appreciate the complexities presented when multiple media types must be integrated in a single database, our focus is specifically directed towards video. However, given the inherent multimedia nature of video (moving and still images, sound, text and graphics), we feel that the solutions offered in this paper will be applicable to databases consisting of a broad range of media types.

In sections 2 and 3 we discuss the indexing problem as it exists in case-based reasoning systems and present an approach to indexing stock video clips. In sections 4 and 5 we discuss the retrieval problem and previous solutions that have been suggested and present the new retrieval systems that we have developed.

## 2. Conceptual Indexing

Video, in combining sound, image, and motion, is among the richest mediums for the communication of ideas. Video is conceptually dense; it is used not only to show the viewer new places, people, activities, and objects, but also to teach lessons, make points, and evoke emotions. One of the great challenges in video indexing is finding ways to embody these concepts in symbolic descriptions that can be searched for matches to users' queries. Ideally, indexes for video should be composed from a language that is expressive enough to represent the breadth of concepts that can be conveyed by video while being restrictive enough to support realistic retrieval mechanisms. When it comes to symbolically describing video, nothing approaches the completeness and flexibility of natural language. However, natural language descriptions have a pair of significant drawbacks which cause serious headaches for designers of retrieval algorithms. Natural language descriptions are neither canonical (there are an enormous number of ways to say exactly the same thing) nor unambiguous (a single word can have an enormous number of meanings).

As an alternative to using natural language as indexes, developers of case-based reasoning systems have, almost without exception, utilized some form of conceptual indexing to organize cases in memory. Conceptual indexes are defined by an organizational scheme and a specific vocabulary that is used to label the entries in a case library. Conceptual indexes have the advantage of canonical representation, i.e. there is a one-to-one mapping between concepts and their representations, which creates many opportunities for simplified retrieval algorithms. However, the costs of using conceptual indexes are substantial; before any concept may be assigned to a particular case, it must have been specified as part of the indexing vocabulary of the system. This commitment to a specified vocabulary restricts the indexes of cases to only those domains that have been analyzed and organized by the system designers and indexers. Accordingly, developers of systems that use conceptual indexing must devote a substantial portion of their efforts to developing the ontologies of the domains that define the index vocabulary. Using conceptual indexing for video requires that developers have the means to organize and create vocabularies to cover the enormous range of concepts which video can be about.

The challenge of developing organizations and vocabularies for the concepts that can be conveyed in video is formidable, but there is a substantial amount of previous AI research that addresses these issues directly. In particular, our extended community of researchers, largely consisting of Roger Schank and his students, has pursued a consistent approach to the development of content theories, rich representational schemes designed to organize the conceptual components of a domain. The domain of choice in this research has been the everyday social world of normal people rather than specialized fields of expertise such as medicine or engineering. This is primarily due to the types of applications that directed this work, especially natural language understanding and

storytelling. This research has furthered our understanding of several important conceptual indexing concerns. It has provided us with an understanding of the broad classes of conceptual entities, which include actions, plans and goals, explanations, stories and points. For each of these classes, representational structures have been developed and taxonomies have been designed to organize and catalog the breadth of the conceptual space (Schank & Abelson, 1977; Domeshek, 1992). Since stock video clips are typically about the everyday social world of normal people, this research should provide a sound basis for developing stock video indexing schemes.

A significant drawback to the indexing schemes based on the content theories mentioned above is that, given the complexity of the representations, only highly trained content analysts and AI researchers could properly compose indexes for cases. If intelligent retrieval systems are to scale up to solve real world problems such as those presented by Andersen Telemedia, we must move in the direction of end-user and automated indexing (Goldstein, Kedar, & Bareiss, 1993; Cleary & Bareiss, 1995; Domeshek & Kolodner, 1994). While significant advances have been made in the fields of computer vision and video analysis that may give assistance to human indexers (Hampapur, Jain, & Weymouth, 1994), the state of the art is far from providing the sort of rich conceptual descriptions most useful for retrieval of stock footage, or most other video types for that matter. For the foreseeable future, our efforts must be directed towards developing simple end-user indexing tools (e.g. Davis, 1994) and, where necessary, simplifying the indexing schemes that we have available without compromising the effectiveness of the retrieval process.

