Advances in Model-based Traffic Vision

A. D. Worrall, G. D. Sullivan, K. D. Baker

Intelligent Systems Group,
Department of Computer Science
University of Reading, RG6 2AY, U.K.
Anthony.Worrall@reading.ac.uk.

Abstract

Model based vision allows use of prior knowledge of the shape and appearance of specific objects to be used in the interpretation of a visual scene; it provides a powerful and natural way to enforce the view consistency constraint [1]. A model based vision system has been developed within ESPRIT VIEWS: P2152 which is able to classify and track moving objects (cars and other vehicles) in complex, cluttered traffic scenes. The fundamental basis of the method has been previously reported [2]. This paper presents recent developments which have extended the scope of the system to include (i) multiple cameras, (ii) variable camera geometry, and (iii) articulated objects. All three enhancements have easily been accommodated within the original model-based approach.

1 Review of methods

The models used consist of 3D geometrical representations of known objects (vehicles) together with calibrated camera and scene models [3]. Using the known camera and scene geometry, and given a provisional position and orientation (derived from data-driven detection of temporal change in the image), a 3D object can be instantiated into the 2D image plane and a "goodness-of-fit" score obtained by comparing the modelled features with the image. An iterative search in position-space and orientation-space is then used to maximize this evaluation score. At each step in the search the model is re-instantiated into the scene and a new goodness-of-fit score evaluated.

1.1 Evaluation of the "goodness-of-fit"

The vehicle model comprises a set of line features specified in a 3D object-centred coordinate reference frame. On instantiation the model is translated and rotated to the appropriate position in the world coordinate frame and finally all lines which are visible from the given camera position are projected onto the image. Each visible line is evaluated using methods similar to the one that was originally reported [6][7][8]. The scores from the individual lines are aggregated to give an overall goodness-of-fit score for the model in the given position. Evidence from

BMVC 1993 doi:10.5244/C.7.56

each line is assumed to be independent, and the aggregate score is given by:

$$E = \sqrt{-4\sum_{\text{Vis}} ln\,(p_i)} - \sqrt{4N_{\text{Vis}} - 1} \qquad \text{(EQ 1)}$$

where p_i is the probability that a score at least as high would have been obtained by randomly placing a similar feature in the scene (as determined by empirical trials, using an image of the scene with no vehicle), and N_{Vis} is the number of visible lines.

The main advantage of this approach is that under assumptions of independence and "large" numbers E has a χ^2 distribution, so that, assuming a correct position for the model, the value obtained is largely independent of the pose of the object and the number of visible lines. This technique was previously reported in [2], with a number of recent changes reported in [4] and [5].

1.2 Pose recovery

The evaluation score defines a scalar function of 6 dimensions - in world coordinates these are most simply defined as the three cartesian coordinates of the object's position and the three angles needed to specify its orientation. In general, we expect that peaks in the 6 df 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 a vehicle.

A considerable computational saving can be made by limiting the object's position to the ground plane, thus only permitting 2D translation and rotation about the vertical axis. Using this simple but realistic physical "ground-plane constraint", only three independent dimensions remain. Even so, an exhaustive search of three dimensions is computationally too expensive, and a number of alternative methods have been used, including: simplex, gradient ascent, and a method which successively performs three local one dimensional searches in the object coordinate system, determined by the best value yet found.

In order to use the ground-plane constraint we need to determine the extrinsic camera parameters, as well as the intrinsic parameters needed in the more general case. This camera calibration determines the pose of the camera with respect to a fixed world coordinate system. The calibration is achieved by matching points in the image with points in a 3D model of the static scene by eye.

