Exploiting Captions for Web Data Mining by Neil C. Rowe

Rowe, Neil C.

Monterey, California. Naval Postgraduate School

Exploiting Captions for Web Data Mining

Neil C. Rowe

Abstract

We survey research on using captions in data mining from the Web. Captions are text that describes some other information (typically, multimedia). Since text is considerably easier to index and manipulate than non-text (being usually smaller and less ambiguous), a good strategy for accessing non-text is to index its captions. However, captions are not often obvious on the Web as there are few standards. So caption references can reside within paragraphs near a media reference, in clickable text or display text for it, on names of media files, in headings or titles on the page, and in explicit references arbitrarily far from the media. We discuss the range of possible syntactic clues (such as HTML tags) and semantic clues (such as meanings of particular words). We discuss how to quantify their strength and combine their information to arrive at a consensus. We then discuss the problem of mapping information in captions to information in media objects. While it is hard, classes of mapping schemes are distinguishable, and segmentation of the media can be matched to a parsing of the caption by constraint-satisfaction methods. Active work is addressing the issue of automatically learning the clues for mapping from examples.

This article is to appear in *Web Mining: Applications and Techniques* ed. A. Scime, 2004.

Introduction

Non-text media are an important asset of the World Wide Web. Most of the world prefers to communicate without written text, using audio, images, and video, because much of the world is illiterate and the increasing ubiquity of television is hurting literacy. The Web is the first information technology that permits access to media objects with the same ease as to text. Easy access to non-text media benefits everyone. For instance, we can find appropriate pictures of orchids or helicopters or quarterbacks in a few seconds without needing to search books or newspapers. It permits teachers to enliven their lectures with well-chosen images, audio, and video. And it permits a news bureau to find the perfect picture instead of an adequate one.

Captions are valuable tools in data mining from the Web. They are text strings that explain or describe other objects, usually non-text or multimedia objects. Captions both help understand and remember media (McAninch, Austin, & Derks, 1992-1993). Captions are especially valuable on the Web because only a small amount of text on Web pages with multimedia (1.2% in a survey of random pages (Rowe, 2002b)) describes the multimedia. Thus, standard text browsers when used to find media matching a particular description often do poorly; if they searched only the captions, they could do much better. Jansen, Goodrum, & Spink (2000) report that 2.65% of all queries in a sample of over 1,000,000 to the Excite search engine were attempting to find images, 0.74% were attempting to find video, and 0.37% were attempting to find audio, so multimedia retrieval was already important in 2000. It will undoubtedly increase in importance as faster, better search engines become available, and more people provide Web multimedia resources.

Captions are also valuable because content analysis of multimedia often does not provide the information that users seek. Nontext media does not usually tell when they were created or by whom, what happened before or after, what was happening outside the field of view when they were created, or in what context they were created; and nontext media cannot convey key linguistic features like quantification, negation, tense, and indirect reference (Cohen, 1992). Furthermore, experiments collecting image-retrieval needs descriptions from users (Armitage & Enser, 1997; Jorgensen, 1998) showed users were rarely concerned with image appearance (e.g. finding a picture with an orange-red circle in the center) but usually meaning which only captions could provide (e.g. "a pretty sunset" or "dawn in the Everglades"). People seem to have a wide range of tasks for which multimedia retrieval is required, many not needing much understanding of the media (Sutcliffe et al, 1997). This is good because content analysis of media is often considerably slower and more unreliable than caption analysis because of the large number of bits involved, and most useful content analysis of images, video, and audio for access requires segmentation, an unreliable and costly process, as well as preprocessing and filtering that is hard to get correct (Flickner et al, 1995; Forsyth, 1999). So finding a caption to a multimedia object simplifies processing considerably.

But using captions from the Web entails two problems: Finding them and understanding them. Finding them is hard because many are not clearly identified: Web-page formats vary widely and many methods denote captions. Understanding them is hard because caption

styles vary considerably and the mapping from the world of language to the world of visual or aural concepts is not often straightforward. Nonetheless, tools to address and often solve these problems are available. The usually limited semantics of Web page specifications and captions can be exploited.

Commercial multimedia search engines are available on the Web, most of them free, and all primarily exploiting text near media. The major ones currently are images.google.com, multimedia.lycos.com, www.altavista.com/image, multimedia.alltheweb.com, www.picsearch.com, www.gograph.com, gallery.yahoo.com, www.ditto.com, attrasoft.com/imagehunt, www.fatesoft.com/picture, www.ncrtec.org/picture.htm, www.webplaces.com/search, sunsite.berkeley.edu/ImageFinder.htm, www.iconbazaar.com/search, www.compucan.com/imagewolf-E.htm, www.goimagesearch.com, and www.animationlibrary.com. But by no means has this software "solved" the problem of finding media. For one thing, the copyright status of media downloaded from the Internet is unclear in many countries, so search engines are cautious in what they index. For another, media-search engines can potentially retrieve pornographic images, and thus are not accessible with the Web browsers used by most educational and government institutions because of automatic filtering (Finkelstein, 2003). In addition, the accuracy of keyword search for media is often significantly lower than that of keyword search for text. Table 1 shows accuracy, as judged by the author, for the best matches found for sample keywords entered in five image search engines. AltaVista appears to be the winner. But performance deteriorates considerably on larger queries, and much of the success on short queries was by exact matches to image file names, which suggests "cheating" by some owners of images to get better exposure since, for instance, "red-orchid" is a vague and poor image name.

Table 1: Retrieval performance (fraction of answers that are correct) in April 2003 with five image search engines on the World Wide Web.

Keywords	images .google .com	multimedia .lycos .com	www .altavista .com/image	multimedia .alltheweb .com	www .picsearch .com
"moon"	0.75 (15/20)	0.94 (17/18)	0.83 (13/15)	0.20 (4/20)	0.81 (13/16)
"captain"	0.45 (9/20)	0.89 (16/18)	0.40 (6/15)	0.25 (5/20)	0.31 (5/16)
"closing"	0.40 (8/20)	0.50 (9/18)	0.93 (14/15)	0.00 (0/20)	0.07 (1/16)
"red orchid"	0.50 (10/20)	0.16 (3/18)	0.86 (13/15)	0.10 (2/20)	0.56 (9/16)
"quarterback throwing"	0.35 (7/20)	0.16 (3/18)	0.47 (7/15)	0.15 (3/20)	-- (0/0)
"angry crowd"	0.30 (6/20)	0.11 (2/18)	0.40 (6/15)	0.00 (0/20)	0.29 (2/7)
"American truck rear"	0.17 (3/18)	0.16 (3/18)	0.55 (8/15)	0.15 (3/20)	-- (0/0)
"missile on aircraft"	0.15 (3/20)	0.05 (1/18)	0.00 (0/15)	0.20 (4/20)	0.06 (1/16)
"diet soda bottle"	0.50 (1/2)	0.05 (1/18)	0.07 (1/15)	0.05 (1/20)	-- (0/0)
"Rockies skyline sunset"	-- (0/0)	0.17 (1/6)	0.27 (4/15)	0.17 (1/6)	-- (0/0)
"president greeting dignitaries"	0.00 (0/3)	-- (0/0)	0.00 (0/15)	-- (0/0)	-- (0/0)

In general, commercial software needs to look good, and to do so it emphasizes precision (the fraction of answers found in all answers retrieved) as opposed to recall (the fraction of answers found in all possible answers on the Web). This is fine for users who are not particular about what media they retrieve on a subject; but they may be missing much better examples, and high-recall searches are necessary for important applications like checking for policy compliance about content of Web pages. To increase recall, software needs to examine Web pages more carefully using the methods we shall discuss.

Finding captions on Web pages

We define a "caption" as any text that helps explain or describe a media object; their justification is that certain aspects of knowledge are best conveyed in words. The terms "media" and "multimedia" here include images, video, audio, and computer software. Images include photographs, drawings, designs, backgrounds, and icons. Multimedia also includes text, and formatted text like tables and charts that can have captions too. On computer systems multimedia are stored digitally, as bits and bytes, in files of a small number of formats handled by browser software. Table 2 shows some common multimedia formats used with Web pages.

Table 2: Common formats (and file extensions) for media types on Web sites.

Media Type	Common Formats
Text	TXT, TBL, XLS, DOC, PPT
Image	JPG, JPEG, GIF, GIFF, PNG, TIFF, PIC, SVG
Video	MPG, MPEG, MOV, SWF, FLI, AVI
Audio	WAV, RAM, MID, ASX
Software	EXE, VBS

Syntactic clues for captions

Much text on a Web page near a media object is not related to that object. Thus we need clues to distinguish captions, allowing there may often be more than one for an object. A variety of clues and ways to process them have been proposed in the literature (Swain, 1999; Sclaroff et al, 1999; Mukherjea & Cho, 1999; Rowe, 1999; Srihari & Zhang, 1999; Favela & Meza, 1999; Rowe, 2000b; Hauptman & Witbrock, 1997; Watanabe et al, 1999) and we will summarize these ideas here.

