Structuring Indexes for Video Clip

Eric A. Domeshek
Andrew S. Gordon

Institute for the Learning Sciences
Northwestern University
1890 Maple Avenue
Evanston, IL 60201
(708) 491-3500
Email: {domeshek | gordon }@ils.nwu.edu

Submitted to IMMI-1
First International Workshop on Intelligence and Multimodality in Multimedia Interfaces:
Research and Applications

1. The Problem: Providing Access to Video Clips

For the multimedia boom to amount to much, it must become easier for producers of such systems to find, manage, and organize large amounts of source materials in a variety of formats. Video materials are, in many ways, the most demanding component format in multimedia systems, being in effect multimedia presentations all on their own. Video clips combine moving images with sound, and may incorporate text and graphics as well. Everyone knows that the data streams for video tax current computers' low level storage and communication technologies. But the situation is even worse at higher levels -- technology for video analysis and retrieval practically does not exist at all. Machine vision and automated video analysis are still in a primitive state. More fundamentally, however, we do not know how we should describe videos that may contain scenes of nearly anything in the real (or imagined) world.

We are engaged in a research project that aims to improve the state of the art in how we formally describe the content and use conditions for video clips. In addition to descriptive formalism, we are also developing search mechanisms and user interfaces so that the entire package serves as an effective indexing and retrieval system for video clips. The representation, search mechanisms, and user interfaces are all necessary components if we hope to make video easily and appropriately available. Since September 1994, we have been working with a large-scale video and multimedia production shop to develop a system that will provide them easy on-line access to stock video clips accumulating in their library. They use these old clips to hold down development costs for new productions. Ultimately, however, work on video indexing schemes to help producers assemble multimedia systems will contribute to related indexing schemes that help end users find the most relevant material within any particular system.

While our initial explorations are limited to stock video clips, we expect to broaden our scope to consider most other kinds of video. The company we are working with divides its video inventory into four categories: stock clips; interview footage; graphics; and finished productions. Each of these types of content are used in different circumstances, have different internal structure, and may come with different supporting data. Interview footage, for instance, may be accompanied by a full text transcript; graphics will often contain recognizable words or may be generated using software that suggests some semantic content. We will explore these issues more in the future. For now, we are concentrating on ways of describing stock clips that relate to their potential use, with a special focus on the content of the pictured scenes.

2. Needed: Formalisms for Describing Video

We have actually identified six different categories of description for stock video clips, of which the content of the scene is only one. Each of these ways of describing a clip might be germane to the question of whether a clip can be used for a certain purpose in a new production. Each of the six forms of description is briefly summarized below.

2.1. The content of the scene

We expect that the content of a video clip will commonly be the most important index type for retrieval. 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. A more complete set of content indexes would include an interpretation of the goals and plans for the actors, and the outcomes and reactions.

2.2. 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 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 content items of the clip.

2.3. 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.

2.4. 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 classifying a clip's likely functions into categories such as transitions, interludes, prologues, background or hooks might serve as a useful complement to the description of the clip's contents.

2.5. 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.

2.6. 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.

3. Using Formal Descriptions as Indexes for Retrieval

The effectiveness of any strategy for retrieving multimedia information from large libraries depends on the labeling of individual items when they are introduced into the system. When items are accurately described, and those descriptions are well organized, effective search strategies are relatively straightforward to design. Our proposal for retrieving video clips focuses on exactly these issues of description and organization of descriptors.

In our video retrieval system, each video clip is individually analyzed by trained indexers who compose a description that will serve as the clip's index. Each index is a set of formalized terms; those terms correspond to concepts contributing to the several description types listed in the previous section. A large part of our current research effort is aimed at analyzing the conceptual space of each of these description types, deciding on terms that conjunctively cover this space, and organizing these terms according to the semantic relationships that exist between them. In the initial stages of our research, our attention has been focused primarily on the development of vocabulary and organizations for content descriptions, as we believe that this will be one of the most useful description types for stock footage retrieval. Even considering only content descriptions, the semantic relationships among index terms are of several different types and can serve different purposes.

