Rule Based Strategies for Image Interpretation

T. E. Weymouth, J. S. Griffith, A. R. Hanson and E. M. Riseman

Department of Computer and Information Science,
University of Massachusetts at Amherst[1]

ABSTRACT

We present an interpretation system which utilizes world knowledge in the form of simple object hypothesis rules, and more complex interpretation strategies attached to object and scene schemata, to reduce the ambiguities in image measurements. These rules involve sets of partially redundant features each of which defines an area of feature space which represents a "vote" for an object. Convergent evidence from multiple interpretation strategies is organized by top-down control mechanisms in the context of a partial interpretation. One such strategy extends a kernel interpretation derived through the selection of object exemplars, which represent the most reliable image specific hypotheses of a general object class, resulting in the extension of partial interpretations from islands of reliability.

1. Introduction

The use of world knowledge, together with top down control, is beneficial and probably essential in domains where uncertain data and intermediate results containing errors cannot be avoided. Ambiguity and uncertainty in image interpretation tasks arise from many sources, including the inherent variation of objects in the natural world (e.g., the size, shape, color and texture of trees), the ambiguities arising from the perspective projection of the 3D world onto a 2D image plane, occlusion, changes in lighting, changes in season, image artifacts introduced by the digitization process, etc. Nevertheless, even with marginal bottom-up information, in familiar situations human observers can infer the presence and location of objects.

We present a system where convergent evidence from multiple interpretation strategies is organized by top-down control mechanisms in the context of a partial interpretation. The extreme variations that occur across images can be compensated for somewhat by utilizing an adaptive strategy. This approach is based on the observation that the variation in the appearance of objects (region feature measures across images) is much greater than object variations within an image. The use of exemplar strategies using initial hypotheses and other top-down strategies results in the extension of partial interpretations from islands of reliability. Finally a verification phase can be applied where relations between object hypotheses are examined for consistency.

In this paper the interpretation task examined is that of labelling an initial region segmentation of an image with object (and object part) labels when the image is known to be a member of a restricted class of scenes (e.g., suburban house scenes). The systems developed by Ohta [5] for understanding images of buildings in outdoor settings and by Nagao [4] for understanding aerial photographs bear some similarity to the techniques employed here. A review of these and other related work in image interpretation appears in [1].

1. This research was supported in part by Defense Advanced Research Projects Agency under contract N00014-82-K-0464, the National Science Foundation under Grant MCS79-18209, and the Air Force Office of Scientific Research under Contract F49620-83-C-0099.

2. A Knowledge Network and Representation Using Schemata

Description of scenes, at various levels of detail, are captured in a set of schema hierarchies [3]. A schema graph is an organizational structure defining an expected collection of objects, such as a house scene, the expected visual attributes associated with the objects in the schema (each of which can have an associated schema), and the expected relations among them. For example, a house (in a house scene hierarchy) has roof and house wall as sub-parts, and the house wall has windows, shutters, and doors as sub-parts. The knowledge network shown in Figure 1 is a protion of a schema hierarchy as developed in [6]. Each schema node (e.g. house, house wall, and roof) has both a structural description appropriate to the level of detail and methods of access to a set of hypothesis and verification strategies called interpretation strategies. For example, the sky-object schema (associated with the outdoor-scene schema) has access to the exemplar selection and extension strategy discussed below.

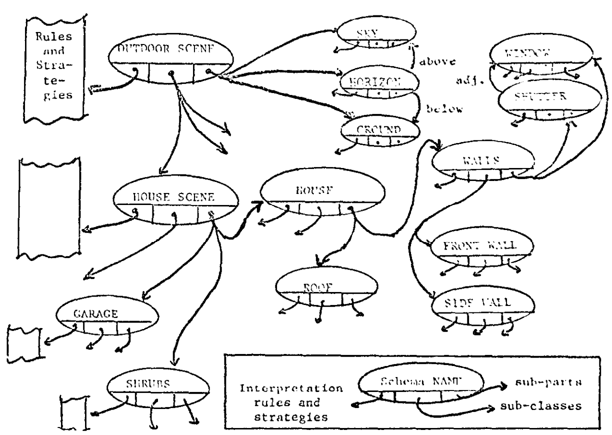

Figure 1. Abstract representation of a portion of the knowledge network. Schema are organized by the component descriptions of subclass and subpart, and by spatial relations. Each schema node has access to a set of interpretation rules which form hypotheses on the bases of image measurements, and interpretation strategies which describe how these hypotheses are combined with information in the network to form a consistent interpretation.