About HTML

The language that built the Web was HTML (Hypertext Markup Language) and it is still very popular today on the Internet. Web pages are text files whose names usually end in ".html" or ".htm" and are displayed by software called Web browsers. HTML defines key aspects of the appearance of Web pages through formatting tags inserted into text, delimited with angular brackets ("<" and ">"). Many tags have opposites indicated by a leading slash character, so for instance italics font is begun by "" and turned off by "". Larger-scale units like titles, headings, quotations, tables, and computer output also have explicit tags in HTML. Even prose paragraphs can be delimited by tags to ensure that a Web browser displays them properly. All these tags are useful in finding text units which could be captions.

Multimedia is indicated in HTML by special tags, usually by the "img" tag with argument "src" being an image file to be displayed at that point on the page. Media can also be the destinations of links in the "src" argument to the "href" tag. Media can be specified with the HTML-4 "object" tag which embeds complex objects in pages. Also used are "embed" for audio and video and "bgsound" for background sound. Media objects typically require considerably more storage space than the text of a Web page; they are generally stored in a separate file, usually in the same directory. Browsers can tell what kind of object a link points to by both the header information associated with the object and the "extension" or end of its file name. Most images are in GIF and JPG formats, so most image file names end with ".gif" and ".jpg". The most popular image formats we found (Rowe, 2002c) from our exhaustive survey of nearly all U.S. military ("*.mil") Web sites in early 2002, after pruning of obviously uncaptioned images (see discussion below), were JPEG (683,404 images), GIF (509,464), PNG (2,940), and TIFF (1,639). The most popular audio formats were WAV (1,191), RAM (664), MID (428), and ASX (238). The most popular video formats were MPEG (8,033), SWF (2,038), MOV (746), FLI (621), AVI (490), and MP3 (397).

Sources of captions in HTML code

Several methods closely associate captions with multimedia on Web pages. "Alt" strings are text associated with media objects that is displayed, depending on the browser, when the user moves the mouse over the object or in lieu of the object on primitive browsers. They can be excellent sources of captions, but many are just placeholders like "Picture here". Clickable text links to media files are also good sources of captions, since the page designer must explain the link. Audio and video take more time to load than images, so they are usually accessed via a clickable link. The "caption" tag can be used to label media objects, but we have rarely seen it used. Finally, a short but tightly coupled caption is the name of the media file itself; for instance "northern_woodchuck.gif" suggests a caption of "Northern woodchuck". All these provide good but not perfect clues as to the content of the media.

But many captions on Web pages are not marked so explicitly, particularly those reflecting traditional print layout. A common

convention is to center caption text above or below a displayed image. Often such text changes font size or style (e.g. italics) to flag its function. So the tags "center", "i", "b", and "font" are moderate clues for captions. Title and heading tags ("title", "h1", "h2", etc. in HTML) can also indicate caption information as their goal is to generalize over a block of information, and they are helpful when other captions are not available. But they do not often give precise details, and tend to omit words so they can be hard to decipher (Perfetti et al, 1987). Paragraphs above or below some media can also be captions (especially single lines just below), but this is not guaranteed.

Figure 1 shows an example Web page and Figure 2 shows its HTML source code. Four photographs, two graphics images, one hyperlink to an audio file, and one hyperlink to another page are specified in the source; both hyperlinks are followed by clicking on associated images. File names here provide useful clues for the five photographs but not so much for the two graphics images; one "alt" string is helpful; the scope of the two "center" commands helps connect text to its associated image; and the names of the link files give clues as to their nature.

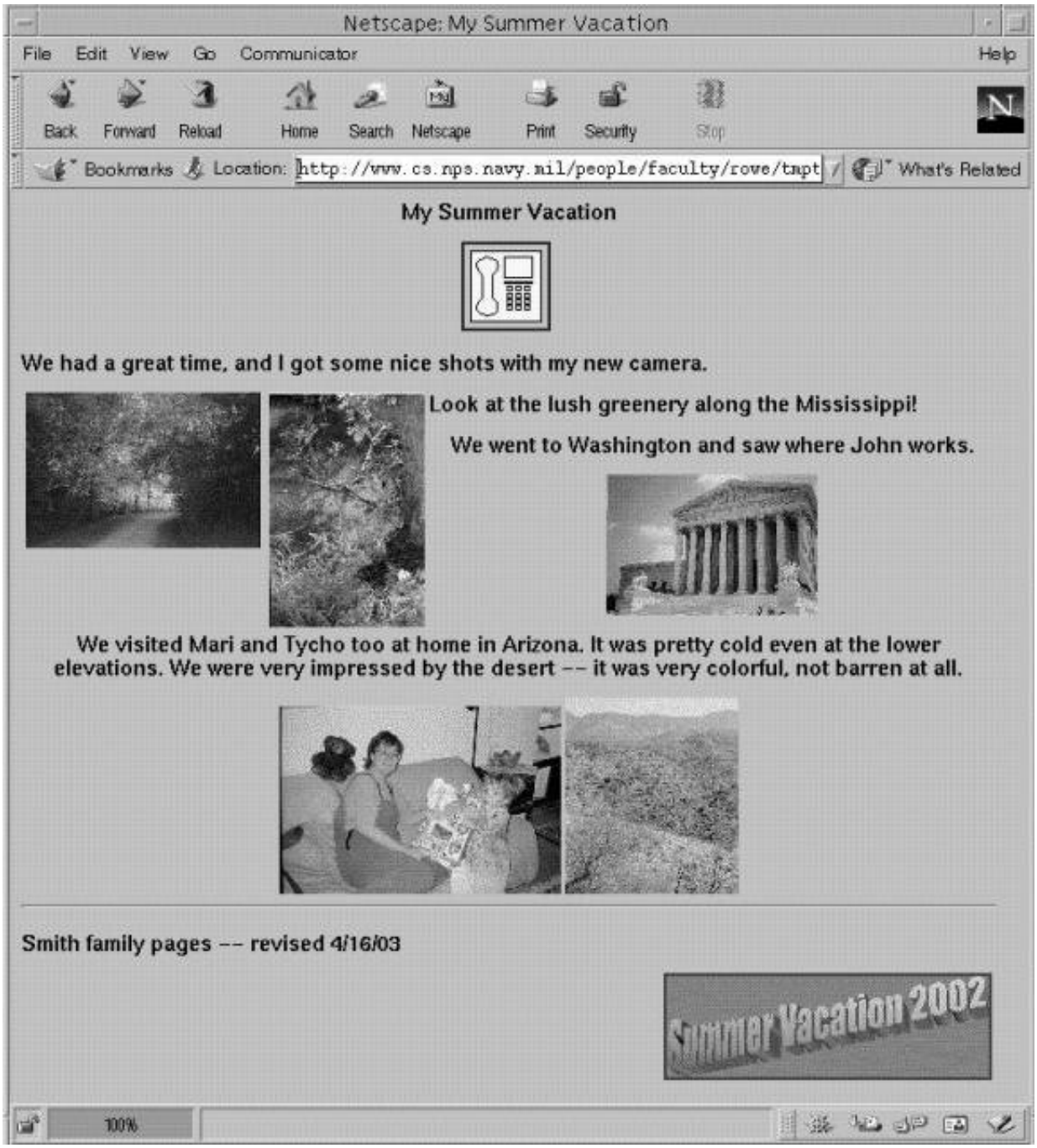

Figure 1: Example Web page.

```
<title>My Summer Vacation</title>
<h1><center>My Summer Vacation</h1>
<a href="reveille.wav"><img src="telephoneicon.gif"></a></center>
<p><i><h2>We had a great time, and I got some nice shots with my
 new camera.</h2></i></p>
<img align=left src="aspenRidgeRoadWheatland.jpg"
alt="Wisconsin where we stayed">
<img align=left src="Mississippi_side_channel.jpg">
<h2>Look at the lush greenery along the Mississippi!</h2>
<p><center><h2>We went to Washington and saw where John works.</h2></p>
<img align=center src="supremecourt.jpg"></center>
<br><center><b><h2>We visited Mari and Tycho too at home in Arizona.  It was pretty cold
even at the lower elevations.  We were very impressed by the desert -- it was very
colorful, not barren at all.</h2>
<img src="tycho_toy_00.jpg">
<img src="desert_saguaro.jpg"></b></center>
<hr><h2><i>Smith family pages -- revised 4/16/03</i></h2></a>
<a href="summer_vacation2.htm"><img align=right src="summervacation.gif"</a>
```

Figure 2: HTML source code for Figure 1.