In developing intelligent retrieval systems for stock video clips, we utilize a simplified conceptual indexing framework. Given the tradeoff between the expressivity of complex index representations and the usability of simple ones, designers must consider the retrieval requirements of the end-users and choose the simplest one that does the job. We feel that it is better to err on the side of simplicity, introducing more complicated structures only if their necessity becomes apparent during system development. We believe we have selected an indexing scheme that meets the representational requirements of the type of retrieval tasks presented by Andersen Telemedia while being simple enough to facilitate the development of end-user indexing tools.

In our framework, indexes for a case are described as a set of disjoint concepts, each of which is represented as a single node in a semantic network. For example, a clip that depicts a professor walking around a college campus may simply be indexed as professor, walking, and college campus, where each of these concepts is a node in the conceptual organizations of people, activities, and places, respectively. The major difference between this style of index representations and those implemented in other conceptual indexing systems is that there is no relational structure between the concepts that make up a case's indexes. Relational structure is an absolute requirement when completely representing cases for the purpose of drawing analogies, analyzing a causal chain, and adapting a case to a new situation, but our investigations to date have suggested that unstructured index representations may be sufficient for many retrieval tasks.

The simplicity of this framework allows the use of very basic matching algorithms. For each concept in the semantic networks of the system, the set of all of the

cases indexed by that concept can be compiled and effectively encoded as a one-dimensional bit vector. At retrieval time, the cases that are indexed by any one particular concept are given in the case-set associated with it. To determine the set of cases that are indexed by the conjunction of two or more distinct concepts, a simple intersection of each concept's case-set can be performed. In this manner, all of the stock video clips that are indexed by both "professor" and "college campus" can be determined by intersecting the case-sets associated with the nodes for these two concepts. This mechanism offer an effective method of matching query representations to index representations during the retrieval process.

In simplifying the representation of indexes, we have been able to focus our efforts on the creation of semantic networks that organize indexes for stock video clips and the development of end-user indexing and retrieval interfaces. In section 3, we present the six classes of indexes that we feel are most functional for stock video clip retrieval. In sections 4 and 5, we compare different interface styles and discuss the implementations we have developed.

## 3. Conceptual Indexes for Stock Video Clips

Given the task of retrieving stock video clips for use in larger video productions, we believe that there are six types of indexes that should be assigned to individual clips. We believe that the first two index types, the content of the scene and the points illustrated, will be the most functional for the majority of video clip retrieval tasks. Accordingly, these two types have received the majority of our attention. Each of these index types is briefly described below, with the most space devoted to describing our progress in developing appropriate indexes for the content of the scene.

### 3.1. The Content of the Scene

We expect that the scene contents of a video clip will commonly be the most important indexes for retrieval. Scene content indexes include information about where the clip is located, the activities that are occurring in the clip, the types of people and the roles they are playing in the those activities, and the salient visible objects. Given a clip library indexed by scene content alone, a user should be able to successfully retrieve clips for many of the kinds of requests typically directed to stock video librarians. Consider these examples of actual stock requests we have collected:

Electrical plants
People in a work area wearing hard hats
Wheat field with wheat blowing in the wind
Traffic jams with cars and bicycles
Polluted streams
Police trying to control a crowd

Perhaps the most salient component of scene content indexes for video is the place where a clip is set. Place indexes make up part or all of several of the requests listed above, including electrical plants, streams, and wheat fields. A casual consideration of the possible places that a clip could take place suggests that a complete

place vocabulary would be very large. The set of possible place terms can be divided into two broad categories. The first category consists of the proper names of specific places, which can be organized by contained-in relationships; for example, Wrigley Field is in Chicago, which is in Illinois, which is in the United States, which is in North America. The second category consists of vocabulary describing the type or function of individual places; example place types include bus stations, legal courts, junkyards, libraries, submarines, and water treatment plants. The rich language of place types extends easily to quite abstract sorts of places, such as natural places versus man-made places. Abstract place types can serve both as organizers for the more specific types and as vague descriptors. A sampling from one branch of a hierarchical organization for types of places illustrates how this can work. Below is a hierarchy for vehicle types, which can serve as the location for many instances of stock video clips:

