2005

Exploiting Captions for Multimedia Data Mining / Chapter in Encyclopedia of Multimedia Technology and Networking

Rowe, Neil C.

Monterey, California. Naval Postgraduate School

Exploiting Captions for Multimedia Data Mining

Neil C. Rowe
U.S. Naval Postgraduate School, USA

ABSTRACT

Captions are essential accompaniments to multimedia data objects as a way to facilitate their data mining. This article describes the kinds of possible captions and the task of recognizing them. It then discusses the forms of the caption-object relationship and the ways in which components of the two can be matched. Automatic generation of captions is also discussed. While no breakthroughs are expected on this subject, captioned multimedia will become increasingly important in future systems as increases in processing speeds enable it to be more easily used.

This is a chapter in the *Encyclopedia of Multimedia Technology and Networking*, ed. M. Pagani, Hershey, PA: The Idea Group, 2005.

INTRODUCTION

Captions are text that describes some other information; they are especially useful for describing non-text media objects (images, audio, video, and software). Captions are valuable metadata for managing multimedia since they help users better understand and remember (McAninch, Austin, & Derks, 1992-1993) and permit better indexing of media. Captions are essential for effective data mining of multimedia data, since only a small amount of text in typical documents with multimedia -- 1.2% in a survey of random World Wide Web pages (Rowe, 2002) -- describes the media objects. Thus standard Web browsers do poorly at finding media without knowledge of captions. Multimedia information is increasingly common in documents as computer technology improves in speed and ability to handle it, and people need multimedia for a variety of purposes like illustrating educational materials and preparing news stories.

Captions are also valuable because nontext media rarely specify internally the creator, date, or spatial and temporal context, and cannot convey linguistic features like negation, tense, and indirect reference. Furthermore, experiments with users of multimedia-retrieval systems show a wide range of needs (Sutcliffe et al, 1997) but a focus on media meaning rather than appearance (Armitage & Enser, 1997). This suggests that content analysis of media is unnecessary for many retrieval situations, which is fortunate because it is often considerably slower and more unreliable than caption analysis. But using captions requires finding them and understanding them. Many captions are not clearly identified, and the mapping from captions to media objects is rarely easy. Nonetheless, the restricted semantics of media and captions can be exploited.

FINDING, RATING, AND INDEXING CAPTIONS

Background

Much text in a document near a media object is unrelated to that object, and even text explicitly associated with an object may often not describe it (like "JPEG picture here" or "Photo39573"). Thus we need clues to distinguish and rate a variety of caption possibilities and words within them, allowing there may be more than one caption for an object or more than one object for a caption. Free commercial media search engines (like images.google.com, multimedia.lycos.com, and www.altavista.com/image) use a few simple clues to index media, but their accuracy is significantly lower than that for indexing text. For instance, (Rowe, 2005) reported that none of five major image search engines could find pictures for "President greeting dignitaries" in 18 tries. So research is exploring a broader range of caption clues and types (Sclaroff et al, 1999; Mukherjea & Cho, 1999).

Sources of Captions

Some captions are explicitly attached to media objects in adding them to a digital library or database. On Web pages, HTML "alt" and "caption" tags also explicitly associate text with media objects. Clickable text links to media files are another good source of captions since the text must explain the link. The name of a media itself can be a short caption (like "socket_wrench.gif"). Less-explicit captions use conventions like centering or font changes to text. Titles and headings preceding a media object can sometimes serve as captions as they generalize over a block of information. Paragraphs above, below, or next to media can also be captions, especially short paragraphs.

Other captions are embedded directly into the media, like characters drawn on an image (Lienhart & Wernicke, 2002) or explanatory words at the beginning of audio. These require specialized processing like optical character recognition to extract. Captions can be attached through a separate channel of video or audio, as with the "closed captions" associated with television broadcasts that aid hearing-impaired viewers and students learning languages. "Annotations" can function like captions though they tend to emphasize analysis or background knowledge.