Another class of captions are embedded directly into the media, like characters drawn on an image (Wu, Manmatha, & Riseman, 1997) or explanatory words spoken at the beginning of audio (Lienhart, 2000). These require specialized segmentation to extract, but this is usually not too hard because they can be (and should be) made to contrast well with background information, permitting localization by texture filtering. Individual characters in images and video can be found by methods of optical character recognition (Casey & Lecolinet, 1996); superimposed audio can be extracted using speech processing.

An easier way to attach captions is through a separate channel of video or audio. For video an important instance is the "closed caption" associated with television broadcasts. Typically this transcribes the words spoken though occasionally describes purely visual information. Such captions are helpful for hearing- impaired people and for students learning the captioned language (Cronin, 1995; NCIP, 2003), but are only available for widely disseminated video. Similar captions are helpful for audio and for special applications of images like museums (The Dayton Art Institute, 2003), consistent with the endorsement by the World Wide Web Consortium (W3C, 1999) of captions to provide enhanced Web access to the disabled.

Finally, "annotations" can function like captions though they tend to emphasize analysis, opinion, background knowledge, or even advertisements more than description of their referent. Annotation systems have been developed to provide collaborative consensus (Sannomiya, 2001), which is especially helpful on the Web. Annotation can also include physical parameters associated with the media such as the time, location, and orientation of the recording apparatus, to get clues as to speaker changes or extent of meetings (Kern et al, 2002). Unlike captions, Web annotations usually are stored on separate pages from their referents with links to them. Many annotation systems impose a limited vocabulary upon the annotator so they can exploit "ontologies" or concept hierarchies to classify pages. An exception is (Srihari & Zhang, 2000) where an expert user annotates in free text by guiding attention successively to parts of an image. Comments on the source code of a computer program are a form of descriptive annotation, and treating them like captions is facilitated by tools such as Javadoc in the Java programming language (Leslie, 2002).

Additional clues for captions

Besides the explicit HTML tags, several other clues suggest captions on Web pages. Length in characters is helpful: Captions usually average 200 characters, and one can fit a likelihood distribution. A strong clue is words in common of the proposed caption and the name of the media file (excluding punctuation), especially uncommon words. For instance, "Front view of woodchuck burrowing"

would be a good match with "northern_woodchuck.gif". Nearness of the caption to the referent is a clue, but not a reliable one: Many "alt" and "filename" text strings embedded in the referent specification are useless placeholders. Segmentation methods from document image analysis (Satoh, Tachikawa, & Yamaai, 1994) can do a better job of determining adjacency when displayed, which can help.

It is helpful to know if an image is a photograph or graphics, since photographs are more likely to have captions. Evidence is the size of the image (graphics is more often small), the ratio of length to width (graphics has larger ratios), the image format (GIF is more likely to be graphics), and the use of certain words in the image file name (like "photo", "view", and "closeup" as positive clues and "icon", "button", and "bar" as negative clues). If image processing can be done, other evidence is the number of colors in the image and the frequency of the most common color. For Figure 1, the small size of telephoneicon.gif and summervacation.gif, "icon" in the first image's name, the length to width ratio of the second image, and the limited number of colors all argue against their being captioned.

The number of occurrences of a media object on a page or site is also a useful clue: Duplicated media objects tend to be decorative and unlikely to be captioned. As a rule of thumb, objects occurring more than once on a page or three times on a site are unlikely (probability < 0.01) to have captions (Rowe, 2002b). Similarly, text that occurs multiple times on a page is not likely to be a caption, though headings and titles can be correctly ascribed to multiple media objects.

Another clue to a caption is consistency with known captions on the same page or on related pages (like those at the same site using the same page-design software). An organization may specify a consistent page style ("look and feel") where, for instance, image captions are always a centered boldface single line under the image. This is helpful in recognizing, for instance, atypically short captions. This effect can be implemented by estimating defaults (e.g. caption size, caption location versus referent, and caption style) from the pages of a site. For instance, *National Geographic* magazine tends to use multi-sentence captions where the first sentence describes the image and subsequent sentences give background.

Table 3 shows the caption candidates for Figures 1 and 2 obtained using these principles.

Table 3: Caption candidates derivable from the Figure 1 Web page.

Type	Caption	Media Object	Comments
title, h1, center	My Summer Vacation	reveille.wav (audio)	Weak candidate since object name different
filename	reveille	reveille.wav (audio)	Possible but not strong
title, h1, center	My Summer Vacation	telephoneicon.gif	Weak candidate since object name different

filename	telephoneicon	telephoneicon.gif	Possible but not strong
title, h1, center	My Summer Vacation	aspenRdigeRoadWheatland.jpg	Weak
p, i, h2	We had a great time, and I got some nice shots with my new camera.	aspenRidgeRoadWheatland.jpg	Possible but not strong since scope does not include media
filename	aspen ridge road wheatland	aspenRidgeRoadWheatland.jpg	Possible but not strong
alt	Wisconsin where we stayed	aspenRidgeRoadWheatland.jpg	Strong
title, h1, center	My Summer Vacation	Mississippi_side_channel.jpg	Weak candidate since object name different
filename	Mississippi side channel	Mississippi_side_channel.jpg	Possible but not strong
h2	Look at the lush greenery along the Mississippi!	Mississippi_side_channel.jpg	Possible but not strong by lack of scoping
title, h1, center	My Summer Vacation	supremecourt.jpg	Weak candidate since object name different
p, center, h2	We went to Washington and saw where John works.	supremecourt.jpg	Medium strength: object name different, but no other good candidate for this image
filename	supremecourt	supremecourt.jpg	Possible but not strong
center, b, h2	We visited Mari....	supremecourt.jpg	Weak because of intervening line break
title, h1, center	My Summer Vacation	tycho_toy_00.jpg	Weak candidate since object name different
center, b, h2	We visited Mari....	tycho_toy_00.jpg	Good because overlap on unusual word "Tycho"
filename	tycho toy 00	tycho_toy_00.jpg	Possible but not strong
title, h1, center	My Summer Vacation	desert_saguaro.jpg	Weak candidate since object name different
center, b, h2	We visited Mari....	desert_saguaro.jpg	Medium: Image reference crosses another image, but no text between images
filename	desert saguaro	desert_saguaro.jpg	Possible but not strong
h2, i	Smith family pages – revised 4/16/03	summervacation.gif	Weak since better reference present
href (link)	summer vacation2	summervacation.gif	Strong since clickable link
filename	summervacation	summervacation.gif	Possible but not strong

Page specification languages beyond HTML

Because HTML is relatively simple, a variety of extensions to it can provide Web-page designers with features in the direction of full programming languages. But generally these provide little additional caption information. JavaScript provides user-interface capabilities, but media references are similar; it does allow file names to be assigned to variables that may then be decoded to find what a caption is referring to. Java Server Pages (JSP), Active Server Pages (ASP), and Java Servlets are popular extensions to HTML that permit dynamically created Web pages. But they do not provide anything new for media or captions.

XML is increasingly popular as a language for data exchange that uses the same protocol as HTML, HTTP. Mostly it is for transfer of textual or numeric data, though the proposed SVG format will extend it to graphics. Media objects can be identified under XML, however, and several kinds of captions can be attached to them.

Page crawlers for multimedia

The Web is large and finding useful information requires a large number of time-consuming page fetches. Thus it is best to index caption-media pairs in advance of queries or data mining, as with the text search engines on the Web. We need a specialized Web "crawler" or "spider" (Heaton, 2002) to search for captioned media. It needs the usual tools of crawlers, including an HTML parser and a queue of page links found, but can ignore most non-HTML pages like text and proprietary document formats as well as media files since its job is to just catalog media links. It may or may not examine files of page markup languages like PDF (Adobe Acrobat) and PS (Postscript) depending on the need for thoroughness. Markup languages are harder to interpret than HTML but do contain captioned images, and major search engines such as Google are now indexing them. A media crawler should also confirm the existence and type of the media it is indexing since there are plenty of mistakes about these things on Web pages; such checking does not require loading the page.

Semantic clues for captions

The above methods will generate many candidate captions for media objects. However, many of these will just be text that accidentally appears next to the media. So at some point one should investigate the candidates more closely and ascertain their meaning or semantics.

Word semantics for captions

Some words in captions suggest depiction, and thus serve both to confirm a caption and help explain its referent. Table 4 shows some example domain-independent words that are good clues for captions. In addition, words relating to space (places and locations) and time (events and dates) are clues. Many of these can be found from training data, as will be discussed. Besides these, there are many domain-dependent clue words; for instance for a library of animal images, animal names are good clues.