Interpretation rules relate image events to knowledge events by providing evidence for or against part/sub-part hypotheses. An interpretation strategy, associated with a schema node, specifies in procedural form how specific interpretation rules may be applied, and how combined results from multiple rules may be used to decide whether or not to "accept" (i.e., instantiate) an object hypothesis. An interpretation strategy thus represents both control local to the node and top-down control over the instantiation process.

Note that the goal is not to have these interpretation rules and strategies extract exactly the correct set of regions. Our philosophy is to allow incorrect, but reasonable, hypotheses to be made and to bring to bear other knowledge (such as various similarity measures and spatial constraints) to filter the incorrect hypotheses. An example of such error detection and correction in the interpretation process will be given in Section 5.

3. Rule Form for Object Hypotheses Under Uncertainty

We will illustrate the form of a simple interpretation rule based on using the expectation that grass is green. The feature used is average "excess green" for the region, obtained by computing the mean of 2G-R-B for all pixels in this region. Histograms of this feature are shown in Figure 2, comparing all regions to all known grass regions across 8 samples of color outdoor scenes. An abstract version is shown in Figure 3. The basic idea is to form a mapping from a measured value of the feature obtained from an image region, say f_I, into a "vote" for the object on the basis of this single feature. One approach to defining this mapping is based on the notion of prototype vectors and the distance from a given measurement to the prototype, a well-known pattern classification technique which extends to N-dimensional feature space [3]. In our case rather than using this distance to "classify" objects in a pure Bayesian approach that is replete with difficulties, we translate it into a "vote".

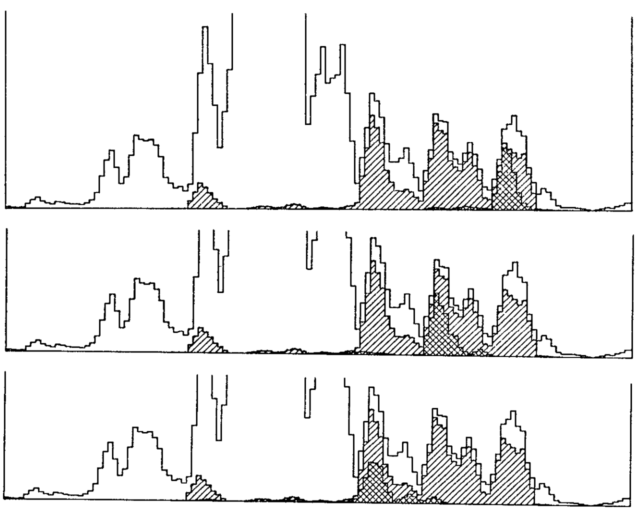

Figure 2. Image histograms of an "excess green" feature (2G-R-B) computed across eight sample images. The unshaded histogram represents the global distribution of the feature. The darkest cross hatched histogram is the distribution of this feature across regions known to be grass (from a hand labeling of the images) in one of three specific images. The intermediate cross hatching represents all known grass regions across the entire sample. Note the shifting (with respect to the full histogram) of the histograms for the individual images.

Let $d(f_P, f_I)$ be the distance between the prototype feature point f_P and the measured feature value f_I. The response R of the rule is then

$$P(f_I) = \begin{cases} 1 & \text{if} \quad d(f_P,f_I) \le \Theta_1 \\ \dfrac{\Theta_2 - d(f_P,f_I)}{\Theta_2 - \Theta_1} & \text{if } \Theta_1 < d(f_P,f_I) \le \Theta_2 \\ 0 & \text{if } \Theta_2 < d(f_P,f_I) \le \Theta_3 \\ -\infty & \text{if } \Theta_3 < d(f_P,f_I) \end{cases}$$

The thresholds Θ_1, Θ_2, and Θ_3 represent a gross mapping from the feature space to a score value that provides an interpretation of the distance measurements. Θ_3 allows strong negative votes if the measured feature value implies that the hypothesized object cannot be correct. For example, fairly negative values of the excess green feature imply a color which should veto the grass label. Thus, certain measurements can exclude object labels; this proves to be a very effective mechanism for filtering many spurious weak responses. Of course there is the danger of excluding the proper label due to a single feature value, even in the face of strong support from many other features. In the actual implementation of this rule form, Θ_1, Θ_2, and Θ_3 are replaced with six values so that non-symmetric rules may be defined as shown in Figure 4. There are many possibilities for combining the individual feature responses into a score; here we have used a simple weighted average.

