Mapping between image regions and caption concepts

# of captioned depictive photographs

Neil C. Rowe

U.S. Naval Postgraduate School
Monterey, CA 93943 USA
rowe@cs.nps.navy.mil

Abstract

We discuss the obstacles to inference of correspondences between objects within photographic images and their counterpart concepts in descriptive captions of those images. This is important for information retrieval of photographic data since its content analysis is much harder than linguistic analysis of its captions. We argue that the key mapping is between certain caption concepts representing the "linguistic focus" and certain image regions representing the "visual focus". The mapping is one-to-many, however, and many image regions and captions concepts are not mapped at all. We discuss some domain-independent constraints that can restrict potential mappings. We also report on experiments testing our criteria for visual focus of images.

1. Introduction

The problem of relating images to their natural-language descriptions ("captions") is a central one in intelligent processing of multimedia. The PICTION project (Srihari, 1995), the INFORMEDIA project (Hauptmann and Witbrock, 1997), (Smoliar and Zhang, 1994), and several Web-retrieval projects (e.g. Smith and Chang, 1996; Frankel and Swain, 1996) have addressed this issue. Our own previous work on the MARIE project (Guglielmo and Rowe, 1996; Rowe and Frew, 1998) has developed methods involving both natural-language processing and image processing for technical photographs. However, despite some promising ideas (Rowe, 1994), our project has not directly addressed the problem of the mapping between images and captions.

This paper reports on a study of 399 captioned images from NAWC-WD, 217 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). NAWC-WD is a Navy test facility for aircraft equipment. All 399 captions have been parsed and interpreted (and processing was forced to backtrack until the best interpretation was found), which permits matching to exploit the semantics of the captions rather than superficial characteristics like the occurrence of particular words (Guglielmo and Rowe, 1996). Our use of real-world data has been helpful keeping our attention focussed on the central problems of multimodal reference rather than only theoretical issues.

Since natural-language processing can be considerably faster than image processing, it is desirable to exploit as much as possible from the caption to understand and index an image. Unfortunately, many important things about an image rarely are mentioned by caption authors: the size of the subject, the contrast, when the image was created, and the background of the image. These things can all be quite important when rating thousands of images retrieved in response to a user query. For instance, the two pictures below both depict Sidewinder missiles and the caption language suggests equally the depiction of Sidewinders, yet the first is a much better response to a query on "Sidewinder" since the second just shows an aircraft carrying Sidewinders.

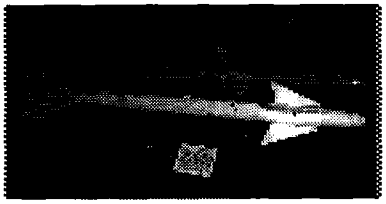

Figure 1: "Sidewinder (AIM-9), the Free World's premier dogfight missile and China Lake's most recognized product."

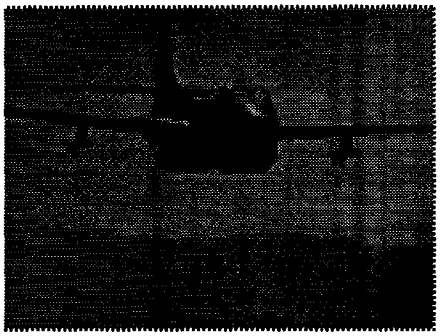

Figure 2: "Sidewinder."

Fortunately, most of the key image features needed for rating the relevance of images do not require extensive amounts of processing. Subject size appears particularly important to users. However, to compute it we must know the subject of the image, and this can be tricky. Clues come from both caption information and the image appearance (placement and contrast of regions of the image, since important regions tend to be at the center with a good color contrast). The challenge is to connect the two domains with mostly domain-independent inferences.

Some other work has investigated the connection between graphical images and their linguistic descriptions (e.g. Pineda and Garza, 1997). There are a variety of referring mechanisms. However, anaphoric references including context-dependent deictic references (Lyons, 1979) are rare because people do not often consider images in order and could get confused by anaphora. Explicit location relationships like "left of" also occur rarely except for images of easily confusable objects (like a group of people or aircraft). (Dale and Reiter, 1995) claims 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 like our NAWC-WD test ones rarely do: Few relate 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.

1. Visual focus

Some captions apply to the image as a whole, particularly those describing a place or time. Here region analysis is of no help in making the mapping. For instance:

Figure 3: "Michelson Laboratory Main Shop, 1948."

