Crossroads
The ACM Student Magazine

Seeing is Believing: Computer Vision and Artificial Intelligence

By **Christopher O. Jaynes**

One of the monolithic goals of computer vision is to automatically interpret general digital images of arbitrary scenes. This goal has produced a vast array of research over the last 35 years, yet a solution to this general problem still remains out of reach. A reason for this is that the problem of visual perception is typically under-constrained. Information like absolute scale and depth is lost when the scene is projected onto an image plane. In fact, there are an infinite number of scenes that can produce the exact same image, which makes direct computation of scene geometry from a single image impossible. The difficulty of this ``traditional goal'' of computer vision has caused the field to focus on smaller, more constrained pieces of the problem. The hope is that when the pieces are put back together, a successful scene interpreter will have been created.

Digital filtering, **motion analysis** , **image registration**, **segmentation**, and model matching schemes are all examples of areas where progress has been made in the field. Other research has focused on the general problem through the use of knowledge and context. The use of external knowledge both about the world and about the current visual task reduces the number of plausible scene interpretations and may make the problem solvable. This approach is referred to as **knowledge-based vision**. Work in the area of knowledge-based vision incorporates methods from the field of AI in order to focus on the influence of context on scene understanding, the role of high level knowledge, and appropriate knowledge representations for visual tasks.

The importance of computer vision to the field of AI is fairly obvious: **intelligent agents** need to acquire knowledge of the world through a set of sensors. What is not so obvious is the importance that AI has to the field of computer vision. Indeed, I believe that the study of perception and intelligence are necessarily intertwined. This article will look at the role that knowledge plays in computer vision and how the use of reasoning, context, and knowledge in visual tasks reduces the complexity of the general problem.

The importance of context and knowledge in vision has been pointed out by psychologists many times, and these ideas have driven computer vision studies as well. Take the example in Figure 1.

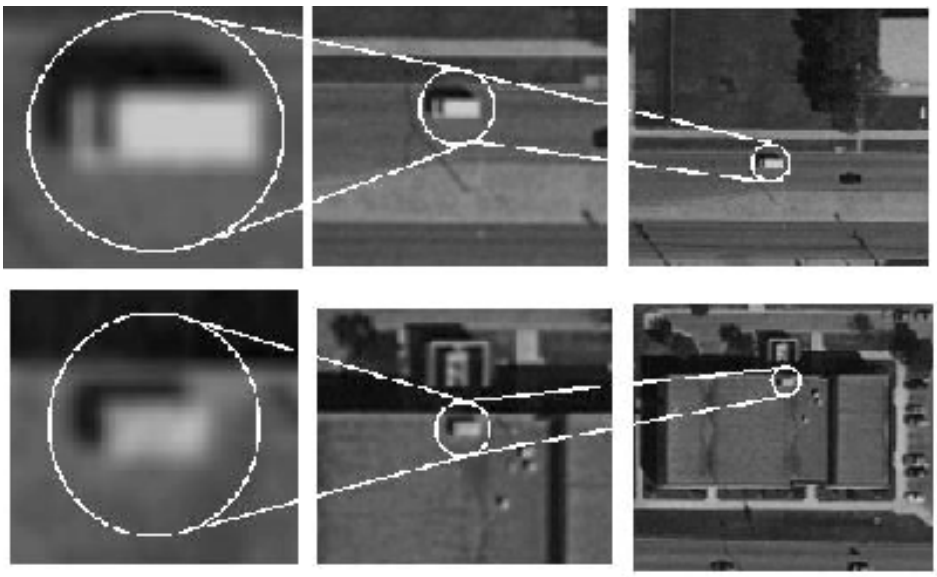

Figure 1

On the far left, the two image patches seem almost indistinguishable. Indeed, one can imagine a classification strategy based on the grey levels in the image to classify the regions as the same object. It is only in the context of the surrounding image and our knowledge about roads, cars, building structures, and how they relate to each other, that the two can be found distinct as a car on a road and a roof vent on top of a building.

Knowledge and context can influence visual interpretations of the scene in two ways. First, knowledge assists in selecting the appropriate image understanding algorithm to apply to the scene. Secondly, the result of applying the algorithm should be interpreted using the knowledge at hand. That is, if a rooftop has been discovered by the interpretation system, then a roof vent detector may be a reasonable algorithm to apply. Furthermore, if a roof vent is discovered without the corresponding roof, the result of the algorithm should be held in doubt.