Table 4: Example word clues for captions.

	Nouns	Verbs	Prepositions
Images	picture, photograph, front, top	shows	above, beside
Audio	sounds, music, excerpt	recorded	after, during, then
Video	clip, movie, video	taped	after, during, then
Software	program, demo, simulation	runs	into

The tense of verbs in captions is important, as generally present tenses are used both to refer to media itself (e.g. "the picture shows") and depicted events in the media (e.g. "President congratulating award winners" using the present participle). This includes present progressive tenses as in "The President is congratulating them". Past tenses often refer to undepicted events before the state or sequence the media depicts, but are also conventionally used for depicted objects in historical media when no present tenses are used, e.g. "The President addressed Congress on January 18, 2003." Future tenses usually give goals unachieved and undepicted in the state the media depicts.

A way to make words more powerful clues is to enforce a limited or "controlled" vocabulary for describing media, much like what librarians use in cataloging books, and this can be mandated in building specialized media libraries. Such a vocabulary can be significantly more unambiguous than unrestricted natural language, and a precise hierarchy can be built that permits generalizations of terms for better keyword matching. Controlled-vocabulary media retrieval systems have been implemented at the U.S. Library of Congress in the process of digitizing their media holdings (Arms, 1999) and at Getty Images Ltd. (Bjarnestam, 1998). But it is not feasible on most World Wide Web pages, where no central control affects what is posted.

The structure of referring phrases

Individual words can be clues, but phrases can be even better. Table 5 gives some examples. The caption can be parsed with a specialized grammar that recognizes these and similar patterns. This is "partial parsing", parsing that need not analyze the whole sentence; it is used frequently in data mining applications. It must guess where a phrase starts, but this can be facilitated by indexing characteristic words of the phrases and their locations in the phrases.

Table 5: Example linguistic referring phrases found in captions.

Phrase	Restrictions	Media type
"the X above"	X is a viewable object	image
"the Figure shows X"	X is a viewable object	image
"Figure N: X"	N is a number, X is a viewable object	image
"view of X"	X is viewable	image
"X beside Y"	X and Y are viewable	image
"Look at the X"	X is viewable	image
"you can hear X"	X is audio	audio
"listen to X"	X is audio	audio
"X then Y"	X and Y are audio	audio/video
"shows X doing Y"	X is viewable and Y is a viewable act	video
"X during Y"	X and Y are viewable acts	image/video

If full parsing of caption candidates is possible (Srihari, 1995; Guglielmo and Rowe, 1996), another clue is that captions are often grammatical noun phrases, e.g. "View of East Wing". Verbs usually occur as participles attached to the head noun, e.g. "East Wing viewed from south", unlike most English prose. But a few captions use imperatives (like "See at the left how we did it") to direct the reader's attention. Captions that represent other syntactic categories can usually be interpreted as examples of ellipsis on the head noun of a previous caption. For instance, "Mounted in a setting" after the previous caption "1 carat diamond on cutting tool" means "1 carat diamond mounted in a setting".

Not all words in a caption are equally likely to be represented in the media, and media search can be more successful if this is exploited. The implicit speech act often associated with descriptive captions is that the grammatical subjects of the caption correspond to the principal objects within the media (Rowe, 1994). For instance, "Stuffed panther in a museum case" has grammatical subject "panther" and we would expect it to correspond to a major region in the picture near the center. "Museum" and "case" have no such guarantee. By contrast, "Museum case containing stuffed panther" implies that all of the museum case can be seen but not necessarily the panther (or clearly, anyway). "Museum case and stuffed panther" has a composite subject and thus both should be visible. Grammatical-subject depiction does require, however, that it is possible to depict the subject. So "Budget trimming in Washington" cannot be mapped to an image because the gerund "trimming" here has an abstract sense. Thesaurus systems like Wordnet (Miller et al, 1990) provide hierarchies of nouns to check automatically if an object is physical.

Some other grammatical constructs also suggest features of the media:

- Present-tense principal verbs of caption sentences, subject gerunds, and participles attached to the principal noun can depict dynamic physical processes in video, audio, and software, e.g. "President greeting dignitaries" for video and "Lawnmower cutting grass" for audio.
- Direct objects of such verbs, gerunds, and participles are usually depicted in the media when they are physical objects, like "computers" in "Students are using computers" and "bins" in "Workers loading bins".
- Objects of physical-location prepositions attached to the principal subject are also depicted in part (but not necessarily as a whole), like "case" in "Stuffed panther in museum case." Example physical-location prepositions are "within", "beside", "outside of", and "with" when used to mean accompaniment.
    - If the subject of a caption sentence is a general term for a media object like "view" or "audio", objects of prepositions or participles attached to it are usually the true subjects of the images, as in "View of North Rim" and "Scene showing damage to aircraft".

Fuzzy deixis

Some special linguistic expressions ("spatial deixis") refer to spatial relationships between captions and media referents or parts of them. An example is "the right picture below" in Figure 1. Such expressions are "fuzzy" in that they describe a vaguely defined region of space that can be modeled with an associated "suitability" at each point (Matsakis et al, 2001). The most common spatial terms are shown in Figure 3. For the example phrase, "below" defines a region of points with bearings roughly within 45 degrees of downwards beneath the word "below" in the text; "right picture" specifies a relationship within 45 degrees of horizontal between the

centers of two adjacent media objects. It is important to distinguish the "deictic center" or the location on the Web page from which the deictic references originate, which is usually the characters of the text itself.

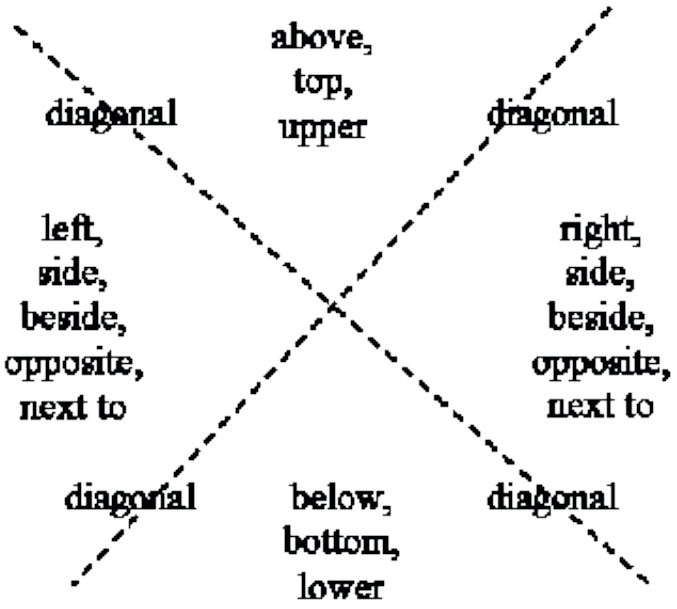

Figure 3: The classic spatial deixis terms.

The language of dynamic media

Video and audio involve changes over time. So verbs and nouns referring to change are especially important in understanding them. Verbs involving physical motion like "go" and information transfer like "report" are common in describing video, as are their participles and gerunds and other related nouns like "song" and "segment". State-change nouns like "destruction" are also common in video to summarize actions or their results. Software media is usually described by a narrower range of dynamic terms. But speech audio is similar to text and can be captioned similarly, and music audio is usually described only as a whole.

Mathematical analysis of clues

Filtering of caption candidates

Which of these many kinds of potential caption information about a media object should we trust? In general, most of them can provide useful information. So we risk serious problems with information overload when analyzing large portions of the Web, and it is helpful to do "information filtering" to rule out obviously useless data. We can do this using especially strong clues as to the unlikeliness of some text being a caption. Several clues mentioned above qualify: large distance of the text on the page from media objects, large size of the text, duplicate occurrence of the media, and use of words very specific to noncaption text like "download", "click", "http", "index", and "pages".

One important class of filtering criteria recognizes markers for scope boundaries on image referents. Lines (like from the "hr" tag in Figure 2) and graphics partition a Web page, and caption-media references rarely cross partitions. Captions themselves almost always bound the scope of other caption-media references, so for instance in Figure 1 the text "We visited Mari..." and the image "supremecourt.jpg" prevent assigning the caption "We had time..." to the images below of "desert_saguaro.jpg" and "tycho00.jpg". Exceptions must be made for heading and title captions, which explicitly have broad scope, and adjacent media objects without intervening text, as the last two mentioned images.

The order of clue filtering is important for efficiency with large amounts of data. Generally speaking, filters should be sorted by increasing values of the ratio of filter cost to filter-test failure probability, but full analysis can be complicated (Rowe, 1996).

