Alvey MMI-007 Vehicle Exemplar:
Performance & Limitations

G. D. Sullivan

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

Introduction.

One of the visual tasks adopted as an "exemplar" for techniques by the Alvey MMI-007 consortium has been the detection and identification of vehicles in natural scenes. We have adopted a knowledge-based approach to the task, which seeks to explain the 2-D image in terms of known context and known geometry of the 3-D object being identified. This approach differs markedly from the more familiar technique of computing a 3-D description of the scene using information derived from stereopsis, or motion, or shading, prior to matching against stored 3-D models.

Part of the motivation for the knowledge-based approach stems from the difficulty of deriving a satisfactory 3-D description of complex natural scenes; large areas of our images often consist of groups of objects such as trees, bushes and buildings which admit no simple 3-D description. A deeper motivation stems from the observation that human visual perception is able to understand single monocular images with little difficulty; it even performs well despite gross distortion of the shading (e.g. poor photocopies) or of line (e.g. satirical cartoon). Indeed when high-level perceptual interpretations are pitted against quantitative low-level 3-D information, the former often dominates. Perhaps the best example is given by the face-mould demonstration (Gregory, 1971), in which a concave casting of a human face is irresistibly seen as a normal convex face, despite stereo, motion and shading evidence to the contrary. The high level percept, based on faint shading details in the image, can even withstand the addition of explicit stereo evidence to the contrary (Georgeson, 1979). A weaker example is given by pseudoscopes (stereoscopes with the images reversed): although the viewer is aware of interocular conflict, familiar objects are still seen normally in front of the background, not as indentations implausibly positioned behind the background. Where possible the human visual system makes use of high-level knowledge in its interpretation of meaningful scenes. It is this ability which we would like to emulate in machine vision.

The Knowledge-based Approach.

The preceding papers in this series have outlined two different contributions towards the vehicle identification exemplar. The work of Godden, Fullwood and Hyde, of Hutber and Sims, and of Morton, is directed towards creating a coarse-grained description of the scene. The attributes of segmented regions of the image are analysed to classify major areas in the image and to provide initial clues to the presence of vehicles. Candidate regions are thereby identified which merit detailed inspection. This type of analysis uses knowledge of the viewing circumstances, such as the camera orientation and the type of scene encountered, and of statistical consistencies between the features of different images of a given type of object. The knowledge used is imprecise and the conclusions drawn are uncertain, but these methods have the important advantage of being fairly robust against changes of viewpoint, and partially independent of the particular objects viewed.

On the other hand, the work reported by Brisdon provides a means for evaluating and refining exact hypotheses about a particular view of a particular vehicle. Detailed knowledge of the 3-D geometry of the vehicle, expressed as an explicit model, is used in making the final decision about the existence, position and type of vehicle. The model defines the exact relationships between object features and hence specifies the features present in an image under any possible view. A given instance of the model can therefore be evaluated with great precision, but only if the viewpoint is first specified.

These two techniques use different kinds of knowledge, and have complementary merits and deficiencies. The main purpose of this paper is to link the two in the context of the vehicle exemplar, and to show that together they form a whole which overcomes the limitations of each.

AVC 1987 doi:10.5244/C.1.6

Hypothesis Generation.

Scene description identifies candidate regions and provides loose bounds on the position of the vehicle in the image and its distance and orientation with respect to the camera. The latter may be deduced from the size and aspect ratio of the candidate regions (Godden, Fullwood and Hyde, 1987, Morton, 1987), together with knowledge of which vehicle rules have been triggered (Hutber and Sims, 1987).

The bounds establish strong constraints on the eventual match between the object model and the image, but still leave a great deal of uncertainty. Two main methods have been developed for refining the match: geometrical reasoning based on assumed correspondences between image details and key features of the object model, and naive search applied to small subspaces of the view transformation. They are used in support of each other.

Our methods of geometrical reasoning have some similarity with the local focus features used by Bolles and Cain (1982), and make use of several methods from the work of Lowe (1987) on reasoning from perceptual groups. Horaud (1987) has also discussed geometrical reasoning based on key groupings of features at junctions.

