Task-Based Representation of Moving Images

Abby Goodrum
College of Information Science & Technology
Drexel University
goodruaa@dunx1.ocs.drexel.edu

What role does task play in the representation of visual information? Does the nature of the task at hand drive the need for either visual or textual representations for moving images? This paper is a report of research conducted to construct a task-based framework for the representation of moving images. This study measured the degree of congruence between moving images and their representations, both text and image based, in a non-retrieval environment with and without task constraints. Congruence in this study was defined as the degree to which human responses to representations mapped to human responses for the moving image being represented.

Introduction

While it is generally agreed that moving images comprise an important source of information that is structured and utilized differently from text, access to most collections of moving images has been grounded on bibliographic conventions; utilizing textual representations of stored items as well as textual queries. Although recent visual information retrieval systems are also capable of utilizing visual surrogates, there has been no systematic examination of the relationship between task specificity and the representativeness of text-based and image-based representations for moving images. Furthermore existing visual information retrieval theories have not incorporated models of the cognitive processing of semantic and pictorial entities. Assumptions of how to best represent moving images have been made without examining the underlying cognitive judgements that users bring to representations of moving images. Therefore, systems for the retrieval of moving images have been designed without an understanding of how different types of representations map to users' pictorial, and semantic processing of moving image information.

Background

Tasks

Tasks are an operationalization of the users' information-seeking problem. Image information needs have been examined by Hastings (1995), Jorgensen (1996), Turner (1995), Keister (1994), and O'Connor (1986). Marchionini (1995), Batley (1988), and Enser (1993, 1995) have also noted the role of task in driving visual information retrieval. Marchionini has characterized tasks according to their specificity and has noted that very specific tasks such as the acquisition of a specific word, date, or image, provides a high degree of confidence in the validity of search results. At the other end of the spectrum are tasks with low specificity such as the interpretation of information, or knowledge discovery tasks. These very general tasks provide for much less confidence in the validity of search results. Similarly, Enser has developed a theoretical model which characterizes visual information needs and representations as either linguistic or visual. Specific information needs are linked to linguistic characteristics, and general information needs are linked to pictorial attributes. Batley's work indicated that keyword searches were more likely to be used to retrieve specific information with a general decline in use with the decline in task specificity

Batley, Marchionini, and O'Connor (1993) have linked the specificity of task to the activity of browsing, and have theorized that very general, non-specific tasks require a greater degree of visual inspection of documents. In all cases, the specificity of task was linked to an increase in cognitive load, as it requires less cognitive effort to scan or browse than to construct complex queries. The need for highly structured representations such as text seems linked to the specificity of the information need.

Representation

An important factor in the successful retrieval of information is the extent to which document representations convey the information content of the documents for which they stand. In order to be effective, representations must convey the content of the original in such a way that users will make the same distinctions between representations that they would make between full documents.

With regard to systems for visual information retrieval, the problem with lack of representational congruency is heightened by the utilization of text for visually directed information needs, and the utilization of images for needs which require the specificity of text. For example, a user seeking images which contain a certain texture, shading, or spatial relationship between objects will find it difficult to specify those attributes using text-based queries. Similarly, users who seek images of Abraham Lincoln may find it difficult to construct a query using image exemplars or image attributes such as shape or color solely. In order to understand how best to represent images for information retrieval, it is therefore necessary to understand something about the cognitive representation of images.

Visual Information Processing

Images are very dense information containers, yet the reading of them by humans is extremely rapid. In order to develop a framework for research into visual information retrieval, it is first necessary to understand how humans perceive images and the process by which they receive visual information and adjust their behavior on the basis of that information. One of the key concepts in understanding visual perception is the difference between the processing of semantic information and the processing of pictorial information. Numerous studies conducted in the areas of cognitive science and psychology have led some researchers to conclude that there are distinct areas in the brain which are used for the processing of visual input and for the processing of verbal input. Of specific interest to cognitive scientists has been the degree to which humans must translate image information into semantic symbols before it can be processed. As most human activity involves interaction between the semantic and pictorial systems, it is assumed that visual information processing and semantic information processing work closely with one another.