We make use of the commonly represented type/subtype, part/whole, and container/contained-in relationships. One standard form of inference works quite well: taxonomic relationships support simple forward chaining subsumption inferences, e.g. if a clip is assigned the feature "pond" then it also gets the feature "body of water". On the other hand, the analogous inference for part/whole relationships does not work, e.g. if a scene shows a bicycle wheel, we cannot assume that the bicycle it came from is also part of the scene. Nor do other common plausible inferences hold with sufficient reliability: if a clip contains a whole we cannot assume it features all the (physical or temporal) parts; if clip contains some part, we cannot know that it also contains other related parts, e.g. if a clip shows a car, we cannot assume that it features views of the steering wheel; likewise if the clip shows a groom putting a ring on his bride, we cannot assume that it also shows the moment when the bride and groom kiss. Actually, the very fact that one item was explicitly coded while the other was not strongly suggests that the clip is not a good exemplar of the uncoded concept.

In addition to these generic relationships which apply to many different index categories, there are also interesting specific relationships that hold only between members of particular index categories. Content descriptions include index terms drawn from categories such as settings, activities, people, and things. There exist significant relationships between settings and activities, between activities and people, and between activities and things. Many activities stereotypically occur in particular settings. A particular activity will generally suggest particular types of people and things that might be involved in the activity. For example, a baseball game stereotypically occurs in such places as a baseball park or a stadium; a baseball game will include people playing roles such as pitcher and batter, and will include props such as bats, baseballs, and baseball gloves. Like most of the earlier generic relationships, these specific relationships do not support computer-controlled inference. However, as described in the next section, all these relationships, can be exploited by the retrieval system to provide support for user-directed search through a large conceptual space.

We should also note an important class of relationships that we are explicitly not attempting to represent: for now, it is a working hypothesis that retrieval need not be sensitive to the detailed relational structure of items composing a description. That is, while users might want to find scenes of a man buying a dog from a woman, they will not be too upset if they also get scenes with other combinations of "buying", "man", "woman", and "dog" (e.g. a woman buying a dog from a man). Discussions with video producers indicate that they are far more concerned about a system returning too few clips than about getting too many.

4. A User Interface to Support Indexed Retrieval of Video

The strategy that we have implemented to retrieve clips from our stock video library can best be characterized as user-directed search. Our approach is to allow users to browse the conceptual space for some type of description (e.g. content description) and incrementally select terms that conjunctively define a query. As each term is added to the query it winnows out all clips that do not include the new term among their descriptors. At any point, the clips

still in the running are those whose assigned index terms contain all the user-requested index features as a subset. The system can give immediate feedback on how the search is progressing, and can, in fact prevent the user from composing queries that will not retrieve any video.

For each term in the conceptual space, we compile a list of all the clips that are described in part by that term. The intersection of lists from two or more terms provides a list of clips whose description includes the conjunction of the terms. When a user begins to construct a query, all clips in the library are initially selected, but as each term is added to the query, the clips that are still in the running can be identified as the intersection of the list of clips for the new term and the previously selected clips. Because the intersection function is relatively inexpensive, even allowing for large sets of video clips, our browsing system can indicate to the user which branches of the conceptual space contain terms that will further limit the number of selected clips without overconstraining the retrieval.

This simple matching algorithm makes it relatively easy to give immediate retrieval feedback and to steer users away from dead ends, but a great deal of the system's usability will be determined by how well we facilitate the user's search for appropriate descriptive terms. This is where the relationships among descriptive concepts sketched in the previous section play their role -- in providing a coherently organized conceptual space of index terms. Terms that bear significant relationships to one another should be easily accessible from one another.

Our current prototype offers a cascading set of choices: first settings, then activities, then people and things. Choosing a setting winnows down the set of possible activities to those that typically occur in the chosen place; choosing an activity winnows down the set of people and things the user might want to consider. Within each of these categories, users can specify the particular concept they are seeking by walking down the part-of, contained-in, and type-of hierarchies formed by the relationships between terms. The effect of this approach is that users are able to explore neighborhoods of concepts to find clips that match or are related to the concepts they are initially seeking.

5. Conclusions

The utility of a multimedia library depends on the quality of its retrieval mechanism, which is in turn depends on the amount of emphasis placed on describing individual entries with useful indexes. For indexing stock video clips we have identified the six types of descriptions which we feel are most useful to video producers: content, point, composition/camerawork, narrative function, source information, and relationships to other clips. Where possible, each of these index types should be specified as a set of index terms which cover its conceptual space and a rich organization which captures the semantic relationships among terms. Carefully constructed organizations of concept terms can be used to support flexible retrieval mechanisms by allowing users to browse a conceptual space and select index terms to form a conjunctive query. Retrieval mechanisms can easily be made sensitive to conjunctions of descriptors and to subsumption relationships among terms, enabling the design of a system that helps users construct successful queries.

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 both the sponsorship and the data 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.