Cues for Rating Captions

A caption candidate's type affects its likelihood, but many other clues help rate it and its words (Rowe, 2005):

- Certain words are typical of captions, like those having to do with communication, representation, and showing. Words about space and time (like "west", "event", "above", and "yesterday") are good clues too. Negative clues like "bytes" and "page" can be equally valuable, as indicators of text unlikely to be captions. Words can be made more powerful clues by enforcing a limited or "controlled" vocabulary for describing media, like what librarians use in cataloging books (Arms, 1999), but this requires cooperation from caption writers and is often impossible.
- Position in the caption candidate matters: Words in the first 20% of a caption are four times more likely to describe a media object than words in the last 20% (Rowe, 2002).
- Distinctive phrases often signal captions, like "the X above", "you can hear X", and "X then Y" where X and Y describe depictable objects.
- Full parsing of caption candidates (Srihari & Zhang, 1999; Elworthy et al, 2001) can extract more detailed information about them, but is time-consuming and prone to errors.
- Candidate length is a clue since true captions average 200 characters, with few under 20 or over 1000.
- A good clue is words in common between the candidate caption and the name of the media file, as for "Front view of woodchuck burrowing" and image file "northern_woodchuck.gif".
- Nearness of the caption candidate to its media is actually not a clue (Rowe, 2002) since much nearby text in documents is unrelated.
- Some words in the name of a media file affect captionability, like "view" and "clip" as positive clues and "icon" and "button" as negative clues.
- "Decorative" media objects occurring more than once on a page or three times on a site are 99% certain not to have captions (Rowe, 2002). Text generally captions only one media object except for headings and titles.
- Media-related clues are the size of the object (small objects are less likely to have captions) and the file format (e.g. JPEG images are more likely to have captions). Other clues are the number of colors and the ratio of width to length for an image.
- Consistency with the style of known captions on the same page or at the same site is also a clue because many organizations specify a consistent "look and feel" for their captions.

Quantifying Clues

Clue strength is the conditional probability of a caption given appearance of the clue, estimated from statistics by c/(c+n) where c is the number of occurrences of the clue in a caption and n is the number of occurrences of the clue in a noncaption. In a representative sample, clue appearances is a binomial process with expected standard deviation $\sqrt{cn/(c+n)}$. This can be used to judge whether a clue is statistically significant. Recall-precision analysis can also compare clues; (Rowe, 2002) showed that text-word clues were the most valuable in identifying captions, followed in order by caption type, image format, words in common between the text and the image filename, image size, use of digits in the image file name, and image-filename word clues.

Methods of data mining (Witten & Frank, 2000) can combine clues to get an overall likelihood that some text is a caption. Linear models, Naive-Bayes models, and case-based reasoning have been used. The words of the captions can be indexed, and the likelihoods can be used by a browser to sort media, for presentation to the user, that match a set of keywords.

MAPPING CAPTIONS TO MULTIMEDIA

Background

Users usually interpret media data as "depicting" a set of objects (Jorgensen, 1998). Captions can be:

- *Component-depictive*: The caption describes objects and/or processes that correspond to particular parts of the media. For instance, a caption "President speaking to board" with a picture that shows a President behind a podium with several other people. This caption type is quite common.
- *Whole-depictive*: The caption describes the media as a whole. This is often signalled by media-type words like "view", "clip", and "recording", as for instance "Tape of City Council 7/26/04" with some audio. Such captions summarize overall characteristics of the media object and help distinguish it from others. Adjectives are especially helpful, as in "infrared picture", "short clip", and "noisy recording"; they specify distributions of values. Dates and locations for associated media can be found in special linguistic formulas.
- *Illustrative-example*: The media presents only an example of the phenomenon described by the caption, as for instance "War in the Gulf" with a picture of tanks in a desert.
- *Metaphorical:* The media represents something related to the caption but does not depict it or describe it, as for instance "WWII novels" with a picture of tanks in a desert.
- *Background*: The caption only gives background information about the media, as for instance "World War II" with a picture of Winston Churchill. *National Geographic* magazine often uses caption sentences of this kind after the first sentence.

