Precise and Efficient Retrieval of Captioned Images: The MARIE Project

NEIL C. ROWE

ABSTRACT

THE MARIE PROJECT HAS EXPLORED knowledge-based information retrieval of captioned images of the kind found in picture libraries and on the Internet. It exploits the idea that images are easier to understand with context, especially descriptive text near them, but it also does image analysis. The MARIE approach has five parts: (1) find the images and captions; (2) parse and interpret the captions; (3) segment the images into regions of homogeneous characteristics and classify them; (4) correlate caption interpretation with image interpretation using the idea of focus; and (5) optimize query execution at run time. MARIE emphasizes domain-independent methods for portability at the expense of some performance, although some domain specification is still required. Experiments show MARIE prototypes are more accurate than simpler methods, although the task is very challenging and more work is needed. Its processing is illustrated in detail on part of an Internet World Wide Web page.

INTRODUCTION

Multimedia data are increasingly important information resources for computers and networks. Much of the excitement over the World Wide Web is about its multimedia capabilities. Images of various kinds are its most common nontextual data. But finding the images relevant to some user need is often much harder than finding text for a need. Careful content analysis of unrestricted images is slow and prone to errors. It helps to find captions or descriptions as many images have them.

Neil C. Rowe, Department of Computer Science, Code CS/Rp, U. S. Naval Postgraduate School, Monterey, CA 93943
LIBRARY TRENDS, Vol. 48, No. 2, Fall 1999, pp. 475-495

Nonetheless, multimedia information retrieval is still difficult. Many problems must be solved just to find caption information in the hope of finding related images. Only a 1 percent success rate was obtained in experiments trying to retrieve photographs depicting single keywords like "moon" and "hill" using the AltaVista search engine on the World Wide Web (Rowe & Frew, 1998). This is probably because most text on Web pages, and even on pages with well-captioned photographs, was irrelevant to the photographs, and the words searched for had many senses. To improve this performance several things are needed:

- a theory of where captions are likely to be on pages;
- a theory of which images are likely to have described content;
- language-understanding software to ascertain the correct word senses and their interrelationships in the caption candidates;
- image-understanding software to obtain features not likely to be in the caption;
- a theory connecting caption concepts to image regions; and
- efficient methods of retrieval of relevant images in response to user queries or requests.

Note that speed is only critical in the last phase; while efficient methods are always desirable, accuracy is more a concern because it is so low for keyword-based retrieval. The time to do careful language and image processing can be justified during the indexing of a database if it can significantly improve later retrieval accuracy.

Consider Figure 1, which shows part of a U.S. Army Web page. Much text is scattered about, but not all of it refers to the pictures. The two formal captions (in italics, but not otherwise identified) are inconsistently placed with respect to their photographs. But the title "Gunnery at Udairi" is a caption too. Next, note that many of the words and phrases in these candidate captions do not describe the pictures. Neither picture shows "U.S. Army Central Command," "power generator equipment," an "Iraqi," "the Gulf War," or "Fort Hood"; matching would falsely retrieve this page for any of these key phrases. Similarly, the words "commander," "senior," "signal," "fire," "target," and "live" are all used in special senses, so this page would be falsely retrieved for queries intending to refer to their most common senses. The only way to eliminate such errors is to parse and interpret caption candidates using detailed linguistic knowledge. Finally, note that many things seen in both photographs are not mentioned in their captions. Only a small part of the left photograph area is devoted to the people, and the right photograph displays many features of the tanks not mentioned in its caption. Thus there are many challenges in indexing the multimedia data on these pages.

Noteworthy current systems for image retrieval are QBIC (Flickner et al., 1995), Virage (Virage Inc., San Mateo, California, USA), and

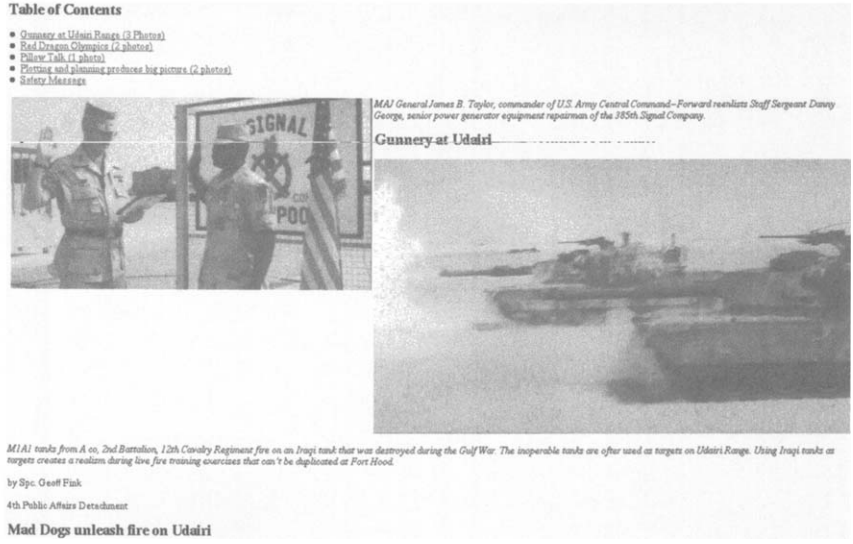
Table of Contents

- Gunnery at Udairi Range (3 Photos)
- Red Dragon Olympics (2 photos)
- Pillow Talk (1 photo)
- Plotting and planning produces big picture (2 photos)
- Safety Message

MAJ General James B. Taylor, commander of U.S. Army Central Command--Forward reenlists Staff Sergeant Danny George, senior power generator equipment repairman of the 385th Signal Company.

