1998

Automatic Caption Localization for Photographs on World Wide Web Pages

Rowe, Neil C.

Monterey, California. Naval Postgraduate School

Information Processing and Management in 1998

Automatic Caption Localization for Photographs on World Wide Web Pages

Neil C. Rowe and Brian Frew

Code CS/Rp, Department of Computer Science, U. S. Naval Postgraduate School, Monterey, CA USA 93943

Abstract

A variety of software tools index text of the World Wide Web, but little attention has been paid to the many photographs. We explore the indirect method of locating for indexing the likely explicit and implicit captions of photographs. We use multimodal clues including the specific words used, the syntax, the surrounding layout of the Web page, and the general appearance of the associated image. Our MARIE-3 system thus avoids full image processing and full natural-language processing, but shows a surprising degree of success. Experiments with a semi-random set of Web pages showed 41% recall with 41% precision for the task of distinguishing captions from other text, and 70% recall with 30% precision. This is much better than chance since actual captions were only 1.4% of the text on pages with photographs.

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.

This paper will appear in *Information Processing and Management* in 1998.

Introduction

Pictures, especially photographs, are one of the most valuable resources available on the Internet through the popular World Wide Web. Unlike text, most photographs are valuable primary sources of real-world data. Unlike conventional copy technology, photographs on the Web maintain their quality under transmission. Interest has increased recently in multimedia technology as its speed has improved by hardware and software refinements; photographs also benefit from these advances. This has meant that multimedia resources on the Web have grown quickly. For these reasons, indexing and retrieval of photographs is becoming increasingly critical.

Other researchers have recently been looking at this problem. Several on-line image archives exist, but their images must be manually described by keywords. Smith and Chang (1997) describes a system that does some image processing and exploits the name of the image file to find images automatically. Frankel, Swain, & Athitsos (1996) describes preliminary work on a similar system, Webseer, that decides which images are photographs and extracts the text near them. Text classification methods like "relevancy signatures" in Riloff and Lehnert (1994) are useful where captions are considered an especially broad class of text. Srihari in the PICTION Project (1995) has investigated the more general problem of the relationship between images and captions in a large photographic library like a newspaper archive. This work mostly emphasized caption processing, with some attention to simple image processing such as face extraction. It assumed captions were already extracted for their pictures, an assumption not applicable on World Wide Web. Our own MARIE Project in the MARIE-1 system, Guglielmo and Rowe (1996), and MARIE-2 system, Rowe (1996), explored similar issues in a large photographic library at the NAWC-WD Navy facility in China Lake, California USA. Rowe and Frew (1997) used knowledge of a technical vocabulary and specialized syntax to aid image processing in categorizing regions in photographs, with the goal of retrieval by content.

Full image understanding is not necessary to index Web images well, because descriptive text is usually nearby. We need to recognize this text and find what it describes. This requires understanding the semantics of layout on Web pages, like where captions are likely to be and how emphasis is marked. It also requires some language analysis, including searches for reference phrases ("the picture above shows") and nouns representing depictable objects as explored in Rowe (1994). It also requires simple image processing like counting the number of colors to recognize photographs. Thus multimodal analysis -- layout, linguistic, and image in synergy -- appears the best way to index pictures, a strategy noted as valuable for several other multimedia information retrieval applications in Maybury (1997). So 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.

To illustrates the problems faced by MARIE-3, Figure 1 shows the top of an example Web page from NAWC-WD. Two image files are referenced for the page, "logotrans.GIF" (the logo) and "page3photo9.GIF" (the photograph). Possible captions for the photograph can be in the title over it, the title under it, and the paragraph at the bottom. (Horizontal lines below the paragraph suggest these are the only possibilities.) "Pacific Ranges and Facilities Department" does caption the photograph, but is not especially helpful because departments are not directly depictable. "Sled Tracks" is more a caption, but not directly because the photograph shows only one track. The real caption information is in the second sentence of the paragraph below, even though it is not adjacent to the picture. So we have several possible captions that can be ranked along several dimensions.

The image neuron

We would like to index millions of pages quickly, so we must avoid complicated image processing, like the shape classification of Rowe and Frew (1997). But one critical decision not requiring much processing is whether an image is likely to be a photograph (and thus valuable to index and likely to have a caption) or non-photographic graphics. Intuitively, photographs tend to be nearly square, have many colors, and have much color variation 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. File format is no clue, since the 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 similarly both kinds.

