Finding Photograph Captions Multimodally on the World Wide Web

Neil C. Rowe and Brian Frew

Code CS/Rp, Department of Computer Science
Naval Postgraduate School
Monterey, CA USA 93943
rowe@cs.nps.navy.mil, http://www.cs.nps.navy.mil/research/marie/index.html

Abstract

Several software tools index text of the World Wide Web, but little attention has been paid to the many valuable photographs. We present a relatively simple way to index them by localizing their likely explicit and implicit captions with a kind of expert system. We use multimodal clues from the general appearance of the image, layout of the Web page, and the words nearby the image that are likely to describe it. Our MARIE-3 system avoids full image processing and full natural-language processing, but demonstrates a surprising degree of success, and can thus serve as a preliminary filtering for such detailed content analysis. Experiments with a randomly chosen set of Web pages concerning the military showed 41% recall with 41% precision for individual caption identification, or 70% recall with 30% precision, although captions averaged only 1.4% of the page text.

1. Introduction

Pictures, especially photographs, are one of the most valuable resources available on the Internet through World Wide Web. Unlike text, most photographs are valuable primary sources of real-world data. And the volume of images available on the Web has grown quickly. For these reasons, indexing and retrieval of photographs is becoming increasingly critical for the Web.

Several researchers have recently been looking at the problem of information retrieval of pictures from large libraries. (Smith and Chang 1996) describes a system that does both simple image processing and simple caption processing. (Frankel, Swain, and Athitsos 1996) provides an interim report on Webseer, a search engine for images on the WorldWide Web. The PICTION Project (Srihari 1995) has investigated the more general problem of the relationship between images and captions in a large photographic library like a newspaper archive. Our own MARIE Project in the MARIE-1 (Guglielmo and Rowe 1996) and MARIE-2 (Rowe 1996) systems explored similar issues in a large photographic library at the NAWC-WD Navy facility in China Lake, California USA.

Full image understanding is not necessary to index Web images well, because descriptive text is usually nearby. We need to recognize relevant text and determine which picture it describes. This requires understanding the semantics of page layout on Web pages, like where captions are likely to occur and how italics and font changes are marked. It also requires some simple linguistic analysis, including searches for linguistic clues like reference phrases ("the picture above shows") and knowledge of which nouns in a sentence represent a depictable objects (Rowe 1994). It also can use some simple image processing like counting the number of colors in the image to guess if it is complex enough for a photograph. Thus multimodal analysis -- layout, language, and image in synergy -- appears the best way to index pictures, a strategy noted as valuable in several other multimedia information retrieval applications (Maybury 1997). Thus MARIE-3 does multimodal indexing of pages of the World Wide Web from their source code. The implementation we describe here is in Quintus Prolog, and test runs were done on a Sun Sparcstation.

2. The image neuron

We would like to index millions of pages reasonably quickly, so we cannot do any complicated image processing like shape classification (Rowe and Frew 1997). But one critical decision we can make without much processing is whether an image is likely to be a photograph (and thus valuable to index) or non-photographic graphics (often not interesting in itself). Intuitively, photographs tend to be close to square, have many colors, and have much variation in color between neighbor pixels. Photographs and non-photographs are not distinguished in the HTML page-markup language used by World Wide Web since both appear as an "img" construct with an embedded "src=" string. The format of the image file is no clue either, since the two most common Web formats of GIF and JPEG are used equally often for photographs and non-photographs. Some HTML images are selectable by mouse, but they are equally photographs and non-photographs too. So distinguishing photographs requires some image analysis. This is a classic case for a linear-classifier perceptron neuron. If a weighted sum of input factors exceeds a fixed threshold, the picture is called a photograph, else not. We used seven factors: size, squareness, number of colors, fraction of impure colors, neighbor variation, color dispersion, and whether the image file name suggested a photograph. These were developed after expert-system study of a variety of Web pages with images, and were chosen as a kind of "basis set" of maximally-different factors that strongly affect the decision. The factors were computed in a single pass through a color-triple pixel-array representation of the image. Size and number-of-colors factors were explored in (Frankel, Swain, and Athitsos 1996), but the other five factors are unique to us.

A nonlinear sigmoid function is applied to the values to all but the last factor before inputting them to the perceptron. The sigmoid function used is the common one of (tanh[(x/s)-c)]+1)/2 which ranges from 0 to 1. The "sigmoid center" c is where the curve is 0.5 (about which the curve is radially symmetric), and "sigmoid spread" s controls its steepness. The sigmoid nonlinearity helps remediate the well-known limitations of linear perceptrons. It also design more intuitive because each sigmoid can be adjusted to represent the probability that the image is a photograph from that factor alone, so the perceptron becomes just a probability-combination device.

