Pose and structure recovery using active models

A D Worrall, J M Ferryman, G D Sullivan and K D Baker
Department of Computer Science
The University of Reading
email: Anthony.Worrall@reading.ac.uk

Abstract

A new formulation of a pose refinement technique using "active" models is described. An error term derived from the detection of image derivatives close to an initial object hypothesis is linearised and solved by least squares. The method is particularly well suited to problems involving external geometrical constraints (such as the ground-plane constraint). We show that the method is able to recover both the pose of a rigid model, and the structure of a deformable model. We report an initial assessment of the performance and cost of pose and structure recovery using the active model in comparison with our previously reported "passive" model-based techniques in the context of traffic surveillance. The new method is more stable, and requires fewer iterations, especially when the number of free parameters increases, but shows somewhat poorer convergence.

1 Introduction

We have previously demonstrated a system for recognising and tracking vehicles in complex traffic scenes, using model-based methods [6]. The system relies on evaluating a "pose hypothesis" by aggregating hypothesis-dependent image evidence. We project a model of a vehicle (under a given pose hypothesis) into the image in the form of a wire-frame drawing (with hidden line removal), and accumulate evidence from image derivatives nearby and perpendicular to the "wires" to form a scalar evaluation score. The pose hypothesis is then refined by carrying out a search for a local optimum in the evaluation function. Thus starting from a "seed pose", cued either by movement analysis [4] or a form of pose voting based on local image derivatives [7] [8], we obtain the best local pose. Once the type and pose of a vehicle has been identified it can be tracked through sequences of video images, by means of simple dynamic filtering [10].

In the case of road-traffic viewed from a static camera, we can calibrate the camera with respect to the scene, and impose the "ground plane constraint" (GPC) - the fact that vehicles usually stand on the roadway. Vehicle models are only allowed to move with three degrees of freedom - translations (X,Y) and rotation (θ) with respect to an arbitrary world coordinate system, having the XY plane coincident with the ground [8]. A search of the evaluation function over these 3 pose parameters then provides an effective and efficient way to obtain the best pose for a given vehicle model.

We report here a more direct way to carry out a similar operation. This uses the seed pose of the object hypothesis to search for prominent edge evidence close to the

BMVC 1995 doi:10.5244/C.9.14

projected wires (as before), but rather than accumulating evidence in favour of the hypothesis, we determine an elemental error term based on the angular distance between the image edge and the predicted model edge. Each elemental error is approximated as a linear equation of the configuration parameters local to the hypothesis. These are solved in the conventional way to find a configuration having minimum least squares residual error. The algorithm then updates its use of image evidence according to the new hypothesis, and iterates.

In the first part of this paper we compare the properties of the new active method with the earlier passive method for recovering the pose of a rigid object, by means of exhaustive perturbation trials about an assumed "ground-truth".

In a companion paper [3] we report the development of a 6 degree of freedom (df) deformable model of vehicles, based on a statistical analysis of different vehicle classes, which is able to account for most of the variation in commonly encountered cars. We show here that both the passive technique, and the new active technique for pose recovery can be modified to operate on the 6 PCA parameters to allow the structure of the initial vehicle hypothesis to be adapted to the prevailing context. We then examine the relative costs and benefits of the active and passive methods for the refinement of both pose and structure.

2 Pose recovery

Both pose recovery techniques use a hypothesis and test strategy similar to that of Lowe [1], but adapted to make better use of external constraints by ourselves [9]. A given pose hypothesis is projected into the image to form a set of lines, which are examined independently. The normal to a line (in the image) defines a direction in which high values of image derivatives would provide evidence for the line. Points of local derivative maxima are found for normals spaced along the line. This evidence may be used in two different ways.

Our earlier work pooled the strengths of the derivative maxima for each projected model line to obtain an evaluation score for the line. The score was expressed as a probability of obtaining such a result, given a random placement of a line of this length, using probability tables previously constructed for that scene by Monte Carlo techniques. The probabilities associated with each visible line of the model instance were treated as independent, pooled, and compared against a chi-squared distribution (see [6]). The final evaluation score obtained proves to be reasonably independent of the pose, the object and the scene. Starting from a seed pose, the pose space may therefore be searched (using the simplex [2], or other algorithm) to determine the local optimum pose, and the evaluation score so obtained gives an absolute measure of the quality of the hypothesis.