Media Properties and Structure

The structure of media objects can be referenced by component-depictive caption sentences. Then valuable information is often contained in the subobjects of a media object that captions do not convey. Images, audio, and video are multidimensional signals for which local changes in the signal characteristics help segment them into sub-objects (Aslandogan & Yu, 1999). Color or texture changes in an image suggest separate objects; changes in the frequency-intensity plot of audio suggest beginnings and ends of sounds; and many simultaneous changes between corresponding locations in two video frames suggest a new shot (Wactlar et al, 2000). But segmentation methods are not especially reliable. Also, some media objects have multiple colors or textures, like images of trees or human faces, and domain-dependent knowledge must group regions into larger objects.

Software can calculate properties of segmented regions and classify them. (Mezaris, Kompatsiaris, & Strinzis, 2003) for instance classifies image regions by color, size, shape, and relative position, and then infers probabilities for what they could represent. Additional laws of media space can rule out possibilities, noting that objects closer to a camera appear larger, and gravity is downward so support of objects can be found (as for people on floors). Similarly, the pattern and duration of speech in audio can suggest what is happening. The subject of a media object can often be inferred, even without a caption, since subjects are typically near the center of the media space, not touching its edges, and well distinguished from nearby regions in intensity or texture.

Caption-Media Correspondence

While finding the caption-media correspondence for component-depictive captions can be difficult in general, there are easier subcases. One is the recognition and naming of faces in an image (Satoh, Nakamura, & Kanda, 1999). Another is captioned graphics since their structure is easier to infer (Elzer et al, 2005).

In general, grammatical subjects of a caption often correspond to the principal subjects within the media (Rowe, 2005). For instance, "Large deer beside tree" has grammatical subject "deer" and we would expect to see all of it in the picture near the center, whereas "tree" has no such guarantee. Exceptions are undepictable abstract subjects as in "Jobless rate soars". Present-tense principal verbs and verbals can depict dynamic physical processes, such as "eating" in "Deer eating flowers", and direct objects of such verbs and verbals are usually fully depicted in the media when they are physical like "flowers". Objects of physical-location prepositions attached to the principal subject are also depicted in part (but not necessarily as a whole). Subjects that are media objects like "view" defer viewability to their objects. Motion-denoting words can be depicted directly in video, audio, and software, rather than just their subjects and objects. They can be translational (e.g. "go"), configurational ("develop"), property-changing ("lighten"), relationship-changing ("fall"), social ("report"), or existential ("appear").

Captions are "deictic", a linguistic term for expressions whose meaning requires information outside the expression itself. Spatial deixis refers to spatial relationships between objects or parts of objects, and entails a set of physical constraints (DiTomaso et al, 1998; Pineda & Garza, 2000). Spatial-deixis expressions like "above" and "outside" are often "fuzzy" in that they do not define a precise area but associate a probability distribution with a region of space (Matsakis et al, 2001). It is important to determine the reference location of the referring expression, usually the characters of the text itself, but can be previously referenced objects like "right" in "the right picture below". Some elegant theory has been developed, although captions on media objects that use such expressions are not especially common.

Media objects also can occur in sets with intrinsic meaning. The media can be a time sequence, a causal sequence, a dispersion in physical space, or a hierarchy of concepts. Media-object sets can also be embedded in other sets. Rules for set correspondences can be learned from examples (Cohen, Wang, & Murphy, 2003).