Vehicles include
- Air vehicles include
  - Airplanes include
    - Military airplanes
    - Commercial airplanes include
      - Cargo airplanes
      - Airline passenger planes
  - Helicopters
- Land vehicles include
  - Road vehicles include
    - Automobiles include
      - Commercial automobiles include
        - Taxis
        - Cargo trucks
      - Private automobiles
  - Rail vehicles include
    - Passenger trains
    - Cargo trains
- Water vehicles include
  - Commercial boats include
    - Fishing vessels
    - Shipping vessels
  - Private boats include
    - Powerboats
    - Sailboats

Part-of relationships also serve as a useful compliment to the type-of links between place indexes illustrated above. For example, there are many parts of an airline passenger plane that are distinguishable places in their own right, including the cockpit, the lavatory, and the passenger seating area.

In addition to place, the scene content indexes of a clip may include several other types. One of the most important features of many clips is activities that are taking place in the scene. These can span a wide range of human actions, including business meetings, weddings, seminars, dating, relaxing, studying, etc. Often it is reasonable to index a clip by the types of people that are visible, e.g. Japanese businessmen,

professional hockey players, and elementary school students. Sometimes the objects in the scene serve as important indexes, such as automated factory machines, notebook computers, and violins. Like place indexes, the number of distinct concepts for these types that can be captured in video clips is enormous. Unlike place indexes, it is difficult to develop compact, comprehensible, and unambiguous organizational structures for very abstract activities, people types, or objects. While developing abstract organizational schemes for these types is an interesting challenge for academic researchers, any scheme aiming to allow relatively untrained users to incrementally add concepts to a conceptual network must be a simple and intuitive.

Our solution to this organization problem is to capitalize on the situated nature of most human activities, people types, and objects. Here the term "situated" is used to mean that these concepts are typically associated with particular places. Much of what happens in the world is fairly conventional, and so the scene content components in a clip are often related in very predictable ways. For example, baseball games, professional baseball players, and baseballs can typically be found in baseball stadiums; you might look for business meetings or businesspeople in conference rooms of office buildings; weddings typically happen in churches and involve brides, grooms, wedding cakes, and wedding rings. Of course, there are often exceptions to these generalizations, but by associating activities, people types, and objects with the places in which they are typically found, the relatively simple organization of places can serve as the backbone of the organization for these types as well. With the addition of activity-in, people-types-in, and objects-in links, the place organization can be extended to incorporate activities, people-types, and objects in a single index network. To illustrate this idea, consider a sampling of links that can originate from the concept of airline passenger planes, which can be found in the example hierarchy above:

`Airline Passenger Planes`
- Is type of: `Commercial Airplane`
- Types: `Jet-engine planes, Propeller planes`
- Is part of: nothing
- Parts: `Cockpit, Passenger seating section, Lavatory`
- Activities-in: `Piloting, Reading, Mobile computing`
- People-in: `Pilots, Businesspeople, Airline attendants`
- Objects-in: `Magazines, Laptop computers, Luggage`

The organizations outlined here provide an effective framework for the incremental development of large index networks to describe scene content. We expect that the types of scene content indexes we have described here will be sufficient to meet the representational requirements of a substantial portion of typical stock video retrieval tasks, including all of the ones presented at the beginning of this section.

### 3.2. The points illustrated by the clip

Video, like all media, often needs to communicate abstract ideas and relationships, and stock clips are often sought to carry that communicative burden. In many cases, it may be appropriate to index clips by the points that they make or support, such as "clients appreciate service that anticipates their needs" or "you can do a good job with fewer

resources if you are creative". Often points are comprised, in part, of concepts found in scene content indexes. For instance, a clip showing a military helicopter rescuing a civilian sailboat during a storm might effectively make the point that "sailboats are risky". However, the same clip may make the more general point that "weapons of destruction can be used for humanitarian purposes", which makes no direct reference to the concrete scene content items of the clip.