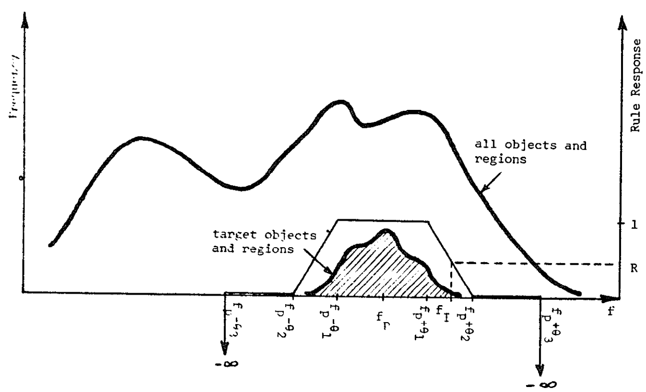

Figure 3. Structure of a simple rule for mapping an image feature measurement f_I into support for a label hypothesis on the basis of a prototype feature value obtained from the combined histograms of labeled regions across image samples. The object specific mapping is parameterized by four values, f_P, Θ_1, Θ_2, Θ_3, and stored in the knowledge network. The use of six values will allow an asymmetric response function.

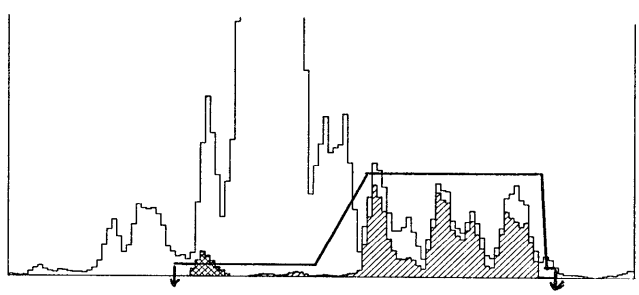

Figure 4. An example grass rule, showing an asymmetrical structure, superimposed on the histogram of Figure 2.

4. Exemplars and Islands of Reliability

The extreme variations that occur across images can be compensated for somewhat by utilizing an adaptive strategy. Variation in the appearance of objects (region feature measures across images) is much greater than object variations within an image (see Figure 2).

In the initial stages, there are few if any image hypotheses, and development of a partial interpretation must rely primarily on general knowledge of expected object characteristics in the image and not on the relationship to other hypotheses. The most reliable object hypotheses, derived from the interpretation rules, can be considered object "exemplars" and form basis of adaptation. One strategy extends the kernel interpretation by using the features of labelled exemplar regions including color, texture, shape, size, image location, and relative location to other objects. This is similar to the method in [4], where "characteristic regions" were used to guide hypothesis formation in the early stages of interpretation. The exemplar region (or set of regions) forms an image-specific prototype which can be used with a similarity measure to select and label other regions of the same identity. A verification phase can be applied where relations between object hypotheses are examined for consistency. Thus, the interpretation is extended through matching and processing of region characteristics as well as semantic inference.

Exemplars can be used more generally than we have presented them here. For example a house wall showing through foliage can be matched to the unoccluded visible portion based upon color similarity and spatial constraints derived from inferences of house wall geometry. The shape and/or size of a region can be used to detect other instances of multiple objects, as in the case of finding one shutter or window of a house, one tire of a car, or one car on a road. Additional spatial and perspective constraints can also be employed in object recognition.

Exemplar hypothesis rules differ from general hypothesis rules in that they are more conservative; they should minimize the number of false hypotheses at the risk of missing true target regions by narrowing their range of acceptable responses. If all regions are vetoed, secondary strategies are invoked; for example, the veto ranges can be relaxed, admitting less reliable exemplars. Figure 5 compares the results of the grass exemplar rule with the general grass hypothesis rule. The strategy can also be used to generate lists of hypotheses ordered by reliability.

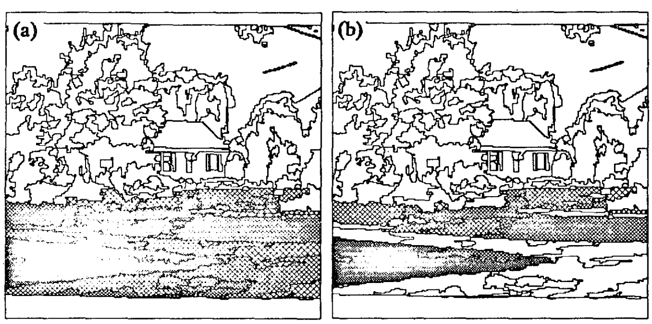

Figure 5. The exemplar hypothesis rule is more selective than the corresponding general interpretation rule (based on a less selective rule form). Figure 5a shows the general grass interpretation rule, while Figure 5b shows the exemplar rule. Note that the general form of the rule results in more incorrect region hypothesis (which could be filtered by constraints from the knowledge network). Although the examplar rule misses some grass regions, those found have high confidence.