So distinguishing photographs requires some image analysis. This is a classic case for the simplest kind of artificial neuron, a linear classifier (see Simpson (1990)). If a weighted sum of quantitative input factors exceeds a fixed threshold, the picture is called a photograph, else not. Figure 2 lists the factors we used. These were developed after study of a variety of Web pages with images, and were chosen as a maximally-different "basis set". The factors are computed in a single pass through a color-triple pixel-array representation of the image. (Size and number of colors are used in WebSeer.)

The second and third columns of Figure 2 give parameters |c| and |s| of a nonlinear sigmoid function that is applied to the values to all but the last factor before inputting them to the perceptron. The sigmoid used is the common one of |(tanh [(x / s) - c)] + 1) / 2| which ranges from 0 to 1. The center of horizontal symmetry is |c|, and the steepness is |s|. The nonlinearity helps remediate the well-known limitations of linear perceptrons. It also makes design easier because each sigmoid can be adjusted to represent the probability that the image is a photograph from its factor alone.

The "name-suggests-photograph" factor examines the name of the image file for common clue words and abbreviations. Example non-photograph words are "bak", "btn", "button", "home", "icon", "line", "link", "logo", "next", "prev", and "return". To find them, the image file name is segmented at punctuation marks, transitions between characters and digits, and transitions from uncapitalized to capitalized characters. From our experience, we assigned a factor value of 0.05 to names with a segmented non-photograph word (like "blue_button"); 0.2 to those whose front or rear is a non-photograph word (like "bluebutton"); 0.9 to those with a photograph word (like "blue_photo"); and 0.7 otherwise (like "blue_view").

Testing methods

Our experiments used a training set of 261 images from 61 Web pages, and a test set of 406 images from 131 (different) pages (see Figure 3). The training set was found by manually searching neighboring sites for pages with interesting images of military research. For the test set we wanted more randomly chosen pages, so we used the Alta Vista Web

Search Engine (Digital Equipment Corp.) to find pages matching these queries:

```
Army laborator* develop* image:* picture*
Nav* laborator* develop* image:* picture* view*
laborator* project Army military photograph image:*
```

("*" means arbitrary characters, and "image:*" means the page has an image.) Each query got over 100,000 answers; we used the first five windows of page pointers found for each query. This gave 44 Web pages for the first query, 46 for the second, and 41 for the third, not counting 8 unfetchable pages (probably because of page updates after indexing), 11 duplicates in different locations, and 6 pages without images. Only half (68/131) the pages concerned military research or development; others included Web help, indexes, results of searches, library guides, conference proceedings, maps, the Presidential election of 1920, historic sites in India, military-base closure memos, a college physics-department newsletter, school catalogs, military-surplus sales, moon landings, a chemistry reference handbook, amateur-radio information, child-development projects, a resume of a surgeon, a real estate guide, a newspaper salute to women in history, art paintings, contests, paranoid rantings, and detective fiction. We judged these pages quite typical of the Web. We thus had a surprisingly broad sampling of Web pages for our subsequent tests.

We dumped the HTML source code of each page to a file, and dumped each image to a separate file. Images not in GIF format were converted to it, and then all were converted to array format. Hyperlinks on the pages were not expanded. We manually identified, for training and testing, all photographs and captions on the pages; a "caption" had to describe the important content of its image.

Neuron training used the classic "Adaline" feedback method as in Simpson (1990) since more sophisticated methods did not perform any better. So weights were only changed after an error, by the product of the error (here the deviation from 1.0), the input associated with the weight, and a "learning rate" factor. After training, the neuron rated each test 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. The curve in the upper right of Figure 4 shows results on the test set. Clearly most photographs are easy to identify.

Figure 5 shows the neuron inputs (after the sigmoid) for the images in Figure 1; the neuron computed likelihoods of a photograph (with 1.0 the threshold) of 0.46 for "logotrans" and 1.21 for "page3photo9" with the weights of Figure 2. So it correctly identified the images, despite the logo having more colors from dithering. Figure 3 shows implementation statistics. Image processing time could be considerably improved since we used simple but inefficient programs.