Our research has only begun to explore the possibility of indexing stock video clips by the points that they make, but we have a firm research base on which to build. At the Institute for the Learning Sciences (ILS), a great deal of attention has been directed to the task of indexing educational stories so they can be found when the lessons they teach are needed. In a typical ILS training application, these stories are presented as video clips of experts telling about their most interesting experiences (although the fact that a story is recorded on video is completely incidental to its indexing or use). So that these stories can be retrieved just when users can most benefit from their lesson, stories are often indexed by the points that they make. Accordingly, there is a significant and growing corpus of research at ILS on how to represent high-level points of the sort exemplified in the previous paragraph (Schank & Fano, 1992; Shafto, Bareiss, & Birnbaum, 1992; Cleary & Bareiss, 1995). We believe that this research can be used to organize the space of points made in stock video and inform the selection of the most appropriate representations of the points in particular clips.

### 3.3. The composition and camerawork of the clip

The composition and camerawork of a clip may also serve as a valuable index for clip retrieval, especially when functioning as a filter to pick out some small number of usable clips from some larger set with potentially relevant content or point. Over the decades, a rich and stable vocabulary has developed in the video and film production communities to describe techniques for manipulating shot characteristics such as camera angle, motion, and focus. One challenge in implementing composition and camerawork indexes is that often these vocabularies are not completely content independent. Sometimes describing the motions of a camera or the composition of a shot must be done with regard to the place, people, activities, and objects in the scene. For instance, video producers use a special set of descriptors for camera angle when the focus of the shot is a person.

### 3.4. The likely functions of the clip in a larger narrative

Video clips do not blandly record something happening; they are crafted for some purpose. One way to index a clip is in regard to the roles it might play in some larger context, most typically in some narrative sequence in a larger production. As with composition and camerawork, we expect that indexing a clip by its likely functions will serve as a useful complement to the description of the clip's scene contents. Narrative function indexes may include specific types of transitions, interludes, prologues, background or hooks, among others.

### 3.5. Information about the source of the clip

The details surrounding the creation of a video clip are likely to be relevant to individuals looking for stock footage. These details may include the date and time that the footage was shot, the history of how it found its way into the stock video database, the physical location of the original video cassette, the clip's start and stop timecodes, and any licensing or copyright restrictions that may apply. Knowing where a clip is located is primarily bibliographic information, but knowing legal restrictions on a clip's use can obviously have a significant effect on whether and how it is used.

### 3.6. The relationships to other clips in the library

In large collections of stock video clips, there will be many sets of clips that are related in important ways. Some will depict scenes of the same place on the same day, or even at the same time. Some will include the same actors or the same objects from different vantage points. Others may have been recorded with the intention of continuity, that is, one shot may be a reaction shot to something in another clip. These are interesting relationships because they create opportunities to do more with stock footage than would be possible with isolated clips. Also, they may provide alternate access paths to clips that may more exactly fit the user's need than the first clip found.

## 4. Retrieval Mechanisms for Conceptual Databases

In designing intelligent multimedia retrieval interfaces, indexing cases is only half the battle; complete systems must incorporate appropriate end-user retrieval interfaces. Regardless of the quality of the case indexes, no system will be successful unless users can easily locate cases that meet their needs.

It is important to understand how the type of conceptual indexing scheme we have chosen frames the retrieval problem. As cases are indexed directly by nodes in a conceptual network, the retrieval process is centered around the selection of some set of concepts which reflect those that are in the minds of the end-users. Then the cases associated with those concepts, determined by the intersection of each concept’s case-set, can be presented to the user for evaluation. In general, two classes of retrieval methods have been developed that are applicable to the concept selection problem. The first method, Query and Search, has been explored in a number of other intelligent multimedia retrieval systems. The alternative that we have implemented, the Zoom and Browse method, is an outgrowth of research in hypermedia browsing systems.

### 4.1. Query and Search

The first approach to intelligent retrieval is Query and Search, which has been developed out of a long tradition of standard database record retrieval and deductive retrieval databases in AI. In Query and Search, users must construct a request to be parsed by the system into some representational form which is matched against the indexes in the database by a search algorithm. Typically Query and Search systems embed flexible, knowledge-rich matching into their search algorithms, allowing them to retrieve cases similar or related to the user’s query when an exact match cannot be made. The best

Query and Search systems draw extensive inferences based on the user's query to find cases that best satisfy their needs, in a manner similar to human librarians.

