PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Visual Interpretation of Known Objects in Constrained Scenes [and Discussion]

G. D. Sullivan and A. Sloman

Phil. Trans. R. Soc. Lond. B 1992 **337**, 361-370
doi: 10.1098/rstb.1992.0114

Visual interpretation of known objects in constrained scenes

G. D. SULLIVAN

Department of Computer Science, University of Reading, Reading RG6 2AY, U.K.

SUMMARY

Recent work on the visual interpretation of traffic scenes is described which relies heavily on *a priori* knowledge of the scene and position of the camera, and expectations about the shapes of vehicles and their likely movements in the scene. Knowledge is represented in the computer as explicit three-dimensional geometrical models, dynamic filters, and descriptions of behaviour.

Model-based vision, based on reasoning with analogue models, avoids many of the classical problems in visual perception: recognition is robust against changes in the image of shape, size, colour and illumination. The three-dimensional understanding of the scene which results also deals naturally with occlusion, and allows the behaviour of vehicles to be interpreted.

The experiments with machine vision raise questions about the part played by perceptual context for object recognition in natural vision, and the neural mechanisms which might serve such a role.

1. INTRODUCTION

High-level vision is usually taken to refer to deliberate, conscious reasoning about images, based on specific knowledge of the perceived objects and events, whereas low-level vision involves involuntary, unconscious processes, based mainly on general knowledge of the three-dimensional world. Object recognition occupies an uncertain place in this dichotomy. One of the most influential texts on computer vision (Ballard *et al.* 1982) observed that 'an important low-level capacity is object perception'. But object recognition necessarily involves specific knowledge (if only to define the objects perceived). Such knowledge need not be limited to the three-dimensional structure of objects, but may also include relationships between objects and the world in which they are seen. A major issue in computer vision concerns the potential role of this high-level contextual knowledge in deriving low-level recognition (see, for example, Riseman *et al.*, 1987).

This paper discusses recent work on object perception by computer vision which depends critically on high-level reasoning, applied at a very low level of image analysis. It turns out that many of the hardest problems in low-level vision, especially those concerning how fragmentary image features become grouped into meaningful evidence, can be circumvented by the early application of rather simple knowledge about the vision task. The visual task we have addressed is that of understanding traffic moving on roads and at airports. Work recently carried out under industrial collaborative projects has shown that elementary high-level knowledge makes the visual understanding of traffic possible in practice. A level of competence for traffic perception has now been achieved which is likely to lead very shortly to major applications of machine vision in natural outdoor conditions.

The techniques we have adopted for machine vision seem 'natural', and they confer conspicuous advantages for object recognition. It is intriguing to enquire if natural vision makes use of similar methods, and at the end of the paper a number of speculations are raised about the possible role of similar model-based processing in human vision.

2. MODEL-BASED VISION

An essential requirement of a system able to recognize an object is the ability to represent the object. Recognition is the act of discovering the appropriate relationships between the sensed data and knowledge brought to the task by the perceiver. In the case of traffic understanding we wish to answer such questions as: How close are the cars, and how fast are they going? How many cars turn right at a junction? How often does the behaviour of one car interfere with other vehicles? How rapidly do queues form and disperse? The knowledge needed to answer questions such as these includes the geometry of vehicles, the layout of the roadways, the nature of vehicular movement, and the ways in which vehicles interact with each other. Such knowledge can be used in two ways: passively, simply to define the classes of objects and events we wish to recover, or actively, to guide the search for perceptual descriptions.

In computer vision it has proved immensely difficult to proceed from general-purpose features, extracted from images without reference to specific objects, to the determination of meaningful sets of features which are related to one another by virtue of their com-

Phil. Trans. R. Soc. Lond. B (1992) **337,** 361–370
Printed in Great Britain

pliance with the known nature of an object. The reason is clear: knowledge-free feature extractors, such as edge- or bar-detectors, respond to strictly local properties of images, and cannot discriminate between relevant and irrelevant detail. The classical strategy in object recognition is to group features together, according to simple image properties (such as proximity, similarity or continuity) to detect feature groups which are invariant to the viewing conditions. But the relationship which exists between, say, the front bumper of a car and its offside rear wheel arch is difficult to characterize as image invariants: the configuration of the parts in the image varies widely as the car is seen from different viewpoints. Attempts to identify such image-based relationships by purely 'bottom-up' means, based on general invariants (such as proximity), generates a confusion of possible relationships between local features. The combinatorial analysis required to sort out the correct groupings from the welter of meaningless 'coincidences' is impractical to compute (Grimson *et al.* 1990).

As an alternative to using object knowledge passively, at the final stage of the recognition process to classify the groupings already assembled, there has recently been a great deal of interest in using object-specific knowledge actively, as part of the evidence accumulation stage. Recognition then becomes a process akin to mathematical induction: we seek to guess the answer, and then to demonstrate that the answer is correct. Initial guesses are provided by 'cues': simple properties of images which often accompany objects, but are not of themselves reliable indicators of an object. The 'model-based' approach to object recognition has been used successfully by a number of authors (Brooks 1983; Hogg 1983; Lowe 1985, 1991; Worrall *et al.* 1991). The key problem is no longer how to group together sufficient image features to reliably deduce the presence of an object. Instead, we need firstly to make intelligent guesses about possible objects, and secondly, to determine an object's likely pose and to verify its presence.

Top-down reasoning may seem to put the cart before the horse: we first need to conclude what is there, then we can gather the evidence in favour of the conclusion. In general-purpose, unconstrained vision this approach is evidently hopeless; there are simply too many possible interpretations. But most practical applications of machine vision are highly constrained and seek only to solve specific problems in specific situations. For example, traffic monitoring systems need only know about vehicles and their movements. Indeed they need only a very limited subset of such knowledge: vehicles are constrained to rest on the ground plane, and this reduces the variability of any single vehicle from having 6 degrees of freedom (d.f.) (of a rigid body) to 3 d.f. (translations on the ground, plus rotation about the normal to the ground plane). Thus, provided that the vision system knows the position of the camera with respect to the ground plane, then the practical problem of understanding vehicle movements becomes very greatly simplified. This is the approach taken within the Esprit VIEWS project.

3. VISION IN CONSTRAINED SCENES

The VIEWS project (the acronym stands for the Visual Interpretation and Evaluation of Wide-area Scenes) is a collaborative research and development venture financed by the Esprit directorate of the European Framework Programme. The objectives are to demonstrate the use of vision systems in a wide range of traffic monitoring applications. The project has mainly concentrated on two types of scene: urban road intersections, and airport ground traffic.

(*a*) *Pose evaluation*

The geometrical properties of the vehicles, and the layout of the roadways, are represented as wire-frame models expressed in individual object-centred coordinate frames. By using simple geometrical reasoning, vehicles can be placed on the ground at arbitrary locations in the world model. We also have a carefully constructed camera model, which allows us to project an instance of the three-dimensional scene onto the two-dimensional image plane. Given an hypothesis of a vehicle in the scene, we can therefore use techniques of computer graphics to derive the two-dimensional image features we would expect to see. Each feature comprises a line segment (the projection of a single wire of the model), which is tagged by a description of its likely image properties, according to its role in the instantiated model (fold- crease- or extremal-edge, bar, mark on a surface, shadow edge, etc.).

The agreement between the prediction and the image data can be tested by a process which we have called 'iconic evaluation' (Brisdon *et al.* 1988; Brisdon 1990). Each linear feature is evaluated by applying simple criteria, based on derivatives of smoothed intensity values in the direction perpendicular to the feature. Thus edge features are scored according to the average strength of the first derivative, and bars according to that of the second derivative. The evidence from all the visible lines of the object is pooled, by expressing each score as a probability, taking account of the lengths of the lines and the type of feature. Probability tables are established empirically, by placing lines of various lengths randomly in a calibration image (ideally the image under investigation, but in practice a previously computed temporal median image), the determining the score. Any given feature score is thereby associated with a probability that a score at least as high would have been obtained by chance from a randomly placed feature of the same type and length.