The advantages of using object exemplars include:

1) an effective means for extending **reliable hypotheses** to regions which are more ambiguous; this is similar to the notion of **"islands of reliability"** [2];

2) a **knowledge-directed** technique for partially dealing with the unavoidable **region fragmentation** that occurs with any segmentation algorithm or low-level image transformation/grouping; regions that are "similar" to the exemplar can be both labelled and merged;

3) exemplars play a natural role in the implementation of an **hypothesize-and-verify control** strategy; hypotheses are formed based upon initial feature information and subsequently can be used in a verification process where the relationship between labelled regions provides consistency checks on the hypotheses and the evolving interpretation.

5. Results of Rule Based Image Interpretation

Experiments are being conducted on a set of eight "house scene" images. Thus far, we have been able to extract sky, grass, and foliage (trees and bushes) from five house images with reasonable effectiveness, and have been successful in identifying houses and their parts, including shutters (or windows), house wall and roof in three of these images. The interpretation strategies use many redundant features, each of which can very often be expected to be present. The premise is that many redundant features allow any single feature to be unreliable. Object hypothesis rules were employed as described in previous sections, while object verification rules requiring consistent relationships with other object labels are under current development. The final results shown in Figure 6 are an interpretation based on coarse segmentations. Further work on segmentation is being carried out, as is the refinement of the exemplar selection and matching rules (that were shown in section 3).

An extremely important capability for an interpretation system is feedback to lower level processes for a variety of purposes. The interpretation processes should have focus-of-attention mechanisms for correction of segmentation

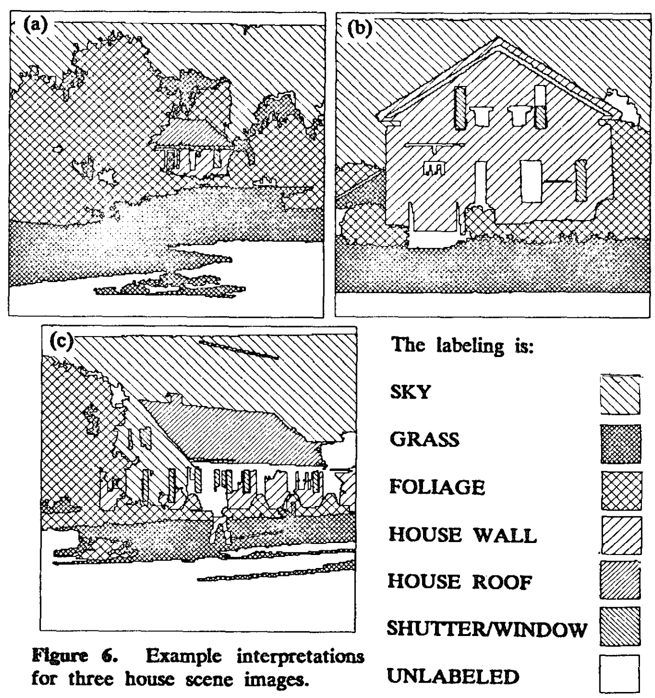

Figure 6. Example interpretations for three house scene images.

errors, extraction of finer image detail, and verification of semantic hypotheses. An example of the effectiveness of semantically directed feedback to segmentation processes is shown in Figure 7. There is a key missing boundary between the house wall and sky which leads to incorrect object hypotheses based upon local interpretation strategies. The region is hypothesized to be sky by the sky strategy, while application of the house wall strategy (using the roof and shutters as spatial constraints on the location of house wall) leads to a wall hypothesis.

There is evidence available that some form of error has occurred in this example: 1) conflicting labels are produced for the same region by local interpretation strategies; 2) the house wall label is associated with regions above the roof (note that while there are houses with a wall above a lower roof, the geometric consistency of the object shape is not satisfied in this example); and 3) the sky extends down close to the approximate horizon line in only a portion of the image (which is possible, but worthy of closer inspection).

In this case resegmentation of the sky-housewall region, with segmentation parameters set to extract finer detail, produces the results shown in Figure 7a. Subsequent remerging of similar regions produces a usable segmentation of this region as shown in 7b. It should be pointed out that in this image there is a discernable boundary between the sky and house wall. Initially, the segmentation parameters may be set so that the initial segmentation misses this boundary. This may occur because of computational requirements (fast, coarse segmentations) or as an explicit control strategy. However, once it is resegmented with an intent of overfragmentation, this boundary can be detected. Remerging based on region means and variances of a set of features allows much of the overfragmentation to be removed. Now, the same interpretation strategy used eariler produces quite acceptable results shown in Figure 8.