Two recent examples of intelligent multimedia retrieval systems that use the Query and Search method are presented in the work of Chakravarthy (1994) and Lenat and Guha (1994). In Chakravarthy's work, user's queries are mapped directly to concepts in a semantic network (organized as sets of synonyms in Wordnet) and matched against indexes for each case. During the matching process, several rules may be used to infer whether a case's index satisfies the user's query based on the semantic relationships between the two. The system described is capable of making many obvious matches, such as returning pictures captioned as "a Dalmatian" in response to the user query "dog". It is also capable of more insightful matches, such as returning "closeup of an arrow hitting the bullseye of target" in response to a query for "shooting". Lenat and Guha take a related approach in their research, utilizing the expansive domain knowledge of the CYC knowledge base to expand both user queries and case indexes to increase the likelihood of successful matches. By drawing reasonable inferences from the captions of video clips, their system is capable of generating the impressive match between the user query "Find images of shirtless young men in good physical condition" and clips captions like "Pablo Morales winning the men's 1992 Olympic 100-meter Butterfly event" and "Three blonde men holding surf boards on the beach".

There are some difficulties associated with the Query and Search method that still need to be addressed. Primarily, by inferring beyond the indexes of a case during the matching process, these approaches lend themselves to false positives, i.e. the retrieval of cases that do not address the needs of the user. While it is reasonable to guess that "Three blonde men holding surf boards on the beach" might include a shot of a shirtless man in good physical condition, it is not necessarily the case. Expanding indexes through inference is somewhat like trying to cleverly service requests for video without actually watching the clips you select; without the means of verifying the legitimacy of inferences, you are bound to make lots of errors. To determine if the level of incorrect inferences is a significant problem, system designers must consider the nature of the retrieval task that the system is supporting. In stock video clip retrieval tasks, video producers consistently argue that more is better; they would prefer to be overloaded with options rather than miss out on a potentially useful clip. But even producers have their limits, and it appears that the index rules used in these systems would generate many indexes that video producers would not likely assign to clips. For retrieval tasks where precision is important, inference-based systems like these may be more useful as intelligent indexing aids that could suggest additional indexes to be verified by a human indexer while storing a case into the database.

There is an important question concerning the Query and Search method that must be addressed: Why are rich inferential search mechanisms necessary in the first place? The answer to this question is that in any system that allows users to formulate arbitrary queries, users will often produce requests for cases that are not in the case library. For stock video clips, where the space of possible descriptions is absolutely enormous, retrieval systems may almost never find cases that exactly matches a complex and specific user query. Rather than returning "No match" in response each time, we would like to be able to say "No exact match, but here are some close matches". Determining what can be considered close to an arbitrary user query requires the kind of

inference that the best of these systems incorporate. There is, however, an alternative to the Query and Search method. In the next section we introduce the Zoom and Browse method, an approach that removes the need for complex inference by prohibiting users from making arbitrary requests for cases that may not be in the case library.

### 4.2. Zoom and Browse

The Zoom and Browse approach takes the position that it is better to offer choices then to play guessing games. That is, rather than having users formulate queries for cases that may not be in the case library, retrieval systems should allow users to browse through the choices that the system has to offer. In this way users can make their own decisions about how closely the system's offerings suit their individual needs. The Zoom and Browse approach is exemplified by Ask Systems (Bareiss & Osgood, 1993), a type of hypermedia navigation application that organizes information, typically stories, in a format based on the flow of human conversations. In Ask Systems all information is linked together according to the questions answered and questions raised by each individual story. When the user is presented with a story, they are shown an organized list of follow-up questions that they can ask that will lead them to other stories in the system. Zooming is used here to refer to the beginning of this conversation process, where users are directed towards a particular starting story in a section of the network that is likely to contain the sorts of stories they will find interesting. Browsing refers to the subsequent process where the users navigate through the stories by following the conversational links provided by the system.

There is a substantial difference between conversational hypermedia and information retrieval, but the style of Zooming and Browsing that exists in Ask Systems can be easily adapted for this task. For this purpose, we introduce a new twist on the previous uses of the Zooming and Browsing interface; rather than stepping through an organization of cases (e.g. stories in an Ask System) we propose that users step through an organization of *case indexes*. In this context, the Zooming process guides the user to some conceptual index that is related to the types of things that they are looking for. During the Browsing process, users navigate through the semantic links that organize the conceptual indexes, consider the options that the system offers, and selects those that most effectively meet their retrieval needs.

