Finding and Labeling the Subject of a Captioned Depictive Natural Photograph

Neil C. Rowe, *Member*, *IEEE Computer Society*

Abstract—We address the problem of finding the subject of a photographic image intended to illustrate some physical object or objects ("depictive") and taken by usual optical means without magnification ("natural"). This could help in developing digital image libraries since important image properties like subject size and color of a photograph are not usually mentioned in accompanying captions and can help rank the photograph retrievals for a user. We explore an approach that identifies the "visual focus" of the image and the "depicted concepts" in a caption and connects them. The visual focus is determined by using eight domain-independent characteristics of regions in the segmented image, and the caption depiction is identified by a set a rules applied to the parsed and interpreted caption. The visual-focus determination also does combinatorial optimization on sets of regions to find the set that best satisfies focus criteria. Experiments on 100 randomly selected image-caption pairs show significant improvement in precision of retrieval over simpler methods, and, particularly, emphasizes the value of segmentation of the image.

Index Terms—Information retrieval, multimedia, caption, subject, photograph, image processing, segmentation, background, natural-language understanding, depiction.

1 Introduction

MULTIMEDIA data is increasingly stored online. However, finding good multimedia data for a user's need can be much harder than finding relevant text data. The matching of shapes or colors is rarely helpful, and some degree of computationally expensive content analysis (examination of the image pattern) is required. To support usable digital image libraries, we need robust methods that will work on a wide range of images.

Here, we investigate a general way of finding the subject of a photographic image to permit automatic analysis of its visual properties for indexing and retrieval. For instance, for Fig. 1, we would like to automatically infer that the gray object in the upper left is doing the loading. Our approach is to segment (partition) a reduced image into a few regions, and check combinations of regions to find a set that best constitutes the "visual focus" of the image. Then, we identify that focus with the subject(s) of the caption obtained from language processing. Our methods apply to normally produced "depictive" photographs, those with clear intended subjects before a background, like most technical photographs.

After some overview, we explain the image properties we use to determine the visual focus on an image and how we combine them. Then, we discuss how we analyze the caption to find its linguistic focus, and report on experiments with an implementation of this theory.

2 Previous Work

The PICTION project [14], the INFORMEDIA project [4], the work of [15], and several Web-retrieval projects like [2] and [13] have emphasized exploitation of image captions also for retrieval. Our MARIE project [3], [12] has done natural-language processing of captions for technical photographs. Certainly, caption information is important in understanding an image. But, many important things about an image are rarely mentioned by caption authors: the size of the subject, the contrast, the color, when the image was created, and the background of the image. Users are often interested in these properties: They help rate images when thousands are returned in response to a user query. With rare exceptions, these properties must be obtained by image segmentation and analysis.

Most of the important features of image subjects, like size and color, do not require extensive processing. But, finding the subject is the challenge. In image processing work, subject extraction is a special case of figure-ground disambiguation. Much of the work has been in controlled environments like factories where one can assume simplifications like that subjects are centered or that all pixels on the sides of the image are part of a single background region [9]. This will not work for natural images where secondary-importance regions often appear, like supports for equipment and variations in terrain cover. Fig. 1 contains several kinds of backgrounds not touching a side of the image, people touching the sides, and important off-center objects. But, clues as to the subjects come from both captions and image region sizes, placements, and contrasts. The challenge is to connect the caption and image information with mostly domain-independent inferences. Then, taking all factors into account, we must solve a combinatorial optimization problem as in [5].

Some work has investigated linguistic references to images [8], but rarely are linguistic descriptions precise enough to help in deciphering an image. Anaphoric references such as deictic references [6] are rare in image descriptions because people do not often view parts of images in a predictable order. Explicit location relationships like "left of" also rarely occur in natural image descriptions except for easily confusable objects (like a group of people). Dale and Reiter [1] claim that referring expressions must contain "navigation" (where the referent is located in the image) and "discrimination" (how the referent can be recognized). But real-world captions in our experience rarely do: Few relate to objects because most illustrate a single object, and few discriminate objects because their intent is to describe significance rather than appearance. Instead, real-world captions generally describe a single object centered in the image.