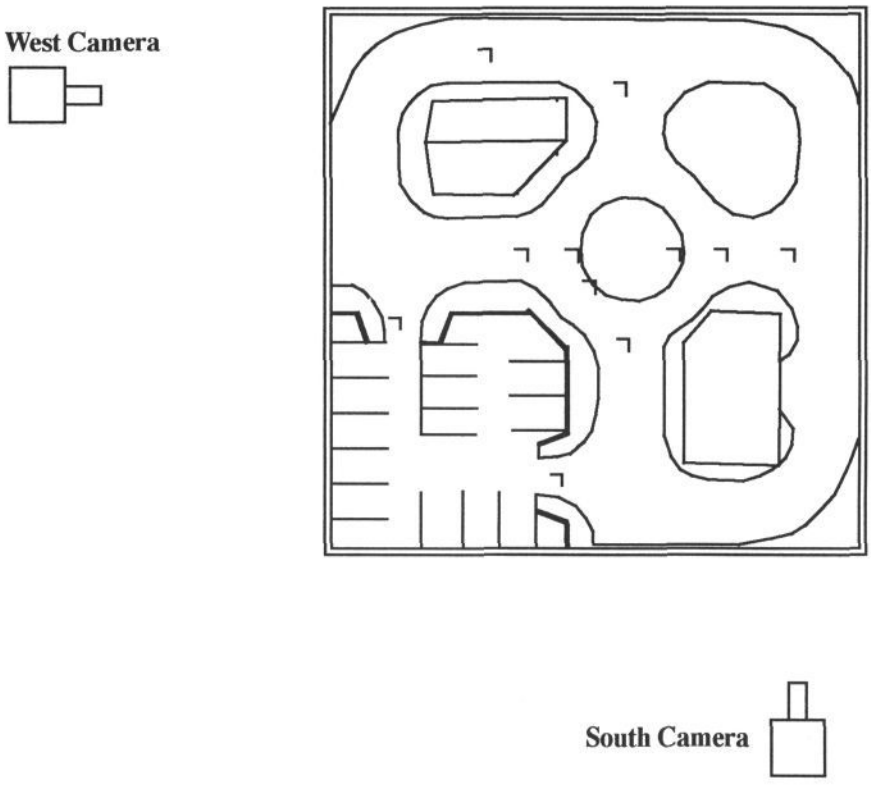


Figure 1: Plan view of the roadway scene. One camera is located to the "south" of the scene and the other camera is located to the "west". The "L" shapes correspond to calibration marks on the roadway.

2 Multiple Cameras

Our previous work only considered a single calibrated camera, but the method is readily extended to multiple calibrated cameras with overlapping fields of view. The classic problem with multiple cameras is how to fuse the information from each image. In the model-based paradigm this problem is extremely simple. We hypothesize a pose for the model and calculate all the visible lines for each of the camera views, and then combine the probabilities from all visible lines as in the case of a single camera (EQ 1).

If the object is outside the field of view of a camera or occluded by a known object (e.g. a building) the final value is largely unaffected (provided the model is visible from at least one view). If the object is visible from two or more substantially different views then accuracy and robustness of the system is greatly increased, since: (i) more image evidence is available to compute E, (ii) the evaluation scores are commonly strongly ridged, in the approximate direction of