The individual model features are treated as independent samples from the empirical probability distributions†, and are pooled using a χ^2 test. The result is a single scalar in χ^2 units, having expectation 0, and typically ranging from -3 (in areas of the image with far less than average detail) to 5 or more when the

† This assumption is obviously false (as the presence of one image feature found on cars is strongly correlated with the presence of others), but it has proved difficult to take conditional probabilities into account. The method provides a convenient way to pool the probabilistic data, which we have found to be robust in practice.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

projected model fits the image very well. Scores over 2 are treated as significant indications of a vehicle.

The resulting evaluation score for a vehicle is reasonably independent of the position and pose of the object, since it measures (to a first approximation) how likely the score was to have been obtained by chance, taking into account the number of visible features, their types and their lengths in the image.

(*b*) *Pose refinement*

In unconstrained viewing of a known object, the evaluation score defines a scalar function of six dimensions: in world coordinates these are most simply defined as the three cartesian coordinates of the object's position and the three angles of orientation. In general, we expect that peaks in the six d.f. function will indicate likely matches between the model and the image. The problem is to locate the peaks, and thereby to determine the pose of the vehicle.

A considerable computational simplification can be made by limiting the object's position to the ground plane, thus permitting only two dimensions of translation and one of rotation about the vertical axis. Using these simple but (normally) realistic physical assumptions, only three independent dimensions remain. Figure 1 illustrates part of the surface of the evaluator function. Only 2 d.f. can conveniently be shown. Pose may be parameterised as pan angle (horizontal position in the image), depth on the ground plane towards and away from the camera, and rotation of the object about its coordinate centre. Pan corresponds approximately to simple translation in the image, so that the evaluation reduces effectively to correlation between a fixed template and the image. The other two parameters introduce nonlinear distortions, due to projection distance and pose angle respectively. These are the two axes shown in figure 1. It can be seen that even for this 'difficult' cross-section, the surface is fairly smooth and well-behaved. The (correct) central peak extends for approximately ± 1 m in depth and $\pm 15°$ of rotation; any initial guess within this range (i.e. which estimates the pose of the vehicle accurately to within these bounds) should be sufficient to discover the correct pose.

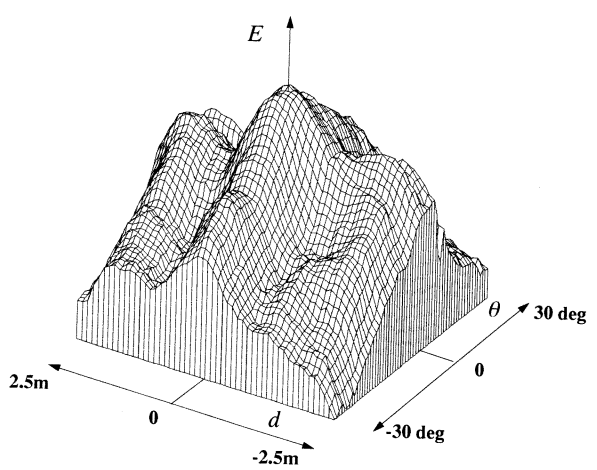


Figure 1. Cross-section of the evaluator function, showing variations in score (E) with depth from the camera (d), and rotation about the vertical (θ), of an object constrained to the ground plane (see text).

On a serial machine, an exhaustive search of the evaluator surface over three dimensions is computationally too expensive. We have therefore developed an iterative method which successively decomposes the problem into three separate one dimensional searches, each based on the current estimate of the object's coordinte frame. This is illustrated for the x-coordinate (the left-to-right axis of the car) in figure 2.

For each sample the model is displaced by an appropriate amount (in the object coordinate frame), instantiated into the image, and its fit to the image is evaluated. The results obtained for a typical search along a single dimension are shown in figure 2*b*, where the abscissa represents the displacement and the ordinate represents the evaluation score obtained (the higher the score the better the fit). Note the subsidiary peak, due to accidental alignment of one side of the car model with the wrong side in the image. The best score and its position are noted and the process is repeated for the other two variables (here, y and rotation about z)—each time starting from the same initial pose.

When all three dimensions have been searched the pose having the highest score found in adopted. It then becomes the initial position for the next iteration, and the search coordinate frame is changed accordingly. If no higher score is found then the three ranges of the search are reduced, to be equal to the previous sampling interval. The process is repeated until all the sampling intervals fall below criterion values (see below).

The initial search range, and the terminating conditions are strongly object- and pose-dependent. To compute them automatically, the object model is approximated as a sphere of diameter equal to the greatest diameter of the model. Simple geometry uses the 'seed' pose (in the world-coordinated frame) to compute the displacements in x and y and the rotations about z (in the object-coordinate frame) which would cause a 1 pixel change in the image for a worst-case object feature. These determine a set of scaling parameters, which are weighted to give the termination conditions required by the application (we typically use 0.5 pixel change). The initial search range is set by similar reasoning, and we typically use bounds corresponding approximately to ± 0.5 times the dimensions of the vehicle and $\pm 10°$. At extreme distances, or in poses in which one of the object coordinate axes is directed towards the camera, the initial ranges may already imply a sampling interval below the terminating conditions, in which case only one iteration (for that axis) is performed.

Figure 3*a* shows a model that was very approximately instantiated near the vehicle by hand. This is fairly typical of an initial estimate of the position obtained from simple image-based cuing processes. The search algorithm was allowed to run, and the result is shown in figure 3*b*. Informally, by eye, it seems that the fit is very good.

To estimate the overall signal-to-noise ratio of the method, we have taken the best fitting position and

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

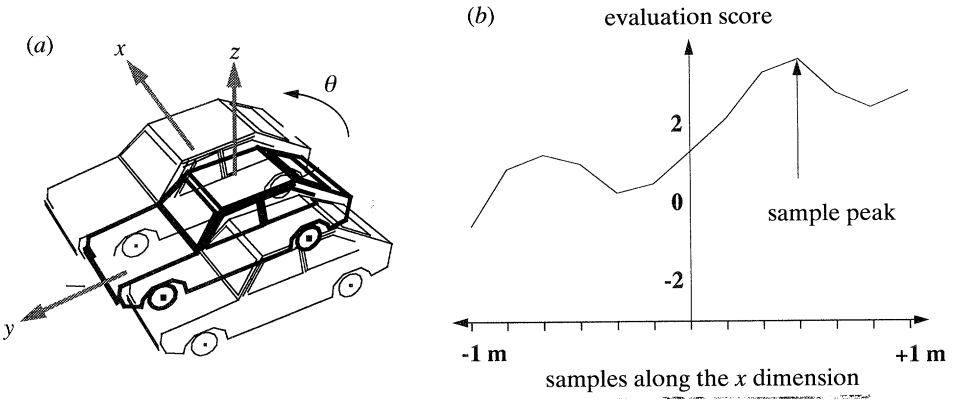


Figure 2. (*a*) For each degree of freedom the model is displaced, instantiated and its 'goodness-of-fit' evaluated. (*b*) Evaluation scores obtained from a model by displacement along a single dimension.

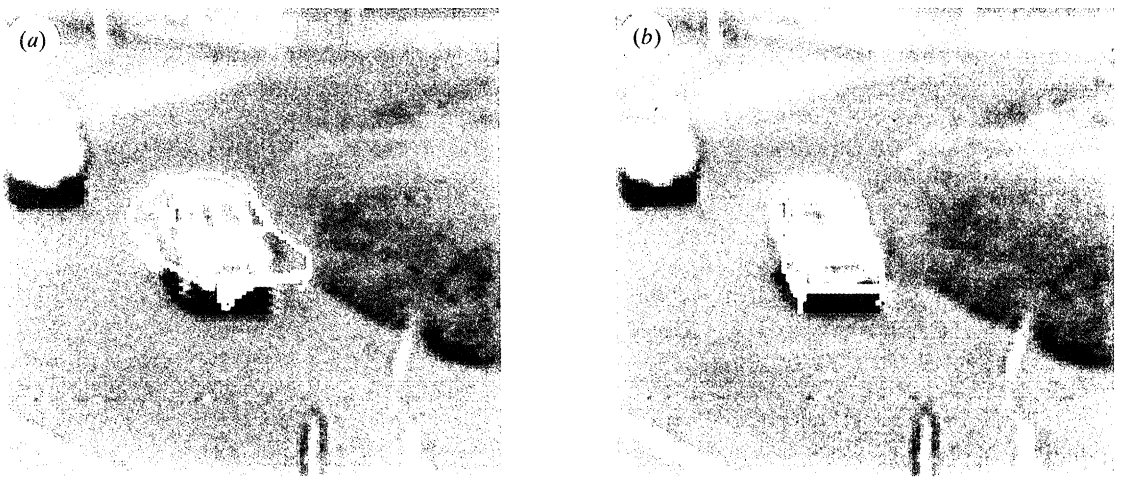


Figure 3. (*a*) Before and (*b*) after gradient ascent (frame 200).