3 Visual Focus

Some captions apply to the image as a whole, particularly those describing a place or time like "Michelson Laboratory Main Shop, 1948." Image analysis is then of little help for indexing. But, usually the caption applies to the largest or most central objects of the image, as in Fig. 1 where the largest region is the loader, or Fig. 2 where the nontrivial region closest the center is the building.

We propose that the subject of a depictive image worth publishing is "visually focused" by several quantifiable indicators. promoted in instructional "how-to" photography books as important principles of good photographs:

1. The subject is relatively large.
2. The subject minimally touches the sides of the image (which rules out the parking lot in Fig. 2).
3. The subject center is near the center of the image (which rules out the clock in the upper right corner of Fig. 1).
4. Its outer edge has good contrast to surrounding regions (which rules out the striations on the right side of Fig. 1), and
5. It is easy to distinguish from nonfocus regions in color and appearance.

• *The author is with Code CS/Rp, US Naval Postgraduate School, Monterey, CA 93943. E-mail: rowe@cs.nps.navy.mil.*

Manuscript received 17 Sept. 1998; revised 8 July 2000; accepted 28 Sept. 2000.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 107411.

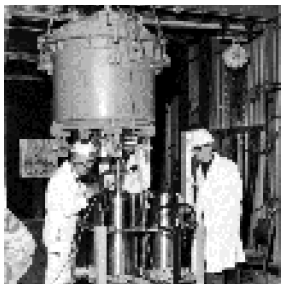

Fig. 1. Example image with caption "Keeping a close eye on the loading process are (from left) Ski Pierczynski, Ed Varnhagen, and Jack Waller."

The last criterion can be further decomposed into:

6. the color difference (which rules out the sky regions in Fig. 2 since they are too alike),
7. the difference in the brightness variation of adjacent pixels within each region (like the texture difference between the upper right side and upper left side in Fig. 1),
8. the collinearity of region edges (like in the horizontal parts of the building in Fig. 2),
9. the similarity of size (like the white parts in Fig. 1), and
10. the average region brightness (since shadows tend to look alike).

Rowe and Frew [11] exploited a subset of these indicators, but its task was the different one of region classification into one of 25 categories, and it only had 25 percent precision in identifying isolated individual regions as part of the subject. We can improve upon this if we use all 10 indicators and allow that a set of regions taken together can make a good subject when each region alone does not. For instance, two off-center regions may have a center of gravity near the picture center and, thus, make a good subject.

Fig. 2. Example image with caption "NWC Range Control Center construction progress. Side view of building with antenna tower completed."

4 Experiments Determining Visual Foci

Our testbed was a sample of 100 captioned images drawn from the US Navy Facility NAWC-WD in China Lake, California. NAWC-WD is a test facility for aircraft equipment, and pictures generally show equipment though some show public-relations events. The 100 images were randomly drawn from 389 of which 217 were drawn randomly from the photographic library and 172 taken from the NAWC-WD World Wide Web pages (and constituting most of the captioned images there in early 1997); the images chosen were distinct from a training set used to develop the