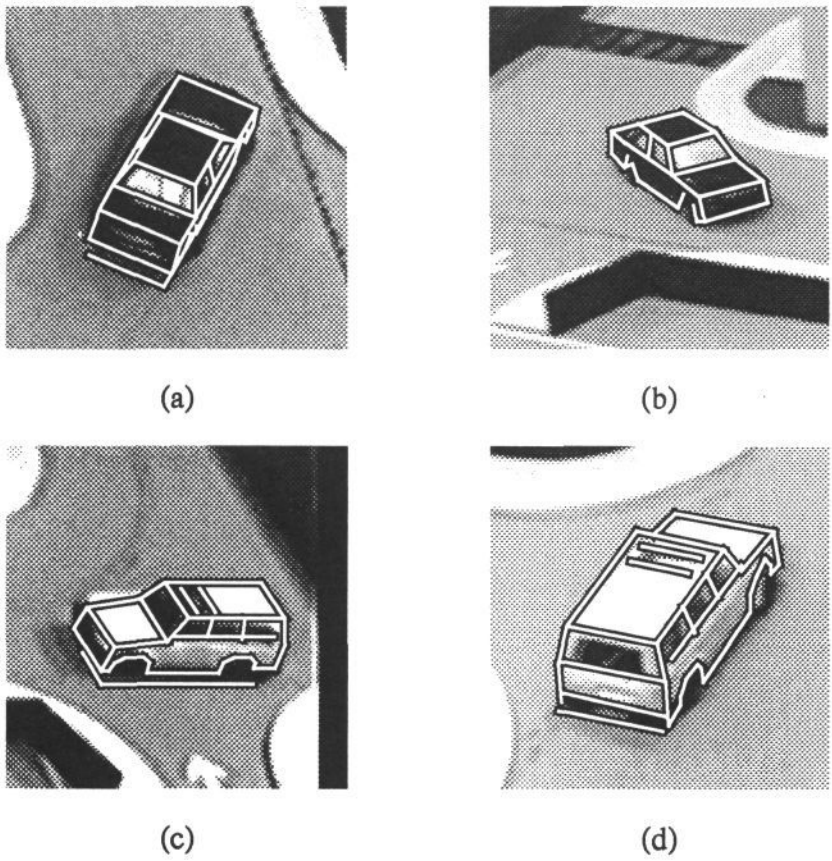


Figure 2: The location of the saloon car from (a) the south camera and (b) the west camera and the estate car from (c) the south camera and (d) the west camera.

the camera and multiple cameras provide scores which are ridged in different directions, and (iii) accidental alignment of the model with distracting features in the scene are unlikely to occur simultaneously in all cameras.

To illustrate the method, we have used an experimental set up consisting of a 24th scale roadway scene containing buildings, a car parking area and a roundabout (see Figure 1). Two radio-controlled cars, a saloon and an estate, can be driven around the scene. The scene is observed by two monochrome cameras connected to the red and green channels of a 24bit frame store. A pair of images were captured with the two cars in view of both cameras. Close-ups of the two cars from each view are shown in Figure 2. The poses of the cars shown in Figure 2 were initially determined by using a voting procedure to identify matches between lines extracted from the south camera and 3D model lines [9].

