Alvey MMI-007 Vehicle Exemplar:

The Knowledge Based Approach

K. D. Baker and G. D. Sullivan

Department of Computer Science
University of Reading, RG6 2AX.

Introduction

Since the seminal work of Roberts (1965), scene-understanding research has continued to be dominated by two main themes: the measurement of local attributes of the image to identify features which characterise objects in the scene, and the use of prior expectations, often explicitly expressed as models, to guide the interrogation of the image. The two themes correspond to the distinction between data-driven methods, and goal-directed methods encountered in many studies in AI. Data-driven methods may be used to express low-level knowledge about image properties, but the goals of vision involve knowledge of objects and their relationships in the scene which demands high-level representations with little direct correspondence in the image data. The central task confronting practical vision systems is to resolve the distinction, and interpret the image data in terms of object abstractions.

The major objective of the MMI 007 Alvey Consortium is the integration of these two approaches and the evolution of a generalised vision system architecture. The programme of work involves a wide range of topics including: the description and representation of 2D features and their relationships; the representation and use of 3D object knowledge; the control of reasoning with multiple sources of information; and the exploration of computing architectures appropriate for a general vision system.

The research involves several groups of collaborating academic and industrial partners working at geographically dispersed sites. To coordinate the efforts of these groups we have opted to study a series of Exemplars. An Exemplar is a demonstration of competence in a prescribed domain to illustrate the feasibility of the underlying concepts. The following papers by Godden, Fullwood and Hyde, Morton, Hutber and Sims, Brisdon, and Sullivan, discuss the work of the Consortium in the realisation of the first such Exemplar: the recognition of a vehicle in a general daylight scene.

This introductory paper presents the philosophy of the exemplar method in the context of the main problems facing the Consortium. The major issues are identified and discussed in terms of the fundamental hypothesis generation and verification strategy we have adopted. In this approach the results of the data-driven descriptive processes are combined with the predictive measures derived from a knowledge of the 3D structure of objects and their context to evaluate quantitatively the existence of the object in the scene.

Understanding 2D Images

Images of scenes contain a wealth of information from many sources including colour, texture, shading, motion etc which, together with geometrical and contextual information and the knowledge of objects and their relationships, determines our understanding of the scene. The representation of this knowledge and the control of the processes that reason over it constitute the architecture of a vision system. It is one of the longer term objectives of this Consortium to investigate architectures which support this form of reasoning. Before this problem can be addressed however, there are a number of more fundamental issues that have to be investigated and these are the subject of current work.

The immediate objective of the Consortium is to seek an explanation of 2D images in terms of prior known characteristics of commonly occurring objects and the contexts in which they are most likely to occur. 2D images of natural scenes contain large areas, such as trees, sky, grass etc, that are not easily defined in terms of geometrical constructs. Instead, identification of these regions may be possible from segmentation methods based on statistical properties such as colour, or region boundaries or texture characteristics, as discussed by Ohta (1985). On the other hand, prior knowledge of objects often includes precise 3D geometrical shape. Our task therefore is to find ways to reconcile the geometrical object knowledge with the often imprecise segmentation analyses to effect the identification of objects in the 2D image data.

AVC 1987 doi:10.5244/C.1.1

The Hypothesis Generation and Verification Cycle

Understanding often follows a cycle of inference (Neisser, 1967). In vision, a description of the image leads to an hypothesis about the existence of the objects in question. Confirmatory evidence is then sought using prior knowledge of the object structure to generate a likely instance of the object which can be matched against the image data. A quantitative evaluation of the match between instance and image confirms or refutes the hypothesis.

In practice we do not expect a vision system to operate on a single hypothesis generation and evaluation cycle. Rather we expect the reasoning to progress on the basis of generating many subsidiary hypotheses and evaluating them using evidence drawn from the different sources of information and using the knowledge of the structure of objects. The principle of the operation of the inference cycles can however be expected to follow a standard pattern, and part of our purpose in demonstrating the principle in a limited context is to expose the problems faced in more complex tasks, where the final hypothesis may depend logically on a set of intermediate hypotheses.

Automating the recognition of objects and their context in natural scenes will therefore require a multitude of intermediate hypotheses to be generated as the 'understanding' of the scene evolves. The reasoning process can be considered as one that seeks to increase the level of confidence in the current interpretation of the scene by progressively increasing the number of recognised objects and their interpreted interrelationships.

The form of a current hypothesis will be dependent on those hypotheses that have previously been verified. A refuted hypothesis may cause backtracking and an alternative to be generated. Because we are dealing with a multitude of information sources, we must expect quasi-independent lines of reasoning to be carried forward in parallel (Barrow and Tenenbaum, 1981). The system architecture to support multiple reasoning processes must therefore also provide for truth maintenance between the components parts (Doyle, 1982, DeKleer, 1986). In the final system we envisage a set of tightly coupled processes working asynchronously with different representations associated with each source of sensory information as discussed by Hanson and Riseman (1978).