Lowe proposed general principles of perceptual grouping which identify viewpoint-independent object features. Several types of perceptual groups, each of which is unlikely to occur by chance, are discussed by Lowe, including proximity of end-points, skewed symmetries, and collinear and coincident lines. One particularly potent grouping, which Lowe used in a successful object identification programme, is that of pairs of parallel lines forming nearly closed parallelogrammes.

The visual purpose of feature groups is to reduce the number of possible matches between object and image features, and to establish the viewpoint. For example, a parallelogramme in an image can be matched successively to each rectangular face on the model, and thereby generate estimates of each of the alternative possible views. Once a viewpoint hypothesis is established, the model can be projected onto the image to predict additional features, and an iterative procedure used to derive a final estimate of the viewpoint parameters. Lowe uses perceptual groups which contain at least three lines, to guarantee sufficient constraints to determine the approximate viewpoint uniquely (given assumed correspondences of features).

We have found it difficult to specify perceptual groups which apply reliably to the vehicle exemplar, since the vehicle images are more complex than those studied by Lowe. The background contains objects which are not modelled by the programme, and groups such as parallelogrammes occur frequently at irrelevant points in the background. In any case, few of the views of vehicles contain clear rectangular surfaces because of the rounded styling favoured by the motor industry. Similar difficulties arise with other general purpose groups, e.g. the edges meeting at a corner of a smoothly styled car are not precisely coincident.

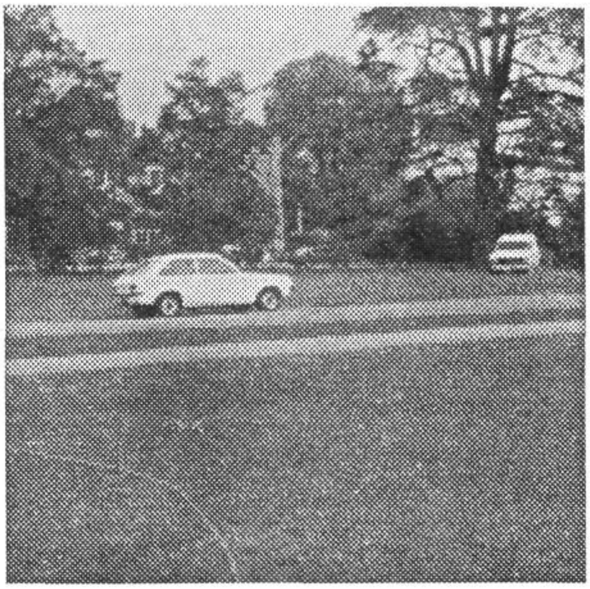

Figure 1. *A typical exemplar image.*

However, the image usually contains object-specific features which can be used fairly reliably, given approximate bounds on the viewing angle. For example if the car is approximately side-on then two wheels are often visible; in front and rear views, clusters of horizontal lines are visible; if the viewing angle is nearly horizontal, then a characteristic roof-line is conspicuous. Note that unlike Lowe's perceptual groups, these feature groups are highly specific to the modelled object and the viewpoint assumptions. The computational burden of testing for all such groups in a data-driven way is prohibitive, and we must rely on partial hypotheses produced by other methods to concentrate the search.

Refinement and verification.

The object specific features may be used to establish partial matches between object and image, but may still be too weak to identify a unique view. For example, knowledge of the centres of two wheels provides 4 constraints on the 6 degrees of freedom of the object with respect to the camera (3 translations and 3 rotations). The remaining freedoms correspond to the angle of tilt of the wheel line away from the camera (interacting with the depth from the camera to preserve the image datum), and rotations of the model about the wheel line.

A further constraint can be introduced from prior knowledge of the camera position with respect to the ground plane and the expectation of upright cars, which together establish the rotation about any chosen wheel line. We are therefore left with one degree of freedom, corresponding to the distance/tilt interaction. Furthermore this freedom is itself constrained since the angle of tilt away from the camera must be within +/-45 deg, for the wheels to be visible. It is straightforward to carry out a 1-D search within these bounds for the best fit, using the full model evaluator reported by Brisdon (in fact two searches are required corresponding to the off-side and near-side views of the car).