Studies conducted by Rosch (1973, 1974, 1975), and by Rosch et al. (1976), indicate that the cognitive processing of image information may occur either in the semantic memory or in the pictorial memory, and that the level of information required determines the direction that processing will take. In experiments where subjects were asked

to judge the similarity of image pairs or word-image pairs, the processing time for judging image pairs was significantly less than the time it took to judge word-image pairs. This indicates that a translation function must occur in order to process images semantically. Further research conducted by Rosch led her to theorize that the cognitive processing of images by humans occurs at three levels. At the basic object level, images map more quickly to other images when they are not constrained by semantic translation. At the subordinate level, processing occurs more quickly when it is translated first into semantic attributes. This level corresponds to the cognitive representation of images by class and hierarchy, as well as by proper names. At the third level, the superordinate level, image information is conceptualized as functionally pictorial in nature, and images map more quickly to other images without semantic translation. This level operates when images are classed by physical attributes such as shape and size, or spatial characteristics such as proximity and direction of several objects together. From this cognitive perspective comes the indication that there are levels of image object coding which map either to the pictorial or the semantic memory. Thus, the level of abstraction or specificity required determines the processing of image information in the brain. The cognitive processes involved in classifying object attributes seem to favor semantic level coding for attributes of class membership and naming of specific objects. Physical and spatial characteristics seem to favor a pictorial level coding which is distinct from the semantic level.

The question, then, is how to represent moving images such that the representation is capable of supporting particular types of tasks that lend themselves to different cognitive approaches to information retrieval? For tasks which are semantically driven, text-based representations for moving images should prove more useful than image-based representations. For tasks which are visually driven, representations for moving images will be perceived as most useful if they are image-based.

The tasks used in this study were modified from actual user queries and were chosen to exploit the contrast between information needs which can be represented with text, and information needs which cannot. The specific task was constructed to correspond to what

Rosch (1974) terms the subordinate level and Panofsky (1962) terms the pre-iconographical level of object description. The general task was constructed to correspond to what Rosch terms the superordinate level and what Panofsky terms the iconographical level of object description

Research Objectives

The purpose of this study was to evaluate the representativeness of text-based and image-based representations in order to construct a framework for the representation of moving images which would incorporate task effects. This investigation was done to test 1) whether image-based representations scale with greater congruence to moving images than text-based representations overall, 2) whether a semantically driven (specific) task forces increased congruence between scaling of text-based representations and moving images, and 3) whether a visually driven (general) task forces increased congruence between scaling of image-based representations and moving images.

Research Method

Two experimental variables were arranged in a 5x3 factorial design. The first variable, type of representation, consisted of (1) Full moving image document, (2) Titles (3) Index terms (4) Keyframes and (5) Salient stills. The second variable, task, consisted of (1) No task, (2) General task, and (3) Specific task. This design yielded 15 cells/groups.

150 participants for this study were recruited from the population of students enrolled for at least one course (during the study) at the University of North Texas, and were randomly assigned to one of the attribute variable groups. Table 1 presents the organization of the participants and the attribute variable groups.

Table 1. Research Design Matrix

	Moving Images	Keywords	Titles	Keyframes	Salient Stills
No task	10	10	10	10	10
Specific/Semantic task	10	10	10	10	10
General/Visual task	10	10	10	10	10

Videotape selection

Twelve B-roll videotapes were selected from Cable News Network Image Source. (B-roll footage is the raw footage stored and used repeatedly to create edited packages). In order to narrow the selection, the broad subject category " Environment" was chosen and the search parameters were limited to just those tapes containing scenes with water. The sampling ratio was randomized, and the tapes were obtained with their full catalog record, description and index terms from CNN Image Source.

Preparation of Research Tape

The research tape was prepared from 10-second clips taken from each of the 12 CNN Image Source tape selections. The order of the presentation of each clip was randomized with regard to all other clips, and each clip was edited onto the research tape with a 5 second blue screen prior to and after each clip. This portion of the research tape provided initial exposure to the entire stimulus set of moving image documents. The second portion of the research tape presented all possible pairs of the clips in random order. The order of presentation within the pairwise grouping was also randomized. Each pair of clips was edited onto the research tape without audio. Each pair of clips was preceded by a 15 second blue screen "fixation field" (Carroll 1976). There was also a 5-second fixation field between each clip in the pair. The fixation field serves both to mark the end of each clip and to ready the viewer for the next stimulus set. Total viewing time for the initial exposure tape and all pairs and fixation fields was 56 minutes. This is well within the guidelines established for scaling visual stimuli by Young (1987). The 15-second fixation field between pairs of stimuli also provided time for the recording of judgments.

Preparation of Image-based Representations

Key Frames

NTSC standard VHS videotape is composed of 30 frames per second. Each of the 300 frames in each clip was digitized and analyzed for the following features using an image processing program: edge

intensity, edge slope, line length, line distance from the origin, and angles. Feature values summed into a single value representing the structure of each frame and the mean and standard deviation were calculated for all frames. For each video clip, four frames were taken from the tails of the curves and one frame was taken from the mean in order to obtain a set of 5 keyframes for each video clip.