Examples like the one above have encouraged researchers to explore knowledge-based vision extensively. Early in the 1970s, the University of Massachusetts set the stage for knowledge-based image understanding with the VISIONS system, a computer vision system for interpreting New England road scenes. The system used 2D color images to produce a labeled, segmented image of the scene. Labels were semantic like ``dirt road'' and were based on a knowledge base relevant to the domain. It was an ambitious attempt to integrate algorithms that had been developed without regard to context into a system that could apply them in an intelligent and context sensitive way. Because algorithms had been developed for highly specific visual tasks, from tree detectors to line extraction routines, for example, the system used a distributed blackboard architecture to invoke the algorithms according to current scene knowledge and context. Although the system was a success, it still did not solve the general scene interpretation problem. Evidence combination and reasoning were based on heuristic evidence and not the more recent AI work in belief maintenance. The system was also restricted to using single 2D color images rather than data from several

differing views of the same scene, which has proven to be useful.

Knowledge for Aerial Image Interpretation

Recent research has looked at methods to improve **image understanding** (IU) strategies with the use of context [**5**], knowledge about the world, and through machine learning methods [**2**]. I have been concerned with the knowledge representations that will allow scene understanding in a general sense. Given a set of images of a single scene, a system should produce as an interpretation the best possible model. By model, I mean not only the 3D geometric structure of the scene but also semantic labels of features in the scene and information relevant to the scene interpretation. For example, a New England tobacco farm may be imaged and reconstructed geometrically. The model should also contain information that the main building is a tobacco barn and that the depression labeled ``drainage ditch'' is currently empty. That is, the model should be as extensive as possible given the set of images and knowledge.

I have been applying my work to the aerial image domain simply because there are a vast number of aerial sensors that are able to produce data relevant to my research. I also work with urban areas, built-up regions with human-made structures often referred to as ``cultural sites'' rather than empty landscapes. Each site may contain many buildings, roads, rivers, parking lots, and other features. Independently, the vision community has been successful in constructing IU algorithms that perform well under certain conditions and for particular parts of a site. For example, building detection has been studied extensively [**6**, **7**, **8**, **9**], and I have recently used perceptual organization techniques to find rooftop polygons from aerial images [**3**]. However, a system that is able to integrate these results to produce a consistent model from arbitrary sets of data has not been created.

I believe that explicit representation of knowledge and context will allow a more purposeful vision system to be constructed by using the knowledge for intelligent control. In order for this to be possible, the knowledge representation should be capable of representing a wide range of information types. These include:

- **World Knowledge:**
  Physical knowledge independent of the sensors and specific site,
- **Site Knowledge:**
  Current set of beliefs about the site,
  Reconstructed geometry,
  Functional areas,
- **Contextual Information:**
  Acquisition conditions for sensor data,
  Available computational resources,
  IU algorithm requirements, and
- **Control Information:**
  Given knowledge and Context invoke appropriate IU algorithm.

Bayesian Networks

There are at least three different mathematical theories of evidence that can help provide reasoning and control to an IU system: ***probability theory***; *Dempster-Shafer's theory of evidence;* and the *theory of* ***fuzzy logic***. An overview of the three different approaches can be found in [**1**]. I base my reasoning mechanisms on probability theory and use a **Bayesian network** representation to allow inference and site model reconstruction. Recent advances in Bayesian networks allow sophisticated reasoning over large networks (over 10,000 nodes) [**4**] and seem feasible for my domain. Nodes in a typical Bayesian network represent degrees of belief about a particular aspect of the site, and edges in the graph represent causal dependencies. Degrees of belief are computed directly from prior probabilities using Bayesian inference.

For example, a simple network representing knowledge about clothing and the weather is shown in Figure 2. Beliefs about weather conditions can be inferred from other factors such as whether people are carrying umbrellas and wearing raincoats. Therefore, our network contains edges from the nodes `umbrella` and `clothing type` to `weather`. The network also defines how particular beliefs are propagated through the network through the use of probability tables. A single Bayesian probability is shown for the case in which people are wearing raincoats and carrying umbrellas. The network actually stores a table of probabilities for all possible combinations for the evidence nodes. When one probability value changes, the change is propagated through the network using the probability table at each node. The Bayesian probability shown states that if I have observed that people are wearing raincoats and carrying umbrellas, then I am able to conclude with almost absolute certainty that it is raining.