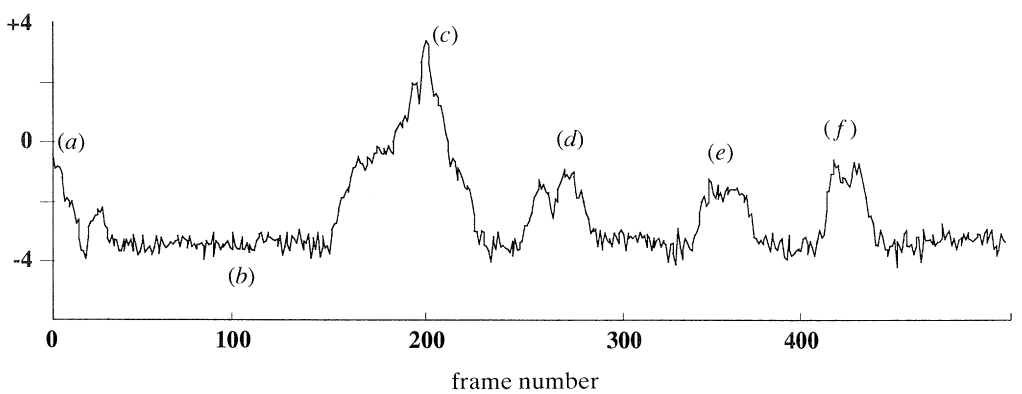


Figure 4. Evaluation function for a fixed position on different images (*a–f*) refer to figure 5).

orientation for the model in this image (as shown in figure 3*b*), and then evaluated this particular model instance in a sequence of 500 images taken at 25 Hz. The results are shown in figure 4. We see one conspicuous peak, with several subsidiary peaks. The images corresponding to the points on the function labelled (*a*) to (*f*) are shown in figure 5. The response of the evaluator is typically about -3.5 in the absence of vehicles. The minor peaks correspond to accidental alignment between the model and a 'wrong' vehicle, but these scores never exceed 0, and the peaks are fairly flat and ragged. The score for the 'correct' vehicle reaches 3.7, and is well-localized. In this simple case, where there is little distracting background image detail, the evaluator provides an intuitively acceptable measure, and has good signal-to-noise ratio.

(*c*) *Equivalence classes*

The performance of iterative algorithms for solving nonlinear maximization problems such as this is

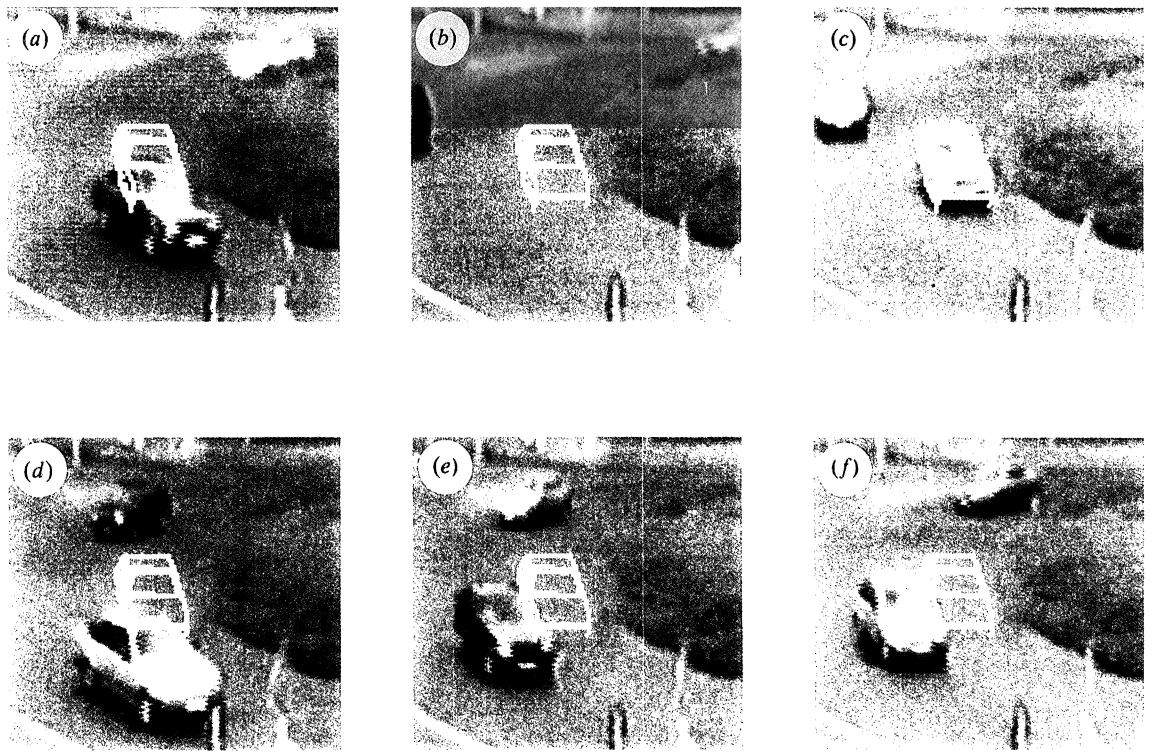


Figure 5. Fixed model in different frames ((*a*–*f*) refer to figure 4). (*a*), frame 0; (*b*), frame 100; (*c*), frame 200; (*d*) frame 273; (*e*) frame 351; and (*f*), frame 418.

mainly determined by two issues: the smoothness of the function and the conspicuousness of the peak. Our local maximization algorithm defines an equivalence relationship on the domain of initial poses, in which each pose can be classified by the local 'attractor pose' to which it moves as a result of the search process. The performance of the algorithm can therefore be esimated by charting the equivalence class for the 'correct' position. Figure 6 illustrates typical results. An exhaustive analysis was carried out close to a known correct pose, determined by eye, using the parameterization shown in figure 2. At each of a sample of (x,y) positions, the orientation about $z(\theta)$ was varied in small intervals, to specify an initial seed pose. The pose refinement process was run, and the resulting values of x,y, and θ (relative to the corect pose) are plotted in figure 6, for two different errors in the (x,y) positions (see legend). The regions of the graphs where the three recovered values approximate 0, indicate 'correct' performance. In this case, figure 6*a* shows good stability over $\pm 12°$; figure 6*b* shows poorer performance, with the model becoming caught up on a local maximum (which was only slightly weaker than the global maximum), located about 7° from the true position for seed angle errors between 0 and $-15°$. Results such as these are typical. Performance is usually good, and the system converges to the correct solution from a wide range of initial

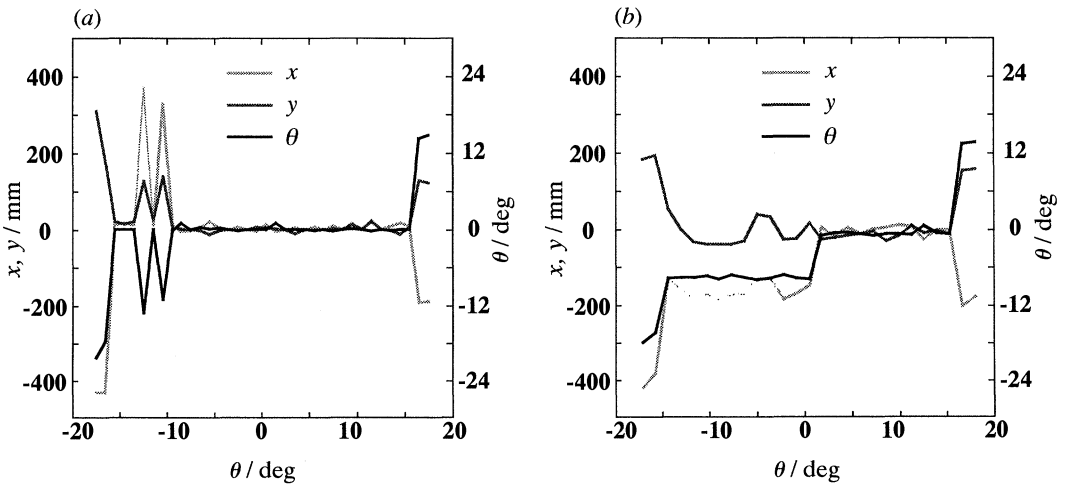


Figure 6. Examples of the stability of the pose refinement algorithm. The graphs show the final pose parameters (x,y,θ) starting from seed poses of: (*a*) (100, 200, θ), and (*b*) (-300, 200, θ); n.b.: (x,y,θ) are measured with respect to the 'correct' pose, determined by eye.

conditions, but nearby false maxima in the evaluation function make it important to control the predicted seed positions accurately as the vehicle is tracked over time.

