Study of User Behavior in Image Retrieval and Implications for Content versus Concept Based Access

Leah Schultz
lschult@tarleton.edu
Computer Information Systems
Tarleton State University
Stephenville, Texas 76402, USA

Abstract

This paper explores the terms assigned by users to images for retrieval purposes in image databases. In order to determine how users conceptualize meaning for image retrieval, sixty-one participants provided potential retrieval terms for 40 images divided into 4 types of images. The categories include landscape, portrait, news, and city photography. The terms provided were analyzed for levels of meaning and relationships between terms supplied and the type of image described are explored. Results indicate significant findings in the level of meaning of terms assigned to images and relationships existed between the type of image viewed and the number and levels of terms provided. The implications for content and concept based retrieval schema are discussed.

Keywords: image retrieval, context based retrieval, concept based retrieval, databases

1. INTRODUCTION

With the proliferation of digital cameras and camera phones combined with the reduction in cost of producing images, the number of images being produced today will likely continue to grow. How these images are being stored and retrieved from databases has been a subject of interest for a number of professional fields from traditional library science to computer science. Originally, two camps emerged as image databases became prevalent. One side championed the manual assignment of retrieval terms by a trained professional using controlled vocabulary while the other faction researched the abilities of computers to determine the means of retrieval to images. In the recent growth of social media, other models have emerged that allow groups or communities to determine how images will be accessed. Key to this argument is the user experience and how the users actually interact with the databases and think about their approaches to retrieval. This research examines user interaction with images and how their behavior fits into these models of retrieval.

2. BACKGROUND

The first group embraces the traditions of library science and assigns text descriptors for the image, relying on traditional text-based information retrieval systems. As the field grew, best practices began to emerge and systems or rules such as the *Anglo-American Cataloging Rules, Revision 2*, *Art and Architecture Thesauraus,* and *Library of Congress Thesaurus of Graphic Materials* were developed to guide indexers in preparing images for retrieval (Jörgensen , 1999).This approach can be time consuming, expensive, and ineffective in many cases as demonstrated by studies of relevance in image retrieval (Shatford, 1984).

The second approach to image indexing, known as content based indexing, focuses on the use of computers to index images based on content. The computer can index images based on many of its physical characteristics such as shape, color, and texture. There have been many advances in this area and content-based retrieval has been particularly effective in scientific fields such as medicine and astronomy (Goodrum, Rorvig, Jeong, and Suresh 2001). Advances in facial recognition have increased greatly in the past few years and these algorithms are routinely used to determine meaning in images, in this case identity.(Moudani, Shahin, Chakik, Sayed, & Mora-Camino, 2011).

Many researchers are examining ways to find common links between the two fields and use the physical characteristics of an image to convey higher levels of meaning. Colombo, Del Bimbo, and Pala (1999) combine the fields of semiotics with the automatic detection of characteristics to convey meaning. The combination of colors and the resulting emotional states and the implementation of line slope by artists to denote meaning are being explored as possible advances in content based image indexing. However, the discussion of meaning in an image is one that continues to plague both camps and warrants a closer look in any discussion of image retrieval.

What an image is "of" maybe be a different discussion than what an image means. One implies a list of the contents of the image or the subject of the photograph, and one implies a higher level of meaning assigned either by the photographer or the viewer. A simple example that would demonstrate the difference between "of" and "about" might be an image of a young man, a young woman, and a small child posed together. The image is "of" three people of various ages, wearing different clothes, in a certain setting. However, many would also say that the image might be "of" a family, although there are assumptions made to get to this designation. Whether those assumptions are reasonable may depend on a wide variety of factors ranging from perceptual abilities to cultural differences and personal experience. This difference in meaning many times is the line in the sand between the two approaches to retrieving images from a database.