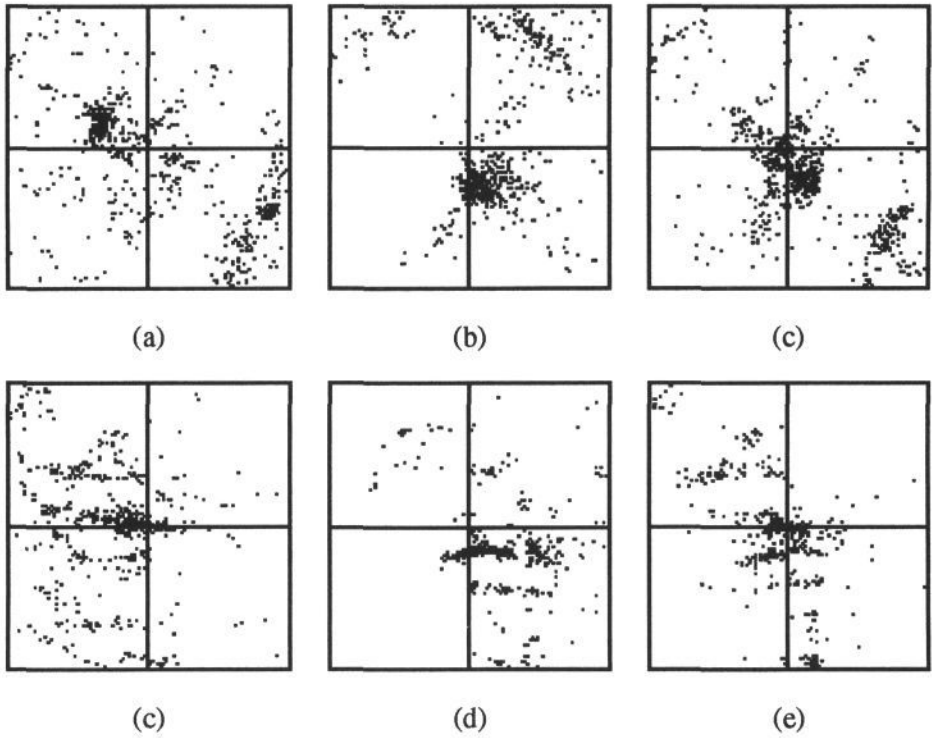


Figure 3: The X-Y location of the pose refinement: Saloon car using (a) the south camera, (b) the west camera and (c) both cameras. Estate car using (d) the south camera, (e) the west camera and (e) both cameras.

These initial poses were used as the basis for tests on the pose refinement. A grid of size ±2.5m in X and Y and ±25 degrees in orientation was centred about the initial poses. Each point on the grid was used as the start pose for the pose refinement under three conditions; using the south camera, the west camera or both cameras. The final X-Y location of these pose refinements are show in Figure 3. Ideally we would hope that all starting poses converge on the origin in X and Y. In practice, the points are scattered within the X-Y plane, with clusters occurring at many stable poses. In these examples the clusters for the two camera case seems significantly better than for either camera alone.

3 On-line Camera Calibration

The transformation between the model and the image is controlled by three matrices; the transformation between the model and the world, the transformation between the world and the camera, and the projection from the camera to the

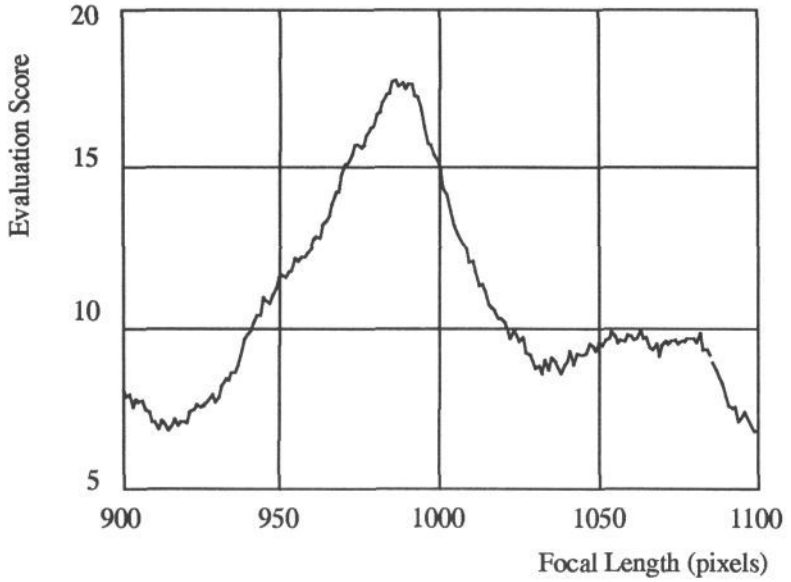


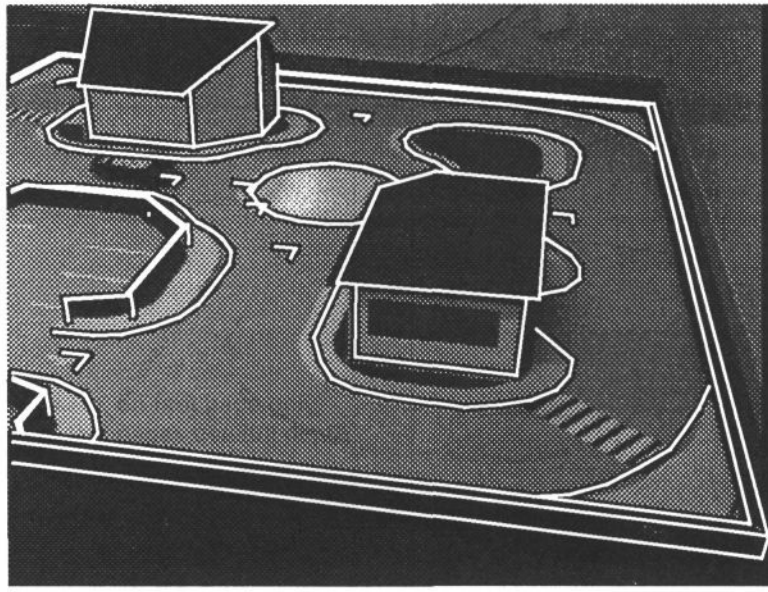
Figure 4: Variation of the evaluation score for the scene model as the focal length of the camera model is changed.