4. DYNAMIC TRACKING

The pose recovered from the iconic search is used as an input measurement to a Kalman filter, which allows us to impose simple dynamic constraints appropriate to vehicles. Instantaneously, a car has only 2 d.f.: its forward (or backward) velocity, v, and its angle about the z-axis (θ). We have implemented a Kalman filter (Marslin *et al.* 1991) which allows freedom in v, $\dot{v}$, θ and $\dot{\theta}$. The three parameters v, $\dot{v}$ and $\dot{\theta}$ are constrained to be within plausible bounds: as the the filter is expressed in the object-centred coordinate frame, these constraints may be estimated on the basis of common knowledge of car dynamics. Note that this characterization of the vehicle's dynamics prohibits it from sliding sideways, but does not (at present) couple v and $\dot{\theta}$. The measurement error assumed in the Kalman filter is also pose-dependent, and is estimated on the same basis as the terminating conditions for the pose refinement process. We typically assume that the error of a recovered pose is gaussian distributed with a standard error equivalent to a 1 pixel worst-case displacement (cf. § 3*b*).

Tracking proceeds as follows. The initial 'cued' pose hypothesis is first refined as described above. Starting with plausible default values, the Kalman filtered estimates of v and θ, together with the newly measured position, provide a prediction for the object's pose in the next frame. This becomes the seed pose for the next search, and the process continues. When the filter parameters become stable, the forward prediction improves in accuracy, and fewer iterations of the separated gradient ascent algorithm are required; in these conditions we have found that a simpler (3D) steepest ascent algorithm is sufficient. It also becomes

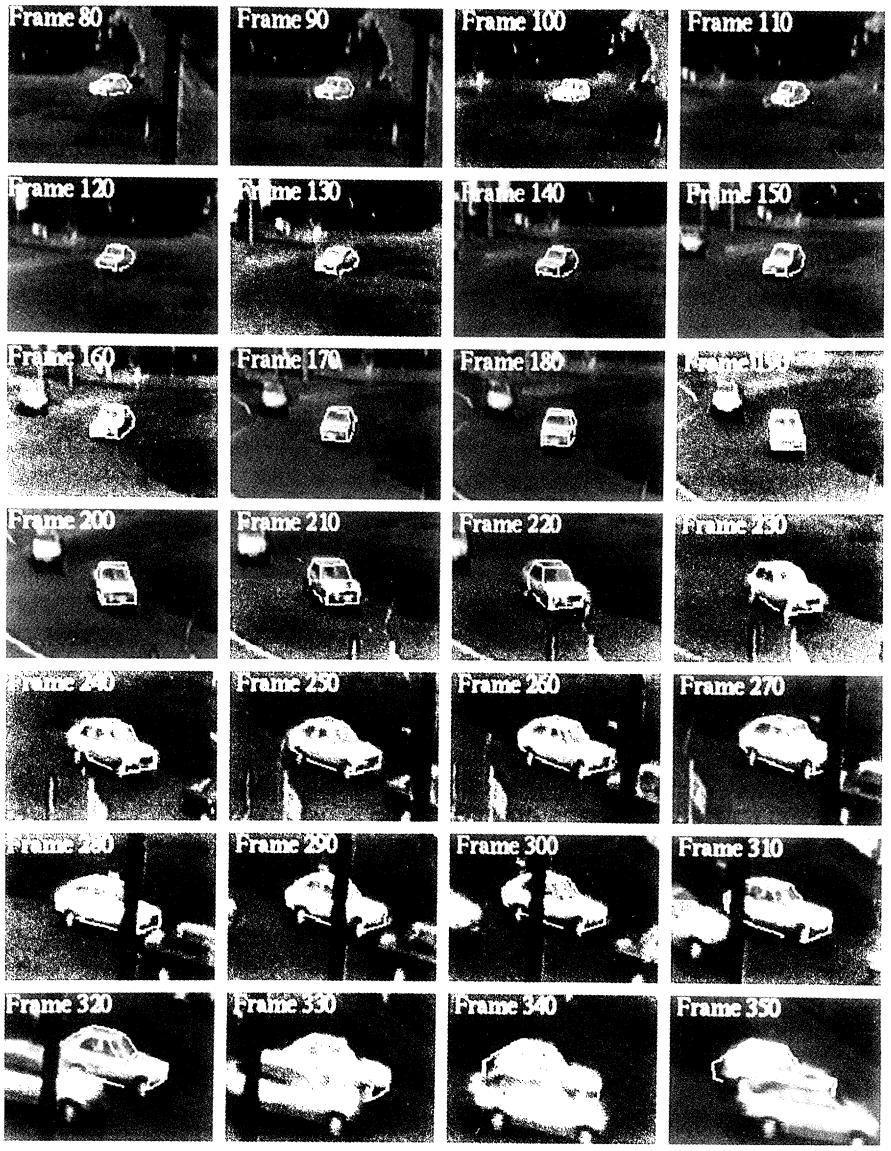


Figure 7. Tracking a single car (every tenth frame from 80 to 350).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

possible to predict several frames in advance and thus reduce the computational burden of the process (though there is a trade-off with the requisite size of the initial search space).

(*a*) *Results of model-based tracking*

The Kalman-filtered iconic search has been used to track vehicles in a number of scenes. Typical results are illustrated in figure 7 for the vehicle illustrated in figures 2–5, for every tenth frame (i.e. every 400 ms). Figure 7 shows the results of the pose-refinement algorithm, i.e. the measurement input to the Kalman filter, not the smoothed result of the filter, which in the present car-centred implementation, is difficult to reconstruct in the world coordinate system. In the main the performance seems very good, and the position and orientation of the vehicle were recovered accurately. Some instability was observed, especially in frames 220–240, where the angle of the vehicle was miscalculated, and 310–350, where the model begins to spin on the spot. Note that at various stages the care was occluded, either by objects in the scene (e.g. the lamp post, frames 270–310) or by other vehicles (frames 320–350). Both types of occlusion are taken into account by three-dimensional reasoning: the lamp post forms part of the scene model, and, though not shown in figure 7, the position of the occluding car was known because all vehicles in the scene were tracked simultaneously.

(*b*) *Initial seeding*

So far no account has been given of how the model-based tracking is initially seeded. In the system being developed, the primary seed hypotheses are provided by the analysis of movement in the image. Two methods have been developed, which offer complementary properties in practical applications. The first (due to work carried out by collaborators at the Fraunhofer Institute, Karlsruhe, Germany) is based on the detection of extremal points in a band-passed image, which move consistently between successive frames as coherent clusters. The second (due to work by Marconi Radar and Control Systems) maintains a running 'temporal median image', and detects regions of change with respect to this. The former method is fast and has already been developed to run on special hardware at video rate (25 Hz). Its main draw-back is that the extrema are clustered together purely on the basis of image velocity vectors, so the process tends to fragment objects which rotate before the camera, and merge adjacent vehicles having similar movement. It also loses vehicles that become stationary. The latter has better ability to segment overlapping vehicles, and to overcome brief periods of immobility, but is more sensitive to sudden overall changes in illumination. In addition, it cannot at present be made to run at video rates.