The current development of interpretation strategies involves the utilization of stored knowledge and a partial model (labelled regions) for hypothesis extension. In these strategies the knowledge network is examined for objects that can be inferred from identified objects, and for relations that would differentiate them. For example, the bush regions can be differentiated from other foliage based on their spatial relations to the house, and front and side house walls can be differentiated using geometric knowledge of house structure (e.g., relations between roof and walls), as shown in Figure 9. In the full system, these rules would not work in isolation as shown here, and the errors made by this type of rule would be filtered by other constraints.

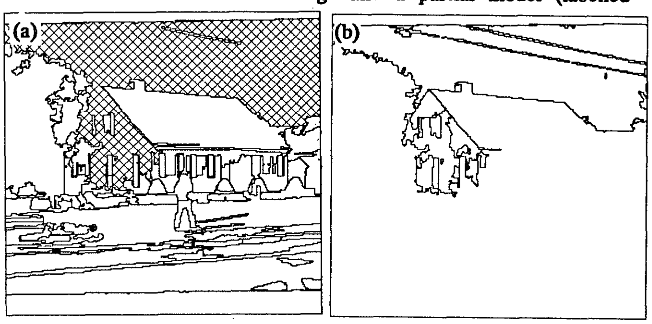

Figure 7. Resegmentation of house/sky region from Figure 6c. Figure 7a is the original segmentation showing the region to be resegmentated; 7b shows the regions resulting from the selective application to the segmentation process to the cross-hatched area in 7a.

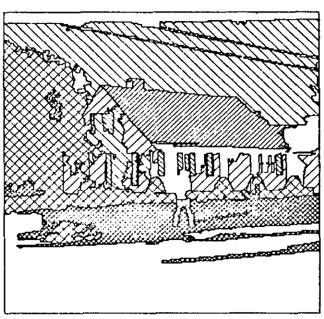

Figure 8. Final interpretation of the house scene in Figure 6c, after inserting resegmented houes/sky regions and reinterpreting the image.

Future work is directed towards refinement of the segmentation algorithms, object hypothesis rules, object verification rules, and interpretation strategies. System development is aimed towards more robust methods of control: automatic schema and strategy selection, interpretation of images under more than one general class of schemata, and automatic focus of attention mechanisms and error-correcting strategies for resolving interpretation errors.

References

[1] Binford, T., "Survey of Model Based Image Analysis Systems," *International Journal of Robotics Research*, Volume 1, Number 1, Spring 1982, pp. 18-64.

[2] Erman, L., Hayes-Roth, F., Lesser, V. and Reddy, D., "The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty," *Computing Surveys*, 12(2), June 1980, pp. 213-253.

[3] Hanson, A. and Riseman, E., "VISIONS: A Computer System for Interpreting Scenes," in *Computer Vision Systems* (A. Hanson and E. Riseman, eds.), Academic Press, 1978, pp. 303-333.

[4] Nagao, M. and Matsuyama, T., *A Structural Analysis of Complex Aerial Photographs*, Plenum Press, New York, 1980.

[5] Ohta, Y., "A Region-Oriented Image-Analysis System by Computer," Ph.D. Thesis, Kyoto University, Department of Information Science, Kyoto, Japan, 1980.

[6] Parma, C., Hanson, A. and Riseman, E., "Experiments in Schema-Driven Interpretation of a Natural Scene," COINS Technical Report number 80-10, University of Massachusetts, 1980. Also in *Nato Advanced Study Institute on Digital Image Processing* (R. Haralick and J.C. Simon, eds.), Bonas, France, 1980.

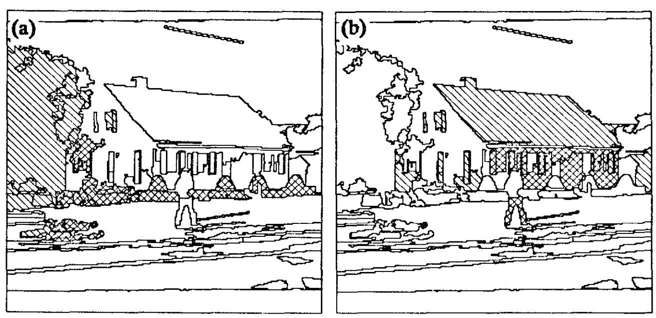

Figure 9. An example of the use of spatial relations to filter and extend region labeling. The geometric relations between house and shrub (in 9a) and between between roof and house front wall (in 9b) are used to refine region hypotheses from the interpretation shown in Figure 6c. Note that there are still ambiguities (the shrub label in the grass area, and the pants labeled as house wall) that require the use of other filters.