Parsing of Web pages

To identify captions, we work on the HTML markup-language code for the page; placement, format, and content are the clues. So we first "parse" the HTML to group the related parts. Figure 6 shows parser output for Figure 1. Angular brackets denote HTML commands, and square brackets and carriage returns are parser groupings.

We examine the text near each image reference (the "img" constructs) for possible captions. "Near" means within a fixed number of lines (usually three, though we tried other numbers) in the parse. (In the test set, 127 true captions were on the same line, 80 were one line away, 44 were two, 14 were three, and 14 were four.) We exclude obvious noncaptions from a list (e.g. "Figure 17", "Return to... ", and "Updated on..."). Figure 7 shows the caption candidates found for Figure 6.

There is a exception when another image reference occurs within the three lines. In Figure 6, "Sled Tracks" could caption "page3photo9.GIF" but not "logotrans.GIF" because the first image is between them. This exemplifies 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.*

This principle was never violated in the 614 captions of the training and test sets. It suggests that the space between caption and image is thought an extension of the image.

Captions often are marked to appear differently from ordinary text. "Sled Tracks" in Figure 6 is marked with "<h2>" and "</h2>", which specify a level-2 (large) boldface header font, noted in the fifth line of Figure 7. HTML text markings include font family (like Times Roman), 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 types to special sources of captions, like "alt" strings used for an image on nongraphic terminals, names of Web pages accessed by clicking on the image, and the name of the image file itself (also used by Smith and Chang (1997)), all of which provide clues about the image. But not all HTML markings suggest a caption. Italicized words among nonitalicized words probably indicate word emphasis, and whether the text is clickable or "active" is independent of captioning. So we eliminate unhelpful markings from the parsed HTML; this requires establishing the scope of multi-item markings like "<h3>" in Figure 6.

The caption neuron

Sometimes HTML code 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 character-recognizing image processing, but we did not explore this. All four ways can provide caption information, though it can often be codes or dummy placeholders: Only 6 of the 131 "alt" strings in the test set were captions.

But most image-caption relationships are not explicit. So in general we must consider carefully all text near images to find captions. Unfortunately, Web pages show much inconsistency in captioning because of their variety of authors. A caption may be below the image or above; it may be in italics or larger font or not; it may be signalled by "the view above" or "the picture shows" or "Figure 17:" or not at all; it may be a few words, a sentence, or a paragraph. So there can be many candidate captions. Full linguistic analysis (parsing and semantic interpretation) of them as in MARIE-1 and MARIE-2 would reveal the true captions, but this would require advance knowledge of word senses of every word 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 chooses the best ones for each image.

We use a seven-input "caption" neuron like the image neuron to rank candidate caption-image pairs. From our browsing on the Web, we identified the seven factors in Figure 8 as the most helpful; Figure 9 shows their values for the candidates in Figure 7. The factors assess, in order, distance; confusability with other text; highlighting; length; use of particular signal words; use of words related to the image file; and use of words denoting physical objects. Again, sigmoid functions convert the factors to probabilities of a caption given that 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 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 sampling, and Bayes' Rule used as |p ("caption" | "factor") = p ("factor" | "caption") * p ("caption") / p ("factor")|. F3 covers both explicit text marking (e.g. italics) and implicit like the "relevancy signatures" in Riloff and Lehnert (1994) (e.g. surrounding by brackets, mentioning "the picture shows", and beginning with "Figure", 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 (138 words, e.g. "page", "return", "welcome", "bytes", "gif", "visited", "links", "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 "StennisPic1") or 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 (numbers and 154 special words including "and", "the", "ours", "of", "without", "when"), with exceptions for spatial and temporal prepositions. F6 also checks for abbreviations within the file name of words or pairs of words in the caption with methods from Rowe and Laitinen (1995).

Figure 3 includes statistics on caption processing, and the lower-left curve in Figure 4 is the recall-precision curve for the output of the caption neuron for the test set after training on the training set. Performance is not as good as for images, but the task is harder and image-caption pairs must be assessed rather than captions alone. Performance was around 50% better for the training set, suggesting a danger in overtraining.

Combining image with caption information

The final step is to combine information from the image and caption neurons. This should improve upon caption-neuron performance because an image unlikely to be a photograph is unlikely to have captions. We compute the product of the probabilities that the image is a photograph and that the candidate captions it, and then compare to a threshold. A product is appropriate because the evidence comes from quite different media and is reasonably independent. To obtain probabilities from the neuron outputs, we use sigmoid functions again.