Gunnery at Udairi

M1A1 tanks from A co, 2nd Battalion, 12th Cavalry Regiment fire on an Iraqi tank that was destroyed during the Gulf War. The inoperable tanks are often used as targets on Udairi Range. Using Iraqi tanks as targets creates a realism during live fire training exercises that can't be duplicated at Fort Hood.

by Spc. Geoff Fink

4th Public Affairs Detachment

Mad Dogs unleash fire on Udairi

Figure 1. Example Portion of a World Wide Web Page.

VisualSEEK (Smith & Chang, 1996), which exploit simple-to-compute image properties like average color and texture. The user specifies color and texture patches, or perhaps an entire image, which is then compared to the images in the database to find the best matches. But these systems strongly emphasize visual properties and can confuse very different things of accidentally similar appearance, like seeing a face in an aerial photograph. So these systems would not help for a typical Web page like Figure 1 since color similarity to the images there would not mean much. Another category of current image-retrieval systems like Chabot (Ogle, 1995) primarily exploits descriptive information about the image, but all this information must be entered manually for each image by someone knowledgeable about it, which requires a considerable amount of tedious work.

The most interesting current research has focused on knowledge-based multimodal methods for addressing the limitations of current systems. Work on indexing of video (Hauptman & Witbrock, 1997; Smoliar & Zhang, 1994) has achieved success using knowledge-based multimodal analysis of images, image-sequence segmentation, speech, on-screen text, and closed-caption information. For single-image recognition, Piction (Srihari, 1995) does natural-language understanding of the caption of an image, and combines this information with results of face localization in the image, to provide a deeper understanding of an image. But Piction assumed that captions were already isolated for each image, and there are many interesting image features besides faces.

This article summarizes a promising approach that the MARIE project has explored recently using knowledge-based methods for accurate photograph retrieval in response to English queries by a user. The idea is to consider image retrieval in a broader perspective than that of Piction. The subtasks are finding the image, analyzing all relevant text, analyzing the image, mapping the results of the text analysis to the results of the image analysis, and efficient subsequent retrieval of this information. By considering these subtasks as parts of a larger context, we will see important issues not addressed by piecemeal efforts.

The methods of MARIE were tested in three partial prototype systems of MARIE-1, MARIE-2, and MARIE-3. These systems primarily address photographs since many users consider them the most valuable multimedia objects, but most of the methods generalize easily. Both explicit photograph libraries (especially the Photo Lab of NAWC-WD, China Lake, California, USA, with its images depicting a wide range of activities at a naval aircraft test facility) and implicit libraries (especially the World Wide Web) were investigated. Most of the code is in Quintus Prolog with some key sections in C. The remainder of the article discusses in turn each of the main problems that MARIE faced.

Locating Indexable Images and Their Captions

Identifying images and their captions is a significant problem with book-like multimedia data (as Figure 1). Web images are easy to identify by the HTML page-markup language used. But symbolic graphics, of no value to index, are generally stored the same way as photographs, as files in GIF or JPEG format, so a useful system must distinguish them. Recent work with the MARIE-3 system (Rowe & Frew, 1998) has shown that seven quickly-found parameters of images are sufficient to distinguish photographs with 70 percent recall (a fraction of all photographs found) and 70 percent precision (a fraction of items found that are photographs) on a test set of random Web pages. The parameters are size, squareness, number of distinct colors, fraction of pure colors (white, black, pure gray, red, green, and blue), color variation between neighbor pixels, variety of colors, and use of common nonphotograph words (like "button" or "logo") in the name on the image file. The parameters are converted to probabilities by applying "sigmoid" (S-shaped) functions of the form yi = tanh ((xi - mu) / sigma) where mu and sigma are constants chosen to set the center and steepness of the sigmoid curve. The probabilities are then input to a "linear-classifier neuron" calculating w0 + w1 y1 + w2 y2 + ... + w7 y7 for a set of weight constants wi determined by training. If the calculation results in a positive number, the image is considered a photograph. For Figure 1, MARIE-3 rated the left image as 0.244 and the right image as 0.123 after training, so both were correctly classified as photographs.

MARIE-3 then looks for captions around each possible photograph.

Captions are not often easy to identify because they take many forms. It is best to work on the HTML source code of the Web page, parsing it to group related things. Another seven-input linear-classifier neuron with sigmoid functions on its inputs can rate the caption candidates. Its input parameters were easy-to-calculate properties of text: distance in lines from the candidate caption to the image reference, number of other candidates at the same distance, strength of emphasis (e.g., italics), appropriateness of candidate length, use of common (counted positively) or uncommon (counted negatively) words of captions, number of identical words between candidate and either image file name or its nongraphics substitute, and fraction of the words having at least one physical-object sense. Figure 2 shows the caption candidates for Figure 1 with their ratings.

The caption neuron by itself showed 21 percent recall for 21 percent precision in matching caption-image pairs. This can be improved by combining its rating with the photograph rating since photographs are much more likely to have captions. Neuron outputs can be converted to probabilities and multiplied, with three refinements. Since a survey of random Web photographs showed that 7 percent had no visible captions, 57 percent had one caption, 26 percent had two, 6 percent had three, 2 percent had four, and 2 percent had five, it is reasonable to limit each image to its three best captions. If a caption can go with more than one image, rule out everything but the strongest match since a useful caption should be precise enough to describe only one image. For example, the "MAJ General" caption candidate in Figure 1 goes better with the left image "image4" (since captions are more often below than above) so the match to "image1" was ruled out; and "Gunnery at Udairi" goes better with the right picture from examination of the HTML code even though it is displayed similarly to the "MAJ General" candidate. Finally, consider possible "invisible" captions—the image file name, the name of any Web page pointed to by the image, any text-equivalent string for the image, and the Web page title—when their likelihoods exceed a threshold.