image. This combination can be written using left acting operators (matrices) as:

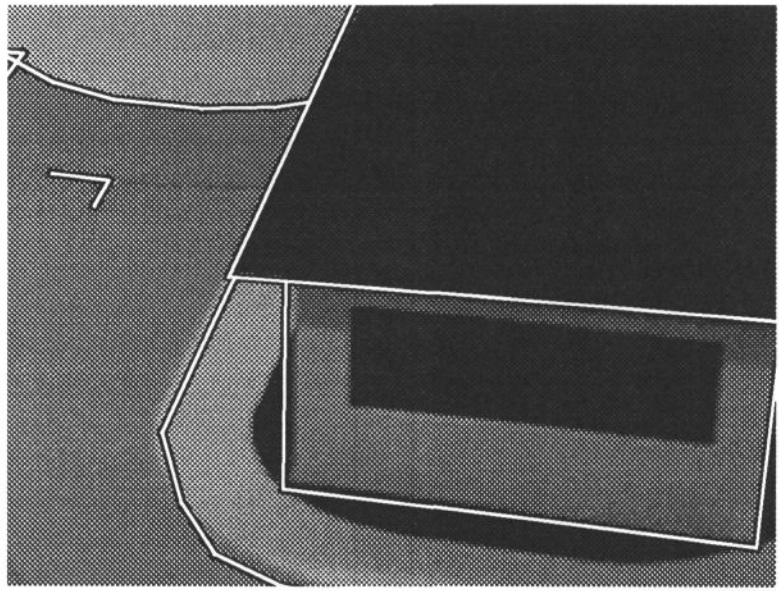
$$M_{mw}M_{wc}P_{ci}. \qquad \text{(EQ 2)}$$

Our earlier work assumed that only the model-to-world transformation varies and the other two are fixed. But the model-based approach also allows us to fix the model-to-world transformation and vary the other two. The scene model represents buildings and marks on the ground that are not expected to change in the world coordinate frame. The scene model can be evaluated in the same way as for the vehicle models in order to determine a "goodness-of-fit" for the camera model. The intrinsic or extrinsic parameters can then be changed in order to maximise this score. This allows the camera parameters for a "fixed" camera to be refined or the parameters of moving cameras to be recovered on-line. Typical results for the variation in the evaluation score for the scene model when the focal length of the camera model is changed are shown in Figure 4.

In order to obtain sufficient resolution with current video images, practical surveillance systems need to use cameras with variable pan, tilt and zoom. These parameters (or rather the matrix they represent) must be known for the model-based system to work. The pan, tilt and zoom camera parameters can determined, by searching over the evaluation score obtained for the stationary scene model. The results obtained when using this method to track the camera focal length



(a)



(b)

Figure 5: Dynamic tracking of the scene model to recover focal length of a camera with variable zoom. (a) The scene model superimposed on the initial wide-angle view and (b) the scene after the camera has been zoomed in.