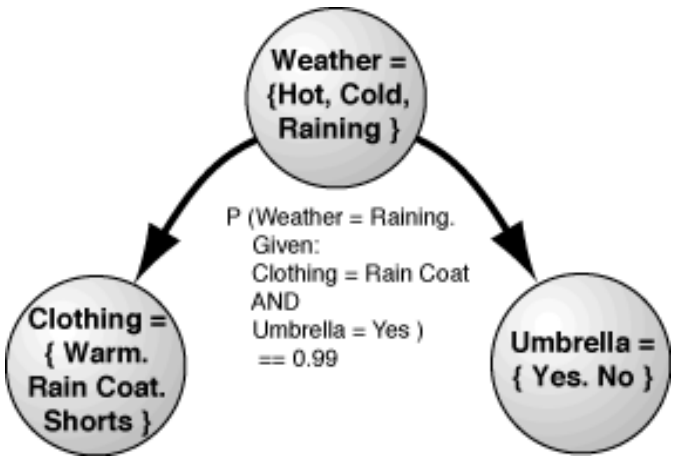

Figure 2

Bayesian networks allow the above kind of **diagnostic inference** as well as inference in the other direction called **causal inference**. That is, if it is raining, I can conclude that people will be wearing raincoats and carrying umbrellas.

When applied to the domain of aerial image interpretation, a Bayesian network can provide us with an implicit control strategy. When searching for the validity of a particular hypotheses (``a building is present'', for example), the edges in the Bayesian network should be followed in order to gather more evidence for the conclusion. When we arrive at a node for which no evidence exists, we should invoke an appropriate IU strategy to gather evidence directly from the data. Each node contains not only a belief and the probability tables, but also information about how beliefs can be computed directly from the data using a set of existing IU algorithms. An example network for detecting and reconstructing a building is shown in Figure 3.