But usually the caption applies to the central objects of the image, as in Fig. 1 where the object whose center of gravity is closest to the picture center (and brightest) is the Sidewinder. Sometimes the central object is not so easy to distinguish, as in Fig. 4 where the test vehicle can be distinguished by its color although it is off-center one-third of the way down the picture and slightly to the left.

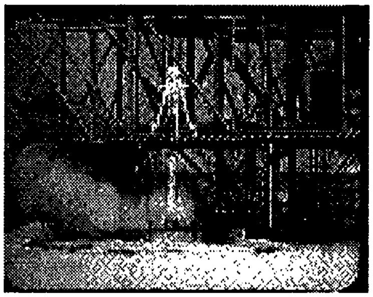

Figure 4: "China Lake's Soft-Landing Vehicle (SLV) during control testing, 1961 (from data film)."

We propose the principle that the subject of good depictive images is "visually focussed" by several quantifiable indicators: it is large, its center is near the center of the image, it minimally touches sides of the image, its edge has good contrast to surrounding regions, and it is especially distinguishable from non-focus regions in color and appearance. These are promoted in instructional photography books as important principles of good photographs.

(Rowe and Frew, 1997) explored a simplified form of some of the focus criteria, but got only 20% accuracy in identifying focus regions in isolation. So we now search for a set of regions taken as a whole, and the focus indicators apply to the union of the regions. In Fig.4 for instance, the union of the vehicle region with its smoke plume is better centered vertically than either alone. The subject of Fig. 5 consists of a white patches with black stripes. These can be grouped together by border contrast, border collinearity, and brightness. Taken together they constitute a strong candidate for the visual focus.

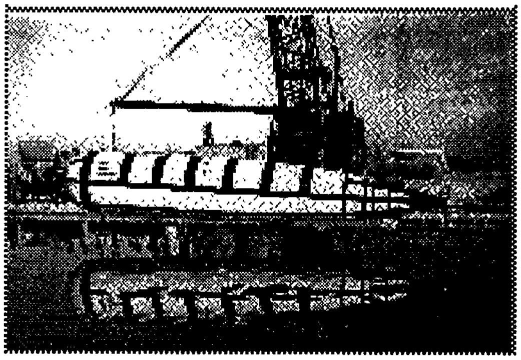

Figure 5: "Moray test vehicle (TV-1A) during testing at China Lake."

2. Experiments determining visual foci

We have begun experiments to test our theory using a random subsample from our test images. We segmented using the program of (Rowe and Frew, 1997) but now updated it to work in the hue-saturation-intensity color space instead of red-green-blue because it generally gave us fewer (although still some) segmentation errors. We used the color-vector difference in hue-saturation-intensity space as in (Smith and Chang, 1996) to measure color difference. Merging continued on each image until it contained less than 100 nontrivial (multiple-pixel) regions.

Then for each image, we compute properties as in (Rowe and Frew, 1997) of the 40 largest regions in the image, and do a heuristic search to find the best subset of these to make the visual focus of the picture. This used an evaluation function with five factors: the square root of the number of pixels in a region; the fraction of the region-edge cells on the picture border; the ratio of the distance of the center of gravity of the region to the center of the image to the distance of the corner of the image from the center of the image; the average color difference along the region boundary; and the difference in average color, size, and border collinearity of the closest-matching region in the image (to model "discriminability"). Nonlinear sigmoid functions are applied to these factors to scale them keep them between 0 and 1, which permits interpreting them as probabilities of being nonfocus regions. Heuristic search tries to find the focus set that minimizes the weighted sum of these nonlinear measures; it must be heuristic because the factors interact, and it must involve search because a greedy algorithm does poorly.

Figures 6-8 illustrate the early performance of our program, on the images of Figures 5, 9, and 10 respectively. The shaded regions are the computed visual focus assuming there were a maximum of ten focus regions. Clearly our segmentation methods need work, but the focus assignments are still encouraging. A closed contour around the regions selected does include much of the main subjects of these pictures. That included a substantial part of the missile (though also its reflection) and crane in Fig. 6, key parts of the aircraft in Fig. 7, and much of the furnace in Fig. 8 (although missing the hard-to-segment person). The search examined 1570, 1375, and 709 region sets respectively for the images before choosing the focus sets shown. We apparently need additional factors weighting against focus fragmentation, however.