In either case, it is possible to obtain fairly reliable tracks in the image, assumed to be due to single objects moving in the scene. The centroid of the moving image data, projected back onto the known ground plane, provides a very approximate indication of the position of the object. By noting how this evolves over time, the likely pose of the object can be estimated. In turn, an analysis of the distribution of extrema, or of the moving region, gives an indication of the broad class of object (e.g. large, small, long) taking into account the distance from the camera and the approximate pose, given by the detection of movement. In highly constrained traffic scenes, this information has proved sufficient to initiate the model-based methods. Further classification, into the precise type of vehicle, can be carried out by examining the relative scores of the iconic evaluator (after pose-refinement) for different models.

Once started, the model-based method is usually capable of running without further input from the movement detection systems. It has good immunity to rotation and partial occlusion, and does not fail when the vehicles are stationary.

5. PERFORMANCE

The traffic understanding system has been successfully applied to a number of scenes at complex road intersections, and at airports. It results in a recovery of the full three-dimensional position and pose of vehicles in the known scene, which can be passed on to the other major component of the VIEWS system, the Situation Assessment Component. This maintains a long-term history of the vehicles in view, and detects events and behaviours of significance to the end-user of the system. These include unitary events (such as a vehicle entering a zone of special significance), binary relations (such as one vehicle causing another to give way, or one vehicle overtaking another), as well as higher order interactions (such as queue formation and dissipation). An application area receiving special attention is that of monitoring traffic at stand-areas of airports, where the airport authority needs to verify that aircraft are visited in the correct order by ancillary vehicles such as baggage handlers, water tankers, fuel tanker, cleaners, service engineers etc. Figure 8 illustrates a typical frame from one of our test sequences; the recognized vehicles are superimposed as wire-frame models on the image (*a*), and shown from overhead view in the scene representation (*b*).

6. DISCUSSION

The primary merit of the model-based method for object perception is that it works and seems reasonably robust. The outcome of the system is a full three-dimensional interpretation of vehicles, moving in the known three-dimensional scene. We have applied the method without significant changes to a wide variety of image sequences, taken under different environmental conditions, involving more than a dozen different vehicle types, partially occluded vehicles, different viewing distances and camera angles, in bright sunlight (giving heavy shadows), at night, in rain and in fog. The method deals naturally with the type of variation to be expected in practical traffic monitoring applications, and is currently being deve-

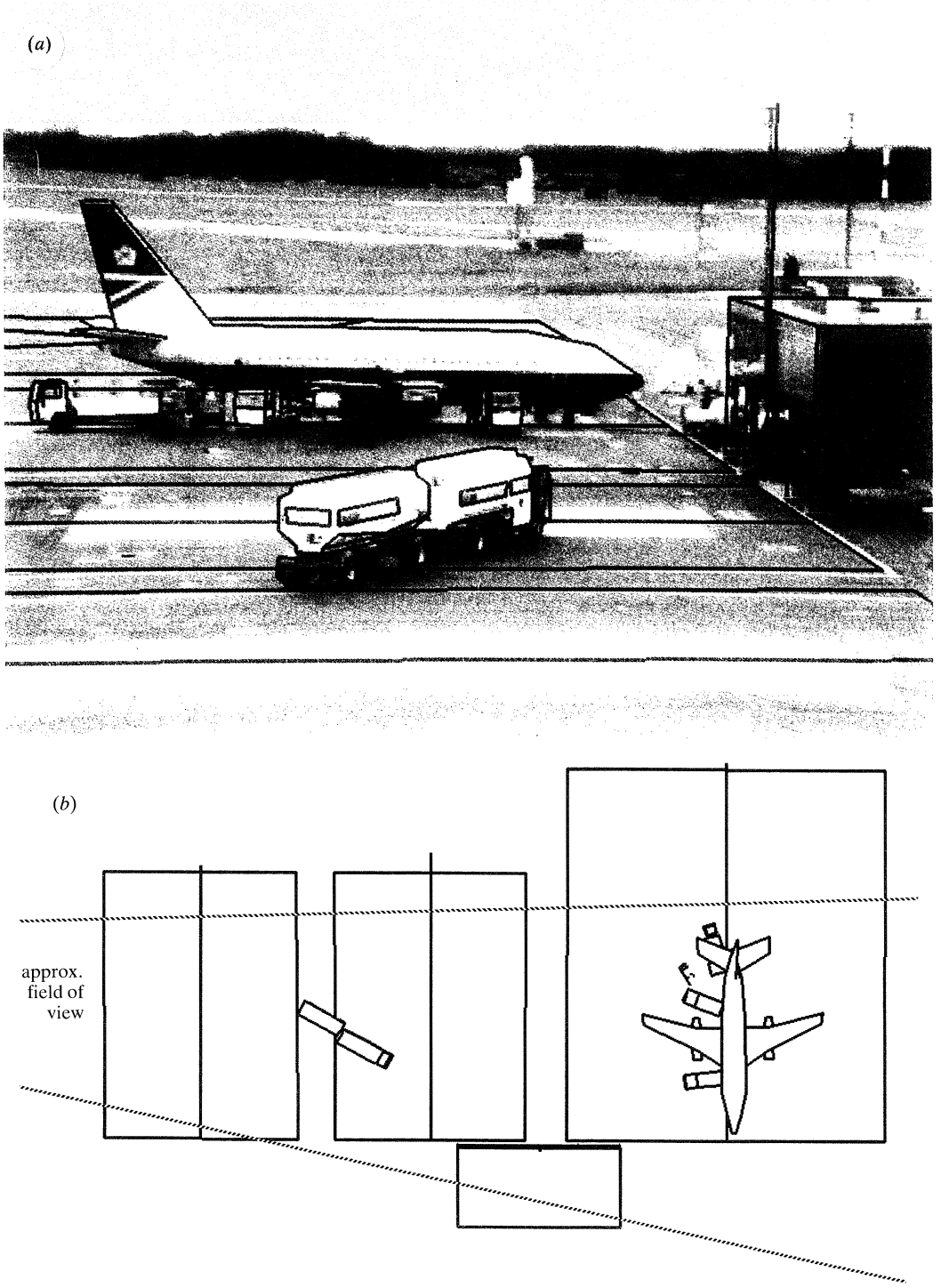


Figure 8. Interpretation of multiple objects in an airport scene. The aircraft, two baggage handlers, a fork-lift truck, a water cart, and a fuel-tanker and trailer, were all tracked successfully in the known scene: (*a*) shows the camera view, (*b*) the view from above.

loped within the Esprit VIEWS project as the perceptual input to systems which monitor vehicle movements in real-time for safety and scheduling purposes.

The key to success is clearly our ability to make use of detailed knowledge of the expected objects and their behaviour. The perceptual problem then becomes sufficiently constrained to be tractable. The 'top-down' use of *a priori* knowledge permits an hypothesis-driven approach to image analysis, and finesses many of the traditional problems in object recognition. In particular, all feature analysis and all grouping is carried out under the control of explicit perceptual hypotheses. In part, this has been an engineering expedient, since it avoids the need for unnecessary low-level vision, and on serial computers this leads to very significant savings in computing time. However, there would be major advantages if massively parallel computing power were available to perform more of the low-level processing without cost. For example, when carrying out the pose refinement process, the present system makes repeated evaluations of individual features which vary little in

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

position, length or orientation in the image. These calculations are one of the main computational costs of the system, and are highly redundant. It would clearly be preferable to pre-process the image to create an intermediate representation of the image data within which the individual hypothesised features could be evaluated more efficiently. Ideally we require a general-purpose representation of the first and second local derivatives in the image, having minimal nonlinearities, so that the data may be efficiently tested for specific features predicted by the hypothesis. Such a computationally intensive front-end to a machine vision system is not yet feasible, though developments in this direction are imminent.

It is interesting to note that our ideal low-level representation matches rather precisely the characteristics of the simple cells in the primary receiving areas of the natural vision system. Could this indicate that these cells play a similar role as an intermediate representation, intended to be interrogated by hypothesis-driven processing? The central problem in object recognition is surely the same for both natural and artificial vision: the discovery that myriad, ambiguous local measurements of the image are mutually consistent with one view (from all possible views) of one object (from all possible objects). A hierarchical, data-driven approach to vision such as was advocated by Marr (1982) must solve the problem of representing the object in a way that can be indexed reliably by means of 'knowledge-free' clusters of features. This presupposes universal invariants in vision (e.g. Forsyth *et al.* 1991), which can be extracted effectively under all viewing conditions, and for all different poses. One suggestion has been to represent objects as sets of characteristic views (Koenderink *et al.* 1979), each of which can be associated with specific feature invariants. However, without further reasoning such a recognition process can only yield a single estimate of pose for each characteristic view. Furthermore, it is difficult to see how this scheme could deal successfully with partial occlusion between models.