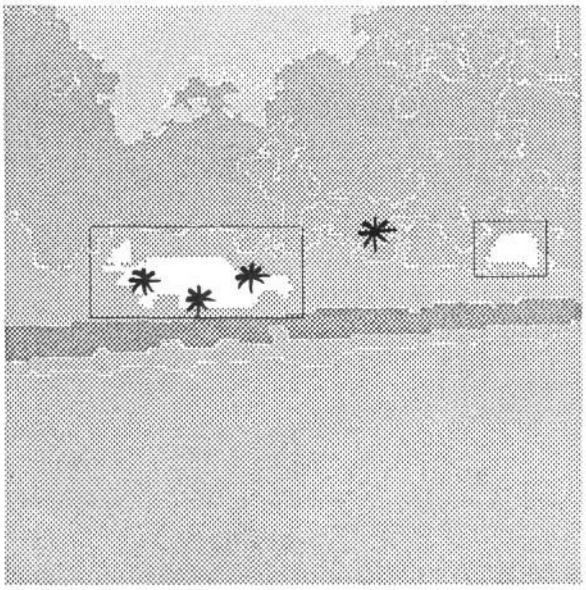

Figure 2. *Classified segmented version of Figure 1. Candidate regions, containing potential cars are shown boxed. The approximate location where sideview vehicle rules are activated are marked by *.*

The overall process applied to the exemplar is illustrated in Figures 1-3. Major regions of the original image (Figure 1) have been segmented and described by colour, shape and relational attributes (Godden, Fullwood and Hyde, 1987). These have been labelled as tree, grass, sky, tarmac, and buildings by use of evidential and contextual reasoning (Morton, 1987) and candidate areas which conform to the vehicle schema have been outlined by boxes (Figure 2). The results of the analysis of feature groupings (Hutber & Sims, 1987) indicate potential sideviews of vehicles at locations marked by asterisks (*). Evidence from these two knowledge sources is combined to eliminate all but the most indicated candidate regions, which are labelled according to the expected view.

Figure 3 shows the results of detailed analysis of the candidate regions, using a multi-scale operator, based on the method of Canny (1983), to identify edge points, which are grouped and classified as straight lines, arcs and closed forms. In the illustrated case we expect a side view of the car, so a wheel-finding strategy is invoked, which searches for small closed, smooth curves. It finds three such candidates, two of which have approximately the same size and are acceptable as wheel hypotheses. Figure 4 indicates the range of car models that are consistent with the hypothesised wheel line, and Figure 5 plots out the evaluation scores obtained by applying the full car model, using a stringent value of the precision parameter (see Brisdon, 1987).

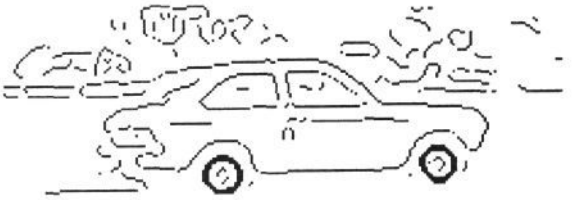

Figure 3. *Analysis of edges present in the larger boxed region of Figure 2. Each connected edge is labelled by its curvature and type (line, arc, closed curve). Acceptable wheel instances are emphasised.*

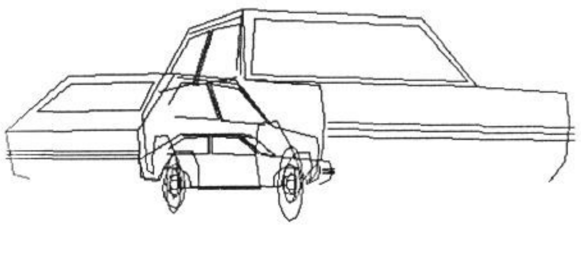

Figure 4. *Model Instances consistent with the wheel-line inferred from Figure 3.*