The "name-suggests-photograph" factor is the only discrete factor used by the perceptron. It examines the name of the image file for common clue words and abbreviations. Examples of non-photographic words are "arrw", "bak", "btn", "button", "home", "icon", "line", "link", "logo", "next", "prev", "previous", and "return". To find such words, the image file name is segmented at punctuation marks, transitions between characters and digits, and transitions from uncapitalized characters to capitalized characters. From our experience, we assigned a factor value of 0.7 to file names without such clues (like "blue_view"); 0.05 to those with a non-photograph word (like "blue_button"); 0.2 to those whose front or rear is such a non-photograph word (like "bluebutton"); and 0.9 to those with a photographic word (like "blue_photo").

The experiments described here used a training set of 261 images taken from 61 pages, and a test set of 406 images taken from 131 (different) pages. The training set examples were from a search of neighboring sites for interesting images. For the test set we wanted a more random search, so we used the Alta Vista Web Search Engine (Digital Equipment Corp.) to find pages matching three queries about military laboratories.

Neuron training used the classic "Adaline" feedback method (Simpson 1990) since more sophisticated methods did not perform any better. So weights were only changed after an error, by the product of the amount of error (here the deviation from 1.0), the input associated with the weight, and a "learning rate" factor. After training, the neuron is run to generate a rating for each image. Metrics for recall (fraction of actual photographs classified as such) and precision (fraction of actual photographs among those classified as photographs) can be traded off, depending on where the decision threshold is set. We got 69% precision (fraction of correct photographs of those so identified) with 69% recall (fraction of correct photographs found). Figure 1 shows implementation statistics.

Statistic	Training set	Test set
Size of HTML source (bytes)	13,269	1,241,844
Number of HTML pages	61	131
Number of images on pages	261	406
Number of actual photographs	174	113
Image-analysis time	13,811	16,451
Image-neuron time	17.5	26.8
HTML-parse time	60.3	737.1
Caption-candidate extraction time	86.6	1987
Number of caption candidates	1454	5288
Number of multiple-image candidates	209	617
Caption-candidate bytes	78,204	362,713
Caption-analysis and caption-neuron time	309.3	1152.4
Final-phase time	32.6	153.9
Number of actual captions	340	275

Figure 1: Statistics on our experiments with the training and test sets; times are in CPU seconds.

3. Parsing of Web pages

To find photograph captions on Web pages, we work on the HTML markup-language code for the page. So we first "parse" the HTML source code to group the related parts. Image references are easy to spot with their "img" and "src" tags. We examine the text nearby each image reference for possible captions. "Nearby" is defined to mean within a fixed number of lines of the image reference in the parse; in our experiments, the number was generally three. We exclude obvious noncaptions from a list (e.g. "photo", "Introduction", "Figure 3", "Welcome to... ", and "Updated on..."). Figure 2 lists the caption candidates found for the Web page shown in Figure 3.

Image field01 line 5 captype filename distance 0: 'field 1'

Image field01 line 5 captype title distance -3: 'The hypocenter and the Atomic Bomb Dome

Image field01 line 5 captype h2 distance 1: 'The hypocenter and the Atomic Bomb Dome'

Image field01 line 5 captype plaintext distance 2: 'Formerly the Hiroshima Prefectural Building for the Promotion of Industry, the "Atomic Bomb Dome" can be seen from the ruins of the Shima Hospital at the hypocenter.'

Image field01 line 5 captype plaintext distance 3: 'The white structure is Honkawa Elementary School.'

Image island line 13 captype filename distance 0: island

Image island line 13 captype plaintext distance -1: 'Photo: the U.S. Army.'

Image island line 13 captype plaintext distance -2: 'November 1945.'

Image island line 13 captype plaintext distance -3: 'The tombstones in the foreground are at Sairen-ji(temple).'

Image island line 13 captype h2 distance 0: 'Around the Atomic Bomb Dome before the A-Bomb'

Image island line 13 captype plaintext distance 0: 'Photo: Anonymous (taken before 1940)'

Image island2 line 14 captype filename distance 0: 'island 2'

Image island2 line 14 captype h2 distance -1: 'Around the Atomic Bomb Dome before the A-Bomb'

Image island2 line 14 captype plaintext distance -1: 'Photo: Anonymous (taken before 1940)'

Image island2 line 14 captype h2 distance 0: 'Around the Atomic Bomb Dome after the A-Bomb'

Image island2 line 14 captype plaintext distance 0: 'Photo: the U.S. Army'

Image ab-home line 17 captype filename distance 0: 'ab home'

Figure 2: Caption candidates generated for the example Web page. "Ab-home" is an icon, but the other images are photographs. "h2" means heading font.