On the other hand, we have shown that the ability to manipulate analogue models of objects, provides an efficient (although implicit) means to reason about the image characteristics of any pose. It allows lighting conditions and occlusions, as well as vehicle dynamics, expected behaviours, and scene context to be taken into account. The concomitant problem is the need to invoke an approximately correct model initially; in known traffic scenes, simple unstructured movement in the image has proved sufficient.

Our experiments with traffic monitoring show that high-level concepts can assist the interpretation of dynamic scenes. It seems probable that natural vision makes use of similar constructs. Many observations point to the crucial role of high-level expectations in visual perception: the influence of figure-ground interpretation on low-level phenomena (Weisstein *et al.* 1987), the hollow face (Gregory 1973), the bias towards behavioural descriptions of movement (Heider & Simmel 1944; Johansson 1975), as well as our everyday perception of cartoons and fragmented line drawings, all point to high-level influences in (apparently) immediate vision. Recent experiments by Biederman reinforce this view: reaction times (RTs) for recognition of familiar objects are longer if the objects are presented out of context; priming effects (where RT is shorter if an object has already been seen) depend on high-level object representation, rather than on low-level image similarity. These effects would be expected of systems using analogue models of objects, dynamics and behaviours. Our demonstration of the active use of analogue models in object recognition by machine raises issues which may provoke a re-appraisal of mechanisms of natural vision. Humans perceive dynamic objects in their relations to the scene. Where, and how, is the geometrical reasoning carried out? How do perceptual hypotheses become evaluated? What cues are used to trigger the initial hypotheses?

The method for iconic evaluation we use to compare instances of models with the image has properties which also echo familiar perceptual phenomena. Each feature score is scaled according to its probability of arising by chance. At present out probability tables are isotropic and global. One of the defects of the system is that there is no competition between different models for image features, and a strong feature which is already explained can disrupt the pose of second model. Examples of this can be seen in figure 5, especially frames 300–350, where the lamp post and the occluding vehicle interfere with the tracked pose. One possible solution would be to introduce inhibitory influences between nearby features, as they become recognized. The competition between objects over a feature could then be implemented by manipulating the probability tables locally, and anisotropically. This would give rise to local interference effects, as in geometrical illusions and orientation-specific masking.

Finally, a continuing mystery in visual physiology concerns the role of the very extensive innervation of the primary receiving cortex from the parietal and temporal areas, providing feed-back from 'higher' to 'lower' visual centres. The crucial stage in object recognition is the act of linking stored concepts to the sensory data. Our experiments with machine vision demonstrate how successful active vision, driven by expectation, can be. Does natural vision make comparable use of 'top-down' methods? Descending innervation is required if perceptual hypotheses are to provide feedback to control the processing of an intermediate representation. Could this play a part in the process whereby an emerging perceptual context directly influences the analysis of low-level features?

This paper draws deeply on work carried out jointly with Professor Keith Baker, firstly, as part of the Alvey MMI-007 project, and currently the Esprit P2157 VIEWS project. Of the many colleagues who have contributed to the research, I wish particularly to acknowledge the invaluable role in both projects played by Dr Anthony Worrall.

REFERENCES

Ballard, D.H. & Brown, C.M. 1982 *Computer vision*. New Jersey: Prentice Hall.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Biederman, I. 1987 Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* **94**, 115–47.
Biederman, I. & Cooper, E. 1991 Priming contour-deleted images: evidence for intermediate representations in visual recognition. *Cogn. Psychol.* **23**, 393–419.
Brisdon, K.S., Sullivan, G.D. & Baker, K.D. 1988 Feature aggregation in iconic model matching. *Proc. Alvey Vision Conference, AVC-88, Manchester*, pp. 19–24.
Brisdon, K.S. 1990 Hypothesis verification using iconic matching. Ph.D. thesis, University of Reading.
Brooks, R.A. 1983 Model-based three-dimensional interpretations of two-dimensional images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 5, pp. 140–150.
Forsyth, D., Mundy, J.L., Zisserman, A., Coelho, C., Heller, A. & Rothwell, C. 1991 Invariant descriptions for 3-D object recognition and pose. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol 13, pp. 971–991.
Gregory, R.L. 1973 The confounded eye. In *Illusion in nature and art* (ed. R. L. Gregory & E. H. Gombrich). London: Duckworth.
Grimson, W.E.L. & Huttenlocher, D.P. 1990 On the verification of hypothesised matches in model-based recognition. *Proc. First European Conference on Computer Vision, Antibes, France.*
Heider, F. & Simmel, M. 1944 An experimental study of apparent behaviour. *Am. J. Psychol.* **57**, 243–259.
Hogg, D.C. 1983 Model-based vision: a program to see a walking person. *Image Vis. Comput.* Vol. 1, pp. 1–20.
Johansson, G. 1975 Visual motion perception. *Scient. Am.* **232**, 76–89.
Koenderink, J.J. & Van Doorn, A.J. 1979 The internal representation of solid shape with respect to vision. *Biol. Cybernet.* **32**, 211–216.
Lowe, D.G., 1985 *Perceptual organisation and visual recognition.* Boston: Kluwer Academic Publishers.
Lowe, D.G. 1991 Fitting Parameterized 3-D Models to Images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 13, no. 5, pp. 441–450.
Marslin, R.F., Sullivan, G.D. & Baker, K.D. 1991 Kalman filters in constrained model-based tracking. *Proc. British Machine Vision Conference*, pp. 371–374. London: Springer-Verlag.
Riseman, E.M. & Hanson, A.R. 1987 A methodology for the development of general knowledge-based vision systems. In *Vision, brain, and cooperative computation*, pp. 285–328. (ed. A. Arbib & A. R. Hanson). Cambridge, Massachusetts: MIT Press.
Weisstein, N. & Wong, E. 1987 Figure-ground organisation affects the early visual processing of information. In *Vision, brain, and cooperative computation*, pp. 209–230. (ed. A. Arbib & A. R. Hanson). Cambridge, Massachusetts: MIT Press.
Worrall, A.D., Marslin, R.F., Sullivan, G.D. & Baker, K.D. 1991 Model-based tracking. *Proc. British Machine Vision Conference*, pp. 310–318. London: Springer-Verlag.

Discussion

A. Sloman (*School of Computer Science, University of Birmingham, U.K.*). All simple slogans about how vision (or any other aspect of intelligence) works are wrong, even if they contain parts of the truth! Although the claim that much vision is driven top-down using prior knowledge must be correct, saying that it is all like that must be wrong, in so far as it cannot account for: (i) the perception of complex novel structures, e.g. turning a corner and seeing a large complex building using unfamiliar materials and architectural designs, or even seeing a new instance of a known type admitting enormous structural variability, e.g. an oak tree; and (ii) the combinatorics of seeing very flexible objects, e.g. a sweater thrown onto a chair.

So it would be more accurate to say that a visual system has both the power to work in data-driven mode, as far as it has to in coping with novel scenes, and also the power opportunistically (to short-circuit interpretation) to turn on a top-down model-driven process when the data permit this. What sort of architecture can facilitate this?

G. D. Sullivan. I accept the points made by Professor Sloman entirely; both data-driven and hypothesis-driven processes are required in vision. For a variety of reasons, recent work in vision has concentrated on data-driven image analysis. The main conclusion from our work is that where high-level knowledge is available many of the very difficult problems in low-level vision become manageable. In artificial vision, the advantages confered by using rather simple knowledge about traffic scene are enormous. The extreme alternative – of context-free interpretation of each object and event – not only seems impossibly difficult, but also fails to make sensible use of the perceptual understanding which is being built up. To me at least, it is unimaginable that natural vision fails to make full use of the perceived context.