Our new approach uses the position of a prominent image derivative to determine a local error term, the sum of which (over all normals along all visible projected lines) is to be minimised. The error term is given by the angle between the ray to the image edge and the plane of projection of the model line. The known structure of the model allows this to be expressed in terms of the model's configuration parameters (X,Y,θ), and after first order approximation, each measurement determines a linear equation. The over-determined system of equations is then solved using standard methods to minimise the least square error over all normals to all lines. The new pose is evaluated (as above) and

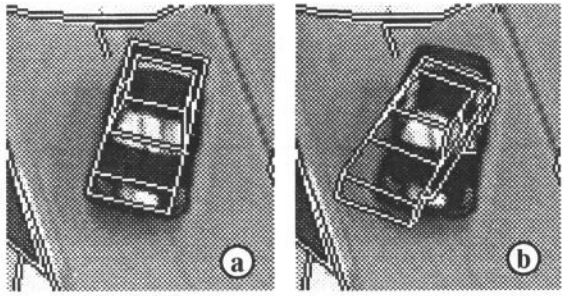

Figure 1 Laboratory test scene showing a close up of a 1:24 scale toy car: (a) with the "true" pose superimposed (b) at a perturbed initial pose displaced from (a) by 12.5°, and 0.5m, 0.5m (scaled).

the whole process is iterated - each time repeating the search for image evidence - until the evaluation score (derived as for the passive model) fails to improve. Brief mathematical details are given in the Appendix.

3 Convergence tests for pose recovery

The performance of the two methods has been compared by using a laboratory test scene. A close up (approximately 100 *120 pixels) of a radio-controlled car is shown in Figure 1(a), with the assumed true pose of the model superimposed. The test involves displacing the model by a given amount (as illustrated in Figure 1(b)) to provide a seed pose, and allowing one or other of the pose recovery algorithms to run to completion.

Figure 2(a, b) show results of the two methods for an exhaustive set of seed poses, using 11 values each for X, Y and θ, ranging over ± 1.0 m, ± 1.0 m (at 1:24 scale) and $\pm 25°$ respectively. The diagram on the left in each pair shows each pose as a short vector of appropriate orientation, at the appropriate location (note that where two recovered poses are identical, then they merge into one vector). The diagram on the right in each pair shows a histogram of poses distributed across (X,Y) and collapsed across θ.

It can be seen that both algorithms converge fairly well to the assumed ground truth at the centre of the diagrams. In general, the passive search seems somewhat superior in its ability to recover from large initial errors, and avoid local aliases; this is no doubt due to the long range searching used in the simplex algorithm. On the other hand, the active method appears more stable (note the stronger tendency to fall within one histogram bin) but is more sensitive to being trapped in local maxima.

However, the main advantage of the active technique is that it involves far fewer iterations. In the case of the passive model, using a simplex search, the actual number of iterations is not directly controlled, but here was typically 60-90. The results for the active method are shown after 10 and 25 iterations. Note that relatively minor changes occur after 10 iterations of the active method, and at that point its performance seems comparable to the passive method. We have not yet had the opportunity to implement the active method with as much care to efficiency as with the earlier, passive method, so a direct comparison of the computing speeds is not possible.

4 Shape recovery

In a companion paper [3] we report a highly parameterised vehicle model, able to describe buses, cars, lorries, etc. We then show how this can be specialised by means of principal component analysis to create a deformable model of the class of cars. The general car

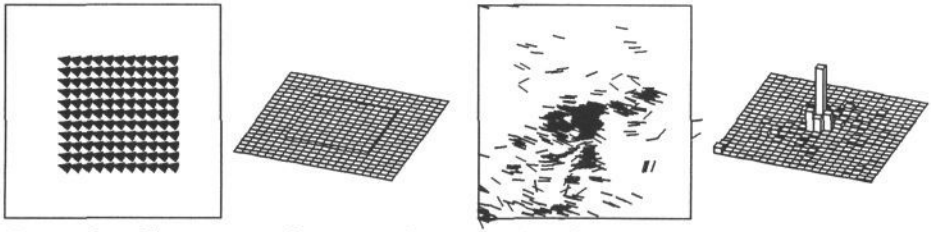

Figure 2a Convergence for pose refinement using the passive method.
Left: Starting positions as needle diagrams (left) and as histograms collapsed across orientation (right).
Right: Resulting poses, after passively optimising the evaluation score, using the simplex algorithm. Typical number of iterations is 60-90.