Two details need to be addressed. First, a caption candidate could be assigned to either of two nearby images if it is between them (like "Pacific Ranges and Facilities Department" in Figure 1); 634 of 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. (This was violated only once in the 275 actual captions of the test set.) We use the product described above to rate the alternatives, and choose the best one, except when we would violate the Caption-Scope Nonintersection Principle with the order Image1-Caption2-Caption1-Image2 where Caption1 goes with Image1 and Caption2 goes with Image2, in which case the caption of the weaker caption-image pair is reassigned.

Second, our training and test sets showed a limited number of captions per image: 7% of images 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 compute for an image (though we will show experiments with other numbers), the best three as per the product values. However, this limit is only applied to the visible captions, not to the file-name, pointer page-name, "alt", and page-title caption candidates usually invisible to the Web user. We do allow that a picture has no caption, which can occur for theme-establishing pictures. Page titles can sometimes be global captions for images on their page, but we ignored this since most are only weakly relevant.

The middle curve of Figure 4 is recall versus precision for the final phase with the best observed parameter values: sigmoid centers of 0.8 and 1.0 for the image and caption neurons respectively, sigmoid spreads of 0.5, a maximum of three visible captions, and a maximum distance of three lines. Figure 10 gives results of experiments on different parameters, showing that performance is reasonably consistent. Recall is for captions within a window of nine lines centered on the image reference. Figure 10 shows a recall limit with our methods (due to the caption distance limit), and that precision for higher recall values trades off with precision for lower recall values. Figure 11 shows the final captions proposed for Figure 1; all but the second are "visible" types. Figures 12 and 13 show a more difficult example, and Figure 3 has relevant statistics.

To increase the usefulness of a found caption, we should index its words with pointers to the associated photographs, ignoring common words. We also explored best-first search for the final phase, evaluating the probabilities of caption sets as a whole, but it did not perform as well. Relaxation algorithms are not appropriate here because interactions are a minor influence on the caption suitability.

As a final test, we did photograph retrieval using keyword searches, comparing performance of a word index on our found captions with Alta Vista's index on all the words of Web pages. We created queries by appending, to the three queries of Section 3, a requirement for the presence of one of the twelve most common depictable nouns in the captions of the test set, for 36 queries total. The queries required at least one image on the page. We then checked how many of the pages selected by our index, or the ten pages top-rated by Alta Vista, had at least one picture showing the noun in question (see Figure 14). Though there was much variation with search word, our index averaged 0.42 in precision with 0.91 recall over the twelve words, while Alta Vista averaged 0.097 precision with approximately 0.35 recall (taking the test-set pages as the universe for the last number, although they are not strictly comparable since the queries were done a year later).

Conclusions

Photographs are one of the most valuable resources on World Wide Web, but it is hard to index them effectively. Since 1.4% of the text on Web pages were photograph descriptions in our sample of pages with images, standard Web keyword indexes have trouble finding them. Our research has shown that some simple text, format, and image analysis can markedly increase (to 41% precision and recall) the success rate in finding captions describing photographic images on representative pages, and the subsequent success rate in finding photographs matching keywords (by a factor of 3-4).

Our methods should work well on Web pages of a technical, historical, or informational nature like our sample. Since pages often imitate books, our methods should also work well for digitized books with described photographs (since books give many obvious clues as to their formatting tags). We cannot handle pages and books in which pictures establish themes discussed only indirectly in the text, as often with illustrations to fiction. But we suspect that our methods will not become obsolete as new features are added to HTML because of, judging from our test sample, the great variety of captioning methods and that the obvious captions (like the titles in Figure 1) were not often the best.

References

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

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

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

Riloff, E. & Lehnert, W. (1994, July). Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems, 12*(3), 296-333.

Rowe, N. C. (1994). Inferring depictions in natural-language captions for efficient access to picture data. *Information Processing and Management, 30*(3), 379-388.

Rowe, N. C. (1996, April). Using local optimality criteria for efficient information retrieval with redundant information filters. *ACM Transactions on Information Systems, 14*(2), 138-174.