Clue strengths

How do we systematically find clues and assess their strengths? One model is that of "knowledge system" or "expert system" development (Stefik, 1995): We examine many Web pages and recognize clues on them using our intuition. This works because many clues are quite obvious. Expert-system methodology is a well-developed branch of artificial intelligence, and a number of techniques from it can facilitate our task here such as functional analysis of Web pages, use-case analysis, and validation and verification. A weakness of this approach is that experts must quantify clues, and such assessment is often the most difficult part of expert-systems development, prone to numerous errors.

Another approach is to get statistics on a large and representative sample of pages on which captions are marked, and infer conditional probabilities of captions given the clues. This requires, for each of a set of Web pages, that people identify the caption-media pairs that are present. Then for each potential clue k, calculate $c_k/(c_k + n_k)$ where c_k is the number of occurrences of the clue in a caption and n_k is the number of occurrences of the clue in a noncaption. Clue appearances tend to be somewhat independent on Web pages because the pages often contain a good deal of variety. Thus clue appearance can be modeled as a binomial process with expected standard deviation $\sqrt{c_k n_k/(c_k + n_k)}$. The standard deviation can be used to judge whether a clue is statistically significant. A conditional probability that is two standard deviations away from the statistically observed fraction is 83% certain to be significant, and one three standard deviations away is 96% certain to be significant. This standard-deviation criterion rules out many potential clues, especially the numerous possible word clues. Table 6 gives some example clues and their conditional probabilities in a sample of Web pages (Rowe, 2002b) with 2,024 possible text-media pairs where the text was adjacent to the media. Of these 21.1% were confirmed to be caption-media pairs during exhaustive manual inspection, and statistical significance was defined to be more than one standard deviation away from 0.211.

Table 6: Example clues and their strengths in a test set, from (Rowe, 2002b)

Clue	Probability	Significant?
italics (<i>)	0.40	no
table datum (<td>)	0.47	yes
largest heading font (<h1>)	0.33	no
second largest heading font(<h2>)	0.49	yes
positive image words	0.31	yes
negative image words	0.19	no
positive caption words	0.31	yes
negative caption words	0.25	no
caption > 112 chars.	0.36	yes
caption < 31 chars.	0.15	yes
image diagonal > 330	0.28	yes
image diagonal < 141	0.09	yes
JPEG-format image	0.39	yes
GIF-format image	0.09	yes
digit in image filename	0.25	no
4 digits in image filename	0.40	yes
image in same directory as Web page	0.23	no
Web page name ends in "/"	0.14	yes
.com site	0.19	no

.edu site	0.21	no
.org site	0.16	no
.mil site	0.33	yes
centered text ( )	0.06	yes
title (<title>)	0.34	yes
alternative text (<alt>)	0.35	no
text-anchor link (<a>)	0.65	yes
image filename	0.04	yes
caption-suggestive wording	0.47	yes

(Rowe, 2002b) did recall-precision analysis of the importance of nine major categories of clues for image captions using a random sample of Web pages. Results showed that text-word clues were the most valuable in identifying captions, followed in order by caption type, GIF/JPEG 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; distance of the caption from the image was unhelpful in identifying a caption. Similar study of just the easily-calculated features of images, including number of colors, saturation of colors, frequency of the most common color, and average local variation in color, showed that only image size was significantly helpful.

A complication is that some clues interact. For instance, the conditional probability of caption given a particular word in it depends on its relative location: Depictive words are significantly more likely to occur early in the caption. (Salton and Buckley, 1988[1]) provides a useful survey of some other general clue-strength factors such as term frequency, inverse document frequency, term discrimination, and formulas including them.

Combining evidence for a caption-media pair

Best performance in identifying captions can be obtained by combining evidence from all available clues. This is a classic principle in data mining (Witten & Frank, 2000) and a spectrum of methods are available. One popular way is a "linear model" where we take a weighted sum of the clue probabilities, and label the text as a caption if the sum exceeds fixed threshold; linear regression can be used to find the best weights from a training set. But this tends to overrate candidates that are strong on one clue and poor on others.

A popular alternative is the multiplicative approach called Naive Bayes, which in this case would estimate the probability of a caption as:
$p(caption \mid clues) = p(clue1 \mid caption)\,p(clue2 \mid caption)\,p(clue3 \mid caption)\ldots p(caption) / p(clues)$ where p means probability, p(X|Y) means the probability of X given Y, and "clues" means all clues together. We again test whether this number exceeds a fixed threshold to decide if some given text is a caption. The probabilities on the right side can be obtained from statistics on a training set and are reliable given sufficient instances of the clues. This approach has the disadvantage of penalizing captions for one weak clue, which can be unfair for the less-certain clues. So the linear model idea can be embedded in the calculation by taking a weighted sum of several related weak clues, then using that sum in the product. This seems to work well for many applications.

Setting the threshold for both approaches is a classic problem of trading off precision and recall. Here precision means the fraction of the captions identified by software that were actually captions, and recall means the fraction of the captions identified by software of all the actual captions in the test set. If we value precision, we should set the threshold high; if we value recall, we should set the threshold low. In general, we must choose a weighted average of the two.

Adjusting caption-media probabilities from their context

Context should also affect the probability of a caption-media pair. For instance, in Figure 1 the paragraph "We visited Mari..." could go with either the images above or below. But the presence of stronger caption candidates for the images above -- the text beside the vegetation pictures and above the Supreme Court picture -- argues against further captions for those images. In general, strong candidate captions for some media object decrease the likelihoods of other candidates. This can be modeled mathematically by normalizing likelihoods over all candidate captions on an image, or dividing the likelihoods by their sum.

Another context issue is that media objects on the same page or at the same site tend to be captioned similarly, as mentioned earlier. Examples of such consistency are whether captions are in italics or boldface, whether they are centered with respect to the media display, whether they are above, below, left, or right of the media, and whether headings and titles on the page are useful captions for the media therein. We can model this by increasing the likelihoods of caption-media pairs that are consistent with those of their neighbors on the same page or same site. Likelihood changes can be done gradually in several cycles as a form of "relaxation" process from which a consensus gradually emerges.

Obtaining training and test data from the Web

A big problem for a statistics-based approach, however, is that it needs a random sample of the Web. This is difficult because the Web is deliberately decentralized and has no comprehensive index. Browsers index the contents of pages, but do not provide listings of page links. One approach is to pick some known pages and do a random search from there, preferably a depth-first search on randomly chosen links to traverse the maximum number of pages (Rowe, 2002b). But pages with many links to them are more likely to be visited this way, especially index pages that are not representative of a site. It would be better to observe the pages visited by a random set of users over a period of time.

An alternative requiring less initial data could be to use "bootstrapping", a technique helpful in large data mining projects. Using an initial set of clues and their rough probability estimates obtained from an expert, one can rank caption-media pairs found by a semi-random search. The strongest candidates in those pairs (say those ranking in the upper 20%) can be assumed to be captions, and conditional probabilities calculated on them. Continue searching randomly, now ranking new pages with the revised clue probabilities for those clues shown to be statistically significant. Repeat the process of revising the probabilities (and perhaps weights) many times, adding and dropping clues as we proceed in a kind of feedback process. Eventually one has examined a large number of pages and obtained reliable clue probabilities.

Mapping captions to multimedia

Once we have identified a caption and its associated media object, we have a task of matching or finding the mapping between the two. The justification for captioned media is that two modalities often provide information more efficiently than one modality could (Heidorn, 1997). Evidence suggests that images are valuable in understanding natural-language utterances, independent of captions, by creating mental pictures and making predictions about how it happened (DiManzo, Adorni, & Giunchiglia, 1986). Studies have shown eye movements in people comprehending descriptions of imaginary things (Spivey, 2000).

Studies of users have shown they usual consider media data as "depicting" a set of objects (Armitage and Enser, 1997; Jorgensen, 1998) rather than a set of textures arranged in space or time. There are thus three corresponding entities in any caption situation: Real-world objects, the media representation of the objects in terms of bits, and the caption that describes the media objects and relates them to the real-world objects. It should be noted that media access is usually just one aspect of information that users seek: A user must be properly assigned to appropriate media with an appropriately chosen mix of methods (Srihari, Zhang, & Rao, 2000).

"Deictic" is a linguistic term for expressions whose meaning requires assimilation of information from a referent outside the expression itself. Thus captions are deictic. Several theories have been proposed as to how the linguistic structures map to characteristics of their referent for various kinds of linguistic expressions. Good work in particular has been done on the physical constraints entailed by spatial linguistic expressions (Giuchiglia et al, 1996; Mukerjee, 1998; Di Tomaso et al, 1998; Pineda & Garza, 2000). Unfortunately, few captions on the Web use spatial expressions. In the 1716 training examples of Web captions found by random search in (Rowe, 2002b), there were only 123 occurrences of any the terms "right", "left", "above", "below", "beneath", "beside", "opposite", "next to", "side of", and "diagonal"; of these, only 33 referred to pictures as a whole and 36 referred to components of pictures. So Web caption interpretation must focus on analysis of less direct correspondences.