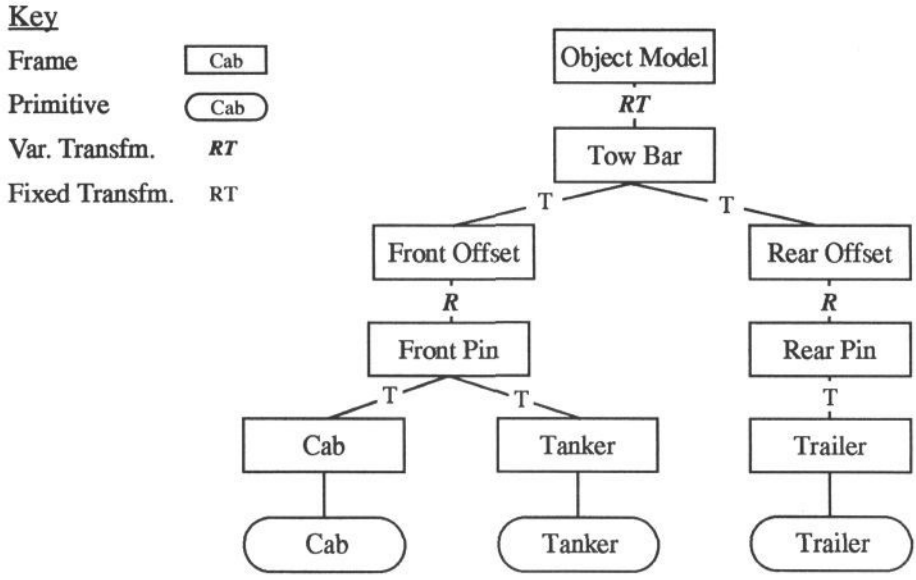


Figure 6: An example of the construction tree for a model of a fuel tanker and it's trailer. The links are the transformations of Rotation (R) and Translation (T). The leaf nodes are model primitives and the other nodes represent coordinate frames.

under conditions of variable zoom are shown in Figure 5. The tracking was started with a wide-angle setting and the camera was then manually zoomed in to the final position, whilst the focal length was tracked automatically.

4 Articulated Models

The vehicle models used in the system are comprised of primitives expressed as leaf nodes in a tree structure where the other nodes of the tree are separate coordinate frames, and the arcs are transformations between these coordinate frames (see Figure 6). In earlier work, these internal transformations were kept fixed, to produce a rigid model of an object. However, the system is easily adapted to take account of articulated vehicles, such as a lorry and trailer. For example, to model an articulated lorry, separate rigid models for the cab and trailer can be produced and then connected with a transformation which has one degree of freedom of rotation about the vertical axis. The pose of the articulated lorry can then be obtained by searching over the extra parameter of articulation as well as the three previous parameters (x, y, θ). In the airport example considered in VIEWS the fuel tanker and its trailer are connected by a tow bar with two articulated couplings. This means the vehicle has two extra degrees of freedom.