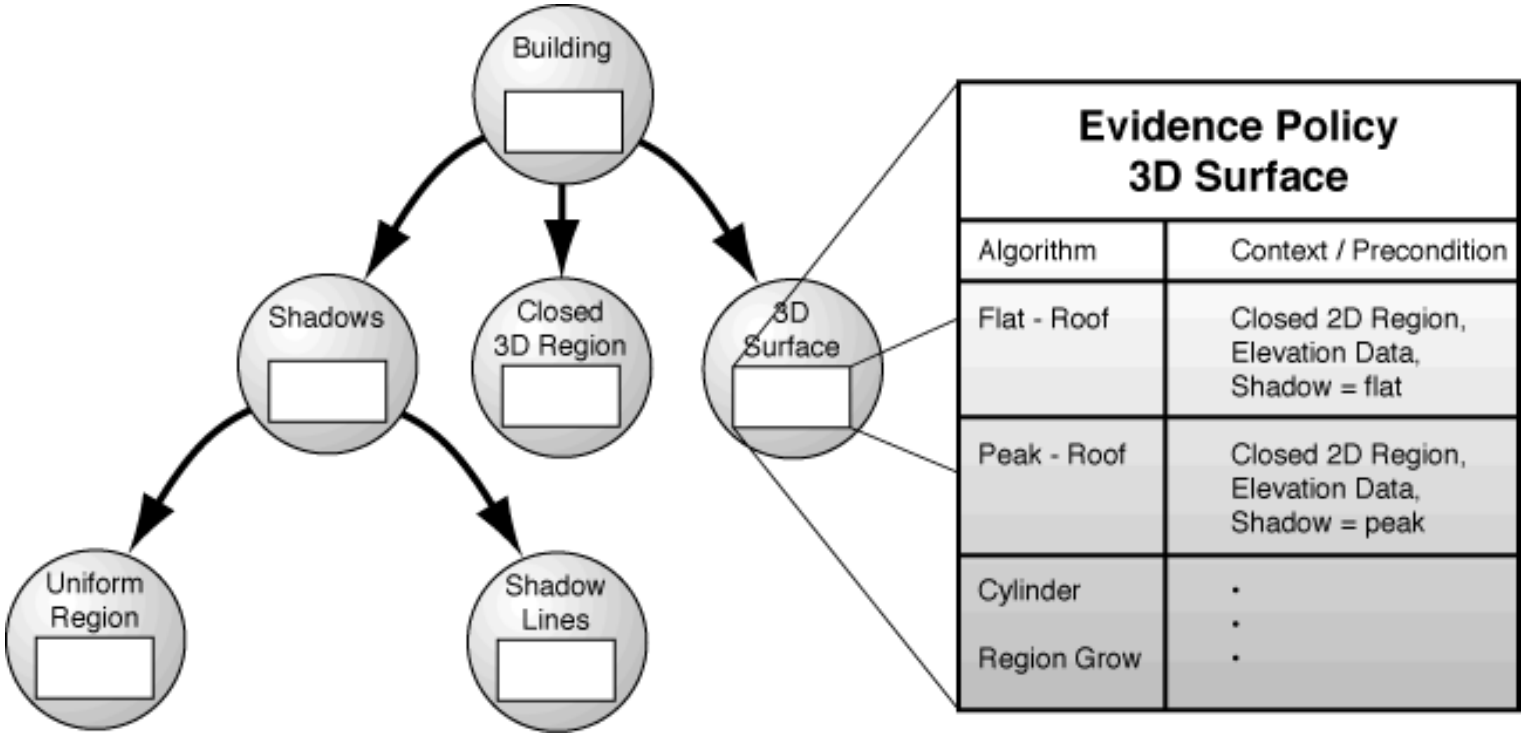

Figure 3

Building confidence depends on the system's ability to find shadows cast by the building in optical images, a closed region denoting the building rooftop, and a good surface fit in corresponding 3D elevation data. The set of algorithms for gathering evidence for the node `3D Surface` is shown. The choice of the particular algorithm depends on available data, computational resources needed, and the value of other nodes in the network.

Example Knowledge-Based Interpretation

Suppose that we have several views of a local site, and a corresponding digital elevation map that may have been computed from radar. The system has various types of IU algorithms at its disposal, and there are many different features that can be grouped into 3D models. Therefore, it is

important that the reconstruction proceeds in a purposeful manner. We will use the network from **Figure 3** to drive the selection of algorithms that will help to accumulate evidence for a particular local scene interpretation, shown in Figure 4.

Because the presence of a building depends upon evidence of a rooftop, the rooftop detector is invoked. Lines are extracted from the image and grouped together based on a perceptual compatibility measure. A 2D polygon detector is used to compute the position and shape of possible building rooftop boundaries. The region circled and labeled ``A'' has been detected as a possible rooftop. The network then, propagates the belief that a building exists, based on the fact that a closed polygon is present. Because evidence from other sources has not been found, the belief at node `Building` remains low. Further evidence lays in the fact that buildings cast shadows, and that they have regular height in the elevation data.