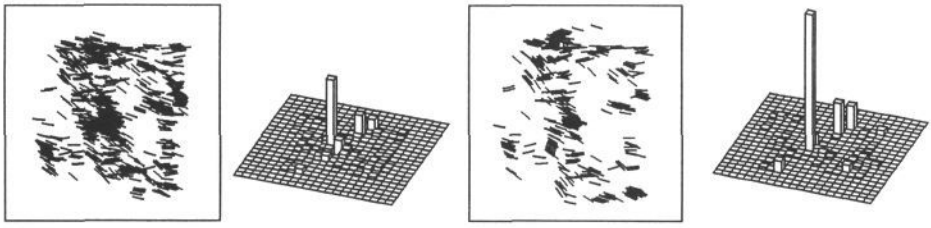

Figure 2b Convergence for pose refinement using the inverse method.
Left: After 10 iterations. Right: After 25 iterations.

model comprises a mean vehicle, which can be deformed along 6 PCA coordinates to describe sub-classes of estate, hatchback and saloon cars.

Both the passive and active methods of pose refinement can readily be adapted to the task of recovering the PCA structure parameters. In the case of the passive model, we merely use the simplex algorithm to search over the 6 structure parameters. In the case of the active model, the local error measurements can be expressed in terms of the structure parameters (see Appendix). Both methods may also combine pose and structure refinement in one system, but it is not clear that this is the best strategy. Instead, we have found it more satisfactory to alternate stages of pose and structure refinement, so that the parameters controlling the search for image evidence of each stage can be controlled explicitly.

Figure 3 illustrates the performance of the two methods. For each of three vehicles (top row), the second row shows the mean generic car model placed by hand onto the image near the vehicle. The third row shows the results after refining the pose and structure of the model using the passive method (using simplex search), and the bottom row shows the results using the active model. It can be seen that both methods converge towards a better description of the object, though the results are a little variable. The companion paper [3] reports the use of the recovered structure parameters to classify vehicles as hatchback, saloon or estate car.

5 Pose refinement using adapted shape

A major objective of the present work is to develop a generic car model whose structure can be adapted to fit the image. Such a "bespoke" model should provide greatly superior

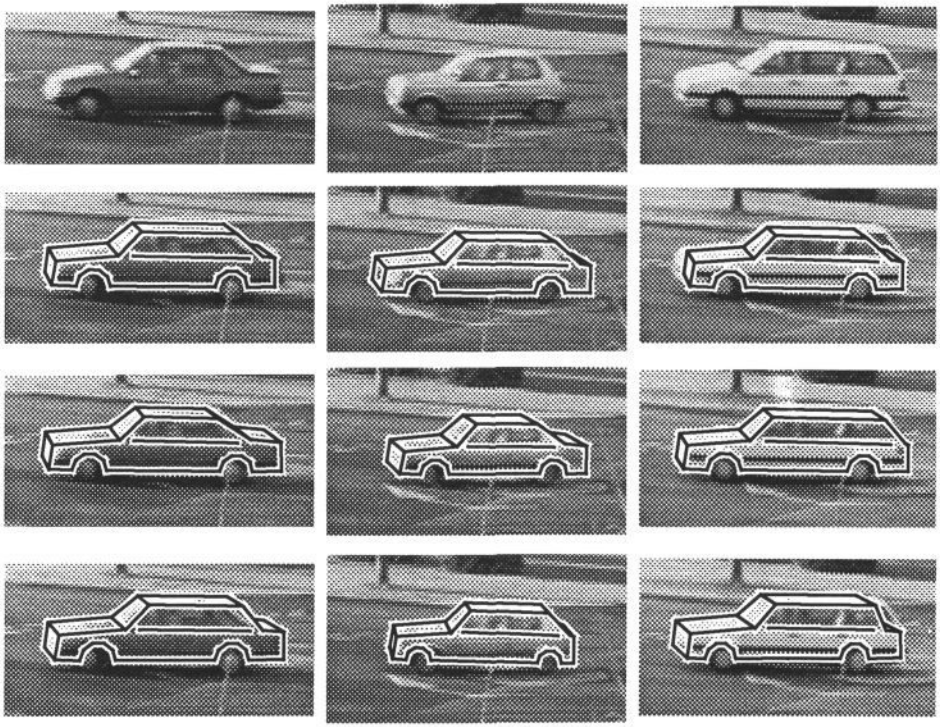

Figure 3 Pose and structure refinement for three different car sub-classes (top) from the generic car mean model (row 2) using either the passive method (row3) or the active method (bottom).

performance in model-based tracking. It is difficult to compare tracking systems, since performance is very sensitive to many minor factors and uncontrolled sources of noise. However, an essential requirement is that the pose refinement process should be stable.