For deeper understanding of media, the words of the caption can be matched to regions of the media. This permits calculating the size and contrast of media sub-objects mentioned in the caption, recognizing the time of day when it is not mentioned, and recognizing additional unmentioned objects. Matching must take into account the properties of the words and regions and constraints on them (Jamieson et al, 2006). Statistical methods can be used, except that there are many more categories, entailing problems of obtaining enough data. Some help is provided by knowledge of the settings (Sproat, 2001). Machine-learning methods can learn the associations between words and types of image regions (Barnard et al, 2003; Roy, 2000/2001).

Generating Captions

Since captions are valuable, it is important to obtain good ones. The methods described above for finding caption candidates can be used to collect text for a caption when an explicit one is lacking. Media content analysis can also provide information that can be paraphrased into a caption; this is most possible with graphics images. Discourse theory can help make captions sound natural by providing "discourse strategies" such as organizing the caption around one media attribute that determines all the others, like the department in a budget diagram (Mittal et al, 1998). Then guidelines about how much detail the user wants, together with a ranking of the importance of specific details, can be used to assemble a reasonable set of details to mention in a caption. Automated techniques can find keywords for captions, if not captions, by clustering media objects with known captions (Pan et al, 2004). Captions can also be made "interactive" so changes to them cause changes in corresponding media.

FUTURE TRENDS

Future multimedia-retrieval technology will not be dramatically different, although multimedia will be increasingly common in many applications. Captions will continue to provide the easiest access via keyword search, and caption text will remain important to explain media objects in documents. But improved media content analysis (aided by speed increases in computer hardware) will increasingly help in both disambiguating captions and mapping their words to parts of the media object. Machine-learning methods will be increasingly used to learn the necessary associations.

CONCLUSIONS

Captions are essential tools to managing and manipulating multimedia objects as one of the most powerful forms of metadata. A good multimedia data-mining system needs to include them and their management in its design. This includes methods for finding them in unrestricted text as well as ways of mapping them to the media objects. With good support for captions, media objects are much better integrated with the traditional text data used by information systems.

REFERENCES

Armitage, L. H., & Enser, P. (1997). Analysis of user need in image archives. *Journal of Information Science, 23*(4), 287-299.

Arms, L. (1999, Fall). Getting the picture: observations from the Library of Congress on providing access to pictorial images. *Library Trends, 48*(2), 379-409.

Aslandogan, Y., & Yu, C. (1999, January-February). Techniques and systems for image and video retrieval. *IEEE Transactions on Knowledge and Data Engineering, 11* (1), 56-63.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., & Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research, 3*, 1107-1135.

Cohen, W., Wang, R., & Murphy, R. (2003, August). Understanding captions in biomedical publications. *Proceedings of International Conference on Knowledge Discovery and Data Mining*, Washington, DC, 499-504.

DiTomaso, V., Lombardo, V., & Lesmo, L. (1998). A computational model for the interpretation of static locative expressions. In Oliver, P., & Gapp, K.-P. (Eds.), *Representation and processing of spatial expressions* (pp. 73-90). Mahwah, NJ: Lawrence Erlbaum.

Elworthy, D., Rose, T., Clare, A., & Kotcheff, A. (2001, June). A natural language system for retrieval of captioned images. *Natural Language Engineering, 7* (2), 117-142.

Elzer, S., Carberry, S., Chester, D., Demir, S., Green, N., Zukerman, I., & Trnka, K. (2005, June). Exploring and exploiting the limited utility of captions in recognizing intention in information graphics. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,*, Ann Arbor, Michigan, 223-230.

Jamieson, M., Dickinson, S., Stevenson, S., & Wachsmuth, S. (2006, June). Using language to drive the perceptual grouping of local image features. *Proceedings of the Conference on Computer Vision and Pattern Recognition*, New York, New York, 2102-2109.

Jorgensen, C. (1998). Attributes of images in describing tasks. *Information Processing and Management*, *34*(2/3), 161-174.

Lienhart, R., & Wernicke, A. (2002). Localizing and segmenting text in video, images, and Web pages. *IEEE Transactions on Circuits and Systems for Video Technology, 12* (4), 256-268.