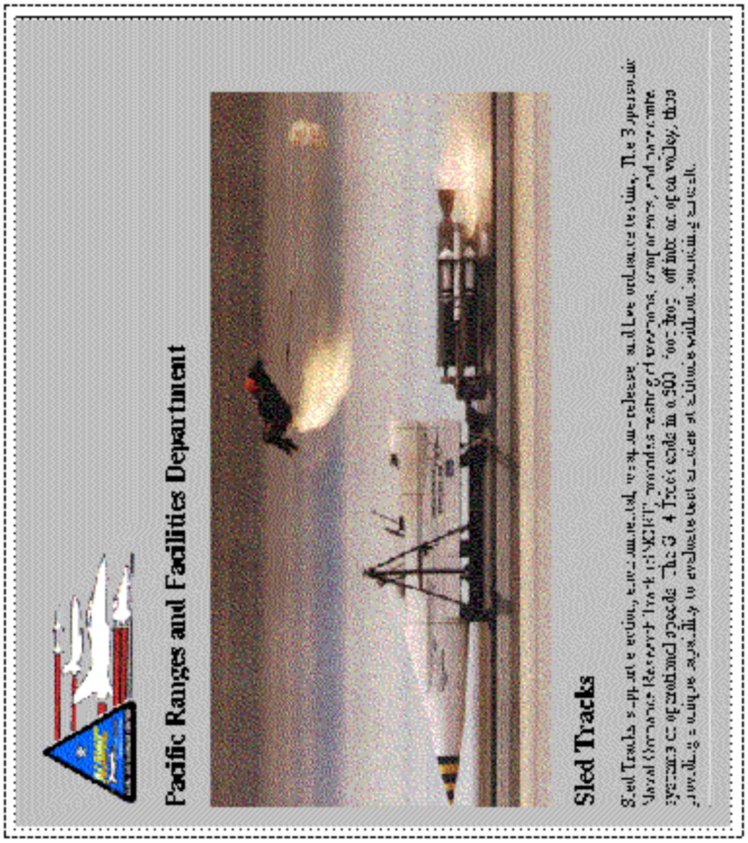
Rowe, N. C. & Frew, B. (1997). Automatic classification of objects in captioned depictive photographs for retrieval. In *Intelligent Multimedia Information Retrieval*, ed. M. Maybury (pp. 65-79). Palo Alto, CA, USA: AAAI Press.

Rowe, N. C. & Laitinen, K. (1995). Semiautomatic disabbreviation of technical text. *Information Processing and Management, 31*(6), 851-857.

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

Smith, J. R. & Chang, S.-F. (1997, Fall). Visually searching the Web for content. *IEEE Multimedia*, to appear.

Srihari, R. K. (1995, September). Automatic indexing and content-based retrieval of captioned images. *IEEE Computer, 28*(9), 49-56.



Pacific Ranges and Facilities Department

Sled Tracks

Sled Tracks support ejection, environmental, weapon-release, and live ordnance testing. The Supersonic Naval Ordnance Research Track (SNORT) provides testing of weapons, components, and parachute systems at operational speeds. The G-4 Track ends in a 500-foot drop-off into an open valley, thus providing a unique capability to evaluate test articles at altitude without launching aircraft.

Figure 1: Example World Wide Web page.

```
Image neuron    Sigmoid Sigmoid Weight
factor  center  spread  after training
Size    100     50      0.20
(Geometric mean of height and width)


Squareness      0.5     0.2     0.34
(Minimum of height/width and
width/height, except 0.5 when height
equals width)


Number of colors        4       6       0.10
(Number of distinct
red-green-blue triples used)


Fraction of impure colors       0.05    0.05    0.32
(Fraction of cells not pure white,
black, grey, red, green, or blue)


Neighbor variation      0.25    0.2     0.22
(Fraction of horizontally-adjacent
pixels of the same color)


Color dispersion        0.1     0.1     0.23
(Fractional distance of
mean color in the sorted
color histogram)
```

Name suggests photograph (true or false) (Whether the name of the image file contains common words in icon names)	-	-	0.20

Figure 2: Inputs to the image neuron.

Statistic	*Training set*	*Test set*
Size of HTML source (bytes)	*213,269*	*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*
Number of invisible candidates	*434*	*1107*
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*	*274*

Figure 3: Statistics on our experiments (times are in CPU seconds).