Initial Hypothesis Generation. A first-stage hypothesis can be inferred from the results of low-level segmentation algorithms together with the contextual reasoning processes that label the segmented areas (Godden, Fullwood and Hyde 1987, Morton, 1987). Further support for the presence of an object is obtained from a statistical analysis of groups of regions (Hutber & Sims, 1987). Both methods provide only imprecise evidence for the existence of an object but do strongly constrain the search space of the more computationally expensive verification algorithms. Their output is a set of bounded regions in the 2D image data, each of which is thought to contain an object of interest, together with an initial view-hypothesis.

Hypothesis Verification Process. Verification of an hypothesis involves quantitative evaluation of the projection of the object in the 2D image data. The two major factors affecting the difficulty of this task are the number of different objects that could occur in the scene and the number of different possible viewpoints of each object. Together these present a vast search space and for the problem to be manageable we must find ways to constrain the search and limit the number of possible interpretations of the data.

In the approach adopted here the scene hypothesis provides a significant constraint on the area of the 2D image that has to be considered. We can further limit the search by considering the occurrence of only a single object. Clearly, in a general scene understanding system this restriction must be lifted but during this stage of the investigation it is valid.

Limiting our search to a single object and constraining the position of its occurrence in the image to one or more defined regions assists in reducing the total search space. However the exact location and orientation of the object with respect to the viewer and the viewpoint has yet to be established. Establishing the correspondence between 2D image features and 3D object components is the major problem in model driven vision systems. The difficulties experienced in the solution of this non-linear spatial correspondence problem has impeded the application of model-based methods for industrial vision systems.

Spatial Correspondence and Perceptual Organisation

Psychological evidence suggests that the human visual system relies heavily on perceptual organisation when interpreting a scene (e.g. Kubovy and Pomerantz, 1981). Groups of image features, defined by end point proximity, co-linearity, or parallelism, have been used as clues in automatic recognition processes (Lowe, 1985, 1987). The discovery of a specific feature group can trigger a search for related features according to the structure contained in the three dimensional object model. In a complex scene a partially recognised object then helps to control the interpretation of adjacent structures. The search thus becomes progressively more constrained as each additional piece of evidence contributes to the emerging scene description.

Perceptual organisation has played an important role in the model driven approach to image understanding developed by Lowe. The existence of feature groups which have very low probability of accidental occurrence in the 2D image allow image features to be related to a small subset of the 3D object features. The viewpoint parameters are derived by applying the viewpoint consistency constraint which require that all image feature be projections from the model for a single viewpoint. Back projection from the 2D image space to the 3D model space is a non-linear problem and is solved by an iterative method.

The images Lowe has discussed contain multiple instances of well defined groups of edge segments forming trihedral vertices and parallelogrammes. The images of our Exemplar do not lend themselves well to the use of such simple groups. Our objects are formed from curved surfaces, and the position of edge data therefore becomes somewhat viewpoint dependent. In effect then, although the concept of perceptual groups is important, we cannot rely upon the adequacy of a single group in our work. It is more likely that we will be forced to consider object-specific feature groups. Furthermore, it may be advantageous to go beyond groups based on edge segments and include surface features.

Framework for Integration

The overall objective of our Exemplar task requires successful integration of investigations carried out at various sites in the Consortium. The control of a geographically distributed research programme is a complex task and for it to be successful the work must progress within an agreed framework. In the first instance it was necessary for individual groups to work independently to accumulate a baseline of expertise. The results of these quasi-independent activities are now being brought together to demonstrate a competence for progressively more difficult visual tasks.

The integration of the separate investigations into a working vision system can be approached in well defined experimental stages, each demonstrating additional capabilities. For this reason the progress of the research has been geared to a series of milestones that provide for demonstrations of competence at increasingly complex visual tasks. As the complexity of the scene increases a greater number of constraints must be utilised to achieve recognition of the objects. The consequence of this however is that greater emphasis falls on the need to coordinate and control the use of the information available to the reasoning processes. To reduce the complexity of the search it is usually necessary to increase the complexity of the control of the reasoning processes.

Each of our Exemplars has been chosen to span both increasing levels of complexity and several domains of applications. It is by means of an Exemplar that a uniform focus of attention can be achieved throughout the Consortium providing a common understanding of the problems to be addressed by the dispersed research groups. By defining an Exemplar it is possible for the individual groups to work almost independently with the loose interaction between the algorithms being achieved by passing the results between the various sites. At a later stage, when the architectural issued are investigated further, it will be necessary to achieve a tight coupling of the processes. At that time it will be necessary for all processes to be available on the same processor.