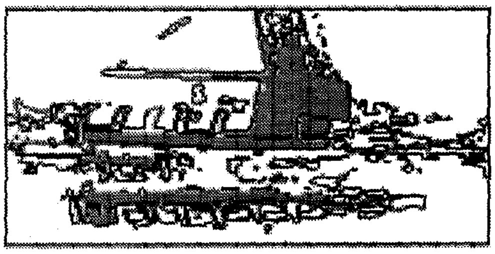

Figure 6: Segmentation and analysis of Fig. 5.

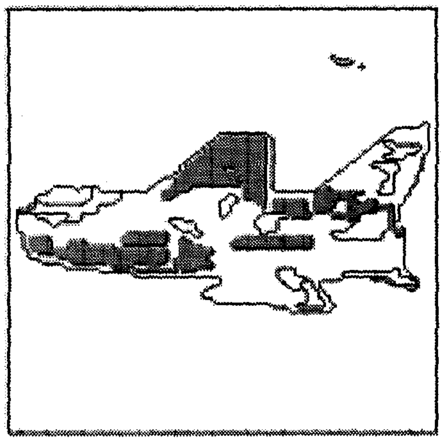

Figure 7: Segmentation and analysis of Fig. 9.

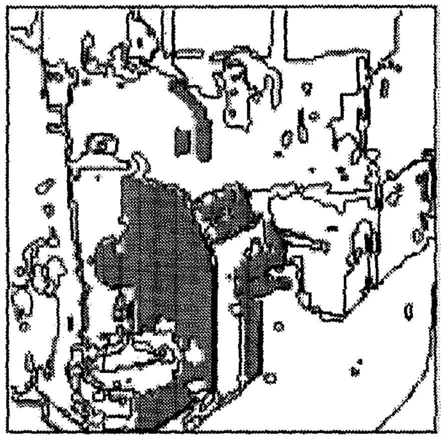

Figure 8: Segmentation and analysis of Fig. 10.

3. Linguistic focus

The other source of focus information is the linguistic focus of a caption, as discussed in (Rowe, 1994). In the image and caption below, "Corsair" is the subject and is the main depicted object. However, it has a participial phrase involving the verb "carry" which typically links a depictable object to the caption subject, so the ERDL also must be depicted (and is under the wing, in a distinctive bright red color in the original photograph). But furthermore, the Walleye must be depicted too because a correct case analysis of the caption should infer that the ERDL is part of the Walleye. Hence three things must be depicted: Corsair, ERDL, and Walleye. This illustrates why full linguistic analysis is important for technical captions since there are numerous kinds of case relationships that should be distinguished. It also suggests the importance of corpus-based linguistic methods (Charniak, 1993) because the preferred cases differ considerably between applications.

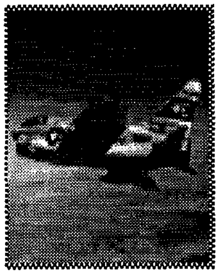

Figure 9: "A-7 Corsair of tenant OPTEVFOR squadron VX-5 carrying Walleye II (AGM-62) ERDL (extended-range data link)."

Shown below is the semantic interpretation computed by our MARIE-2 software for the caption of Fig. 9. (Sense numbers come from the Wordnet thesaurus system (Miller, 1990) except for sense 0 for NAWC-specific concepts defined by us.) The first concept in our meaning lists is the principal subject.

[a_kind_of(v3,'A-7'-0), owned_by(v3,v29), a_kind_of(v29,'VX-5'-0), owned_by(v29,v26), a_kind_of(v26,'OPTEVFOR'-0), a_kind_of(v26,renter-1), agent(v52,v3), a_kind_of(v52,carry-107), tense(v52,prespart), object(v52,v109), a_kind_of(v109,'extended-range data link'-6), part_of(v109,v2), a_kind_of(v2,'Walleye Ii'-0), a_kind_of(v2,'AGM-62'-0)].

When a caption subject or verb is a priori nondepictable, it permits its direct and indirect objects to have depictability guarantees if they are of the right types. For instance, "analysis" below is a mental action that is not depictable. "Using" is a verb that specifically links to more-precise objects, which are fully depicted if they are depictable; so "furnace" is the only guaranteed-depicted concept below:

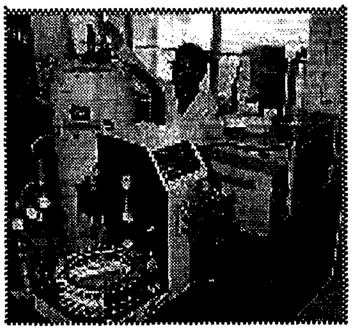

Figure 10: "Analysis using graphite furnace."