Of course, discovering the semantic context in the first place remains a major problem, especially for the perception of novel or unpredictable structures. In the present state of artificial vision, such unconstrained tasks are impossible; they also continue to present unanswered challenges for theories of natural vision.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

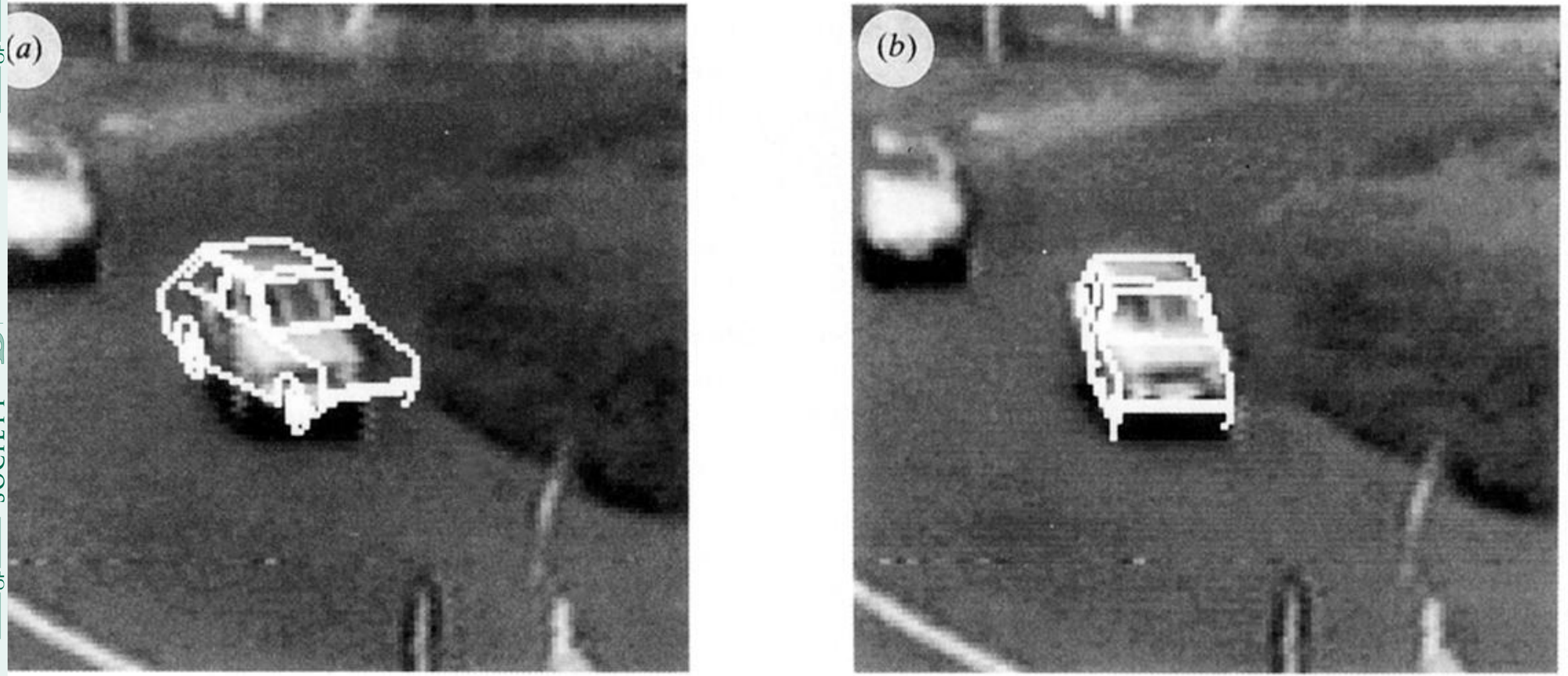
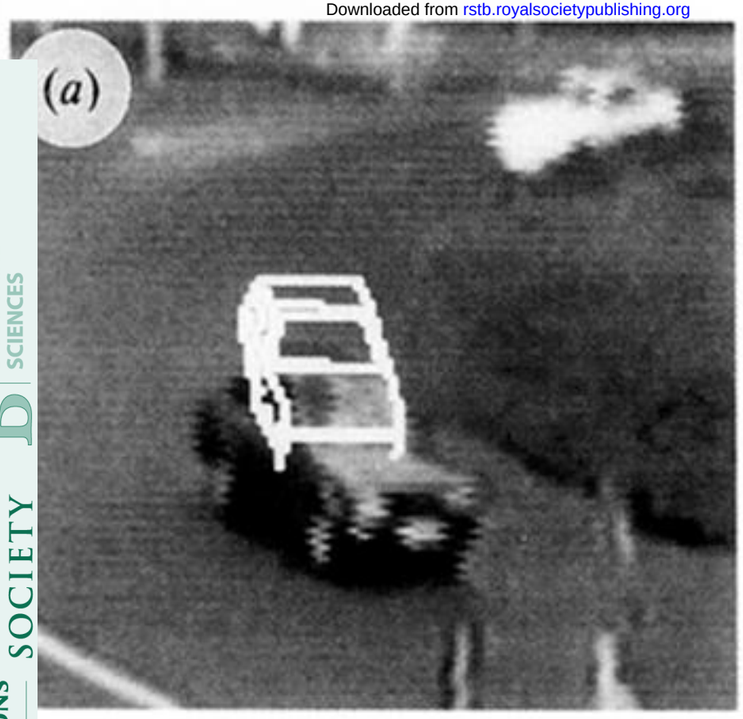
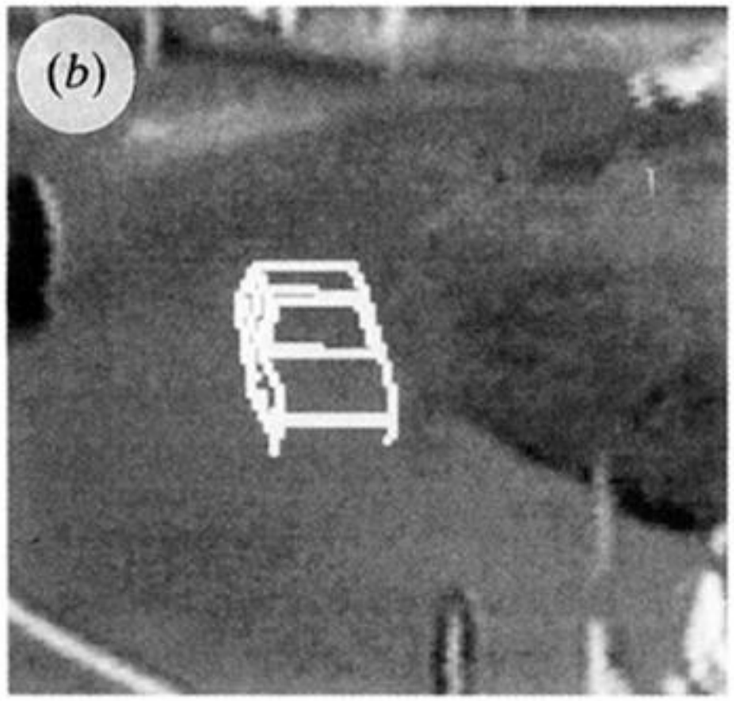
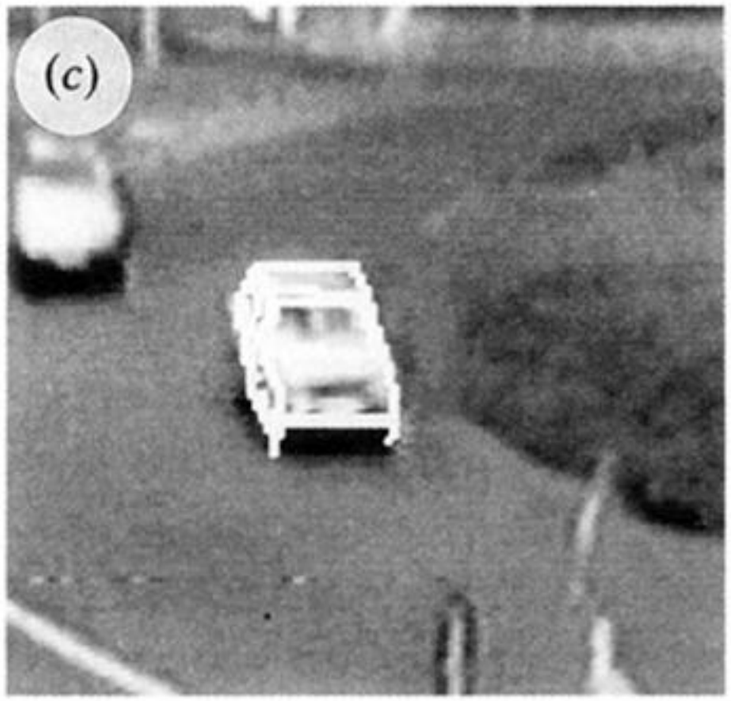
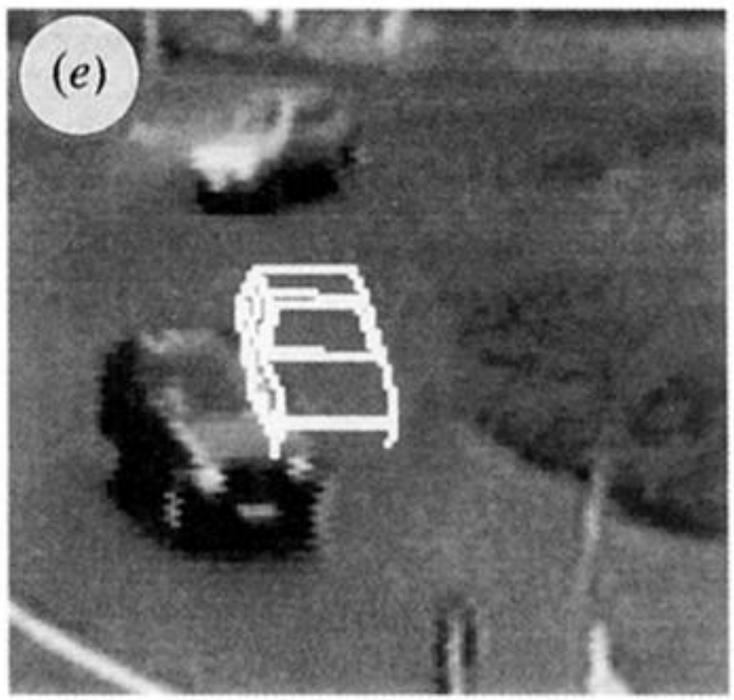