Correspondence in general between captions and media

To be more precise, several relationships are possible between media and their captions. These can apply separately to each sentence of a multi-sentence caption.

- *Component-depictive*: The caption describes objects and/or processes that map to particular parts of the media. This is the

traditional narrow meaning of "caption". For instance, a caption "President greeting dignitaries" with a picture that shows a President, several other people, and handshaking. Such captions are common, and are especially helpful when analogy and contrast of parts is a good way to distinguish media objects (Heidorn, 1999).

- *Whole-depictive*: The caption describes the media as a whole. This is often signalled by media-type words like "view", "clip", and "recording". For instance, a caption "The scene on the trading floor" with a picture of the New York Stock Exchange.
- *Illustrative-example*: The media presents only an example of the phenomenon described by the caption. This occurs for captions too abstract to be depicted directly. For instance, "Labor unrest in the early 20th century" with a picture of a labor demonstration from 1910.
- *Metaphorical:* The media represents something analogous to the caption but does not depict it or describe it. For instance, the caption "Modern American poetry" with a picture of some trees. Variations on this kind of relationship occur with media that evoke emotional responses, like the caption "Stop police brutality" with a picture of police standing in a line.
- *Background*: The media represents some real phenomenon, but the caption only gives background information about it. For instance, the caption "The Civil War caused inflation" with a portrait of Abraham Lincoln. The house style of *National Geographic* magazine is generally to have all sentences of a caption except the first be of this kind.

These distinctions are important because only with the first two does the text describe the media, and only with the first can prediction of the components of the media be done from the caption. So most research has focused on component-depictive captions. For these there is considerable variation in content. The level of detail can vary; the World Wide Web Consortium recommends using together both short captions and long captions (WTC, 2003). Some captions focus more on the objects within the media, and some focus more on relationships or processes involving the objects. Some captions focus on small details that are of special importance. So we cannot in general assume much about what a caption covers.

Media properties and structure

Whole-depictive caption sentences refer to the aggregate characteristics of the associated media object. Table 7 gives some examples.

Table 7: Example aggregate properties of media objects

Property	Media Type	Example Caption Phrase
size	image, video	big picture
real-world scale	image, video, audio, software	closeup video
duration	video, audio, software	longer clip
color	image, video, software	infrared image
brightness	image, video	dark view
loudness	video, audio	busy excerpt
texture	image, video, audio	high-contrast image
shape	image, video, software	narrow photo
sound quality	video, audio	tinny recording
valuation	image, video, audio, software	difficult game

Many of the caption words that refer to such features are fuzzy, associated with a degree of suitability that is a function of position in a range of values. For instance, a picture is "big" with a probability that is monotonically increasing with the length of its diagonal in pixels, say $d^2/(10000+d^2)$. Fuzzy terms are often context-dependent, so for instance a big icon is smaller than a big photograph. But this is less a problem in the restricted world of Web pages, with its narrow range of dimensions, durations, display parameters, etc., than it is for the real world; typical values of fuzzy quantities can be obtained by tests with users.

Media objects also have a structure that can be referenced by component-depictive caption sentences. Although we have argued here that captions are the easiest way to understand media on the Web, there will always be valuable information in the structure of a media object that captions do not think to convey. Images, audio, and video represent multidimensional spaces, and one can measure

intensity of the signal at locations in those spaces. The gradient magnitude, or local change in the intensity, is an excellent clue to the boundaries between phenomena in the data. Color changes along edges in an image suggest separate objects; changes in the frequency-intensity plot of audio suggest beginning and ends of sounds (Blum et al, 1997); and many simultaneous changes between corresponding locations in two video frames suggest a new shot (Hauptmann and Witbrock, 1997). Partition of a media object into semantically meaningful subobjects based on such changes is "segmentation" (Russ, 1995), and is central to many kinds of content analysis. But despite years of research, segmentation is not especially reliable because of the variety of the phenomena it must address. And when we have objects of multiple colors or textures, like a panther with spots or a beach with varying textures of gravel, domain-dependent knowledge must be used to group segmented regions into coherent objects regardless of their differences. Human faces are a classic example: Eyes, mouths, and hair contrast strongly with flesh.

Figure 4 shows an example image segmentation, for the bottom left image in Figure 1. The Canny algorithm (Russ, 1995) was applied separately to the red, green, and blue image components of the original color image, and a union taken of the resulting region-partition edge cells. Dilation and erosion operations were used to simplify the segmentation and reduce noise. White represents areas in which significant changes in color are present for adjacent pixels; both white and black regions of the image can correspond to objects. It can be observed that faces were extracted but not all their features, so the faces would be hard to identify. The widely varying colors of the child's clothes and toy caused difficulties for the segmentation as it merged them with furniture.

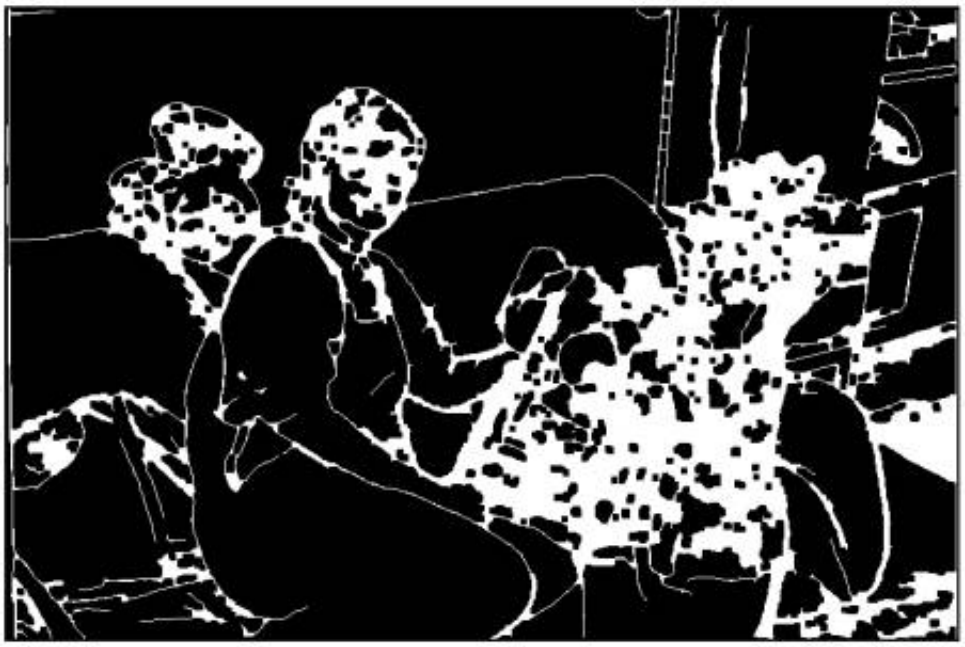

Figure 4: A segmentation of the lower left picture in Figure 1.

Figure 5 shows an example natural audio spectrum, a sonar recording from deep-sea sensors. The horizontal axis is time (the whole picture covers several minutes) and the vertical axis the frequency in the range 1 hertz to 128 hertz. Segmentation could extract the whale calls (the angled shapes in the upper frequencies) from the earthquakes (in the lower frequencies) and the calibration signals (the straight horizontal lines).

Figure 5: Example sonar recording; time is horizontal axis and frequency is vertical axis.

Table 8 shows the structure of some video, the CNN Headline News broadcast from 5:30 PM to 6:00 PM EST on May 4, 2003. Television video is hierarchically structured with strict timing constraints, and this structure can be inferred and exploited to understand what is going on (Gao, Xin, & Xi, 2002).

Table 8: The video structure of 15 minutes of a television newscast.