This approach to index selection lends itself to an effective form of incremental case discrimination by controlling which indexes are available for selection. If the conceptual network only contains indexes that have cases associated with them, then all indexes should be selectable at the beginning the retrieval process. To allow users to look for cases that contain a conjunction of multiple indexes, the system must dynamically update the selectability of each available index or risk allowing the users to select a conjunction of indexes for which there are no available cases. Fortunately, it is simple to dynamically determine the selectability of any index given our indexing framework; an index is selectable if the intersection of its case-set and all of the case-sets of the previously selected concepts is not empty.

The Zoom and Browse method of index selection can also be used as an interface for end-user indexing of new cases. As stated in section 2, we are currently using a very simple representation for storing cases in our systems, namely an unstructured set of individual conceptual indexes. If the indexes already in the conceptual index

organizations are sufficient to describe a new case, then by disabling the incremental discrimination mechanisms it is possible to use a Zoom and Browse interface to select each of the indexes for a new case and update each concept's case-set accordingly. However, because we believe that the development of index organizations should be data-driven, then end-user indexers must be given tools to incrementally add new conceptual indexes to the system. While the development of these tools has been set aside for future work, we believe that the their success will depend highly on the quality of the conceptual index organizations that end-users have to work from.

To use the Zoom and Browse method for the purpose of index selection, the Zooming process must be effective enough to place the user into a section of the conceptual organization that is close to the concepts they are looking for. Likewise, the browsing process requires that the conceptual organization of indexes can be navigated by people unfamiliar with semantic networks as used in AI research. Determining the feasibility of these requirements for stock video retrieval has been one of the central focuses of our research. We are currently in the process of developing a large-scale video retrieval system and accompanying indexing tools based on the Zoom and Browse method. These systems, which are in the early stages of development, are described in section 5.

## 5. Retrieval systems for stock video

In September of 1994 we began our project to construct a stock video library for Andersen Telemedia. We started by collecting sample video clips, consulting with producers, and reviewing the video production literature. Design work focused on analyzing these sources to uncover an initial set of high-level conceptual categories for indexing, and settling on an interaction style and accompanying interface. Over time, we collected a larger representative corpus of clips, and extended and refined selected parts of the indexing scheme. We now have a first implementation of the interface, the underlying storage and retrieval architecture, and an initial corpus of 200 indexed video clips. The system is constructed in Visual Basic running on a Windows PC, and makes use of an Access database to manage both the library contents and the conceptual indexing scheme.

In this first version of our stock video retrieval system only scene content indexes are supported. As described in section 3.1, scene content indexes can be effectively organized by using a hierarchy of places as the backbone of a larger conceptual network that includes indexes for activities, types of people, and objects. This organization is used directly to support both the Zoom and Browse mechanisms in our system. For Zooming, we capitalize on the utility of places to organize a neighborhood of related indexes. During the Zooming process, the user identifies a particular place index which is related to the sort of clips they are looking for. During the Browsing process, users are presented with the conceptual neighborhood around the place index and permitted to incrementally select indexes that correspond to clips in the case library. In this version of our stock video retrieval system, both of these processes are seemlessly integrated into one interface.

Figure 1 shows the retrieval screen in its initial state ready to begin the selection of scene content indexes. The cartoon map that dominates the screen graphically depicts many of the higher-level concepts in the underlying organization of place indexes, e.g. manufacture-related places, water-related places, and roadway-related places. This map is actually the top level of a shallow hierarchy designed to organize place indexes in a simple and intuitive manner. When a user double-clicks on one of the labeled items, the top-level cartoon map is replaced by a subsequent map that offers a selection of more specific or related place indexes to choose from. For example, double-clicking on the item labeled "Water Places" brings up a new graphic containing items for oceans, rivers, lakes, islands, beaches, fishing piers, and harbors, among others. In this manner, the user can Zoom to any particular place index that is present in the system.