There are exceptions when size differences are involved. "Measuring" is a similar nondepictable gerund below, but the difference in the size of measuring equipment and the sample means that the sample cannot be seen clearly. Nonetheless, the place where the sample resides in the image is within the boundaries of the image.

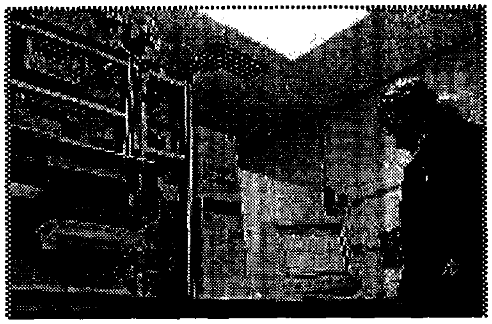

Figure 11: "Measuring residual explosives in soil sample."

So we propose the following logical constraints for the mapping from descriptive captions to depictive images, extending the criteria of (Rowe, 1994):

1.The only depictable objects are physical objects that are not geographical locations.
2. Actions are depictable if they involve physical motion or a change to a visible property.
3. Actions are potentially-depictable if some instances involve physical motion or a change to a visible property.
4. Depictable subjects of all caption sentences and clauses (including separate components of compound subjects) are inferred to be depicted. (Example: "Corsair" in Fig. 9.)
5. Depictable present participles or present-tense verbs are depicted if they attach to a depictable subject. (Example: "carrying" in Fig. 9.)
6. Depictable objects of depictable or potentially-depictable participles or verbs are depicted. (Example: "ERDL" in

Fig. 9.)

7. Depictable objects of physical-relationship prepositions are depicted at least in part. (Example: "F-18" in "Sidewinders on F-18".)

8. Depictable objects of time-relationship prepositions are depicted in part if they represent depictable events. (Example: "firing" in "Vehicle during firing").

4. Mapping between linguistic and visual foci

There remains a matching problem between caption concepts and image regions. This can often be done by a relaxation process. For instance for Fig. 12, the ship is the grammatical subject and hence depictable; but also "firing" is depictable (as the smoke generated), and when a verb is depictable, often its direct object is too, as is the BOMROC missile in this case. Hence there are three things to find in the image. There are five main regions in the image: ship, missile with plume, water, land, and sky. The last three can be excluded for foci since they touch both sides of the image. The remaining two do center somewhat close to the center of the image. This and customary relative-size information suggests the right matches.

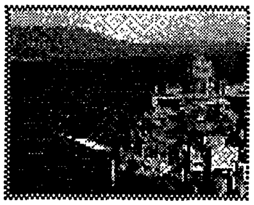

Figure 12: "U. S. S. Clarion River firing BOMROC, 1966."

In general, we propose that the mapping between most descriptive captions and their corresponding depictive images is one-to-many from each of certain focused caption concepts to a set of regions. If c represents caption concepts and r represents image regions, then we postulate that the mapping can be written as:

$$f_i(r_{j1}, r_{j2}, r_{j3}, \ldots) = c_i .$$

However, since figuring each such f is often impossible without domain-dependent knowledge (like that people exhibit shades of pink), a domain-independent approach must generally make do with a relation rather than a function, and model the situation as pairing two sets, linguistic focus (concepts) and visual focus (image regions). The linguistic focus can be determined by the rules given in section 3. But the visual focus involves satisficing criteria and many candidate sets may be possible. We then must use a metric based on the criteria of section 1 to evaluate sets of image regions and find the best (minimum) candidate.

5. Action-confirmatory objects

Some objects in the visual focus may not be in the linguistic focus if a photograph is taken hastily (as images of test flights of aircraft) and the photographer did not have time to get the subject centered and close. More commonly, some important unmentioned object may balance the visual focus. In Fig. 13, the flatbed that held the missile in transport helps convey the meaning of "arriving", so it is part of the visual focus; "awaiting" is not depictable. (In the implementation, however, the flatbed was too dark to segment well, so it was ignored except for its top during focus assignment.) In general, if a physical-motion action is in linguistic focus, postulate that the agent or instrument of the physical-motion action is also in visual focus even if not in linguistic focus.

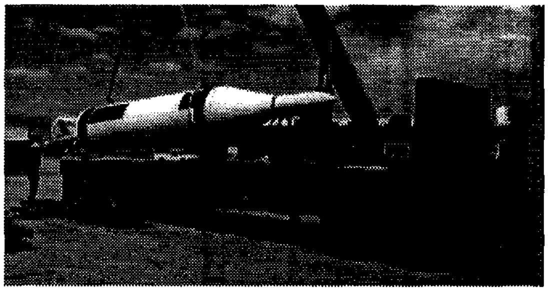

Figure 13: "Awaiting painting and placement: Polaris missile arriving."