Many studies have been dedicated to determining how people perceive images, assign meaning, and search for images (Colombo, Del Bimbo and Pala,1999; Hastings, 1995; Choi & Rasmussen, 2003; Jörgensen ,2004; Jörgensen and Jörgensen; 2005). From these studies, models have emerged that try to assist in determining levels of meaning in images which can ultimately be useful in deciding how to provide access to images in databases. Shatford Layne (1986) expanded on Panofsky's (1955) work in art history for the purpose of developing a model to enhance indexing of images. She loosely adapts Panofsky's three levels in a model that includes, "generic of", "specific of", and about. These three levels of interpretation are then applied to the facets of who, what, where and when. For example, an image at the generic of level might list the contents of an image: a baby, a mother. The specific of might add additional levels of cultural interpretation to include the virgin Mary and the baby Jesus. The highest level of interpretation deals in more abstract concepts which might lead to assigning terms such as salvation. In the discussion of approaches to retrieval, many would argue that content based access would be more effective at the generic level and would be more problematic at the higher levels when the computer was required to provide more interpretation.

Ultimately, however, an understanding of how the user interacts with the database, the data, and sets about the task of retrieving images is an important factor that cannot be overlooked. Using a previously developed model of meaning, search terms provided by participants were analyzed for trends in levels of meaning, from named objects to abstract ideas. In addition to the terms provided, relationships were examined to determine if certain types of images were more likely to generate terms at different levels in Shatford Layne's model. This information could then be used to evaluate improvements in content based retrieval algorithms and their effectiveness in providing access in image databases.

3. METHODOLOGY

In this study, 61 undergraduate students at a public university participated in the data collection phase of the study. Students were not required to have any specific subject knowledge to participate and students were recruited from all backgrounds. All of the students who participated were undergraduate students with an average age of 22.4 years of age (*SD*=4.9). Forty-eight of the students were Caucasian, five African American, four Hispanic, three Asian, and one participant did not disclose their ethnicity. The group was very closely distributed by gender with 29 females, 31 males, and one participant who declined to identify gender.

The students represented a variety of academic backgrounds with the largest number coming from the computer information systems department (*n*=19). Participants rated their computer familiarity with more than half rating themselves at least somewhat familiar with computers. Specifically, 25 students rated themselves as very comfortable, 14 somewhat comfortable, 7 neutral, 5 somewhat uncomfortable, and 10 rated themselves as very uncomfortable.

Participants were presented with 40 images on the web, one at a time, and were instructed to provide terms that would be used to retrieve the image being viewed. The entire task took students 30 minutes to 1 hour to complete. Images were selected for this study from a variety of sources encompassing government collections, news agency images, stock photography resources, and images from the researcher's personal collection. The images did not focus on one specific subject area and did not require expert subject knowledge in a field such as art history. Images were selected to represent different categories to determine if there may be relationships between the type of image being viewed and the user's approach to retrieval. The categories used for this study include landscape photography, portrait photography, cityscapes photographs, and news photography. The selection of these categories represents two general types of images, detailed and non-detailed images. Generally speaking, the portrait and landscape images have fewer details included as compared to the news photography and the city scenes.

Once data collection sessions were complete, the data was downloaded into a spreadsheet and two coders working independently analyzed each term provided and assigned it a level of meaning according to Shatford's model. Data that coders did not agree on were not included in the data analysis. However, very few terms were excluded due to the high level of agreement between coders (98.7%).

4. RESULTS

The mean of terms submitted by participants was 170.97 terms per student. There was a mean of 260.73 terms submitted per image viewed. In order to determine if there was a difference in the means of terms supplied for the different image types, a one-way analysis of variance was performed. The lowest mean number of terms supplied was for the images in the portrait category followed by images in the landscape category. The two highest means belonged to the news and the cityscape categories of images. The results of the ANOVA indicate that there is significant difference between the means of the categories being studied, $F(3,36)=6.24$, $p<.01$.
Using the Tukey HSD post-test to determine the nature of the relationships between the means of the groups, it was determined that the mean of the city scene images was significantly different than the two groups of portrait and landscape ($p<.05$) There was no significant difference between the means of the remaining groups.