High level structure	Medium-level structure	Low-level structure (partial)
Introduction	30.0: Headlines	
	30.5: Short summaries	
Stories	31.5: First story: Iraqi weapons	Shot 1, 7 sec.
		Shot 2, 7 sec.
		Shot 3, 11 sec.
		Shot 4, 30 sec.
		Shot 5, 12 sec.
		Shot 6, 8 sec.
		Shot 7, 7 sec.
	33.0: Second story: Secretary of Defense discussing detainees	Shot 1, 21 sec.
		Shot 2, 57 sec.
		Shot 3, 9 sec.
	34.5: Third story: Democratic Presidential debate	Shot 1: 8 sec.
		Shot 2: 3 sec.
		Shot 3: 5 sec.
		Shot 4: 3 sec.
	35.0: Fourth story: Shuttle astronauts	Shot 1: 6 sec.
		Shot 2: 9 sec.
		Shot 3: 8 sec.
		Shot 4: 2 sec.
Advertisements	35.5: Teasers	Shot 1: 5 sec.
		Shot 2: 6 sec.
		Shot 3: 13 sec.
		Shot 4: 8 sec.
	36.0: Commercial advertisements	
Stories	38.0: Weather	
	39.5: Fifth story	
	40.0: Sixth story	
	40.5: Seventh story	
	41.0: Sports story 1	
	41.5 Sports story 2	

	42.0 Sports story 3 42.5 Sports stories 4 and 5 43.0 Sports story 6	
Advertisements	43.5 Teasers	
	44.0 Commercial advertisements	

Additional constraints can be exploited to understand media because media space obeys additional laws. For instance, objects closer to a camera appear larger, and gravity is usually downward, so inferences can often be made about distance to objects and support relationships between objects from this. For video, objects in motion can be assumed to continue that motion unless something like a collision prevents it. For audio, many natural sounds decay in intensity exponentially when their cause is removed.

The subject of a media object can often be inferred even without spatial or temporal references in a caption (Rowe, 2002a). Media on the Web are rarely selected at random but are intended to illustrate something. The subject must be clear to accomplish this purpose. That means it is typically near the center of the media space, not "cut off" or bordering edges of the media space. It also usually means it is well distinguished from nearby regions in intensity or texture. These criteria allow us to focus content analysis on the most important parts of the media object and thereby better match any caption information we have. For instance, for pictures of aircraft, we can use subject identification to rate images by aircraft size.

Special cases of caption-media correspondence

While finding a caption-media correspondence can be difficult in general, there are several easier subcases. One is the recognition and naming of faces in an image using its caption (Srihari, 1995; Satoh, Nakamura, & Kanda, 1999; Houghton, 1999). Names are distinctive capitalized structures in text, often associated with special words like "Mr.", "Prof.", and "President". Faces are distinctive visual features in images, with a distinctive range of flesh tones, and distinctive eye, mouth, and hair features in a rather narrow range of geometric patterns. So good candidate names and faces can be extracted from captions and images without much trouble. Name-face associations can be found using any available spatial-reference terms in the caption and by looking for consistent evidence between different modalities or different images mentioning the same name. For instance, video of a television newscast showing a face often contains the spoken name, the name in the closed-caption, and the name written in graphics on the screen.

Similarly, dates and locations for associated media can often be extracted from captions with simple rules (Smith, 2002). Another important special case is captioned graphics since the structure of graphics is easier to infer than the structure of camera images (Rajagopalan, 1996; Anderson & McCartney, 2003). Graphic objects like lines, polygons, and textures can map to particular caption words, and common sense or experience can often give match criteria. For instance, thin solid lines on a map often map to roads, circles to towns, numbers to numbers of roads, and characters to names of towns. The proposed SVG image format has received considerable interest recently as a Web vector-graphics format that should be faster to download; at the same time, it is easier to map to captions than the standard image formats (Arah, 2003).

Learning connections between captions and media

In the general case, we must match segmented regions of an image to words of a caption, obeying the constraints discussed so far. This must take into account the properties of the regions and the properties of the words, and try to find the best matches. Statistical methods can be similar to those used for identifying clues for captions in text, except that we must now consider many different categories to which to assign words, one for each kind of region, entailing problems of obtaining a large enough sample for reliable evidence. Graph matching methods can be used when relationships between a set of regions are known or are likely (Gold & Rangarajan, 1996). Relaxation techniques, useful for many problems in vision, are helpful. Also helpful are preformulated knowledge about the settings and backgrounds of things described in captions (Sproat, 2001). For instance, when a caption says "President speaking" we expect to see a man in a suit with a microphone and a podium with insignia, and people seated or standing behind or to the side.

To reduce the problem of manually specifying detailed knowledge about object appearance under many different conditions, machine-learning methods may be fruitful. (Barnard et al, 2002) propose a process for learning associations between words of a caption and

image regions by taking statistics over a training set of image-caption pairs, analogous to learning of word associations in parsing language. (Roy, 2000/2001) learns associations from spoken descriptions including syntactical and semantic constraints. Case-based or "best-match" reasoning can help with these learning tasks because instances of an object often cluster in multidimensional feature space (Barnard, Duygulu, & Forsyth, 2001). This work is promising because it directly addresses the key problem of dealing with a large amount of information, but much more work needs to be done. For instance for Figure 4, associating the words "tycho" (the child) and "toy" in the name of the image file with regions in the segmentation will be difficult because neither corresponds to a clear region of the image.

Dynamic phenomena in media and media sequences

Additional phenomena occur with media that extend over time (video, audio, and software). Physical-motion verbs and event nouns can be depicted directly rather than just their subjects and objects, as with "Growth of a flower in time-lapse photography". As a start, dynamic phenomena can be categorized as translational (e.g. "go"), configurational ("develop"), property-change ("lighten"), relationship-change ("fall"), social ("report"), or existential ("appear"). Each needs domain-dependent knowledge to recognize in the media object.

A related phenomenon is that media objects often occur in sets on a page that are displayed as sequences or structures where position is important. Some examples are:

- The media are in increasing order of time. For instance, pictures may show key steps of a process like a ceremony.
- The media represent different locations in some physical space. For instance, pictures may show New Year's celebrations.
- The media illustrate a hierarchy of concepts. For instance, pictures may show a vehicle and its components.
- The media represent functional or causal sequences. For instance, a visual repair manual for a vehicle may show the parts before and after disassembly.

And media sets and sequences can be embedded in other sets and sequences. Such structures provide additional implicit semantics that can be exploited for retrieval. For instance, recognition that a set of pictures is in time sequence is helpful for retrieval even when the captions say nothing.

Conclusions

Multimedia is an increasing presence on the World Wide Web as a more natural match to the multimodality of humans than text. Captions usually provide a much easier way to automatically understand media than attempting media content analysis, which is often problematic and inconsistent. On the Web, something like a caption is usually nearby on the page if we can recognize it. But there are ambiguities in what constitutes a caption, problems in interpreting the caption, problems in understanding multiple captions on the same object and multiple objects for the same caption, problems in understanding the media at least superficially, and problems in matching the caption to the media. A key issue is how deep an understanding is necessary of captions and media to build useful systems; the survey of this chapter suggests that much can be done without detailed understanding, but high-accuracy systems require more.

We envision future multimedia-retrieval technology as not much different from that of the present, just better. Keyword search will continue to provide the easiest access, and text will remain important to explain media objects on Web pages. As this chapter has shown, good progress has been made in recent years on the technical issues needed to improve performance. Now a variety of results and tools are available for those building useful systems that permit us to illustrate articles, books, and presentations with perfectly matched pictures, sounds, video clips, and demonstration programs, enhancing their information content.

References

Armitage, L. H., & Enser, P. (1997). Analysis of user need in image archives. *Journal of Information Science, 23* (4), 287-299.
Anderson, M., and McCartney, R. (2003). Diagram processing: computing with diagrams. *Artificial Intelligence, 145,* 181-226.
Arah, T. (2003). SVG: the future Web format. Retrieved from www.designer-info.com/Writing/ svg_format.htm.
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.