It can be seen that there is a clear region in the evaluation function, which is well above the expected noise level of 1, and we take the position of the maximal peak to identify the most likely instance of the car. The results obtained under the wrong viewpoint assumption (i.e. viewing the near-side of the car) are also shown; the maximal value is far weaker so this hypothesis can be rejected. As a final refinement we allow all 6 degrees of freedom to vary within tight constraints, to overcome any slight error in camera assumptions, or measurement of the wheel-line. Using the most stringent value of the precision parameter for the evaluator (see Brisdon, 1987), we seek the (independent) local maxima for small changes of translation relative to the camera, and rotation relative to the model axes. A new "best estimate" is obtained, and the refinement is iterated until stable. We obtain a peak along each of the 6 coordinate axes, illustrated in Figure 6, which identifies the viewpoint to within approximately +/-5deg and +/-10cm (at a viewing distance of approximately 30m). When projected onto the image, the fit appears extremely good (to the eye) but we have yet to carry out formal measurements.

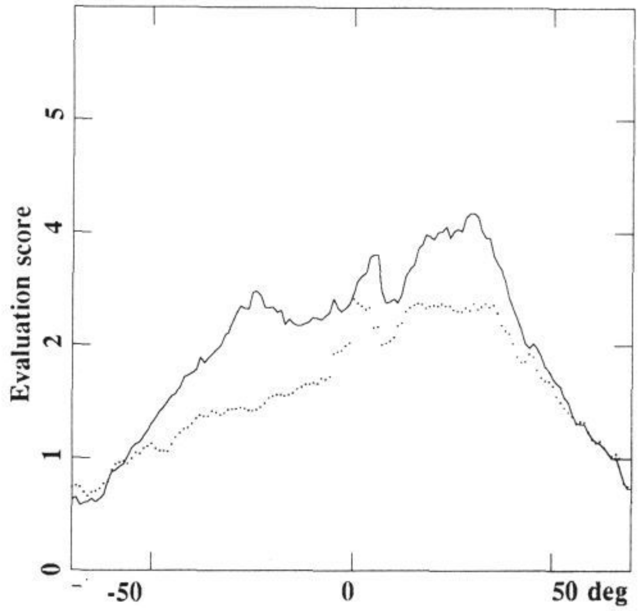

Figure 5. *Evaluation scores, as a function of the angle of tilt of the wheel-line away from the camera, for off-side (continuous) and near-side (dotted) views.*

A second example is shown in Figures 7-9. In this case a candidate region has been isolated as before, and the evidence from the feature grouping rules (Hutber and Sims, 1987) indicates that it corresponds to either a rear or a front view of a vehicle. An alternative strategy to that based on the wheel-line is therefore invoked which looks within the candidate region for a characteristic inverted U shape made by the roof-line. Patterns of image features which conform to the definition are picked out in Figure 8. In all there are 16 possible views of the roof which could match the roof outline, but several of these collapse under the assumption of near horizontal viewing. This feature group contains three data lines which may be tentatively related to 3 edges of the model; the 6 viewpoint parameters can therefore be solved uniquely, for each possible correspondence assumption.

The exact solution still requires inversion of a non-linear system of equations. We adopt an iterative approach similar to Lowe's, which minimises the perpendicular separation between the ends of the modelled lines and the imaged lines. Our method differs from that reported by Lowe in using a parameterisation of the cost function in the camera coordinate system, which simplifies the geometrical analysis and allows further 3-D constraints to be introduced in the search more conveniently (e.g the vehicle is upright, or on a specified plane, or at a certain distance, etc.).