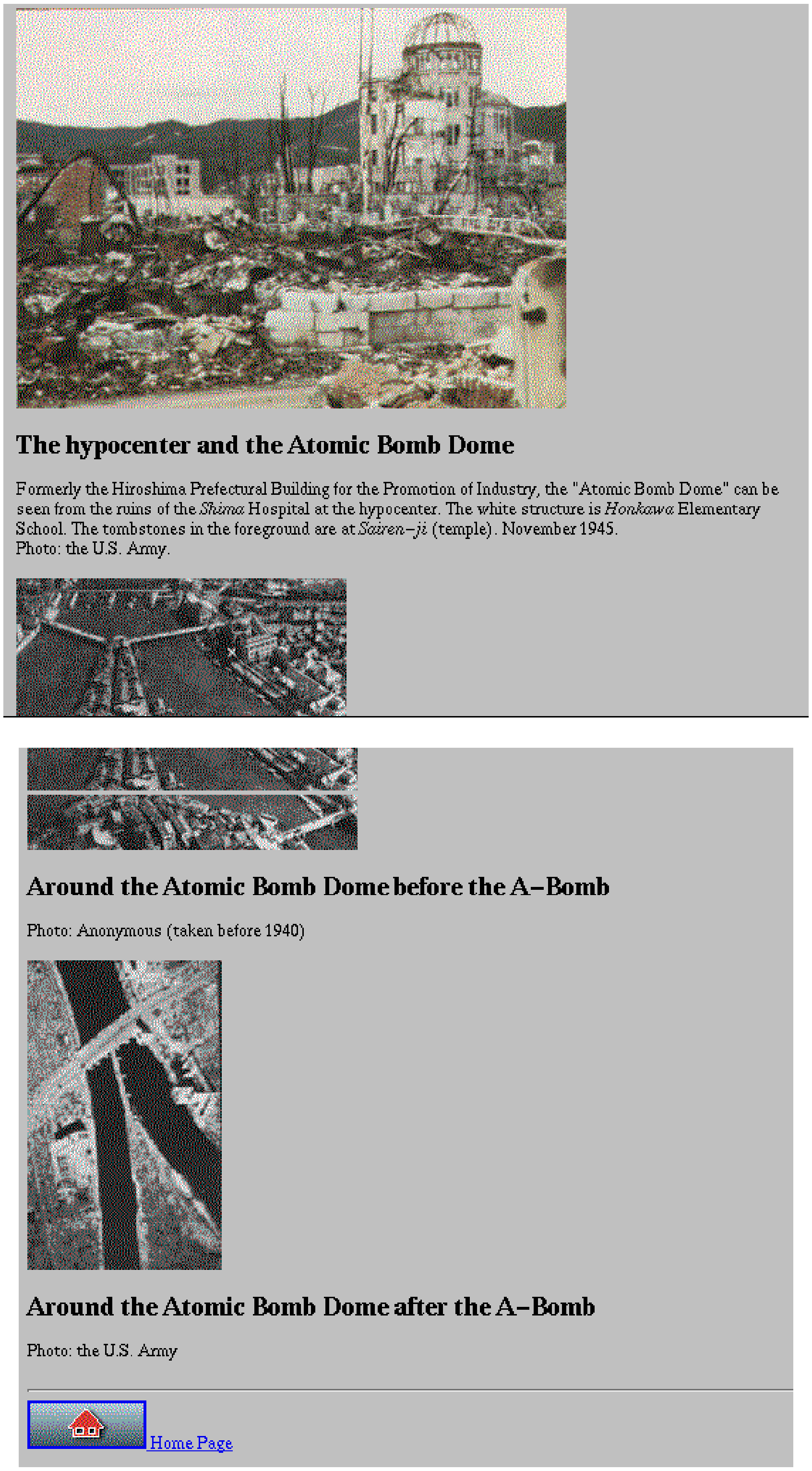

Figure 3: An example page from World Wide Web.

There is a exception to captioning when another image reference occurs within the three lines. This is a case of a principle analogous to those of speech acts:

The Caption-Scope Nonintersection Principle: Let the "scope" of a caption-image pair be the characters between and including the caption and the image. Then the scope for a caption on one image cannot intersect the scope for a caption on another image.

The justification for this is that the space between caption and image is thought an extension of the image, and the reader would be confused if this were violated.