When a user single-clicks on a particular place, the system displays the lists of activities, people, and objects that are linked to the selected place index in their respective panes along the right edge of the screen. For example, single-clicking on the label "fishing pier" brings up lists which may include indexes for fishing, fishermen, and fish in the panes for activities, people, and objects. Double-clicking on one of these labels replaces items in the pane with a list of more specific types or parts of the selected index. For example, double-clicking on the label "fish" offers a new selection of object indexes, which may include sharks, goldfish, or tuna, depending on the existence of clips for these types of fish in the case library. In this manner the user can quickly browse through lists of related indexes to see what the system has to offer.

At any point, the user has the option of selecting an index that is presented on the screen indicating the desire to retrieve the cases that are in the case-set of the index.

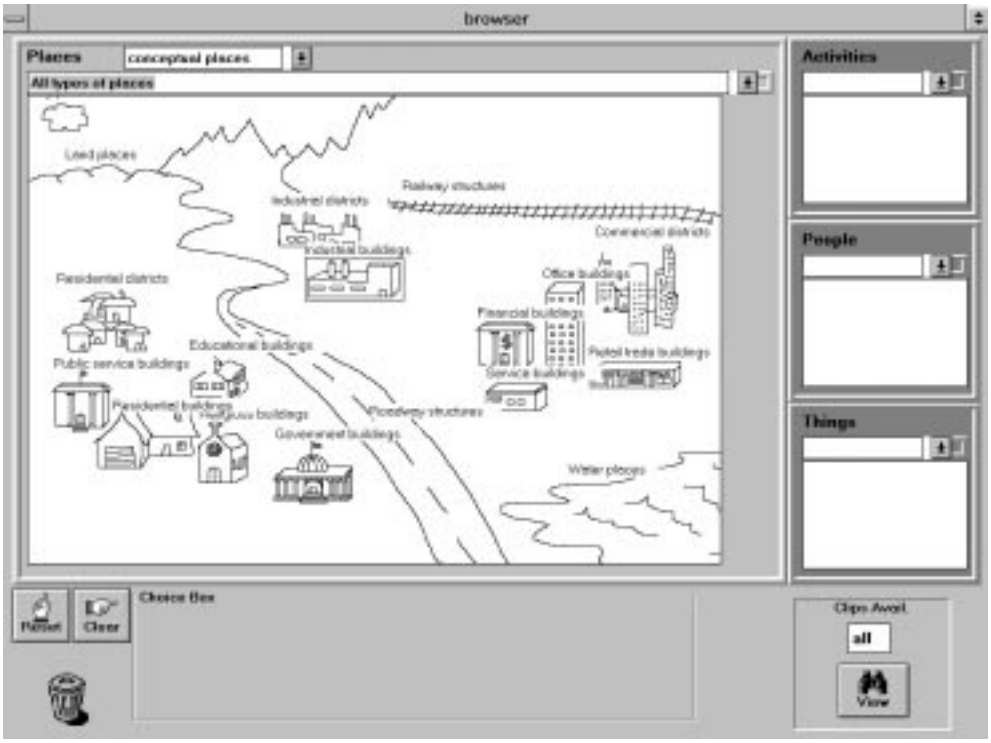


**Figure 1.** The retrieval system at the beginning of the index selection process.

When a user drags an index label from any of the four panes into the Choice Box, the system examines the index's case-set and informs the user of how many cases were retrieved. If the user has dragged more than one item into the Choice Box in order to request cases that contain multiple indexes, the number of cases given by the intersection of each index's case-set is displayed. To prevent the user from forming a request for multiple indexes for which there are no applicable cases, the selectability of each available index is calculated by the methods described in section 2. As the user moves their mouse pointer across the various indexes that are presented on the screen, the system visually informs the user of whether it can be dragged into the Choice Box given the indexes that are already there and prevents them from making an illegal selection.

As this retrieval system continues to develop we intend to improve both the organization and graphics of the shallow place hierarchy to find the simplest possible means of Zooming to specific places. As our conceptual index organizations develop and our case library grows, we intend to expand this approach to all of the categories of video indexes presented in section 3. We believe that the current version of our retrieval system provides an effective framework for the development of large-scale systems that meet the real-world needs of users of stock video databases.