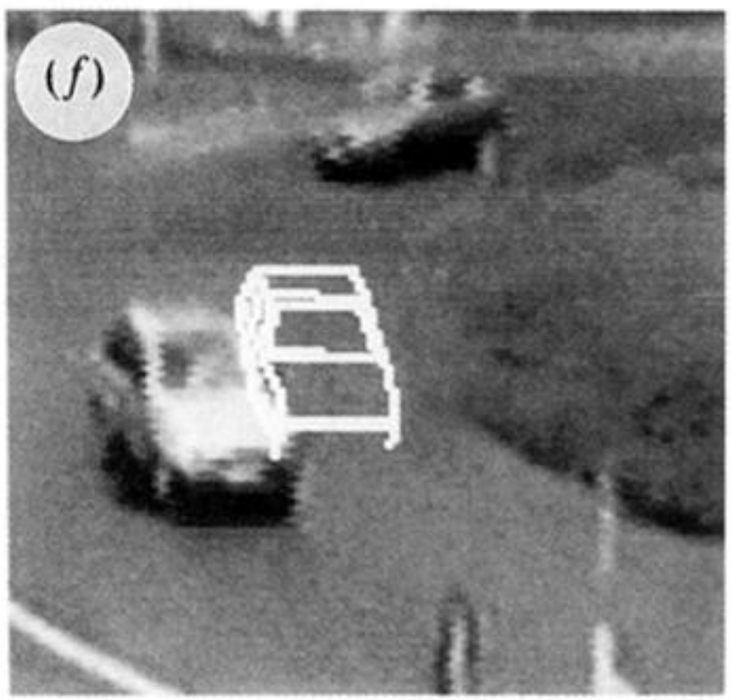
Figure 3. (*a*) Before and (*b*) after gradient ascent (frame 200).











gure 5. Fixed model in different frames ((*a–f*) refer to figure 4). (*a*), frame 0; (*b*), frame 100; (*c*), frame 200; (*d*) frame 273; (*e*) frame 351; and (*f*), frame 418.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Figure 7. Tracking a single car (every tenth frame from 80 to 350).

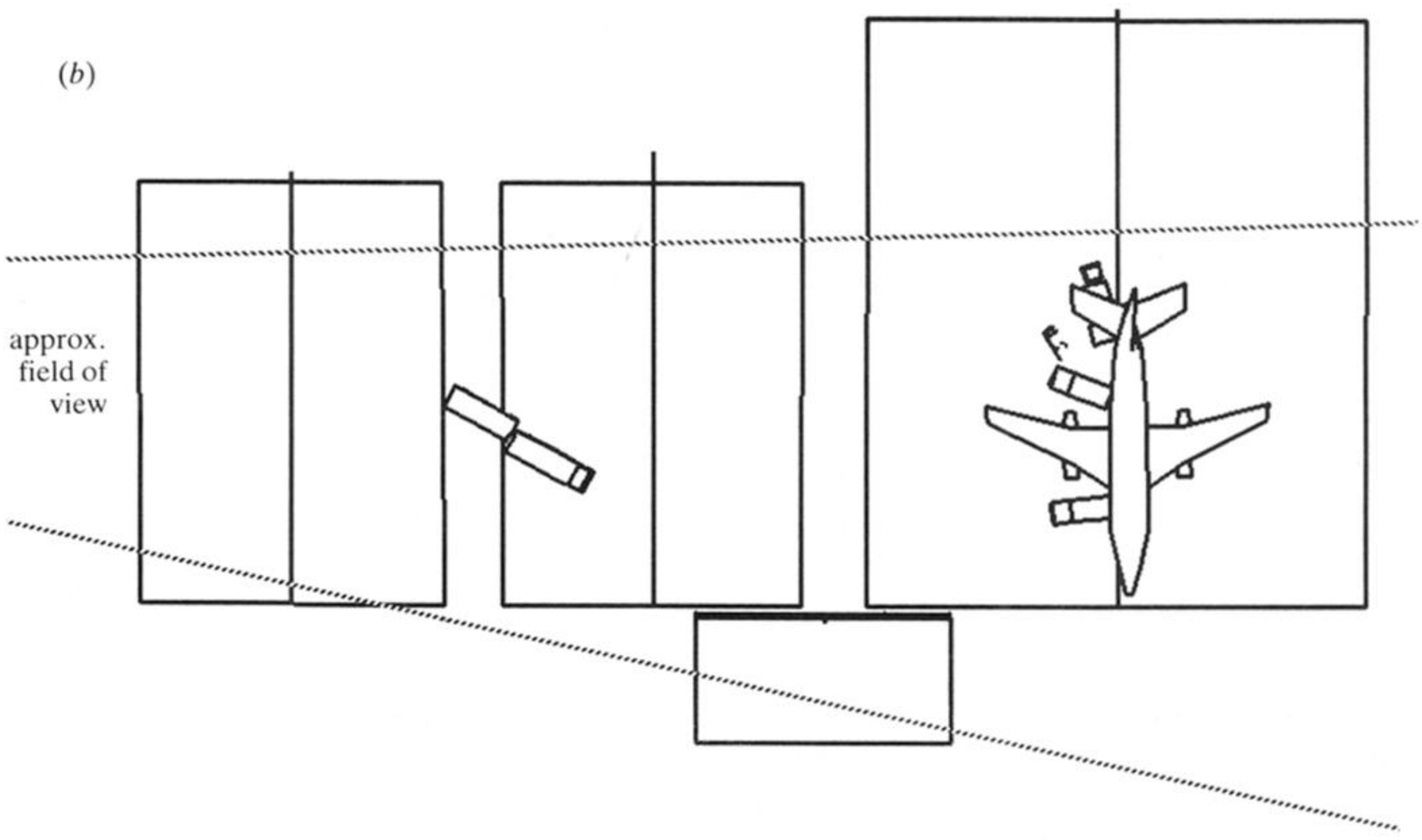


Figure 8. Interpretation of multiple objects in an airport scene. The aircraft, two baggage handlers, a fork-lift truck, a water cart, and a fuel-tanker and trailer, were all tracked successfully in the known scene: (*a*) shows the camera view, (*b*) the view from above.