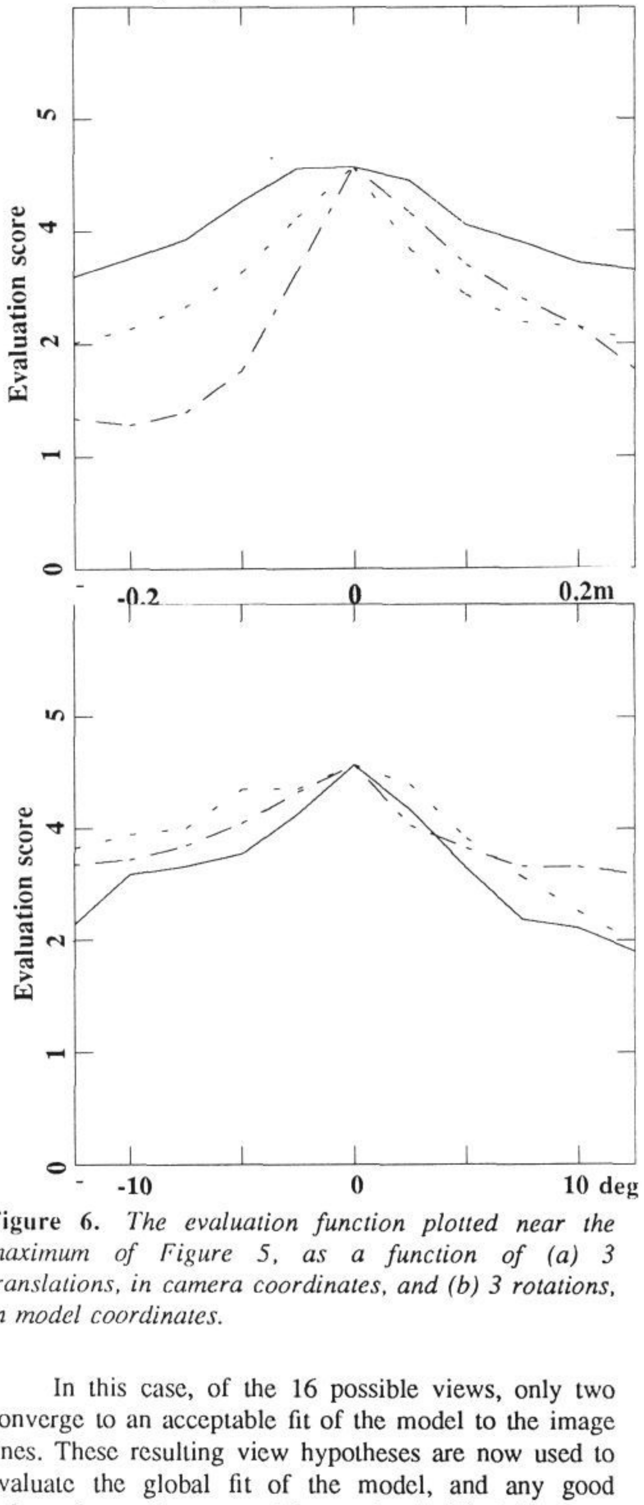

Figure 6. *The evaluation function plotted near the maximum of Figure 5, as a function of (a) 3 translations, in camera coordinates, and (b) 3 rotations, in model coordinates.*

In this case, of the 16 possible views, only two converge to an acceptable fit of the model to the image lines. These resulting view hypotheses are now used to evaluate the global fit of the model, and any good values (exceeding an arbitrary threshold) initiate an iterative search for the optimal fit, as before. The results of the evaluation near the best fit are shown in Figure 9.

Other critical features for solving the view-point have also been studied, including the rear window trapezoid, the wheel arches, and the characteristic _/‾ shape of the bonnet-windscreen lines. In each case we seek object features that are fairly robust against viewpoint and easily defined as clusters of image features.

Assessment.

A major objective of this research has been to investigate a vision system capable of performing a realistic exemplar task using images of natural scenes. Our approach adopts a hypothesise-and-test strategy in which a variety of different types of knowledge constrain the hypothesis generation stage. Scene knowledge and characteristic groupings of features trigger a detailed analysis of candidate areas, which uses specific knowledge of the geometry of the car. As with the work of Lowe, we use viewpoint-invariant methods to draw attention to plausible initial hypotheses, which are progressively refined to the point where a view-specific evaluation can be carried out.