The Exemplar

The first Exemplar, which is the subject of the papers that follow, has been defined to exercise the feasibility of the hypothesis generation and verification strategy. We have set ourselves the task of recognising a vehicle, in particular a hatchback car (based on the Vauxhall Chevette), in a general daylight scene containing grass, trees, roadways, buildings etc. The initial hypothesis concerning the presence of the vehicle is generated from information derived from low level segmentation algorithms together with the results of contextual reasoning. Critical features are used to constrain the spatial correspondence problem and allow instances to be projected from the 3D object model onto the 2D image data. The verification of the hypothesis follows from an evaluation of an instance of the vehicle. The Exemplar therefore exercises and integrates the following set of activities:

i) the development of effective edge, line, colour and region segmentation algorithms applied to natural scenes with a commonly occurring non-trivial object.

ii) the provision of evidential reasoning mechanisms that work from imprecise information to establish plausible hypotheses of the presence of a vehicle in the 2D scene.

iii) the construction of an effective three dimensional geometrical model of a particular vehicle from actual measurements.

iv) the solution of the spatial correspondence problem which links the 2D data available in the image to the 3D object model and provides an estimate of the viewing parameters.

v) the evaluation of a hypothesis by matching a projected instance against the image data to produce verification or rejection of the presence of the vehicle.

vi) the critical appraisal of the results of these algorithms on a series of images to assess the limitation of the methods.

The detail of the algorithms supporting the implementation of this Exemplar will be the subject of the papers by Godden, Fullwood and Hyde, Morton, Hutber and Sims, Brisdon, and by Sullivan.

Assessment and Future Direction

The implementation of the first Exemplar, the recognition of a vehicle in a natural daylight scene, has provided the first experimental evidence of the viability of our hypothesis generation and verification strategy. It has demonstrated that the search space can be made manageable by contextual reasoning on the results of the low level image segmentation analysis. This allows the computationally expensive verification process to be directed at limited numbers of bounded regions of the image data. Using object-specific focus features it has been possible to solve the spatial correspondence problem and obtain a good approximation to the viewing parameter. The evaluation algorithm provides a response characteristic that allows an unambiguous confirmation of the existence of the particular vehicle in the given bounded region.

Although the Exemplar has provided the necessary mechanism to integrate the work from several sites, it is nevertheless only a limited demonstration of the principle of the cycle of inference applied to the recognition of objects in 2D images. A discussion of the performance and the limitations of the work is given in the paper by Sullivan. What has been established is that the method is viable and has provided a framework on which to build our understanding of knowledge based machine vision.

We are now in a position to refine the methods developed under the spur of the vehicle exemplar, and to build towards a more general treatment of the problem of recognising known 3-D objects in constrained scenes, and to develop a system architecture capable of supporting the many levels of reasoning required.

Acknowledgement

Part of the work reported here was carried out at the Department of Electrical and Electronic Engineering, Plymouth Polytechnic.

References

Barrow, H.G. and Tenenbaum, J. M. (1981) Computational Vision. Proc IEEE 69, 572-595.

Bolles, R.C. and Cain, R. A. (1982), Recognising and locating partially visible objects, the local-feature-focus method. Int. J. Robotics Res., 1, 57-82.

Brisdon, Kay, (1987), Evaluation and Verification of Model Instances, AVC-1987

Doyle, J. (1982) A Truth Maintenance System. Artif. Intell. 12, 231-272.

Hanson, A. R. and Riseman, E. M. (1978) *Computer Vision Systems.* New York: Academic Press.

DeKleer, A. J. (1986) An assumption-based TMS, Artificial Intelligence, 28

Hutber, D. and Sims, P.F. (1987), Use of Machine Learning to Generate Rules, AVC-1987

Godden, R.J. Fullwood, J.A. and Hyde, J. (1987), Image Segmentation and Attribute Generation, AVC-1987

Kubovy, M. and Pomerantz, J. R. (1981) *Perceptual Organisation* Lawrence Erlbaum Associates.

Lowe, D. G. (1985) *Perceptual Organisation and Visual Recognition*, Kluwer, Boston MA.

Lowe, D. G. (1987) Three-dimensional Object Recognition from single two-dimensional images. Artif. Intell. 31, 233-395.

Morton, S.K. (1987), Object Hypothesis by Evidential Reasoning, AVC-1987

Neisser, U. (1967) *Cognitive Psychology*, New York: Appleton-Century-Crofts.

Ohta, Y. (1985) *Knowledge-based Interpretation of Outdoor Natural Color Scenes.* Pitman.

Roberts, L. G. (1965) Machine perception of three-dimensional solids. In *Optical and elecro optical infromation processing.* ed. Tippett. 159-197, MIT Press.

Sullivan, G. D. (1987) Performance and Limitations, AVC-1987.