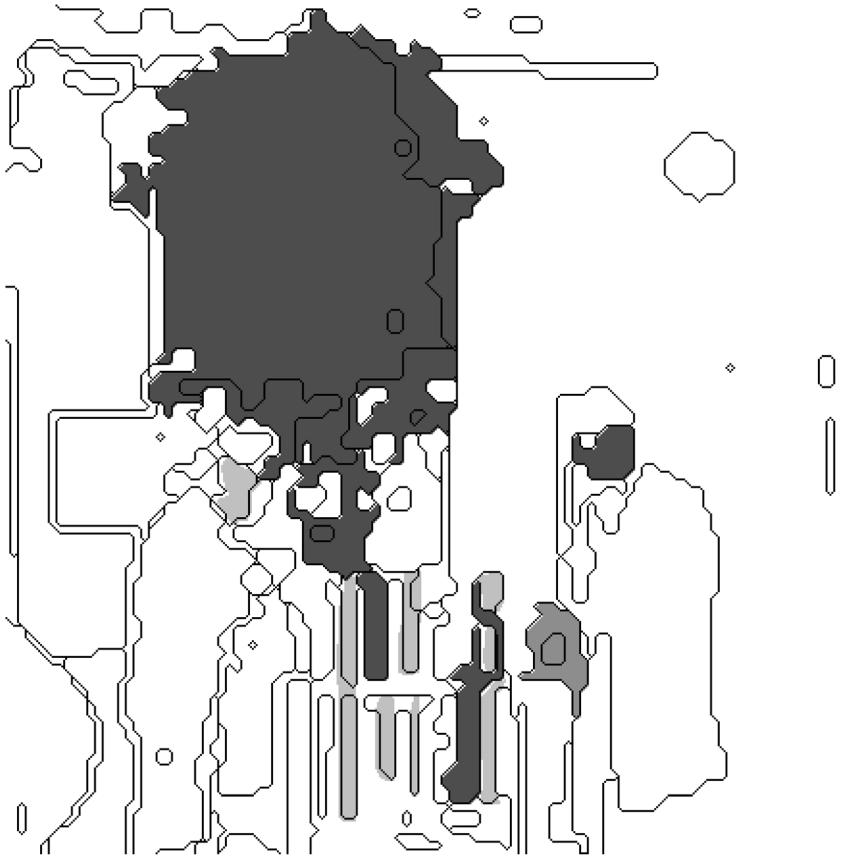

Fig. 3. Focus analysis of Fig. 1.

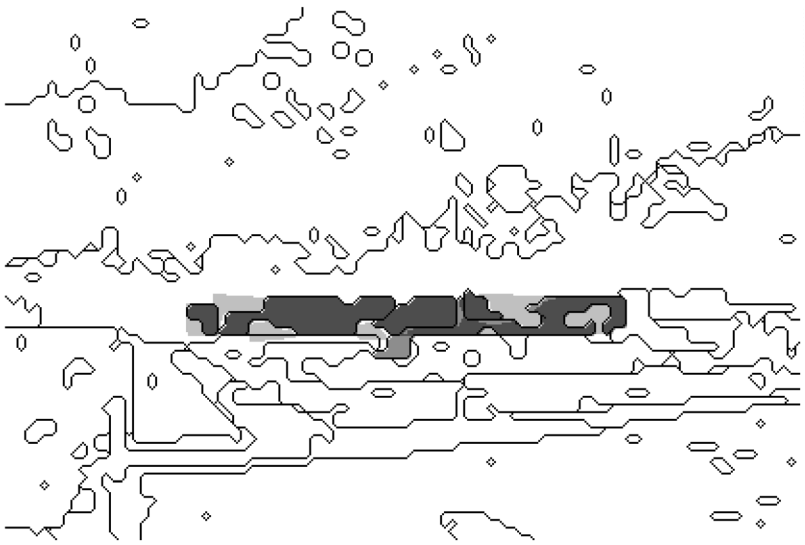

Fig. 4. Focus analysis of Fig. 2.

methods. Captions were parsed and interpreted with MARIE's natural-language understanding software [3]; processing was forced to backtrack until the best interpretation was found. We reduced the 100 sample images to thumbnail size (about 12,000 pixels per image, a data reduction of about 100 to 1) to permit faster processing since focus identification only need find large features of the image and we could still see such features at this size. We converted all thumbnails to GIF format, created arrays of the pixel values, and smoothed the arrays with four-cell sums (since GIF format is dithered). All but 14 of the 100 images were in color.