Salient Stills

Salient video stills were derived by sampling each clip at the rate of one frame per second. No claim is made that this represents an optimal sampling ratio, however good results have been obtained using a lower sampling rate. (Rorvig 1993) The stills were submitted along with the video clips to two professional video indexers and three media librarians, who selected a single still image to represent each video clip. This activity was extended to three iterations in some cases until majority agreement was reached. The still image with highest agreement was retained as the salient still for each video clip.

Preparation of Text-based Representations

Titles

Titles for each of the clips were taken from the videotape titles provided by the cataloging records from CNN.

Keywords

Keywords were taken from the terms provided by the cataloging record provided by CNN. Five terms were used for each item. Where more than five terms existed in the catalog description, the clips and their terms were submitted to two professional video indexers, and three media librarians for ranking of the most representative terms. The five terms with highest agreement for each clip were used.

Construction of Tasks

Two tasks were chosen to represent a continuum from specific information needs such as those which can be expressed using key-

words in a precise search statement, to subjective information needs which are difficult to express in a search statement, and are dependent on characteristics of a scene as interpreted by an individual. These two extreme levels of the continuum were chosen in order to exploit this contrast between information needs which can be represented easily with text, and information needs which cannot. The tasks were adapted from existing user queries to the CNN database for environmental images. The tasks were necessarily modified after the research tape was edited in order to insure that images for the task were still present after editing.

Specific Task:

"You have been asked to gather information for a documentary about geysers. The director has asked you to find information on the Old Faithful Geyser in Yellowstone National Park."

General Task:

"You have been asked to gather information for a documentary about the environment. The director has asked you to find information that illustrates the fragility of our water resources."

Recording Similarity Judgments

Similarity judgments were obtained for the moving images and their representations where no task was specified. Subjects were asked to indicate the degree of similarity between the stimuli in each paired set by making a mark on a 5-inch line. The right-hand end of the line was labeled "Completely Different." The left-hand end of the line was labeled "Exactly the Same".

Recording Task Judgments

Task judgments were obtained for the moving images and their representations where a task was specified. Subjects were asked to indicate the degree of usefulness of each stimuli in a paired grouping by placing a mark on a 5 inch line The right-hand end of the line was labeled "Stimulus One" The left hand end of the line was labeled "Stimulus Two." The middle of the line was marked "Exactly the

Same." Where both stimuli were deemed equally useful or useless for the task, subjects were instructed to mark the middle of the line.

Recording Data for Analysis

Following the collection of all participants' judgments, their marks were converted to numeric values and entered into matrices for each group. The average values for each judgment within each group were scaled multidimensionaly using the ALSCAL program in SPSS. The resulting solutions were then analyzed for goodness of fit, and were plotted. Congruence of judgments for each type of representation by each task was analyzed for congruence against the judgments for the moving image documents

Results

For all multidimensional scaling in this study, 3 dimensions provided a good fit with the data. For each group, the first three dimensions described the bulk of the raw data.

The dispersion of judgments across groups was plotted and the congruity values were calculated as the differences between the root of the sum of the three squared axis points for each representation stimulus subtracted from the root of the sum of the three squared axis points for the moving image document stimulus points. As hypothesized, there was greater congruence between moving image documents and image-based representations overall. The introduction of a specific task, however, forced congruence between the text-based representations and the moving image document configuration. As can be seen from Table 2, image-based representations scaled with greater congruity overall. Keyframes exhibited the greatest number of congruent points (12), followed by Salient Stills (9), Keywords (8), and Titles (7). The two image-based representations, Salient Stills and Keyframes, demonstrated the highest degree of congruency when no associated task had been assigned.

Table 2. Representational Congruity by Task

Representation	No Task	Specific Task	General Task	Total
Salient Stills	6	1	2	9
Keyframes	4	4	4	12
Keywords	0	5	3	8
Titles	2	2	3	7

The image-based representations scaled with greater congruity for all No-Task groups with a total of 10 stimulus points, compared to 2 text-based stimulus points displaying congruity. The Specific Task resulted in increased congruity between the text-based representations and the Moving Image Documents. The Task-One Keywords and Titles account for a high degree of congruity on 7 stimulus points compared to 5 points for image-based congruity. General Task shifts congruity to a point of equilibrium between the text-based and image-based representations. Although Keyframes account for slightly more points of congruity, the difference is not great.