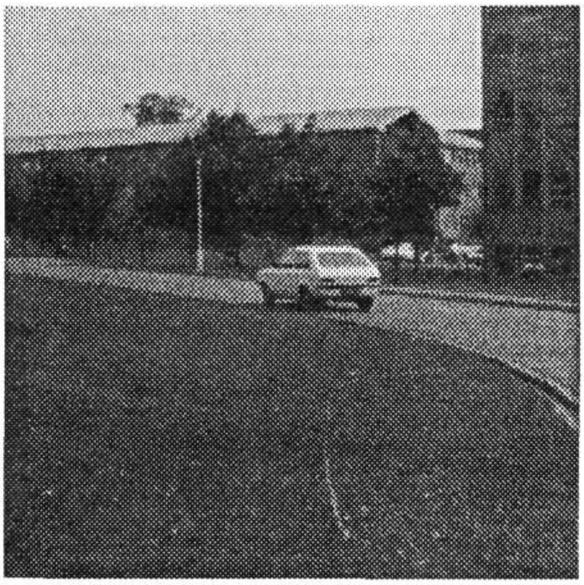

Figure 7. *A Rear-view example image.*

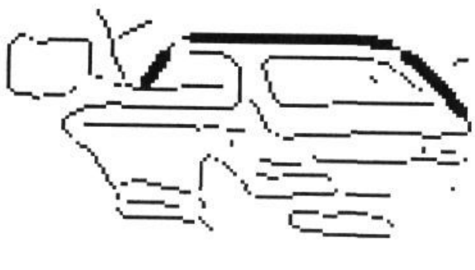

Figure 8. *Analysis of edges in the candidate region of Figure 7. Patterns conforming to the roof outline definition are emphasised.*

The main novelties in our methods have been introduced to overcome difficulties due to the more demanding vision task. Lowe has studied a form of the bin-picking problem, where examples of an object have to be recognised from a jumble of identical objects which have no expected relationships within the scene; furthermore the objects - e.g. a disposable plastic razor - are well defined as wire-frame models, and the important edge features can be found reliably by low-level analysis. In our exemplar task, we have to take account of objects in the scene which are not explicitly modelled, but which do imply likely relationships with the modelled object; furthermore the modelled object contains curved surfaces, which frequently cause low-level features to be mislocated due to specularities, or view-dependent extremal boundaries.

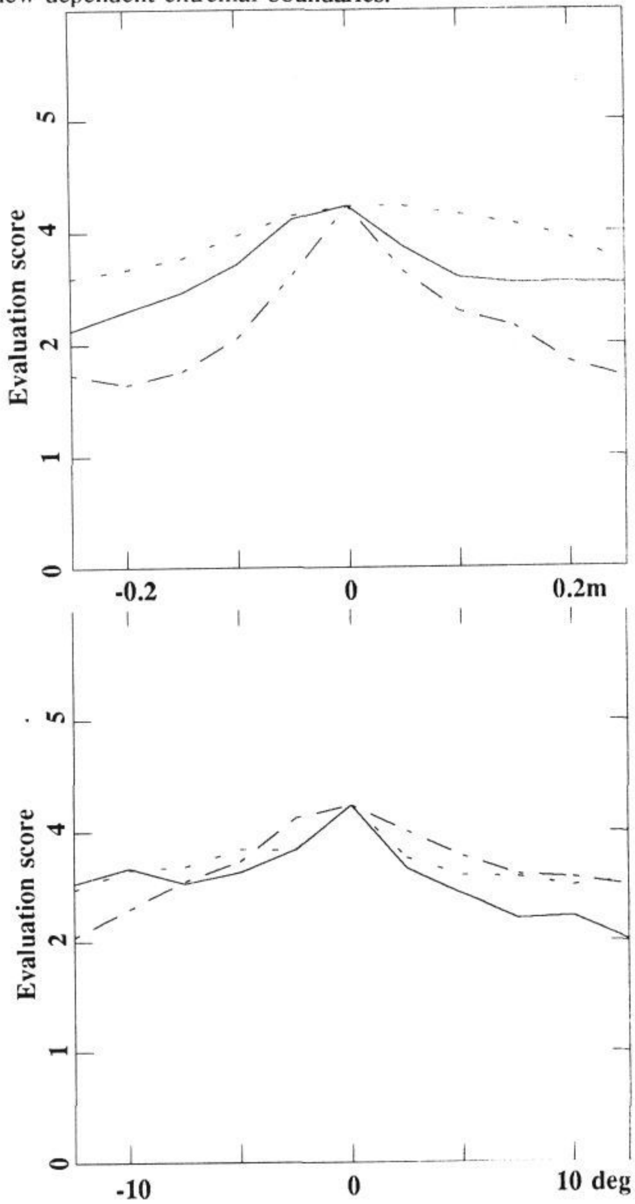

Figure 9. *Evaluation function plotted near the best solution of the roof line (as Figure 6).*