Since good region segmentation is critical to success in subject identification and we must handle a wide range of images, we do a careful segmentation using self-adjusting thresholds. We used an improved version of the split-merge program of [11], updated to work in the hue-saturation-intensity color space because it gave fewer errors than red-green-blue in segmentation of our training images. Split-merge methods partition an image into regions of strongly homogeneous characteristics and then merge neighboring regions based on a variety of similarity criteria; we used this since it generally works well for color images [16]. Initially, the image is split into regions averaging about 20 pixels in size by adjusting a clustering threshold. Careful merging is then done by a best-first search (meaning that the strongest merge candidate pair at any time is merged), starting with single-pixel regions. We used the color-vector difference in hue-saturation-intensity space, as in [13] to measure color difference. To get the region difference, we multiply the color difference by the square root of the average intensity of the two regions (a way to combat noise in saturation and hue at low levels of intensity), add a factor for the difference in local region color variation (to discriminate against merges of regions of very different textures), and subtract a factor for pixel density of the resulting region within its bounding box (to discriminate against merges that create long narrow regions). Merging continues until regions are large enough to provide good candidates for focus analysis. After experimentation with the training images for optimizing the average number of incorrect splits and merges in a segmentation, this was defined by statistics on the set of multipixel regions S that do not significantly touch the region boundary (meaning they do not touch it at all, or else are more than 50 pixels and touch it with no more than five pixels), and its subset T of the regions of 10 or more pixels:

1. S has no more than 250 regions,
2. T has no more than 50 regions,
3. T has no less than five regions,
4. T covers at least 500 pixels, and
5. the weighted area of the nearest region to the center is more than 100, where weighting is by the square of the fractional distance from the center (to discriminate against segmentations having no large region near the center).

To reduce merges of meaningful objects into background regions because of accidental color coincidences along their borders, we consider splits of regions in a final step. Splits are postulated between pairs of points on the region boundary having a local minimum of the ratio of their straight-line distance to the distance along the region boundary between them with a maximum allowed of 0.065 for this ratio as set by experimentation.

Then for each of the 40 largest regions of the image, we compute 26 statistics [11] covering size, elongation, symmetry, color, color variation, boundary smoothness, and boundary contrast. We do best-first search to find the region subset that is the best choice for the visual focus of the picture. This search uses an evaluation function which computes metrics for the factors listed in Section 3: the number of pixels in the region set; the fraction of cells in the set on the picture border; the relative distance of the center of gravity of the region set to the center of the

Fig. 5. Example image with caption "Apple Computer, Inc., Crada on computer network and communications development."

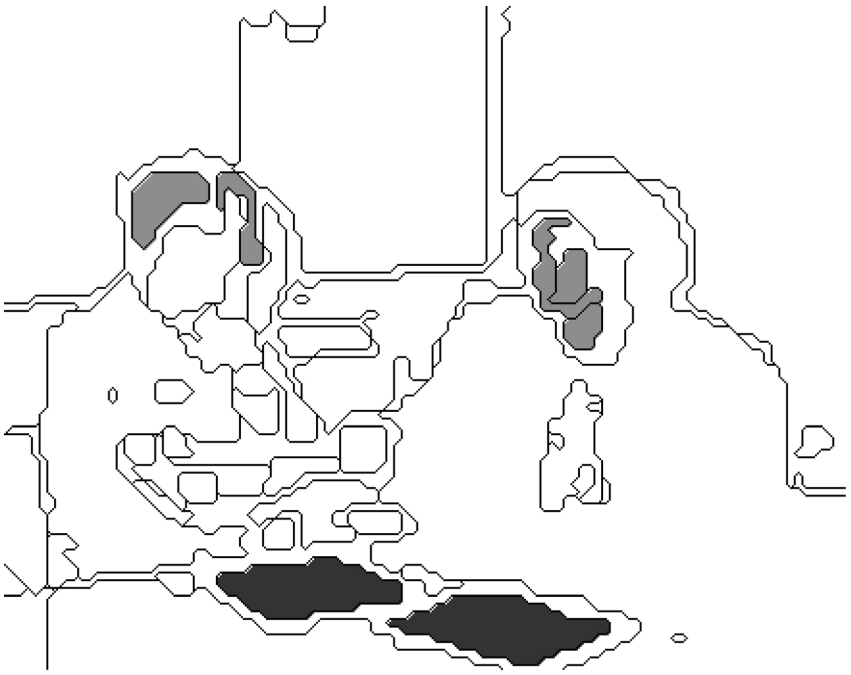

Fig. 6. Focus analysis of Fig. 5.