All this gave 41 percent recall with 41 percent precision on a random test set, or 70 percent recall with 30 percent precision, demonstrating the value of multimodal evidence fusion. Processing required 0.015 cpu seconds per byte of HTML source code, mostly in the image analysis, and the program consisted of 83 kilobytes of source code. Figure 3 shows the final captions found by MARIE-3 for Figure 1, one for the first photograph and two for the second.

Linguistic Processing

Lexical Processing

Once likely captions are found, their words can be indexed as keywords for later retrieval (excluding words that are not nouns, verbs,

Referred Image	*Caption Type*	*Caption Distance*	*Caption Candidate*
image4	plaintext	-1	"gunnery at udairi range (3 photos)"
image4	plaintext	-1	"red dragon olympics (2 photos)"
image4	plaintext	-1	"pillow talk (1 photo)"
image4	plaintext	-1	"plotting and planning produces big picture (2 photos)"
image4	plaintext	-1	"safety message"
image4	bold	-2	"table of contents"
image4	emphasis	1	"maj general james b. taylor, commander of u.s. army central command-forward reenlists staff sergeant danny george, senior power generator equipment repairman of the 385th signal company."
image4	heading2	2	"gunnery at udairi"
image1	heading2	0	"gunnery at udairi"
image1	emphasis	-1	"maj general james b. taylor, commander of u.s. army central command-forward reenlists staff sergeant danny george, senior power generator equipment repairman of the 385th signal company."
image1	emphasis	1	"m1a1 tanks from a co, 2nd battalion, 12th cavalry regiment fire on an iraqi tank that was destroyed during the gulf war."
image1	emphasis	1	"the inoperable tanks are often used as targets on udairi range."
image1	emphasis	1	"using iraqi tanks as targets creates a realism during live fire training exercises that can't be duplicated at fort hood."
image1	plaintext	3	"4th public affairs detachment"

Figure 2. Candidate Captions Inferred by MARIE-3 for the Web Page in Figure 1.

Image	*Caption*	*Final rating*
image4	"MAJ general james b. taylor, commander of u.s. army central command-forward reenlists staff sergeant danny george, senior power generator equipment repairman of the 385th signal company."	0.937
image1	"Gunnery at Udairi"	0.954
image1	"m1a1 tanks from a co, 2nd battalion, 12th cavalry regiment fire on an iraqi tank that was destroyed during the gulf war."	0.929

Figure 3. Final captions inferred by MARIE-3 for the Web page in Figure 1, with their final ratings.

adjectives, or adverbs). Several information-retrieval systems and Web search tools do this. But retrieval precision, as a result, will not be high because a word can have many meanings and many relationships to neighboring words. Nonetheless, most captions are unambiguous in their context. "View of planes on taxi from tower" is unambiguous in our NAWC-WD Navy-base test captions since "planes" are always aircraft and not levels, "tower" is always a control tower when aircraft are mentioned, "taxi" has a special meaning for aircraft, aircraft are always taxiing and not on top of a taxicab, and the view (not the aircraft or taxiing) is from the tower. Keyword-based retrieval will thus get many incorrect retrievals with the words of this caption. They would furthermore usually miss captions having synonyms or generalizations of the words, like for the caption "Photograph of 747's preparing to takeoff as seen from control" and the query "View of planes on taxi from tower."

Fortunately, true caption language understanding is easier than most text understanding (like automatic indexing of journal articles) since captions must describe something visible. Caption language heavily emphasizes physical objects and physical actions; verbs usually appear as participles, with a few gerunds and past tenses; and words for social interactions, mental states, and quantifications are rare. All this simplifies analysis. Also, many valuable applications involve technical captions, whose accessibility could be valuable to enhance, but whose difficulty primarily resides in code words and unusual word senses that are nonetheless unambiguous, grammatically easy to classify, and often defined explicitly somewhere (e.g., "zeppo" of "zeppo radar").

Many caption words are familiar words of English, and MARIE's parser needs only their parts of speech and superconcepts, obtained from the Wordnet thesaurus system (Miller et al., 1990). For the remaining words, caption writers often try to be clear and follow simple recognizable lexical

rules like "F-" followed by a number is a fighter aircraft and any number followed by "kg" is a weight in kilograms. In developing MARIE-2, such rules covered 17,847 of the 29,082 distinct words occurring in 36,191 NAWC-WD captions (see Figure 4). Person names, place names, and manufacturer names were obtained in part from existing databases for a total of 3,622 words. Of the remaining words, 1,174 were misspellings, of which 773 were correctly deciphered by our misspelling-detection software (including misspellings of unknown words, by examining word frequencies). Of the remaining words, 1,093 were abbreviations or acronyms, of which 898 were correctly deciphered by our abbreviation-hypothesizing and misspelling-fixing software (Rowe & Laitinen, 1995) using context and analogy. Of the remaining equipment names, 1,876 were not important to define further. That left 1,763 words needing explicit definition; almost

Number of captions	36,191
Number of words in the captions	610,182
Number of distinct words in the captions	29,082
Subset having explicit entries in Wordnet	6,729
Number of word senses given for these words	14,676
Subset with definitions reusable from MARIE-1	770
Subset that are morphological variants of other known words	2,335
Subset that are numbers	3,412
Subset that are person names	2,791
Subset that are place names	387
Subset that are manufacturer names	264
Subset that have unambiguous defined-code prefixes	3,256
Unambiguous defined-code prefixes	947
Subset that are other identifiable special formats	10,179
Subset that are identifiable misspellings	1,174
Misspellings found automatically	713
Subset that are identifiable abbreviations	1,093
Abbreviations found automatically	898
Subset with definitions written explicitly for MARIE-2	1,763
Remaining words, assumed to be equipment names	1,876
Explicitly used Wordnet alias facts of above Wordnet words	20,299
Extra alias senses added to lexicon beyond caption vocabulary	9,324
Explicitly created alias facts of above non-Wordnet words	489
Other Wordnet alias facts used in simplifying the lexicon	35,976
Extra word senses added to lexicon beyond caption vocabulary	7,899
Total word senses handled (includes related superconcepts, wholes, and phrases)	69,447

Figure 4. Statistics on the MARIE-2 lexicon for the NAWC-WD captions.

all of these were nouns. MARIE-2's 29,082-word lexicon construction required only 0.4 of a man-year, and much of the work can be reused unchanged for other technical applications. So porting MARIE-2 requires some work but not much.