Besides the distance, we must also identify the type of caption. Captions often appear differently from ordinary text. HTML has a variety of such text markings, including font family (like Helvetica), font style (like italics), font size (like 12 pitch), text alignment (like centering), text color (like blue), text state (like blinking), and text significance (like a page title). Ordinary text we call type "plaintext". We also assign caption types to special sources of text, like "alt" strings associated with an image on nonprinting terminals, names of Web pages that can be brought up by clicking on the image, and the name of the image file itself, all of which can provide useful clues as to the meaning of the image.

Not all HTML text markings suggest a caption. Italicized words among nonitalicized words probably indicate word emphasis. Whether the text is clickable or "active" can also be ignored for caption extraction. So we eliminate such markings in the parsed HTML before extracting captions. This requires determining the scope of multi-item markings.

4. The caption neuron

Sometimes HTML explicitly connects an image and its caption. One way is the optional "alt" string. Another is a textual hypertext link to an image. A third way is the "caption" construct of HTML, but it is rare and did not occur in any of our test and training cases. A fourth way is text on the image itself, detectable by special character-recognizing image processing, but we did not explore this. All four ways often provide caption information, but not always (as when they are undecipherable codes), so they still must be evaluated as discussed below.

But most image-caption relationships are not explicit. So in general we must consider carefully all the text near the images. Unfortunately, Web pages show much inconsistency in captioning because of the variety of people constructing them and variety of intended uses. A caption may be below the image or above; it may be in italics or larger font or not; it may be signalled by words like "the view above" or "the picture shows" or "Figure 3:" or not at all; it may be a few words, a full sentence, or a paragraph. So often there are many candidate captions. Full linguistic analysis of them as in MARIE-1 and MARIE-2 (parsing, semantic interpretation, and application of pragmatics) would reveal the true captions, but this would require knowledge of all word senses of every subject that could occur on a Web page, plus disambiguation rules, which is impractical. So MARIE-3 instead uses indirect clues to assign probabilities to candidate captions, and finds the best matches for each picture image.

We use a seven-input "caption" neuron like the image neuron to rank possible caption-image pairs. After careful analysis like that for an expert system, we identified seven factors: (F1) distance of the candidate caption to the image, (F2) confusability with other text, (F3) highlighting, (F4) length, (F5) use of particular signal words, (F6) use of words in the image file name or the image text equivalent ("alt" string), (F7) and use of words denoting physical objects. Again, sigmoid functions convert the continuous factors to probabilities of a caption based on the factor alone, the weighted sum of the probabilities is taken to obtain an overall likelihood, and the neuron is trained similarly to the image neuron.

The seventh input factor F7 exploits the work on "depictability" of caption words in (Rowe 1994), and rates higher the captions with more depictable words. For F3, rough statistics were obtained from surveying Web pages, and used with Bayes' Rule to get $p(C|F) = p(F|C) * p(C) / p(F)$, where F means the factor occurs and C means the candidate is a caption. F3 covers both explicit text marking (e.g. italics) and implicit (e.g. surrounding by brackets, beginning with "The picture shows", and beginning with "Figure" followed by a number and a colon). F5 counts common words of captions (99 words, e.g. "caption", "photo", "shows", "closeup", "beside", "exterior", "during", and "Monday"), counts year numbers (e.g. "1945"), and negatively counts common words of noncaptions on Web pages (138 words, e.g. "page", "return", "welcome", "bytes", "gif", "visited", "files", "links", "email", "integers", "therefore", "=", and "?"). F6 counts words in common between the candidate caption and the segmentation of the image-file name (like "Stennis" for "View of Stennis making right turn" and image file name "StennisPic1") and any "alt" string. Comparisons for F5 and F6 are done after conversion to lower case, and F6 ignores the common words of English not nouns, verbs, adjectives, or adverbs (154 words, e.g. "and", "the", "no", "ours", "of", "without", "when", and numbers), with exceptions for physical-relationship and time-relationship prepositions. F6 also checks for words in the file name that are abbreviations of words or pairs of words in the caption, using methods of MARIE-2.

5. Combining image with caption information

The final step is to combine information from the image and caption neurons. We compute the product of the probabilities that the image is a photograph and that the candidate is a caption for it, and then apply a threshold to these values. A product is appropriate here because the evidence is reasonably independent, coming as it does from quite different media. To obtain probabilities from the neuron outputs, we use sigmoid functions again. We got the best results in our experiments with a sigmoid center of 0.8 for the image neuron and a sigmoid center of 1.0 for the caption neuron, with sigmoid spreads of 0.5.