Chi-square was used to determine whether the frequency of occurrence for each representation was significant. The obtained x2 was significant at the level .001 $<$.01 for 2 df.

Discussion

Although the configuration of points for any one group do not match the moving image document configuration exactly, there is sufficient evidence to suggest that image-based representations scale with greater congruence to moving image documents than text-based representations overall. Twenty-one out of thirty- six possible points of congruity were image-based representations. Of these, 9 were salient stills and 12 were keyframes. Although only 7 out of 12 text-based same stimulus points scaled with proximity to the moving image document - specific task configuration, this is a notable increase in the number of text-based representations which shared congruence with the no-task group. Furthermore, this level of congruity is not achieved by text-based representations in any other group. This supports the hypothesis that a specific task forces increased

congruence between the text-based representations and moving image documents. It is important to note that the stimulus object of interest in the specific task, (Old Faithful Geyser, stimulus number 8) exhibits the greatest congruity between a text-based representation and the moving image document. Thus, the title: "Yellowstone," scaled with the greatest congruence to the moving image document for the task of identifying images of Old Faithful Geyser. This indicates support for the underlying theoretical basis for the hypothesis that humans process visual information semantically at the basic object naming level. Although keyframes performed slightly better than other representations for the general task, there is not sufficient support for the third hypothesis that given a general task, image-based representations scale with greater congruence to moving image documents than text-based representations.

Conclusions

Results of this study show that task specificity drives the representativeness of text-based representations for moving images. The findings indicate that the role of task in representing moving images constitutes a potentially important element in the visual information retrieval process. Further research is needed to explore the activity of browsing and its relationship to task specificity. Further research is also needed to model interactions between information seekers and representations provided in a visual information retrieval system (VIR).

References

Batley, S. 1988. Visual information retrieval: Browsing strategies in pictorial databases. *Proceedings of the 12th International Online Information Meeting. December 6-8,1988.* 373-381. London: Learned Information.

Carroll, J.M. 1976. Segmentation in cinema perception. *Science* 191: 1053 - 1054.

Chang, S. and R. Rice. 1993. "Browsing: A multidimensional framework." In *Annual Review of Information Science and Technology* 28: 231-276.

Enser, P.G.B. 1995. "Pictorial information retrieval." *Journal of Documentation*, 51: 126-170.

Hastings, S. 1995. "Query categories in a study of intellectual access to digitized art images." *Proceedings of the 58th Annual Meeting of the American Society for Information Science, Chicago, IL.* 32: 3-8.

Jorgensen, C. 1996. "Indexing images: Testing an image description template." *Proceedings of the 59th Annual Meeting of the American Society for Information Science, Baltimore, MD*: 209-213.

Keister, L.H. 1994. "User types and queries: Impact on image access systems." In *Challenges in Indexing Ilectronic Text and Images*, ed. R. Fidel and T. Bellardo. Medford, NJ: American Society for information Science.7-22.

Marchionini, G. 1995. *Information seeking in Electronic Environments.* Cambridge: Cambridge University Press

O'Connor, B. 1993. "Browsing: A framework for seeking functional information." *Knowledge: Creation, Diffusion, Utilization* 15: 211-232.

O'Connor, B. 1985. "Access to moving image documents: Background concepts and proposals for surrogates for film and video works." *Journal of Documentation*, 41: 209-20.

Panofsky, E. 1962. *Studies in the visual arts.* New York: Harper & Row.

Rosch, E. 1973. "Natural categories." *Cognitive Psychology* 4: 328-350.

Rosch, E. 1974. "Linguistic relativity." In *Human communication: Theoretical Explorations*, ed. A. Silverstein.Hillsdale, NJ: Lawrence Erlbaum. 95-121.

Rosch, E. 1975. "Cognitive representations of semantic categories." *Journal of Experimental Psychology: General*, 104: 192-233.

Rosch, E., C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. 1976. "Basic objects in natural categories." *Cognitive Psychology* 8: 382-439.

Rorvig, M.E. 1993. "A method for automatically abstracting visual documents." *Journal of the American Society for Information Science*, 44: 40-56.

Turner, J.I.M. 1995. "Comparing user-assigned terms with indexer-assigned terms for storage and retrieval of moving images: Research results." *Proceedings of the 58th Annual Meeting of the American Society for Information Science, Chicago, IL*,32: 9-12

Young, F.W. 1987. *Multidimensional Scaling. History, Theory, and Applications.* Hillsdale, NY: Lawrence Erlbaum Assoc.