Statistical Parsing

Although captions are usually unambiguous in their subject areas, effort is needed to determine the word senses and word relationships used in a subject area because many of these are atypical of natural language. This information could be laboriously defined manually for each subject area, but a better way is to learn it from context. This can be done with a statistical parser that learns word-sense frequencies and sense-association frequencies from a corpus of examples. A bottom-up chart parser (Charniak, 1993) will suffice. This is natural-language processing software that builds up interpretations of larger and larger substrings of the input word list by always trying to combine the strongest substring interpretations found up to that point. The strength of an interpretation can reflect the relative frequencies of the word senses used, the frequencies of the word-sense associations made, and the frequencies of the parse rules used.

To simplify matters, restrict the grammar to unary and binary parse rules (rules with only one or two symbols as the replacement for some grammatical symbol). The degree of association between two word strings in a binary parse rule can be taken as the degree of association of the two headwords of the strings. Headwords are the subjects of noun phrases, the verbs of verb phrases, sentences, and clauses, the prepositions of prepositional phrases and so on. So the headword of "F-18 aircraft on a runway" would be "aircraft," and the headword of "on a runway" would be "on." Then a particular caption interpretation can be rated by combining the degrees of association of the headword pairs at every node of its parse tree with a priori frequencies of the word senses for leaves of the tree. This is a classic problem in evidence fusion, and the simplest solution is to assume independence and multiply the probabilities. It also helps to weight the result by an adjustable monotonically-increasing function of the sentence length to encourage work on longer subphrases.

However, word-sense association frequencies are often sparse and statistically unreliable. So use "statistical inheritance" to estimate frequencies from more general ones. For instance, to estimate the frequency of "on" as a preposition and "runway" as the subject of its prepositional phrase, look for statistics of "on" and any surface or, if the sample size is not enough, "on" and a physical object. When sufficiently reliable frequency statistics for a modified pair are found, divide by the ratio of the number of occurrences of the substitute word to the number of occurrences of "runway," since a more general concept should associate proportionately more with

"on." Similarly, generalize "on" to the class of all physical-relationship prepositions, or generalize both "on" and "runway" simultaneously. Generalize even to the class of all prepositions and the class of all nouns (i.e., the parse-rule frequency) if necessary, but those statistics are less reliable; one should prefer the minimum generalization having statistically reliable data. Statistical inheritance is self-improving because each confirmed interpretation provides more data.

This statistical approach to parsing addresses the two main linguistic challenges of captions—interpretation of nouns modifying other nouns ("nominal compounds") and interpretation of prepositional phrases. Both involve inference of case relationships. Figure 5 lists important cases for nominal compounds in the captions studied in Rowe and Frew (1998). Distinguishing these cases requires a good type hierarchy which Wordnet can supply as well as providing synonym and part-whole information for word senses.

Example	*Case relationship*
B-747 aircraft	subtype-type
aircraft wing	whole-part
F-18 Harrier	object-alias
aircraft BU#194638	type-identifier
lights type iv	object-type
rock collection	object-aggregation
mercury vapor	material-form
10-ft pole	measure-object
infrared sensor	mode-object
wine glass	associate-object
Delta B-747	owner-object
Captain Jones	title-person
Commander NWC	job-organization
aircraft closeup	object-view
foreground clouds	view-object
Monterey pharmacy	location-object
sea site	object-location
NRL Monterey	organization-location
Benson Arizona	sublocation-location
assembly area	action-location
arena test	location-action
reflectivity lab	subject-location
assembly start	action-time
July visit	time-action
training wheels	action-object
parachute deployment	object-action
air quality	object-property
project evaluation	concept-concept
night-vision goggles	concept-object
ECR logo	object-symbol

Figure 5. Important Cases of Nominal Compounds for Captions.

The result of parsing and semantic interpretation of a caption is a meaning representation. Since captions so rarely involve quantification, tenses, and hypothetical reasoning, their meaning is generally expressible with "conjunctive semantics"—as a list of type and relationship facts. Each noun and verb maps to an instance of its word-sense type, and instances are related with relationship predicates. Figure 6 shows the semantic interpretations found by MARIE-3 for the inferred captions of Figure 1. Multiple captions found for the same image were treated as separate sentences. The "v" symbols are existentially quantified variables for type instances, and hyphenated numbers are sense numbers. Sense

Inferred caption on left photograph: "maj general james b. taylor, commander of u.s. army central command-forward, reenlists staff sergeant danny george, senior power generator equipment repairman of the 385th signal company."