Matsakis, P., Keller, J., Wendling, L., Marjarnaa, & Sjahputera, O. (2001, August). Linguistic description of relative positions in images. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, 31*(4), 573-588.

McAninch, C., Austin, J., & Derks, P. (1992-1993, Winter). Effect of caption meaning on memory for nonsense figures. *Current Psychology Research & Reviews, 11*(4), 315-323.

Mezaris, V., Kompatsiaris, I., & Strinzis, M. (2003, September). An ontology approach to object-based image retrieval. *Proceedings of International Conference on Image Processing*, II: 511-514.

Mittal, V., Moore, J., Carenini, J., & Roth, S. (1998, September). Describing complex charts in natural language: a caption generation system. *Computational Linguistics, 24* (3), 437-467.

Mukherjea, S., & Cho, J. (1999). Automatically determining semantics for World Wide Web multimedia information retrieval. *Journal of Visual Languages and Computing*, *10*, 585-606.

Pan, J.-Y., Yang, H.-J., Faloutsos, C., & Duygulu, P. (2004, August). Automatic multimedia cross-modal correlation discovery. *Proceedings of the Conference on Knowledge Discovery in Data*, Seattle, Washington, 653-658.

Pineda, L., & Garza, G. (2000, June). A model for multimodal reference resolution. *Computational Linguistics, 26*(2), 139-193.

Rowe, N. (2002, July/August). MARIE-4: A high-recall, self-improving Web crawler that finds images using captions. *IEEE Intelligent Systems*, *17*(4), 8-14.

Rowe, N. (2005). Exploiting captions for Web data mining. In Scime, A. (Ed.), *Web mining: applications and techniques* (pp. 119-144). Hershey, PA: The Idea Group.

Roy, D. K. (2000/2001). Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication, 4*(1), 33-56.

Satoh, S., Nakamura, Y., & Kanda, T. (1999, January-March). Name-It: naming and detecting faces in news videos. *IEEE Multimedia, 6*(1), 22-35.

Sclaroff, S., La Cascia, M., Sethi, S., & Taycher, L. (1999, July/August). Unifying textual and visual cues for content-based image retrieval on the World Wide Web. *Computer Vision and Image Understanding*, *75*(1/2), 86-98.

Sproat, R. (2001). Inferring the environment in a text-to-scene conversion system. *Proceedings of International Conference on Knowledge Capture*, Victoria, British Columbia, Canada, 147-154.

Srihari, R., & Zhang, Z. (1999, Fall). Exploiting multimodal context in image retrieval. *Library Trends, 48(*2), 496-520.

Sutcliffe, A., Hare, M., Doubleday, A., & Ryan, M. (1997). Empirical studies in multimedia information retrieval. In Maybury, M. (Ed.), *Intelligent multimedia information retrieval* (pp. 449-472). Menlo Park, CA: AAAI Press / MIT Press.

Wactlar, H., Hauptmann, A., Christel, M., Houghton, R., & Olligschlaeger, A. (2000, February). Complementary video and audio analysis for broadcast news archives. *Communications of the ACM, 43* (2), 42-47.

Witten, I., & Frank, E. (2000). *Data mining: Practical machine learning with Java implementations*. San Francisco, CA: Morgan Kaufmann.

DEFINITIONS OF TERMS

"alt" string: An HTML tag for attaching text to a media object.
caption: Text describing a media object.
controlled vocabulary: A limited menu of words from which metadata like captions must be constructed.
data mining: Searching for insights in large quantities of data.
deixis: A linguistic expression whose understanding requires understanding something besides itself, as with a caption.
HTML: Hypertext Markup Language, the base language of pages on the World Wide Web.
media search engine: A Web search engine designed to find media (usually images) on the Web.
metadata: Information describing another data object such as its size, format, or description.
Web search engine: A Web site that finds other Web sites whose contents match a set of keywords, using a large index to Web pages.