## 6. Summary of Positions

Rich multimedia databases are useless without the ability to access the right piece of information at the right time. The need for intelligent information retrieval systems presents a new opportunity to apply research in artificial intelligence to large-scale real-world problems. By viewing the multimedia information as cases in a case library, we can attempt to apply the indexing and retrieval solutions developed in existing case-based reasoning systems to the problems facing users of multimedia databases. In particular, research on developing rich conceptual indexing frameworks for stories of the everyday social world seems to be directly applicable to the problem of indexing stock video clips. We present a simplified conceptual indexing framework that lacks the relational structure between index terms that exists in many other case retrieval systems, but meets the representational requirements of the stock video retrieval task. In simplifying the indexing framework, we have been able to focus our efforts on the creation of semantic networks that organize and catalog stock video indexes and the development of end-user indexing and retrieval interfaces. We offer six categories of indexes that are appropriate for stock video clips. One category, scene content indexes, has received the majority of our attention because of its importance to stock video users. In describing our early efforts at developing effective retrieval interfaces, we compare two different approaches to case retrieval: the Query and Search method and the Zoom and Browse method. Previous researchers have developed impressive, knowledge-rich retrieval systems that follow the Query and Search approach, but we argue that it is better to offer choices than to play guessing games. We present an early version of the Zoom and Browse interface that we are developing for the retrieval of stock video clips based on scene content indexes.

## Acknowledgments

Thanks to Jacob Mey, Lon Goldstein, Eric Lannert, Linda Wood, Raul Zaritsky, Andre van Meulebrock, and Anil Kulrestha who all contributed to this research. Special thanks to Andersen Telemedia for providing information and materials for this work. The Institute for the Learning Sciences was established in 1989 with the support of Andersen Consulting. The Institute receives additional support from Ameritech and North West Water, Institute Partners.

## References

Chakravarthy, A. (1994) Toward Semantic Retrieval of Pictures and Video. *AAAI-94 Workshop Program for Indexing and Reuse in Multimedia Systems* (pp. 12-18).

Cleary, C., & Bareiss, R. (1995) Automated Linking in Structured Hypermedia Systems. Unpublished draft.

Davis, M. (1994) Knowledge Representation for Video. *AAAI-94 Workshop Program for Indexing and Reuse in Multimedia Systems* (pp. 19-28).

Domeshek, E. (1992) *Do the Right Thing: A Component Theory for Indexing Stories as Social Advice*. PhD Thesis. Yale University.

Domeshek, E., & Kolodner, J. (1994) End-User Indexing of Design Lessons. *AAAI-94 Workshop Program for Indexing and Reuse in Multimedia Systems* (pp. 119-125).

Goldstein, E., Kedar, S., & Bareiss, R. (1993) Easing the Creation of a Multipurpose Case Library. *Papers from the 1993 workshop on Case-Based Reasoning* (pp. 12-18) Menlo Park, CA: AAAI Press.

Hampapur, A., Jain, R., & Weymouth, T. (1994) Digital Video Indexing in Multimedia Systems. *AAAI-94 Workshop Program for Indexing and Reuse in Multimedia Systems* (pp. 187-198).

Kolodner, J. (1993) *Case-Based Reasoning.* San Mateo, CA: Morgan Kaufman.

Lenat, D., & Guha, R. (1994) Strongly Semantic Information Retrieval. *AAAI-94 Workshop Program for Indexing and Reuse in Multimedia Systems* (pp. 58-68).

Riesbeck, C., & Schank, R. (1989) *Inside Case-Based Reasoning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R. (1982) *Dynamic Memory: A Theory of Reminding and Learning in Computers and People.* Cambridge, England: Cambridge University Press.

Schank, R., & Abelson, R. (1977) *Scripts Plans Goals and Understanding: An Inquiry into Human Knowledge Structures.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R., & Fano, A. (1992) A Thematic Hierarchy for Indexing Stories in Social Domains. Technical report no. 39. The Institute for the Learning Sciences, Nortwestern University, Evanston, IL.

Schank, R., Kass, A., & Riesbeck, C. (1994) *Inside Case-Based Explanation*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shafto, E., Bareiss, R., & Birnbaum, L. (1992) A Memory Architecture for Case-Based Argumentation. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 307-312).