When looking at the number of terms that fell into the levels of meaning in Shatford Layne's model, 9,924 provided were at the generic level, 244 at the specific level, and 21 at the abstract level. A chi-square test indicates a relationship between the type of image being viewed and the level of meaning supplied in Shatford Layne's model ($x^2(6, N = 10279) = 224.89, p < .01$) Looking at the relationships, the cityscape images were the only category that showed a lower than expected number of generic terms. In the specific category, the cityscape category showed more than twice the expected count while landscape and portrait did not even receive half the number of expected specific terms. The news category received very close to the expected count in each category. Because there were so few abstract terms provided, it is difficult to determine if a relationship might exist in this category.

Limitations

Because the research is based on a small sample size taken from a convenient sample, the results of the study are limited. The exclusive use of college students as participants could also affect the generalization of results to a wider population.

5. CONCLUSIONS

The results from the study indicate a heavy reliance on named objects in terms provided by students. The overwhelming number of generic level terms supplied indicate that participants tended towards answering the question "What is the image of?" more than they indicated "What the image is about?" The promise in these results indicate that as improvements in computer analysis of images holds great promise in providing less expensive and less time consuming access to images than the traditional approaches of library science. As computer scientists continue to improve in areas such as facial recognition or pattern recognition in medical imagery, these concepts could also be adapted to identify objects in images for retrieval purposes.

However, one cannot ignore the fact that there were still many terms provided at the higher levels of meaning that could be difficult for computers to accurately assign. One of the most common examples would be the assignment of an emotion to an image, particularly those in the portrait category. Could a computer be programmed to analyze the nuances of facial appearances to determine the difference between a grimace of fear and a smile? Many researchers are making progress in this area but in a general subject image database, cost may be prohibitive in using content based approaches (Sarode & Bhatia, 2010).

Where the real promise lies in the marriage between concept and content based image analysis is the availability of resources via the internet and other databases. For example, a news photograph that showed a picture of George W. Bush received many terms with the specific level name and also many participants assigned the generic term "president". In this example, the specific term might be more easily accomplished by facial recognition than the more generic term which indicates some level of interpretation in what position the person holds. However, these obstacles can be easily overcome with integration of other resources. In a generic image database, facial recognition software could identify the individual and retrieve related concepts from other resources that have been created through more traditional routes. So in this example, the person George Bush is indentified automatically and his role of president or governor of Texas might be retrieved from a concept based controlled vocabulary such as the Library of Congress or even something less formal such as Wikipedia.

A similar technology solution might be applied to common designations for some of the results found in the landscape and city images. It was not uncommon for participants in the study to provide a geographical location for the image, even though it was many times incorrect. With inclusion of GPS enabled devices in smart phones and potentially in other cameras in the future, the GPS metadata attached to digital images can easily be tied to a database to locate the actual geographic name of the place where the image was taken. In addition to retrieving this information, databases could again be shared to provide additional retrieval mechanisms. For example, a picture taken at Shea Stadium in New York could easily retrieve New York City as a geographical access term but also could retrieve similar terms such as baseball or Mets without any assistance from a human indexer.

With the differences in terms provided for the type of image, it may also indicate to the database manager what type of access might be best suited to the content. With the higher number of terms provided for the more detailed images, the time and money might be well spent creating recognition algorithms for databases containing these types of images such as the cityscapes or news photography. These might be less useful and it may be more appropriate to apply more traditional means of providing access in databases that contain portraiture or landscape photography.

One area that poses difficulties for both approaches to providing access to images is the user's personal interaction with the image. Arnheim (1969) and Fischler and Firschein (1987) discuss the interaction of perception with other cognitive abilities such as language, memory, as well as culture when discerning meaning in an image. These influences also appeared in the responses to the tasks. This seems to be the case when terms were assigned that seemed to have no apparent reason for its appearance. A student who assigned 'beer bottle' to the image of the beach and one who assigned the term 'viagra commercial' to an image of an older couple hugging demonstrate the affect of previous memories and culture. Because these elements are based on the experiences and memories of the viewer, providing this level of access would be difficult for either approach