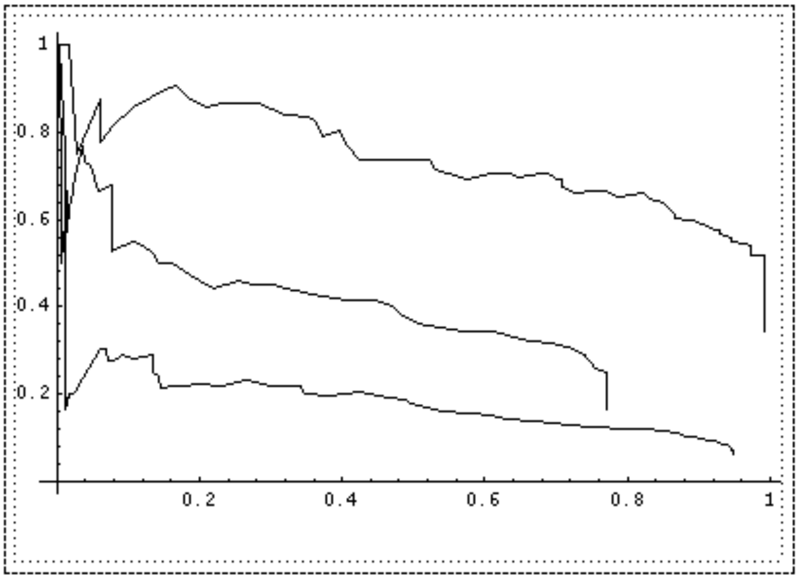


Figure 4: Recall (horizontal) versus precision (vertical) for photograph identification (the top curve), caption identification from text alone (the bottom curve), and caption identification combining both image and caption information (the middle curve).

Image	*logotrans*	*page3photo9*
Size factor	*0.54*	*0.96*
Squareness factor	*0.22*	*0.45*
Number of colors factor	*0.93*	*0.79*
Fraction of impure colors factor	*0.01*	*0.87*
Neighbor variation factor	*0.02*	*0.83*

Color dispersion factor	*0.28*	*0.82*
Name factor	*0.12*	*0.90*
Neuron output	*0.46*	*1.21*

Figure 5: Inputs to the image neuron from analysis of the logo ("logotrans") and photograph ("page3photo9") images in Figure 1. Higher inputs suggest photographs. The inputs together correctly distinguish the two images, despite the logo having a higher number of colors because of dithering.

```
<html>
<head>
[<TITLE>][PRFD Sled Tracks][</TITLE>]
</head>
<blank>
<body>
<img src="./logotrans.GIF">
[<h2>][Pacific Ranges and Facilities Department][</h2>]
<img src="./page3photo9.GIF">
[<p>][<h2>][Sled Tracks][</h2>][Sled Tracks support ejection, environmental,
   weapon-release, and live  ordnance testing.][The Supersonic Naval
   Ordnance Research Track (SNORT)  provides testing of weapons, components,
   and parachute systems at operational  speeds.][The G-4 Track ends in a
   500-foot drop-off into an open valley,  thus providing a unique capability
   to evaluate test articles at altitude  without launching aircraft.]
<hr>
<dir>
[<h3>][<li>][<a href="./index.html">][Return to Pacific Range and Facilities Welcome Page ][</h3>]
[<h3>][<li>][<a href="http://www.chinalake.navy.mil/">][Return to NAWCWPNS Home Page][</h3>]
</dir>
</body>
</html>
<blank>
```

Figure 6: Output of the HTML parser on the input in Figure 1.

```
Image logotrans line 7 captype filename distance 0: logotrans
Image logotrans line 7 captype h2 distance 1: 'Pacific Ranges and Facilities Department'
Image page3photo9 line 9 captype filename distance 0: 'page 3 photo 9'
Image page3photo9 line 9 captype h2 distance -1: 'Pacific Ranges and Facilities Department'
Image page3photo9 line 9 captype h2 distance 1: 'Sled Tracks'
Image page3photo9 line 9 captype plaintext distance 1: 'Sled Tracks support ejection,
   environmental, weapon-release, and live ordnance testing.'
Image page3photo9 line 9 captype plaintext distance 1: 'The Supersonic Naval Ordnance
   Research Track (SNORT) provides testing of weapons, components, and
   parachute systems at operational speeds.'
Image page3photo9 line 9 captype plaintext distance 1: 'The G-4 Track ends in a 500-foot
   drop-off into an open valley, thus providing a unique capability to evaluate
   test articles at altitude without launching aircraft.'
```