Two details need to be addressed. First, a caption candidate could be assigned to either of two nearby images if it is between them; 634 of the 5288 caption candidates in the test set had such conflicts. We assume captions describe only one image, for otherwise they would not be precise enough to be worth indexing. We use the product described above to rate the matches and generally choose the best one. But there is an exception to prevent violations of the Caption-Scope Nonintersection Principle when the order is Image1-Caption2-Caption1-Image2 where Caption1 goes with Image1 and Caption2 goes with Image2; here the caption of the weaker caption-image pair is reassigned to the other image.

Second, our training and test sets showed a limited number of captions per image: 7% had no captions, 57% had one caption, 26% had two captions, 6% had three, 2% had four, and 2% had five. Thus we limit to three the maximum number of captions we assign to an image, the best three as per the product values. However, this limit is only applied to the usual "visible" captions, and not to the file-name, pointer page-name, "alt"-string, and page-title caption candidates which are not generally seen by the Web user.

Figure 4 shows example results and Fig. 5 shows overall performance. We got 41% recall for 41% precision, or 70% recall with 30% precision, with the final phase of processing. This is a significant improvement on the caption neuron alone, where recall was 21% for 21% precision, demonstrating the value of multimodality; 1.4% of the total text of the pages was captions. Results were not too sensitive to the choice of parameters. Maximum recall was 77% since we limited images to three visible captions; this improves to 95% if we permit ten visible captions. The next step will be to improve this performance through full linguistic processing and shape identification as in MARIE-2.

[field01,'The hypocenter and the Atomic Bomb Dome'] @ 0.352

[field01,'Formerly the Hiroshima Prefectural Building for the Promotion of Industry, the "Atomic Bomb Dome" can be seen from the ruins of the Shima Hospital at the hypocenter.'] @ 0.329

[field01,'The white structure is Honkawa Elementary School.'] @ 0.307

[field01,'field 1'] @ 0.263

[island,'Around the Atomic Bomb Dome before the A-Bomb'] @ 0.532

[island,'Photo: Anonymous (taken before 1940)'] @ 0.447
[island,'Photo: the U.S. Army.'] @ 0.381

[island,island] @ 0.258

[island2,'Around the Atomic Bomb Dome after the A-Bomb'] @ 0.523

[island2,'Photo: the U.S. Army'] @ 0.500

[island2,'island 2'] @ 0.274

Figure 4: Final caption assignment for the example Web page. One incorrect match was proposed, the third for image "island", due to the closest image being the preferred match.

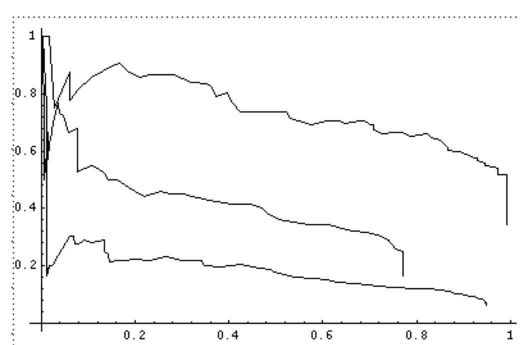

Figure 5: Recall (horizontal) versus precision (vertical) for photograph identification (the top curve), caption identification from text alone (the bottom curve), and caption identification combining both image and caption information (the middle curve).

6. Acknowledgements

This work was supported by the U.S. Army Artificial Intelligence Center, and by the U. S. Naval Postgraduate School under funds provided by the Chief for Naval Operations.

7. References

Frankel, C.; Swain, N. J. P.; and Athitsos, B. 1996. WebSeer: An Image Search Engine for the WorldWide Web. Technical Report 96-14, Computer Science Dept., University of Chicago, August.

Guglielmo, E. and Rowe, N. 1996. Natural-Language Retrieval of Images Based on Descriptive Captions. *ACM Transactions on Information Systems, 14,* 3 (July), 237-267.

Maybury, M. (ed.) 1997. *Intelligent Multimedia Information Retrieval.* Palo Alto, CA: AAAI Press.

Simpson, P. K. 1990. *Artificial Neural Systems.* New York: Pergamon Press.

Smith, J. R., and Chang, S.-F. 1996. Searching for images and videos on the World Wide Web. Technical report CU/CTR/TR 459-96-25, Columbia University Center for Telecommunications Research.

Srihari, R. K. 1995. Automatic Indexing and Content-based retrieval of Captioned Images. *IEEE Computer, 28,* 49-56.

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. and Frew, B. 1997. Automatic Classification of Objects in Captioned Depictive Photographs for Retrieval. To appear in *Intelligent Multimedia Information Retrieval*, Maybury, M., ed. Palo Alto, CA: AAAI Press.