Although the data in this study provides additional information on how users assign descriptions in an image, translating this information can be difficult when considering search strategies. As Markulla and Sormunen (2000) suggest, the selection of search terms is also influenced by the ultimate use of the image. Participants were asked to assign descriptive terms to images without a context for the activity. The potential exists for a similar difference in terms supplied as seen in a search arena. A quick, small sample of terms assigned to images in two different Internet databases were studied to see if a similar phenomena may exist. A quick, non-scientific review of images in popular Internet image databases shows possible support of this idea. At the website Flickr®, where website users upload their photos to share with other users on the site, the creator of the content is asked to assign the terms for their images. Selecting 40 images at random using the recent photos page and studying the number of terms assigned by the user shows that the average number of terms provided by the creator of the photograph was 8.3. Doing a similar sampling from an online stock photography site where users also supply the search terms there are very different results. The purpose of this website is to sell photography through a community website and the easier it is to find a photograph, the more likely it is to be purchased. As might be expected, the average was much higher at 33.6 terms on average assigned to each photograph. This difference in purpose of the database demonstrates that the lack of context of this study may also have an effect on the number of terms being supplied by participants.

As images continue to be created in large numbers, the storage of said data in databases for retrieval continues to pose problems for professionals in many fields. Attempts to understand how users conceptualize images and potentially attempt to retrieve them should be a driving force in the approaches from both traditional fields of library science as well as more technical approaches of computer science. Interconnectivity between resources produced by both camps could potentially bring higher success in image access. Advancements in these areas have wide reaching implications for image retrieval beyond specific subject image databases. Currently, many internet search engines and networking sites depend on the provider of the content to provide retrieval access to image data or depend on textual analysis of the document which includes the image. Again, understanding how users search for images and cooperation between both content and concept based retrieval paradigms to solve these problems potentially can be applied to solve image retrieval problems in small subject specific databases or immense databases such as web search engines.

6. REFERENCES

Arnheim, R. (1969) *Visual thinking*. Univeristy of California Press, Berkeley.

Choi, Y. & Rasmussen, E. (2003). Searching for images: The analysis of users' queries for image retrieval in American history. *Journal of the American Society for Information Science and Technology*, *54*(6), 498-511.

Colombo, C., Del Bimbo, A. & Pala, P. (1999). Semantics in visual information retrieval. *IEEE Multimedia*, *6*(3), 38-53.

Fischler, M. & Firschein, O. (1987). The eye, the brain, and the computer. Addison-Wesley, Boston.

Goodrum, A.A., Rorvig, M.E., Jeong, K., & Suresh, C. (2001). An open source agenda for research linking text and image content features. *Journal of the American Society for Information Science and Technology*, *52*(11), 948-953.

Hastings, S.K. (1995). Query categories in a study of intellectual access to digitized art images. *Proceedings of the 58th Annual Meeting of the Society for Information Science, ASIS '95*, *32*, 3-8.

Jörgensen , C. (1999). Access to pictorial material: A review of current research and future prospects. *Computers and the Humanities*, *33*, 293-318.

Jörgensen, C. (2004). The visual indexing vocabulary: Developing a thesaurus for indexing image across diverse domains. *Proceedings of the 67th ASIS&T Annual Meeting*, *41*, 287-293.

Jörgensen, C. & Joregensen, P. (2005). Image querying by image professionals. *Journal of the American Society for Information Science and Technology*, *56*(12), 1346-1359.

Markkula, M. & Sormunen, E. (2000). End-user searching challenges indexing practices in the digital newspaper photo archive. *Information Retrieval*, *1*, 259-285.

Moudani, W., Shahin, A., Chakik, F., Sayed, A. & Mora-Camino, F.(2011). An Efficient Approach for Image Recognition using Data Mining. *International Journal on Computer Science & Engineering*, 3(1), 55-68.

Panofsky, E. (1955). *Meaning in the visual arts: Papers in and on art history*. Garden City, NY: Doubleday, Inc.

Sarode, N. & Bhatia, S. (2010) International Journal on Computer Science & Engineering, 2(5), 1552-1557.

Shatford, S. (1984). Describing a picture: A thousand words are seldom cost effective. *Cataloging and Classification Quarterly*, *4*(4), 13-30.