Figure 7: Output of the candidate-caption locator on the input from Figure 6.

```
Caption neuron  Sigmoid Sigmoid Weight after
factor  center  spread  training
F1: Distance in lines from      1       3       0.20
caption to image        inverse,
        for caption below
        0.8     2.5
        inverse,
        for caption above
F2: Number of other candidate   1.5     3       0.33
captions at the same distance   inverse
F3: Caption emphasis    -       -       0.10
(0.9 for suggestive format,
0.5 for HTML emphasis,
0.05 for noncaption format,
0.2 else)
F4: Length of caption   300     160     0.32
in characters   inverse,
        for >19 chars.
        10      10
        else
F5: Use of common caption       0       2       0.22
or noncaption words: Compute
number of typical caption
words - number of noncaption
words + number of years
F6: Use of words also in        0.4     0.4     0.23
image file name or in   for file name
"alt" string    0.2     0.2
(excluding common words for "alt"
of English)
F7: The fraction of words       0.15    0.3     0.21
(excluding common words
of English) in the
caption that have at least
one physical-object sense
```

Figure 8: Input factors for the caption-recognizing neuron.

Image	*Candidate caption*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*	*Total*	
logotrans	*"logotrans"*	*0.95*	*0.79*	*0.2*	*0.45*	*0.5*	*0.2*	*0.5*	*0.78*	
logotrans	*"Pacific Ranges"*		*0.5*	*0.66*	*0.5*	*0.96*	*0.27*	*0.12*	*0.77*	*0.93*

	and Facilities Department"								
page3photo9	"page 3 photo 9"	0.95	0.79	0.2	0.69	0.5	0.2	0.5	0.86
page3photo9	"Pacific Ranges" and Facilities Department"	0.46	0.66	0.5	0.96	0.27	0.12	0.77	0.92
page3photo9	"Sled Tracks"	0.5	0.66	0.5	0.55	0.5	0.12	0.77	0.84
page3photo9	"Sled Tracks support ejection, environmental, weapon-release, and live ordnance testing."	0.5	0.34	0.2	0.93	0.5	0.12	0.75	0.83
page3photo9	"The Supersonic Naval Ordnance Research Track (SNORT) provides testing of weapons, components, and parachute systems at operational speeds."	0.5	0.34	0.2	0.88	0.5	0.12	0.56	0.77
page3photo9	"The G-4 Track ends in a 500-foot drop-off into an open valley, thus providing a unique capability to evaluate test articles at altitude without launching aircraft."	0.5	0.34	0.2	0.85	0.5	0.12	0.46	0.74

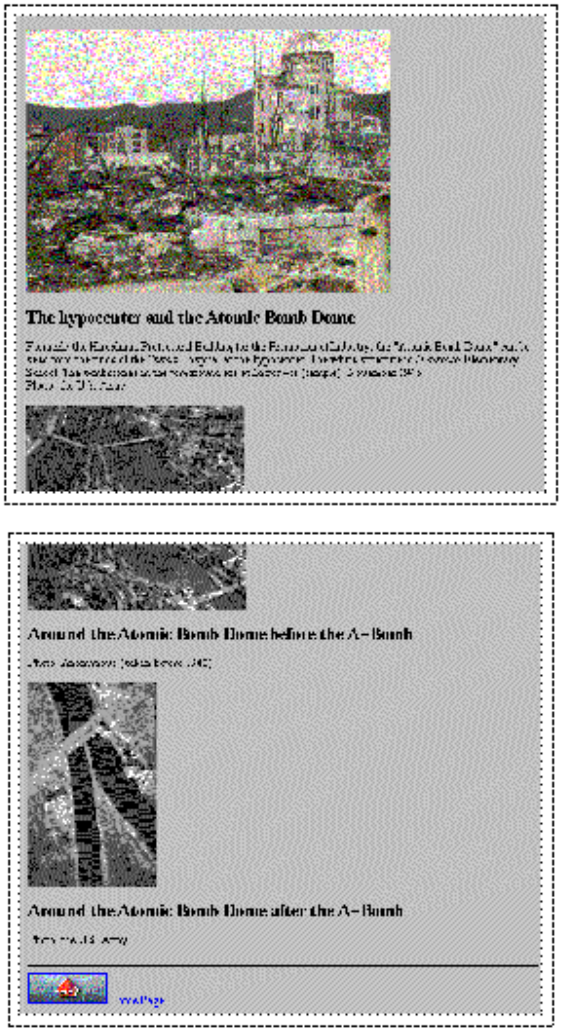
Figure 9: The caption-neuron inputs for the eight candidate captions in Figure 7 (not including the likelihood the image is a photograph).