[a_kind_of(v435,enlist-103), quantification(v435,singular),
property(v435,repeated), object(v435,v14),
a_kind_of(v14, Staff Sergeant-0), a_kind_of(v14, George-0), identification(v14, Danny),
a_kind_of(v14,serviceman-2), subject(v14,v558),
a_kind_of(v558,equipment-1), part_of(v558,v486),
a_kind_of(v486,generator-2), agent(v464,v486), a_kind_of(v464,power-3),
property(v14,senior-51), owned_by(v14,v22),
a_kind_of(v22,Signal Company-0), property(v22,385th-50),
rank(v22,385), quantification(v22,the), agent(v435,v11),
a_kind_of(v11,Major General-0), a_kind_of(v11,James-0),
a_kind_of(v11,Taylor-0), identification(v11,B.-0),
a_kind_of(v11,commander-2), located_at(v11,v212),
a_kind_of(v212,front-1), located_at(v212,v209), a_kind_of(v209,command-0),
property(v212,central-52), part_of(v212,v8),
a_kind_of(v8,United States Army -0)]).

Inferred caption on right photograph: "gunnery at udairi. m1a1 tanks from a co, 2nd battalion, 12th cavalry regiment, fire on an iraqi tank that was destroyed during the gulf war."

[a_kind_of(v1,gunnery-1), at(v1,v15), a_kind_of(v15, Udairi -0), during(v1,v18),
a_kind_of(v18,shoot-109), quantification(v18,plural), object(v18,v148),
a_kind_of(v148,tank-4), property(v148, Iraqi -50),
quantification(v148,a), object(v156,v148),
a_kind_of(v156,destroy-101), tense(v156,past),
quantification(v156,singular), during(v156,v21),
a_kind_of(v21, Gulf War -0), quantification(v21,the), agent(v18,v23),
a_kind_of(v23, M1A1 -0), quantification(v23,plural), from(v23,v49),
a_kind_of(v49,company-4), quantification(v49,a), owned_by(v49,v75),
a_kind_of(v75,battalion-1), property(v75, 2nd -51), owned_by(v75,v89),
a_kind_of(v89,regiment-1), subject(v89,v85),
a_kind_of(v85,cavalry-1), property(v89, 12th -50), rank(v89,12)]).

Figure 6. Semantic Interpretations Found by MARIE-3 for the Inferred Captions of Figure 1.

numbers 0-49 are for nouns (with nonzero sense numbers from Wordnet version 1.5); 50-99 are for adjectives (with the number minus 50 being the Wordnet sense number); and 100-149 are for verbs (with the number minus 100 being the Wordnet sense number). The "a_kind_of" expressions relate a variable to a type, so for instance "a_kind_of(v435,enlist-103)" says the caption refers to some v435 that is an instance of the verb "enlist" in Wordnet sense number 3. Other expressions give properties of variables, like "quantification(v435,singular)" meaning that the enlisting event was singular as opposed to plural, and two-variable expressions relate the variables, like "object (v435,v14)" meaning that the object of the enlisting event was some v14, an instance of "Staff Sergeant."

In general, MARIE-2 (MARIE-3 is not finished) took a median CPU time of 9.7 seconds (and a geometric mean of 10.2 seconds, the antilogarithm of the mean of the logarithms) to parse randomly selected NAWC-WD caption sentences, sentences averaging 7.4 words in length, with more difficulty for a few sentences with rare syntactic constructs. This processing used 226K of source code and 6832K of data (including 1894K of word sense statistics and 1717K of binary statistics). MARIE-2 found the correct interpretation on its first try for the majority of the test sentences, with a geometric mean of 1.9 tries. Figure 7 shows the value of different kinds of linguistic information to the interpretation of some representative sentences. Figure 8 shows statistics on four successive test sets on this particular technical dialect; word-frequency statistics for each set were calculated before going to the next, so "learning" only took place then. Note how the introduction of new syntactic and semantic rules is declining, although significant numbers of new words are still being introduced in this open-ended real-world dialect.

The caption's meaning representation can be indexed in a database under the picture name. The same linguistic processing methods can interpret natural-language queries, look up word senses in the index, and do a subgraph-graph isomorphism match between the query semantic network and the caption semantic network (Rowe, 1996; Guglielmo & Rowe, 1996) to prove that the query is covered by the caption.

Image Processing

While captions are generally simpler to analyze than unrestricted natural language, captioned images are not much easier than most images; they are the most valuable when they show much variety. Linguistic processing also takes much less time than image processing; NAWC-WD captions average twenty-two words long while their pictures need 10,000 pixels for minimal representation. Nonetheless, image processing of captioned images can provide valuable information not in captions. Captions rarely describe the size, orientation, or contrast of important objects in the image. They rarely describe easy-to-see features like whether the

Number	*Sentence*
1	pacific ranges and facilities department, sled tracks.
2	airms, pointer and stabilization subsystem characteristics.
3	vacuum chamber in operation in laser damage facility.
4	early fleet training aid: sidewinder 1 guidance section cutaway.
5	awaiting restoration: explorer satellite model at artifact storage facility.
6	fae i (cbu-72), one of china lake's family of fuel-air explosive weapons.
7	wide-band radar signature testing of a submarine communications mast in the bistatic anechoic chamber.
8	the illuminating antenna is located low on the vertical tower structure and the receiving antenna is located near the top.

Sentence number	*Training*		*Final*		*No binary*		*No unary*	
	Time	*Tries*	*Time*	*Tries*	*Time*	*Tries*	*Time*	*Tries*
1	27.07	13	17.93	5	8.27	5	60.63	19
2	70.27	10	48.77	9	94.62	14	124.9	23
3	163.0	19	113.1	19	202.9	23	2569.0	22
4	155.2	9	96.07	3	63.95	8	229.3	22
5	86.42	8	41.02	3	49.48	6	130.6	30
6	299.3	11	65.78	7	68.08	5	300.4	15
7	1624.0	24	116.5	5	646.0	12	979.3	25
8	7825.0	28	35.02	2	35.60	3	>50000	-

Figure 7. Example sentences and their interpretation times in CPU seconds during training; after training; after training without binary co-occurence frequencies; and after training without unary word-sense frequencies.