Barnard, K., Duygulu, P., & Forsyth, D. (2001). Clustering art. Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, II-434-II-441.
Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., & Jordan, M. (2002). Matching words and pictures. Retrieved from http://www.cs.berkeley.edu/~kobus/research/publications/JMLR/JMLR.pdf.
Bjarnestam, A. (1998, February). Text-based hierarchical image classification and retrieval of stock photography. Retrieved from
Eakins, J., Harper, D., & Jose, J. (Eds.), *The challenge of image retrieval: papers presented at the Research Workshop held at the University of Northumbria*, Newcastle-on-Tyne, BCS Electronics Workshops in Computing, http://www1.bcs.org.uk/homepages/686.
Blum, T., Deislar, D., Wheaton, J., & Wold, E. (1997). Audio databases with content-based retrieval. In Maybury, M. (Ed.), *Intelligent multimedia information retrieval* (pp. 113-135), Menlo Park, CA: AAAI Press / MIT Press.
Casey, E., & Lecolinet, R. (1996, July). A survey of methods and strategies in character segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18* (7), 690-706.
Cohen, P. (1992, November). The role of natural language in a multimodal interface. Proc. Fifth Annual Symposium on User Interface Software and Technology (UIST 92), Monterey, CA, 143-149.
Cronin, B. (1995, Fall). The development of descriptive video services. *Interface, 17* (3), 1-4.
The Dayton Art Institute (2003). Access art. Retrieved from http://tours.daytonartinstitute.org /accessart/index.cfm.
DiManzo, M., Adorni, G., & Giunchiglia, F. (1986, July). Reasoning about scene descriptions. *Proceedings of the IEEE, 74(7)*, 1013-1025.
Di Tomaso, 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.
Favela, J., & Meza, V. (1999, September/October). Image-retrieval agent: integrating image content and text. *IEEE Intelligent Systems, 14* (5), 36-39.
Finkelstein, S. (2003). BESS vs. image search engines. Retrieved from www.sethf.com/anticensorware/ bess/image.php.
Forsyth, D. A. (1999, Fall). Computer vision tools for finding images and video sequences. *Library Trends, 48* (2), 326-355.
Flickner, M., Sawhney, H. Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., & Yanker, P. (1995, September). Query by image and video content: The QBIC System. *Computer, 28* (9) 23-32.
Gao, X., Xin, H., & Ji, H. (2002). A study of intelligent video indexing system. Proc. Fourth World Congress on Intelligent Control and Automation, Vol. 3, pp. 2122-2126.
Giunchiglia, E., Armando, A., Traverso, P., & Cimatti, A. (1996, August). Visual representation of natural language scene descriptions. IEEE Transactions on Systems, Man and Cybernetics, Part B, 26 (4), 575-589.
Gold, S., & Rangarajan, A. (1996, April). A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18* (4), 377-388.
Guglielmo, E. J. & Rowe, N. (1996, July). Natural-language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems, 14* (3), 237-267.
Hauptman, G., & Witbrock, M. (1997). Informedia: News-on-demand multimedia information acquisition and retrieval. In Maybury, M. (Ed.), *Intelligent multimedia information retrieval* (pp. 215-239), Menlo Park, CA: AAAI Press / MIT Press.
Heaton, J. (2002). *Programming Spiders, Bots, and Aggregators in Java*. San Francisco, CA: Cybex.
Heidorn, P. B. (1997). Natural language processing of visual language for image storage and retrieval. Ph.D. dissertation, School of Information Sciences, University of Pittsburgh. Retrieved from http://www.isrl.uiuc.edu/~pheidorn/pub/VerbalImage/.
Heidorn, P. B. (1999, November). The identification of index terms in natural language objects. Proc. Annual Conference of the American Society for Information Science, Washington, DC, pp. 472-481.
Houghton, R. (1999 September/October). Named faces: putting names to faces. *IEEE Intelligent Systems, 14* (5), 45-50.
Jansen, B. J., Goodrum, A., & Spink, A. (2000). Search for multimedia: video, audio, and image Web queries. *World Wide Web Journal, 3* (4), 249-254.
Jorgensen, C. (1998). Attributes of images in describing tasks. *Information Processing and Management*, *34* (2/3), 161-174.
Kern, N., Schiele, B., Junker, H., Lukowicz, P., & Troster, G. (2002). Wearable sensing to annotate meeting recordings. Proc. Sixth International Symposium on Wearable Computers (ISWC), 186-193.
Leslie, D. (2002, October). Using Javadoc and XML to produce API reference documentation. Proc. ACM-SIGDOC Conference, Toronto, Canada, 109-109.
Lienhart, R. (2000, June). A system for effortless content annotation to unfold semantics in videos. In Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries, Hilton Head, SC, 45-59.
Matsakis, P., Keller, J., Wendling, L., Marjarnaa, and 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.
Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990, Winter). Five papers on Wordnet. *International Journal of Lexicography*, *3* (4).
Mukherjea, S., & Cho, J. (1999). Automatically determining semantics for World Wide Web multimedia information retrieval. *Journal of Visual Languages and Computing*, *10*, 585-606.
Mukherjee, A. (1998). Neat versus scruffy: a review of computational models for spatial expressions. In Oliver, P., & Gapp, K.-P. (Eds.), *Representation and processing of spatial expressions* (pp. 1-36). Mahwah, NJ: Lawrence Erlbaum.
NCIP (National Center to Improve Practice in Special Education) (2003). Video and captioning. Retrieved from http://www2.edc.org/NCIP/library/v&c/Toc.htm.
Perfetti, C. A., Beverly, S., Bell, L., Rodgers, K., & Faux, R. (1987). Comprehending newspaper headlines. *Journal of Memory and Language*, *26,* 692-713.
Pineda, L., & Garza, G. (2000, June). A model for multimodal reference resolution. *Computational Linguistics, 26* (2), 139-193.
Rajagopalan, R. (1996). Picture semantics for integrating text and diagram input. *Artificial Intelligence Review, 10* (3-4), 321-344.
Rowe, N. (1994). Inferring depictions in natural-language captions for efficient access to picture data. *Information Processing and Management, 30* (3), 379-388.
Rowe, N. (1996). Using local optimality criteria for efficient information retrieval with redundant information filters. *ACM Transactions on Information Systems, 14* (2) (April), 138-174.
Rowe, N. (1999, Fall). Precise and efficient retrieval of captioned images: The MARIE project. *Library Trends*, *48* (2), 475-495.
Rowe, N. (2002a, January/February). Finding and labeling the subject of a captioned depictive natural photograph. *IEEE Transactions on Data and Knowledge Engineering*, *14* (1), 202-207.
Rowe, N. (2002b). MARIE-4: A high-recall, self-improving Web crawler that finds images using captions. *IEEE Intelligent Systems*, *17* (4), July/August, 8-14.
Rowe, N. (2002c, July). Virtual multimedia libraries built from the Web. Proc. Second ACM-IEEE Joint Conference on Digital Libraries, Portland, OR, 158-159.
Roy, D. K. (2000/2001). Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication, 4* (1).
Russ, J. C. (1995). *The image processing handbook, second edition.* Boca Raton, FL: CRC Press.
Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, *24*, 513-523.
Sannomiya, T., Amagasa, T., Yoshikawa, M., & Uemura, S. (2001, January). A framework for sharing personal annotations on web resources using XML. Proc. Workshop on Information Technology for Virtual Enterprises, Gold Coast, Australia, 40-48.
Satoh, S., Nakamura, Y., & Kanda, T. (1999, January-March). Name-It: naming and detecting faces in news videos. *IEEE Multimedia, 6* (1), 22-35.
Satoh, T., Tachikawa, M., & Yamaai, T. (1994). Document image segmentation and text area ordering. *IEICE Transactions on Information and Systems, E77-01* (7), 778-784.
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.
Smith, D. (2002, July). Detecting events with date and place information in unstructured texts. Proc. Second ACM/IEEE-CS Joint Conference on Digital Libraries, Portland, OR, 191-196.
Spivey, M. J., Richardson, D. C., Tyler, M. J., & Young, E. E. (2000, August). Eye movements during comprehension of spoken scene descriptions. Proc. of the 22nd Annual Meeting of the Cognitive Science Society, Philadelphia, PA, 487-492.
Sproat, R. (2001). Inferring the environment in a text-to-scene conversion system. Proc. International Conference on Knowledge Capture, Victoria, British Columbia, Canada, 147-154.
Srihari, R. (1995). Use of captions and other collateral text in understanding photographs. *Artificial Intelligence Review*, *8* (5-6), 409-430.
Srihari, R., & Zhang, Z. (1999, Fall). Exploiting multimodal context in image retrieval. *Library Trends, 48* (2), 496-520.
Srihari, R., & Zhang, Z. (2000, July-September). Show&Tell: A semi-automated image annotation system. *IEEE Multimedia, 7* (3), 61-71.
Srihari, R., Zhang, Z., & Rao, A. (2000). Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval, 2* (2), 245-275.
Stefik, M. (1995). *Knowledge systems*. San Francisco: Morgan Kaufmann.
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.
Swain, M. J. (1999, February). Image and video searching on the World Wide Web. Retrieved from Harper, D., and Eakins, J. (Eds.),

CIR-99: The challenge of image retrieval: papers presented at the 2nd UK Conference on Image Retrieval, Newcastle-on-Tyne, BCS Electronics Workshops in Computing, http://www1.bcs.org.uk/homepages/686.
W3C, (1999). Web content accessibility guide. Retrieved from http://www.w3.org/TR/WAI-WEBCONTENT/.
Watanabe, Y., Okada, Y., Kaneji, K., & Sakamoto, Y. (1999, September/October). Retrieving related TV news reports and newspaper articles. *IEEE Intelligent Systems, 14* (5), 40-44.
Witten, I., & Frank, E. (2000). *Data mining: Practical machine learning with Java implementations*. San Francisco, CA: Morgan Kaufmann.
Wu, V., Manmatha, R., & Riseman, E. (1997, July). Finding text in images. Proc. Second ACM Conference on Digital Libraries, Philadelphia, PA, 3-12.

---

[1] Can you briefly summarize or list?