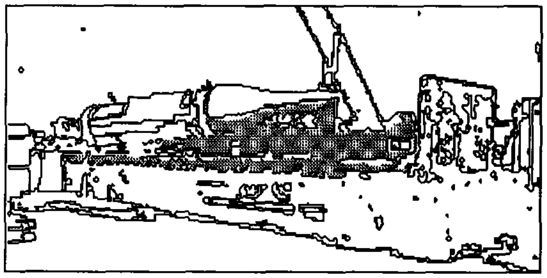

Figure 14: Visual-focus analysis of Figure 13.

For Fig. 15, the best candidate for the visual focus by our criteria is the side of the pool since it is well contrasted and closest to the center. But the set of the person, toy, and porpoise regions is equally well centered, and is the true visual focus: A better caption might be "Notty the porpoise playing with toy presented by trainer". The problem is that "mugging" often involve props (as a kind of "stage acting") and other beings (as a kind of "social act"). Domain-dependent knowledge like the presence of eyes on animals would disambiguate this case, but domain-dependent-rules limit portability, an important design criterion.

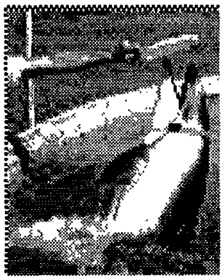

Figure 15: "'Notty' the propoise mugging for the camera, 1962."

The flatbed, human arm, and prop above (and the measuring equipment in Fig. 11) are what we call "action-confirmatory" depicted objects. These unmentioned concepts help confirm the meaning of a relatively vague state-change verb or action noun. They can be inferred by rules like: (1) If a depicted action involves a state change, the visual focus may include agents and instruments of the state change, particularly if the action concept is very general; and (2) If a nondepictable action involves a necessary prop and object, the visual focus must include them.

People and their body parts often appear as action-confirmatory objects, like the arm above the operator in Fig. 10 which balances the furnace on the left., and the operator on the right of Fig. 11 who balances the soil sample on the left horizontally (albeit not vertically). A general principle of photography is to include the "human element", so people may be visually focused for nondepictive purposes.

Future work will need to exploit these auxiliary objects as well as usual size relationships between objects and better criteria for grouping of visual-focus sets.

6. Acknowledgements

This work was supported by the U.S. Army Artificial Intelligence Center, and by the U. S. Naval Postgraduate School under funds provided by the Chief for Naval Operations. All photos are U.S. Navy photos shown in the actual size used for image processing using standard pixel density.

7. References

Charniak, E. 1993. *Statistical Language Learning*. Cambridge, MA: MIT Press.

Dale, R. and Reiter, E. 1995. Computational Interpretation of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science,* 19 (2), 233-263.

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

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

Hauptmann, G. and Witbrock, M. 1997. Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. In *Intelligent Multimedia Information Retrieval,* Maybury, M., ed. Palo Alto, CA: AAAI Press, 215-239.

Lyons, F. 1979. Deixis and Anaphora. In The Development of Conversation and Discourse, T. Myers, ed., Edinburgh University Press.

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

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K., 1990. Five Papers on Wordnet. *International Journal of Lexicography, 3,* 4 (Winter).

Pineda, L. A. and Garza, E. 1997. A Model for Multimodal Reference Resolution. ACL/EACL 1997 Workshop on Referring Phenomena in a Multimedia Context and Their Computational Treatment, Budapest.

Rowe, N. 1994. Inferring Depictions in Natural-Language Captions for Efficient Access to Picture Data. *Information Processing and Management, 30,* 3, 379-388.

Rowe, N. and Frew, B. 1997. Automatic Classification of Objects in Captioned Depictive Photographs for Retrieval. In *Intelligent Multimedia Information Retrieval,* ed. M. Maybury, AAAI Press, 65-79.

Rowe, N. and Frew, B. 1998. Automatic Caption Localization for Photographs on World Wide Web Pages. *Information Processing and Management, 34,* 2.

Smith, J. and Chang, S.-F. 1996. VisualSEEk: A Fully Automated Content-Based Image Query System. Proceedings of ACM Multimedia 96.

Srihari, R. K. 1995. Automatic Indexing and Content-Based Retrieval of Captioned Images. *IEEE Computer, 28,* 9 (September), 49-56.

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