Statistic	*Training set 1*	*Training set 2*	*Training set 3*	*Training set 4*
Number of new captions	217	108	172	119
Number of new sentences	444	219	218	128
Number of total words in new captions	4488	1774	1535	1085
Number of distinct words in new captions	939	900	677	656
Number of new lexicon entries required	c.150	106	139	53
Number of new word senses used	929	728	480	416
Number of new sense pairs used	1860	1527	1072	795
Number of lexical-processing changes required	c.30	11	8	7
Number of syntactic-rule changes or additions	35	41	29	10
Number of case-definition changes or additions	57	30	16	3
Number of semantic-rule changes or additions	72	57	26	14

Figure 8. Overall Statistics on the Training Sets.

image is a daytime view, an outdoor view, or a historical photograph. Nor do they mention things obvious to people familiar with the picture subject, a serious problem for specialized technical images. For instance, captions rarely mention that sky or ground is shown in a picture, and NAWC-WD captions rarely mention that the photographs were taken at NAWC-WD.

MARIE's basic image processing segments the image into regions, computes region properties, and broadly classifies regions. It uses a robust "split-and-merge" method on thumbnail reductions (to about 10,000

pixels each) of the images. The image is split into small irregular regions based on color similarity, and the regions are merged in a best-first way using color and texture similarity until the number and size of the remaining regions simultaneously achieves several criteria. Typically this was when 50-100 regions remained. Then some final splitting criteria are applied, and seventeen key properties of the regions are calculated. The seventeen were developed from an extensive survey of a variety of captioned images. They represent key dimensions for distinguishing regions, including color, texture, density, horizontality and verticality, boundary shape, and boundary strength. Figure 9 shows the region segmentations found for the Figure 1 photographs, and Figure 10 lists properties computed for regions in general.

Reliable identification of objects in unrestricted photographs is very difficult because of the wide range of subjects and photographic conditions. Nonetheless, some experiments on the easier task of trying to classify the regions into one of twenty-five general categories (Rowe & Frew,

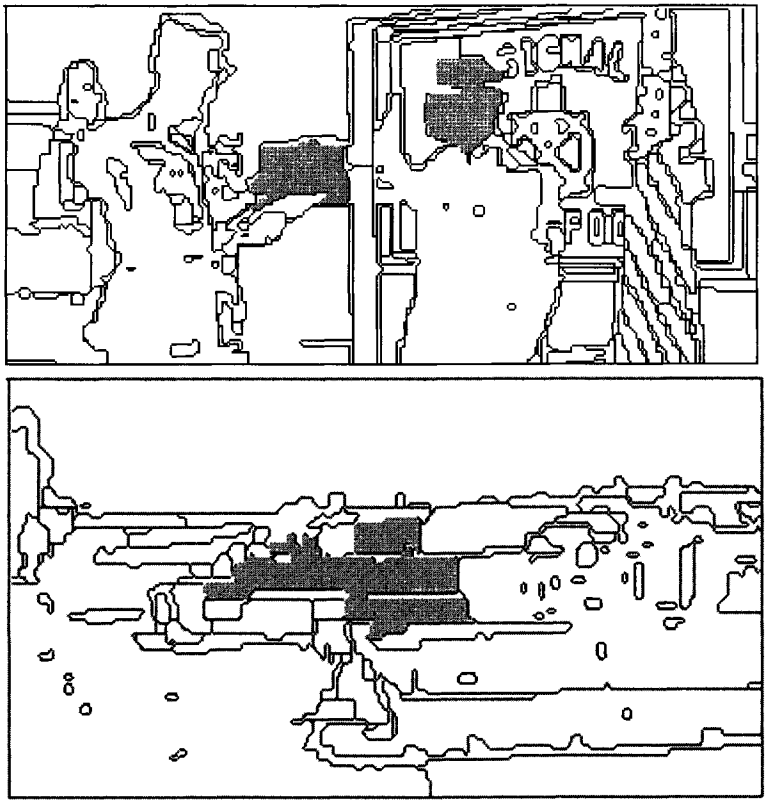

Figure 9. Segmentation and focus assignment for the photographs on the Figure 1 Web page. The shaded areas represent the computed best visual focus.

Name	Definition
circularity	area/(circumference*circumference)
narrowness	height/width of bounding rectangle
marginality	1/(1+(circumference/number border cells))
redness	average red brightness
greenness	average green brightness
blueness	average blue brightness
pixel texture	average brightness variation of adjacent cells
brightness trend	brightness correlation with x or y
symmetry	skew of center of mass from bounding-rectangle center
contrast	average strength of the region edge
diagonality	smoothed-boundary diagonality
curviness	smoothed-boundary curviness
segments	smoothed-boundary number of inflection points
rectangularity	smoothed-boundary number of right angles
size	area in pixels
density	density of pixels in bounding rectangle
height	y-skew (unsigned) of center of mass within image

Figure 10. Properties Computed for Image Regions.

1997) provide hope of helping in matching to the caption (though the next section reports a better approach). These experiments used a random sample of 128 photographs from the NAWC-WD library covering a wide variety of subjects and activities and taken under a variety of lighting conditions. A neural network was trained to classify each region of the picture as one of twenty-five—airplane, airplane part, animal, bomb, bomb part, building, building part, equipment, fire, flower, helicopter, helicopter part, missile, missile part, mountain, pavement, person, person body part, rock, ship part, sky, tank, terrain, wall, and water (some of these are domain-dependent but easily generalizable). Twenty-five classes appear sufficient for many applications because captions usually refine the classifications; if the caption mentions a B-747 and no other airplanes, that is likely to be the airplane shape in the image. The neural network takes the seventeen region properties as inputs and computes the likelihood the region is of a particular class for each of the twenty-five classes. Weights connecting inputs and outputs are learned. So the output for "equipment" exploits a high weight on the number of right angles of the region border, and the output for "sky" exploits a low weight on the brightness variation of adjacent cells.