image; the average strength of the color contrast along the external edges of the region set; the color difference with the most-similar region not in the set (larger differences are more desirable); the texture difference with the most-similar region; the size difference with the most-similar region; and the collinearity of the edges in region set. Nonlinear sigmoid functions are applied to these factors to keep them between 0 and 1, which permits interpreting them as probabilities of the region set being the focus based on that factor alone, and the calculation can be interpreted as a single artificial neuron. Heuristic search tries to find the focus set that minimizes the weighted sum of these measures; it must be heuristic because the factors interact, and it must involve search because greedy algorithms do not always work. We also explored simulated annealing since [5] found it helpful, but it was significantly slower for us: Apparently there is often not much advantage to exploring very-different region sets with our images.

5 Using Linguistic Information to Name the Visual Focus

The other source of focus information is the image caption. Rowe [10] proposed and tested rules for identifying "depictability" of nouns in a caption. We improved upon its performance by writing new rules applying to caption semantic representations instead of raw captions to avoid problems of ambiguous words, and we wrote new rules now permitting verbs as subjects. The semantic representations were obtained by a statistical parser that we previously developed and trained on a set of 616 NAWC-WD captions [3], and that assigns word senses drawn from the Wordnet thesaurus system [7]. Our rules identify "linguistic foci" as:

1. Grammatical subjects of all caption sentences and clauses, including separate components of compound subjects (e.g., "f-16," "f-18," and "pod" for "f-16 and f-18 aircraft from front; radar pod on left").
2. Present participles or present-tense verbs attached to a grammatical subject (e.g., "loading" for "crew loading aircraft").
3. Objects of present participles or verbs attached to a grammatical subject (e.g., "aircraft" for "crew loading aircraft").
4. Objects of grammatical subjects which are kinds of views (e.g., "aircraft" in "closeup of aircraft").
5. Objects of "with" prepositional phrases and "showing" participial phrases attached to a grammatical subject (e.g., "f-16" in "f-18 with f-16").

Not all linguistic foci are depictable in an image, like "department." Other rules restrict depictability to physical objects that are not geographical locations, actions involving physical movement, actions involving a change in a visible property of an object, visual signals, nor sets whose elements are themselves depictable. Depictability is enforced by checking the Wordnet superconcepts of a proposed focus against a short list of approved types. If no linguistic focus for a caption satisfies the depictability requirements, the defaults are the grammatical subjects, like "analysis" for "analysis using methodology."

This results in a set of candidate concepts for the visual focus of a picture. Two additional restrictions not in [11] are now applied. First, we use Wordnet to eliminate redundant candidates (like "vehicle" when "truck" is also a candidate) not caught previously by the resolution of anaphoric references. This also requires checking quantifiers, so "an aircraft" will not be considered redundant with "another aircraft." Second, we eliminate candidates whose relative size is too small compared to others since the visual-focus identification will find only large regions. For instance, we eliminate pilots when aircraft are also candidates since pilots are too small to see in a typical aircraft. To do this, we define inheritable average sizes for classes of objects, plus the standard deviation on the logarithm of the size.

Now, we can map our candidates to the visual focus. Our claim is that the regions in visual focus usually correspond to the remaining concept candidates. This does not tell us which image regions map to which concepts (since, in general, it is a many-to-many mapping), but this should be sufficient enough to establish general properties of the foci like color and size that are important

TABLE 1
Results of Experiments

Experiment	Mean precision	Mean recall	Mean CPU time in seconds per image, not including segmentation
Visual control 1: Nontrivial regions not touching image edges	0.499	0.734	0.17
Visual control 2: Nontrivial region closest to center plus all regions of similar color	0.451	0.302	0.67
Visual control 3: All pixels not of similar color to any boundary pixel	0.012	0.472	380.5 (but segmentation not required)
Visual-focus program described	0.644	0.403	403.6
Linguistic control: All caption words interpreted as nouns or verbs	0.465	1.000	0.001
Linguistic-focus program	0.799	0.616	0.469

for retrieval. So, we index the visual focus with the remaining caption concepts.