This means that there are potentially very many erroneous matches between irrelevant image features and the model, which must be rejected by the initial stage of scene analysis. It also means that a global low-level description of the image is unlikely to capture the information required for hypothesis evaluation. Whereas Lowe verifies his viewpoint hypotheses by matching predicted features against a symbolic description of lines in the image, we return to the original image itself to assess predicted features. The cost of evaluating a single hypothesis is not great. Indeed this approach can be more efficient, since there is no need to generate a complete symbolic representation.

Conclusions.

Our early experiments with the system are encouraging, but we are aware of shortcomings at many of the processing steps and current efforts are directed towards improving the robustness and generality of the methods. In particular we wish to see:

Improved scene description rules, including the ability to use contextual frames corresponding to different terrains, or times of day, or different imaging modes.

The use of other image evidence as primary cues to regions of interest, e.g. locally coherent movement.

The simultaneous analysis of multiple vehicle models, e.g cars, vans, etc.

Improved modelling facilities, to include curved surfaces.

The ability to group mutually related object features together, to allow (e.g.) saloon, estate, or hatchback rear ends of a car to share a definition of its front, and to help identify partially occluded vehicles.

A more principled manipulation of the object specific features, and of the consequent geometrical reasoning, to replace the largely *ad hoc* methods currently employed.

A further major development is needed. At present the processing stages are only loosely coupled. This has been convenient within the multi-site framework of the Alvey collaboration, but represents a major impediment to further development. For example, there is an obvious overlap between the feature groupings derived from machine learning (Hutber and Sims, 1987) and the features used to identify possible correspondences between model and image data (such as wheel or roof lines). It happens that the former uses region information, and the latter uses edge descriptions, but clearly the information derived from both sources should be used to address both sub-problems. The sequential application of separate modules, should be replaced by a system architecture which allows greater freedom of communication between modules, so that hypotheses emerge as a collaboration between modules.

Finally, a long term objective of this work is to develop methods of using high level knowledge about objects and the scenes in which they are likely to appear. In particular we have been concerned to make use of detailed geometrical knowledge at early stages of visual analysis. The strategy succeeds with the vehicle identification exemplar since we have been able to bypass the hardest problem for general vision: that of accessing the appropriate object specific knowledge from amongst the host of possible interpretations of the image. The human ability to do this with very little primary evidence is awesome, but may also be based on discovering characteristic patterns of object specific features. Figure 10 shows a sketch, containing a few of the bonnet, roof-line and wheel clues that our programmes currently use. Even without the bias introduced by the context of this paper, human observers have little hesitation in classifying the object portrayed.

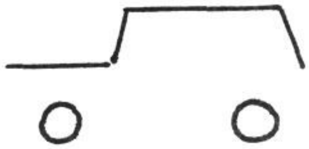

Figure 10. *Illustrating that the object-specific clues used to access the computer's geometrical model seem sufficient for human perception.*

Acknowledgements.

The work reported here was carried out initially at Plymouth Polytechnic, and later at The University of Reading, by the Intelligent Systems Group, under the support of Alvey grant MMI-007.

James Anderson and Anthony Worrall provided the modelling and geometrical reasoning systems, Andrew Rydz coded the edge feature descriptions, and Tim Clarke produced the iterative refinement process.

References.

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, K., (1987), Evaluation and Verification of Model Instances, AVC-1987

Georgeson, M. A. (1979) Random-dot stereogrammes of real objects: observations on stereo faces and moulds. Perception, 8, 585-588.

Gregory, R. L. (1971) *The Intelligent Eye*, Weidenfeld & Nicholson.

Horaud, R., (1987), New Method for Matching 3-D objects with single perspective views. IEEE PAMI-9, 401-412.

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

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