With just this simple approach without relationship constraints, 33 percent precision was obtained in classifying the regions in a random sample. Precision was improved to 50 percent with addition of another level of neural reasoning that used the linguistic-focus ideas explained in

the next section, showing that caption information helps in image analysis. Still better performance could be achieved with appropriate domain-dependent region-relationship constraints (like that the sky is always above terrain), but domain independence is desirable for portability to other image libraries. (The classic alternative of case-based reasoning for region classification only obtained 30 percent precision for twenty-five regions; apparently some classes show too much variation in appearance.) Image processing averaged an hour per image. But again, this time is expended only during database setup when time is not critical.

Caption-Image Reference Semantics

Linguistic Focus

So far caption and image analysis have been considered separately, but their results must eventually be integrated. This requires finding the "linguistic focus" of the caption and "visual focus" of the image because these are implicitly cross-referenced. Rowe's (1994) study showed that captions are a dialect restricted in semantics as well as syntax. The headwords (usually syntactic subjects) of caption sentences usually correspond to the most important fully-visible object(s) in the image (not necessarily the largest). In conjunctions and multi-sentence captions, each headword corresponds to an important visible object. So a pod, pylon, and bracket should be clearly visible in the image for "Radar pod and pylon; front mounting bracket." Physical-action verbs are also depicted in images when appearing as gerunds, as "loading" in "Aircraft loading by crane," as participles, as past tense, or as present tense. Verbs generally correspond to relationships in the image rather than regions.

Other nouns or verbs in a caption are generally visible in part if they are related syntactically to a headword. So "Containers on truck" implies that all the containers are depicted and part but not necessarily all of the truck, while "Truck with containers" implies the opposite. Similarly, "Aircraft cockpit" guarantees only part of the aircraft is visible since "aircraft" is an adjective here. The same is true for direct objects of physical-action verbs like "resistor" in "Technician soldering resistor."

Captions also have several additional conventional forms that signal depiction, like "The picture shows X," "You can see X," and "X with Y" where "with" acts like a conjunction. Also, word forms such as "closeup of X" make X the true headword. This latter is a case of a general principle, that an undepictable headword refers its headword designation to its syntactic object. Depictability means whether a word sense is a physical object in the Wordnet concept hierarchy.

Figure 11 shows the concepts inferred by these and similar rules for the photographs of Figure 1. In tests with random photograph captions, 80 percent precision was obtained with 62 percent recall in identifying concepts shown in the photographs from the captions alone.

Another phenomenon is that some captions are "supercaptions" that refer to more than one picture, using typically-parallel syntactic constructs. For example, in "Sled test: Pretest assembly, close-up of building, parachute deployment, and post-test damage," "sled test" maps to a set of four pictures, but the four contradictory conjuncts map to each of the successive pictures. Negative depictability is also inferable when a caption does not mention something expected for its domain. For instance, NAWC-WD tests equipment for aircraft, so a test not mentioned in a caption is guaranteed not to be shown in its photograph.

Visual Focus

The linguistic focus of a caption corresponds to a subject or "visual focus" of its corresponding image. Captioned photographs alone have special visual semantics, since they are usually selected because they depict their subjects well. From a study of sample photographs, it was observed that the visual foci of captioned images were region sets with generally five characteristics:

1. they are large;
2. their center of gravity is near the picture center;
3. they do not touch the edges of the photograph;
4. their boundary has good contrast; and
5. they are maximally different from non-focus regions of the picture.

A trainable neuron was used to summarize the five factors. The best candidate region set can be found by a best-first search over region sets. This set is then matched to the linguistic focus, excluding redundant terms and those for objects too small compared to the others, like "pilot" when "aircraft" is also in linguistic focus. This requires inheritable average sizes of objects with standard deviations of their logarithms.

Performance of this approach on a random set of images (different from those tested for image processing) was 64 percent precision for 40 percent recall. These figures were computed as ratios of numbers of pixels. So, in other words, 64 percent of the pixels selected as belonging to subjects of the picture were actually part of the subjects. Precision is the challenge since 100 percent recall is easy by just designating the entire picture as the subject. Segmentation was critical for these results since only 1 percent precision was obtained by selecting all pixels whose color was significantly different from the color of any picture-boundary pixel. Altogether, caption-image reference analysis required 29K of source code and less than a second of CPU time per test caption. The shaded areas in Figure 9 represent the best hypotheses found for the subjects of the Figure 1 photographs. The program mistakenly selected a sign in the first picture that was close to the center of the image, but the other regions selected are correct as are their labels (see Figure 11). Region classifica-

image4: [enlist-103], [singular]
image4: [George -0, Staff Sergeant -0,serviceman-2], []
image4: [James -0, Major General -0, Taylor -0,commander-2], [singular]
image1: [gunnery-1], []
image1: [shoot-109], [plural]
image1: [M1A1 -0], [plural]

Figure 11. Results of linguistic focus and depiction analysis showing the terms (with any quantifications) inferred to apply to the subjects of the example photographs (the shaded areas in Figure 9).

tion in the manner of the last section helps avoid such mistakes, but it helps less than the five principles above.

Efficient Image Retrieval at Query Time

Let us now consider what happens once a set of images has been indexed by their caption and image concepts. Such stored information can be queried by parsed English queries, or also by key phrases, in the MARIE systems. Execution of such queries on a single computer processor can be thought of as sequential "information filtering" that successively rules out images on various criteria. Terms in a parse interpretation, or key phrases extracted from them, can each correspond to a separate filter, but there can be many other kinds of useful filters. Efficient query execution strategies are important in implementing these filters because speed at query time is critical to user satisfaction.