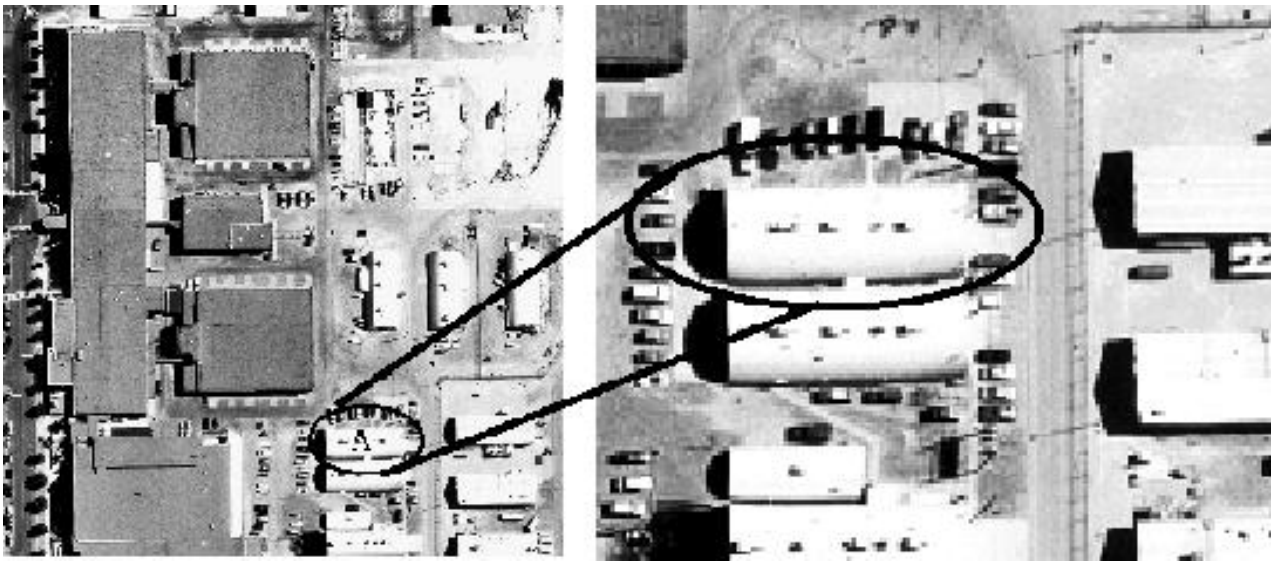

Figure 4

Because the actual surface shape of the building remains in question, two things can occur. Either all possible surface fitting algorithms can be applied to the closed polygon region and the best fit is selected, or one can be selected on the basis of image context. In some cases, the elevation data may be so noisy that several surfaces will fit equally well. In this example, the shadow is clearly visible and can be analyzed. The system determines that the surface is curved by using sun angle information and estimating the shape of the shadow in the ground. Given this information, an appropriate surface is fit to the data. Figure 5 shows the successful reconstruction of the building.

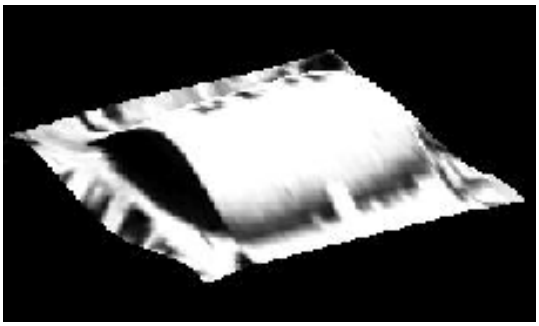

Figure 5

Conclusion

I have presented one way that knowledge can be used to direct and improve the application of lower level IU tasks. As knowledge is gained about a particular site, processing can become more focused. Changes in the data are easily detected as inconsistencies in the system's current set of beliefs. Of course, there are still many issues to be resolved. How much confidence we should have that a particular IU task has given a correct result is still an issue of research. How does the confidence in a particular algorithm effect beliefs that are inferred within the network? This approach, however, is extensible to other domains and other perceptual tasks. It provides a means to reason about a percept and to control perception in an intelligent way. I am confident that as more is understood about the role of context and knowledge in perception, a general scene understanding system will be realizable.

References

1

Bloch, Isabelle. 1995. Information Combination Operators for Data Fusion: A Comparative Review with Classification. *IEEE Transactions on Systems, Man, and Cybernetics*.

2

Draper, Bruce. 1993. Learning Object Recognition Strategies. PhD Dissertation. CMPSCI TR93-50. University of Massachusetts at Amherst.

3

Jaynes, Christopher, Frank Stolle, Howard Schultz, Robert Collins, Allen Hanson, and Edward Riseman. 1996. Three-Dimensional Grouping and Information Fusion for Site Modeling from Aerial Images. In *Proceedings, DARPA Image Understanding Workshop*, Palm Springs, CA.

4

Pearl, J. 1988. ***Probabilistic Reasoning in Intelligent Systems: networks of plausible inference***, Morgan-Kaufmann Publishers.

5

Strat, Tomas and Martin Fischler. 1995. The Role of Context in Computer Vision. In *Proceedings of the International Conference on Computer Vision*, Workshop on Context-Based Vision (June 19), Boston.

6

Venkateswar V. and R. Chellappa. 1991. A Hierarchical Approach to Detection of Buildings in

Aerial Images. Technical Report CS-TR-2720. Center for Automation Research, University of Maryland. August.

7

Lin, Chungan and Ramakant Nevatia. 1996. Buildings Detection and Description from Monocular Aerial Images. In *Proceedings, DARPA Image Understanding Workshop*, Palm Springs, CA.

8

McKeown, David, G. Edward Bulwinkle, and Steven Cochran. 1996. Research in the Automatex Analysis of Remotely Sensed Imagery. In *Proceedings, DARPA Image Understanding Workshop*, Palm Springs, CA.

9

Collins, Robert, Allen Hanson, Edward Riseman, Christopher Jaynes, Frank Stolle, Xiaoguang Wang, and Yong-Qing Cheng. 1996. UMass Progress in 3D Building Model Acquisition. In *Proceedings, DARPA Image Understanding Workshop*, Palm Springs, CA.