image	caption	sigmoid	maximum	maximum	precision		precision		precision		recall
sigmoid	sigmoid	spread	no. of	caption	for 0.2	for equal		for 0.7	maximum		
center	center		captions		distance		recall	recall	recall		
0.2	0.4	0.5	3	3	.450	.404	.305	.766			
0.4	0.4	0.5	3	3	.452	.392	.298	.766			
0.4	0.6	0.5	3	3	.462	.388	.306	.766			
0.6	0.6	0.5	3	3	.438	.402	.298	.766			
0.6	0.8	0.5	3	3	.445	.394	.305	.766			
0.8	0.8	0.5	3	3	.445	.417	.299	.766			
0.8	1.0	0.5	3	3	.452	.408	.303	.766			
1.0	1.0	0.5	3	3	.457	.416	.296	.766			
1.0	1.2	0.5	3	3	.448	.402	.302	.766			
0.8	1.0	0.3	3	3	.453	.382	.280	.766			
0.8	1.0	0.7	3	3	.425	.414	.292	.766			
0.8	1.0	0.5	2	3	.442	.401	--	.653			
0.8	1.0	0.5	4	3	.427	.395	.296	.843			
0.8	1.0	0.5	6	3	.427	.398	.302	.920			
0.8	1.0	0.5	3	2	.459	.412	.313	.785			
0.8	1.0	0.5	3	4	.410	.383	.288	.818			
0.8	1.0	0.5	6	4	.414	.398	.296	.971			

Figure 10: Results of varying the parameters in the final phase of processing on the test set. (The seventh row is the case shown in Figure 4.)

```
[page3photo9,'Pacific Ranges and Facilities Department'] @ 0.370
[page3photo9,'page 3 photo 9'] @ 0.321
[page3photo9,'Sled Tracks'] @ 0.308
[page3photo9,'Sled Tracks support ejection, environmental, weapon-release,
    and live ordnance testing.'] @ 0.293
```

Figure 11: Final caption assignment after best-first search for the example of Figure 8.



The hypocenter and the Atomic Bomb Dome

[illegible]

Around the Atomic Bomb Dome before the A-Bomb

[illegible]

Around the Atomic Bomb Dome after the A-Bomb

[illegible]

Figure 12: Another example Web page.

Image field01 line 5 captype filename distance 0: 'field 1'
Image field01 line 5 captype title distance -2: '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'
******************************************************************************
[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 hypocenter and the Atomic Bomb Dome'] @ 0.316
[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 13: Caption candidates generated, followed by final caption assignment for the example of Figure 12; photograph "field01" is at the top in Figure 12, "island" in the middle, "island2" at the bottom, and "ab-home" is the icon on the bottom. A caption was missed at a distance of +4 (the tombstone one), and one incorrect match was proposed, the third for image "island", due to a preference for the closest image.

Word	Correctly	MARIE-3	Correctly	Query 1	Query 2	Query 3
	matching	proposed	proposed	matches	matches	matches
	pages	pages	pages	(of top 10)		
moon	1	2	1	2	1	4
system	6	13	6	0	0	0
surface	3	5	3	0	0	2
door	1	1	1	1	2	0
laboratory	6	27	4	0	0	2
facility	5	7	4	2	3	3
laser	2	3	2	0	1	1
landing	1	1	1	0	0	0
students	2	6	2	1	0	2
soil	1	2	1	1	2	2
hill	3	3	3	0	0	0
battlefield	2	2	2	0	2	1

Figure 14: Results of tests of photograph retrieval by keyword queries.

Go to paper index