6 Results

Figs. 3 and 4 illustrate the performance of our program on Figs. 1 and 2, respectively, and Fig. 6 shows performance on another picture shown in Fig. 5. The dark shaded regions are the regions correctly identified as the focus of the image; the light gray regions are focus regions missed by the program; and the medium gray regions (like in the lower right of Fig. 3) are incorrectly identified as foci by the program. A maximum of ten focus regions were permitted in these experiments. The linguistic focus analysis labeled the focus areas in Fig. 3 as "loading," "Pierczynski," "Varnhagen," and "Waller," those in Fig. 4 as "building," and those in Fig. 6 as a "CRADA" (a contract). The search examined 246, 403, and 611 region sets, respectively, for the figures before choosing the focus shown.

We compared our program for finding the visual focus with three simpler control methods. For the first control, we interpreted the visual focus as the set of regions after segmentation that are not touching the picture boundary. For the second control, we interpreted the focus as the set of regions similar in color to the nontrivial region closest to the center of the picture. The similarity threshold was the final merging threshold for segmentation. Recall was computed as the ratio of the total area of the regions selected and in the correct focus set to the total area of the regions in the correct focus set; precision was computed as the ratio of the total area of the regions selected and in the focus set to the total area of the regions selected. The correct focus set for each image was created by manual inspection of the segmentation results.

For the third control, we found all pixels whose color difference was not within the final merging threshold for any picture-boundary cell. The idea was to distinguish nonbackground colors. Strictly speaking, it was unfair to use the final merging threshold since the third control does not segment, but it did not help much: We got similar or worse results for every other threshold we tried. To evaluate the results, we looked up region labels after segmentation for each pixel and counted those in correct focus regions. We computed precision as the ratio of the number of correct focus-region pixels selected to the number of pixels selected, and we computed recall as the ratio of the number of correct focus-region pixels selected to the number of correct focus-region pixels.

We tested our programs on the 100 image-caption pairs of the images described in Section 4. Table 1 compares our visual-focus and linguistic-focus programs to the above-mentioned three controls for the visual focus and a simple control for the linguistic focus. "Nontrivial regions" were those of 10 or more pixels. Our linguistic-focus program in the 100 captions eliminated 19 redundant concepts and 14 concepts too large to be shown (e.g., "California"). The linguistic recall and precision were calculated on keyword count ratios with the usual information-retrieval definition. CPU time was measured for a Quintus Prolog implementation on a Sun Sparcstation.

Our visual-focus and linguistic-focus programs definitely perform better on precision than the controls. The precision on the visual focus is more important than recall for our motivating tasks of identifying color and contrast of the image subject; perfect recall is always easy to obtain by retrieving the entire picture. Note that segmentation is critical for good precision, as the only method not using it (control 3) does very badly on that metric; this suggests that in multimedia retrieval systems like [13], features allowing comparison of color similarity of nonsegmented images are of limited value. Our methods take more time than the controls, but this time is expended during one-time indexing of a database and not during access to it when speed is much more important.

Many of the observed errors represented justifiable exceptions to our theory of the visual-linguistic correspondence. Sometimes the first sentence of a multisentence caption establishes a general title for a set of captions, and its objects may not be depicted (as in

Fig. 7. Example image with caption "Awaiting painting and placement: Polaris

"Airframe ordinance and propulsion. Assembly"). Some captions describe the setup or consequences of an action which is not yet depicted in linguistic focus (as in "Pretest view of rocket launch"). Other captions describe the device taking the photograph (as in "Infrared view of model"). Some objects implied by the caption may not be in good visual focus if a photograph is taken hastily or under conditions out of the control of the photographer (as images of test flights of aircraft). Finally, regions in visual focus may not have any counterpart in the caption if they commonly associate with the major caption subject. In Fig. 7, the flatbed that held the missile in transport helps convey the meaning of "arriving" so it is part of the visual focus and balances the missile geometrically ("awaiting" and its objects are not depictable). In general, if a physical-motion action is in linguistic focus, we can postulate that its agent or instrument is also in visual focus. People and their body parts often appear this way since a principle of appealing photographs is to include the "human element."