In a sequence of information filters, it often helps to put the toughest filters first to reduce workload fastest (though filter sequences are conjunctive and give the same results in any order). This is desirable when filters need a constant amount of time per data item, as in hash lookups from a sparse hash table, but not always otherwise. Rowe (1996) showed the criterion for local optimality with respect to interchange of filters i and i+1 in a filter sequence:

$$c(i) / (1 - p(f(i) \mid g(i-1))) <= c(i+1) / (1 - p (f(i+1) \mid g(i-1)))$$

Here f(i) is the event of a data item passing filter i, g(i-1) is the event of a data item passing filters 1 through i-1, c(i) is the average execution cost per data item of filter i, and p means "probability." This is only a local optimality condition since g(i-1) represents context but can be used heuristically to sort the filter sequence, and experiments showed that such sorting nearly always gave the globally optimal sequence. The criterion is especially valuable for placing information filters that do complex processing. An example is the subgraph-isomorphism check mentioned at

the end of the earlier section on "Linguistic Processing." Such matching is valuable but time-consuming, and Rowe (1996) proved it should be done last among MARIE filters.

Another way to improve the efficiency of a filter sequence is to introduce appropriate redundant information filters. Redundant filters can actually help when they are faster than the filters which make them redundant. For instance, the subgraph-isomorphism filter makes redundant the filters that only match noun senses between query and caption, but the latter can quickly cut the former's workload considerably. Another redundant filter used by MARIE-2 broadly classifies the query (for instance, into "test photo," "public relations photo," and "portrait" classes) and rules out matches with incompatible caption classes; this is also much faster but redundant with respect to the subgraph-isomorphism filter. A proven sufficient criterion for local optimality with respect to nondeletion for a redundant filter i in a filter sequence is:

$$c(i) / (1 - p (f (i) | g(i-1))) <= c(i+1)$$

with the same notation as above. Again, experiments showed this led almost always to the globally optimal sequence. Other useful analytic criteria for optimality of information filtering were shown, including those for optimality of disjunctive sequences, negations, and Boolean combinations of filters. Using such optimizations, MARIE-1 took about two seconds per query (the geometric mean of CPU time) for a sample of queries generated by actual NAWC-WD users (which were shorter and simpler than captions). MARIE-2 took about three seconds per query but gave better answers. Optimization took just a few seconds of CPU time and 23K of source code (Rowe, 1996).

Another way to speed up information filtering is by data parallelism—different processors trying different image sets to match to the query. (Other forms of parallelism do not help information filters much.) Load imbalances can occur when some processors finish early on their allocation of work because of random fluctuations in either processing time or the success rate of a filter. But data can be assigned randomly to processors, and the load imbalance estimated quite accurately at each step, which permits judging when the cost of rebalancing the processors is justified. Such parallelism would help applications requiring high-speed retrieval like real-time robots.

Conclusion

The success of text information retrieval for the Internet has obscured the considerably greater difficulty of multimedia information retrieval. Naïve methods, such as searching for keywords likely to be associated with a particular kind of image, encounter low success rates. The MARIE project has explored an integrated solution to multimedia retrieval using knowl-

edge-based methods and clues from linguistics, image analysis, presentation layout, and mathematical analysis. While perfection is not possible for information retrieval tasks, major improvements over the 1 percent success rate of naïve keyword lookup are definitely possible with the ideas presented here. Clear synergism was obtained by using multimodal clues and confirming advantages of multimodal processing (Maybury, 1997). At the same time, MARIE uses mostly domain-independent methods that are relatively easy to extend to new image libraries. The insights from MARIE may prove important in tapping the wealth of multimedia data available on the information superhighway.

Acknowledgments

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.

References

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

Flickner, M.; Sawhney, H.; Niblack, W.; Ashley, J.; Huang, Q.; Dom, B.; Gorkani, M.; Hafner, J.; Lee, D.; Petkovic, D.; Steele, D.; and Yanker, P. (1995). Query by image and video content: The QBIC System. *Computer, 28*(9), 23-32.

Guglielmo, E. J., & Rowe, N. C. (1996). 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 M. T. Maybury (Ed.), *Intelligent multimedia information retrieval* (pp. 215-239). Menlo Park, CA: AAAI Press.

Maybury, M. T. (Ed.). (1997). *Intelligent multimedia information retrieval.* Menlo Park, CA: AAAI Press.

Miller, G. A. (Ed.). (1990). WordNet: An online lexical database (special thematic issue). *International Journal of Lexicography, 3*(4), 235-312.

Ogle, V. E., & Stonebraker, M. (1995). Chabot: Retrieval from a relational database of images. *Computer, 28*(9), 40-48.

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). 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 M. T. Maybury (Ed.), *Intelligent multimedia information retrieval* (pp. 65-79). Menlo Park, CA: AAAI Press.

Rowe, N. C., & Frew, B. (1998). Automatic caption localization for photographs on World Wide Web pages. *Information Processing and Management, 34*(1), 95-107.

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

Smith, J., & Chang, S-F. (1996). VisualSeek: A fully automated content-based image query system. In P. Aigrain, V. Bove, W. Hall, & T. Little (Eds.), *Proceedings of ACM Multimedia '96* (November 18-22, 1996, Boston, MA) (pp. 87-98). New York: Association for Computing Machinery Press.

Smoliar, S., & Zhang, H. (1994). Content based video indexing and retrieval. *IEEE Multimedia, 1*(2), 62-72.

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