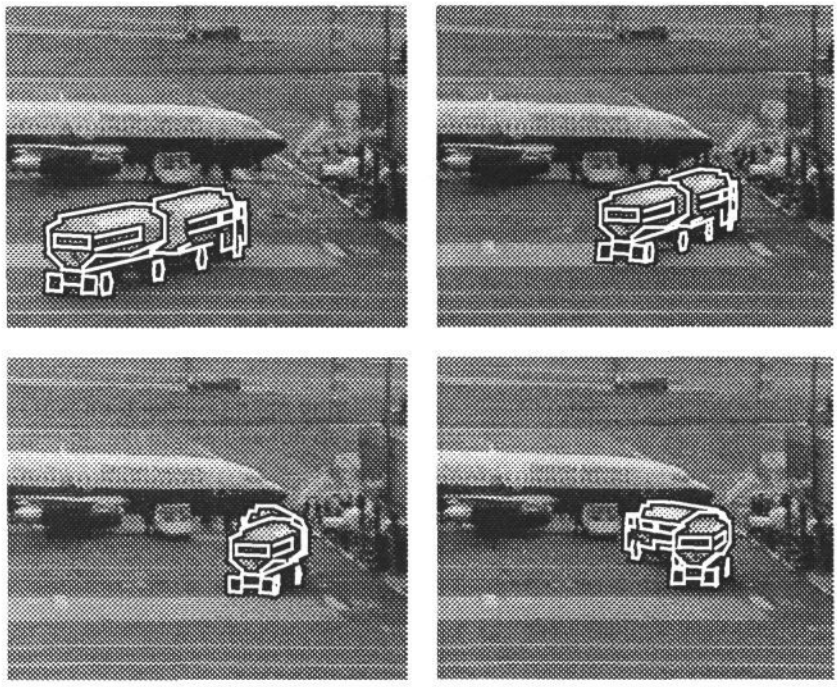


Figure 7: Tracking a fuel tanker and trailer with an articulated model as it turns to come alongside the waiting aircraft.

The result of tracking the tanker and trailer with this articulation is shown in Figure 7.

5 Current status

The model-based approach to traffic understanding adopted in VIEWS has proved rich and powerful. Many of the traditional problems in object recognition prove to be tractable, since the vision system reasons directly in 3D space rather than with the weak invariants of objects found in 2D images. Shape and size constancy, colour and brightness invariance, partial occlusion, dynamic tracking, multiple cameras, moving cameras, and articulated objects have all proved simple to implement, within the one all-embracing paradigm.

The traffic understanding system developed in VIEWS currently exists in several different versions. Complex scenes (airports and road junctions) containing multiple vehicles can, at present, only be processed off-line, using a video disc to synchronise the inflow of images to the processing speed of the system. An initial stage of movement detection and region segmentation,

followed by morphological filtering and tracking, is carried out on a SUN 10/30 at speeds approaching real-time, using images sequences sampled at 5Hz. [Details of this work, carried out by GEC-Marconi, is available as an internal project report.] The model-based pose refinement described here currently takes about 200-500 ms per model, running on a SUN 10/41. A further stage of "arbitration" between the image-based and model-based results introduces an additional minor burden. The entire system, consisting of two SUN 10 workstations with a further SUN acting as the display host, is able to cope with scenes containing about 5 vehicles at a rate only about 10 times slower than real-time (using images sampled at 5Hz).

A second version of the system has been developed to demonstrate the "live" processing, using the purpose built 1/24 scale model road scene, and radio controlled cars previously described. The system has been simplified to optimise processing speed, by using a simpler movement detection system, and minimal vehicle models. The demonstration is capable of tracking a single vehicle in real-time (at ~5 Hz) on a single SUN 10/41 computer.

6 References

[1] Lowe D.G. "The Viewpoint Constancy Constraint" International Journal of Computer Vision Vol. 1 (1987) pp 57-72.

[2] Worrall A.D., Marslin R.F., Sullivan G.D., & Baker K.D. "Model-based Tracking" BMVC-91, Glasgow, 1991, pp310-318.

[3] Sullivan G.D., Worrall A.D. and Marslin R.F. "R121/1: Knowledge-based Image Processing: Model-based Method" VIEWS report PM-03-CEC-TR.R121/1-01 1990.

[4] Sullivan G.D. "Visual Interpretation of Known Objects in Constrained Scenes", Phil. Trans. R. Soc. Lond. B (1992) 337, pp361-370

[5] Baker K.D and Sullivan G.D. "Performance Assessment of Model-based Vision", IEEE Workshop on Applications of Computer Vision, 1992.

[6] Brisdon K. "Alvey MMI-007 Vehicle Exemplar: Evaluation and Verification of Model Instances", Proc Alvey Vision Conference, AVC-87, Cambridge, 1987, pp33-37.

[7] Brisdon K., Sullivan G.D. & Baker K.D. "Feature Aggregation n Iconic Model Matching", Proc Alvey Vision Conference, AVC-88, Manchester, 1988, pp19-24.

[8] Brisdon K., "Hypothesis Verification using Iconic Matching" Doctoral Thesis, University of Reading, November 1990.

[9] Tan T.N., Sullivan G. D., and Baker K. D. "Recognising Objects on the Ground Plane", BMVC-93, Surrey, 1993