7 Conclusions

We have addressed a very difficult and heretofore uninvestigated problem in this paper, that of distinguishing and identifying the subjects of real-world captioned depictive photographs with only general-purpose knowledge. These methods should work well with normally photographed images having single subjects with clear boundaries. We have shown promising results for photos randomly drawn from a technical library, using some robust analysis methods on both the image and its caption. In particular, we have shown that classification of the objects in the image, as in [11], is unnecessary with the right set of focus criteria, and that semantic interpretation of the caption improves the retrieval success over processing of the raw caption words, as in [10]. Our methods should be useful for improving the success rates of multimedia information retrieval on the World Wide Web.

Acknowledgments

This work was supported by the US Army Artificial Intelligence Center, and by the US Naval Postgraduate School under funds provided by the Chief for Naval Operations. All photos are US Navy photos shown at the analyzed resolution.

References

[1] R. Dale and E. Reiter, "Computational Interpretation of the Gricean Maxims in the Generation of Referring Expressions," *Cognitive Science,* vol. 19, no. 2, pp. 233-263, 1995.

[2] C. Frankel, N.J.P. Swain, and B. Athitsos, "WebSeer: An Image Search Engine for the WorldWide Web," Technical report 96-14, Computer Science Dept., Univ. of Chicago, Aug. 1996.

[3] E. Guglielmo and N. Rowe, "Natural-Language Retrieval of Images Based on Descriptive Captions," *ACM Trans. Information Systems,* vol. 14, no. 3, pp. 237-267, July 1996.

[4] G. Hauptmann and M. Witbrock, "Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval," *Intelligent Multimedia Information Retrieval,* M. Maybury, ed., pp. 215-239, 1997.

[5] L. Herault and R. Horaud, "Figure-Ground Discrimination: A Combinatorial Optimization Approach," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 9, pp. 899-914, Sept. 1993.

[6] F. Lyons, "Deixis and Anaphora" *In The Development of Conversation and Discourse,* T. Myers, ed., 1979.

[7] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Five Papers on Wordnet," *Int'l J. Lexicography,* vol. 3, no. 4, Winter 1990.

[8] L.A. Pineda and E. Garza, "A Model for Multimodal Reference Resolution," *Proc. ACL/EACL 1997 Workshop Referring Phenomena in a Multimedia Context and Their Computational Treatment,* 1997.

[9] A.A. Rodriguez and O.R. Mitchell, "Image Segmentation by Successive Background Extraction," *Pattern Recognition,* vol. 24, no. 5, pp. 409-420, 1991.

[10] N. Rowe, "Inferring Depictions in Natural-Language Captions for Efficient Access to Picture Data," *Information Processing and Management,* vol. 30, no. 3, pp. 379-388, 1994.

[11] N. Rowe and B. Frew, "Automatic Classification of Objects in Captioned Depictive Photographs for Retrieval," *Intelligent Multimedia Information Retrieval,* M. Maybury, ed., pp. 65-79, 1997.

[12] N. Rowe and B. Frew, "Automatic Caption Localization for Photographs on World Wide Web Pages," *Information Processing and Management,* vol. 34, no. 2, 1998.

[13] J. Smith and S.-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System," *Proc. ACM Multimedia,* 1996.

[14] R.K. Srihari, "Automatic Indexing and Content-Based Retrieval of Captioned Images," *Computer,* vol. 28, no. 9, pp. 49-56, Sept. 1995.

[15] S. Smoliar and H. Zhang, "Content-Based Video Indexing and Retrieval," *IEEE Multimedia,* pp. 62-72, Summer 1994.

[16] A. Tremeau and N. Borel, "A Region Growing and Merging Algorithm to Color Segmentation," *Pattern Recognition,* vol. 30, no. 7, pp. 1191-1203, 1997.