Figures 4 (top) illustrate three different instances of the 6 df PCA model fitted to examples of the three sub-classes, which were not part of the training set. [These fits were obtained in a slightly different way from those in Figure 3.] Using these values of the deformation parameters to define a rigid object, we then carried out convergence tests as detailed in Section 3, using the active method described in Section 2. The results are shown in Figures 4 (middle row).

For the purposes of comparison, we also carried out equivalent tests using the mean generic model as a rigid object, with the results shown in Figure 4 (bottom). It is clear that pose refinement performance using the structurally fitted model is significantly superior to that using the mean class model. This indicates that superior tracking performance should be possible.

Figure 5 expands on the experiment illustrated in Figure 4 for the saloon car example. The results of the convergence test are shown as a needle diagram (centre), flanked by two views of the same histogram. Four of the attractors (poses to which many initial positions converge) have been picked out, and are illustrated superimposed on the image. The top and bottom needle diagrams show examples of the equivalence classes determined by the pose refinement algorithm, i.e. the starting poses that converge to these

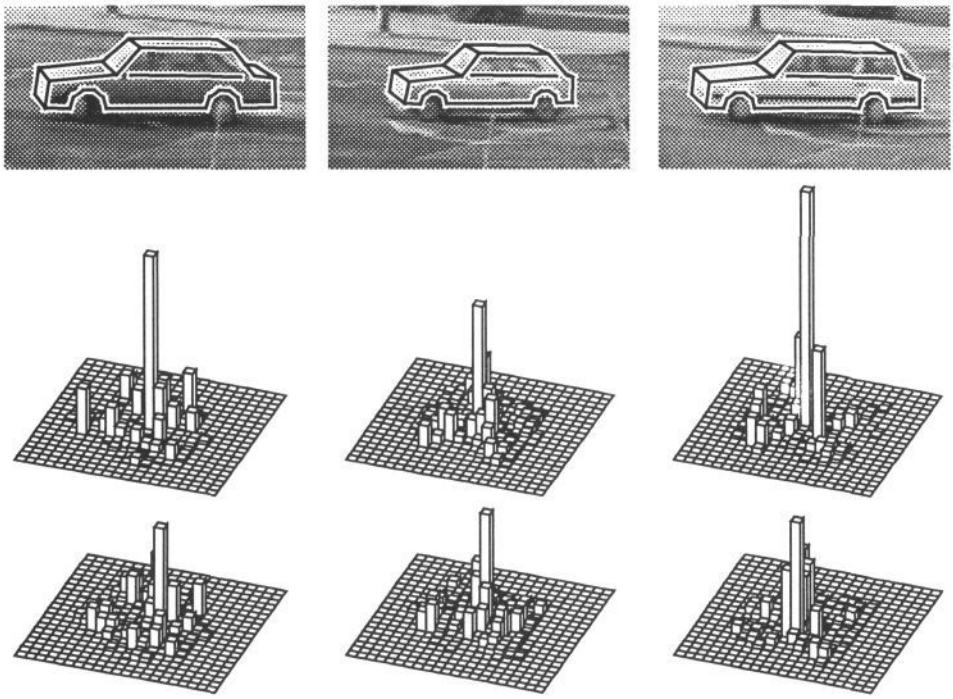

Figure 4 Instances of three sub-classes described by the deformable model (top). Middle: Convergence results using the fitted models. Bottom: Convergence results using the mean generic model.

four attractors. It should be noted that the top right and bottom left cases are actually very close to each other (in X,Y) and mainly differ in θ ; these are too close to be distinguished in the histograms.

6 Relative costs and benefits

The practical utility of pose-refinement depends on four main factors: the stability of the process, the area over which convergence is obtained (i.e. the equivalence classes illustrated in Figure 5), the cost per application of the image interrogation stage, and the number of iterations needed for adequate performance. Our experience suggests that the active technique requires fewer iterations and is more stable, but is inferior in other respects. The following experiment was carried out to compare these factors as they might affect a tracking task.

A short video of a saloon car was selected; to improve temporal resolution, separate fields of the video were analysed, by duplicating alternate lines. The best-fitting PCA model was determined (as in Section 5) in field 0. The recovered structure was used as a rigid model, and the recovered pose as an initial position. Poses in successive fields were then determined (i) by tracking through the sequence, using the pose recovered in field n as the seed pose for field $n+1$ (dotted lines in Figure 6), and (ii) using the pose in field 0 as the initial pose for each of the subsequent fields (continuous lines in Figure 6).

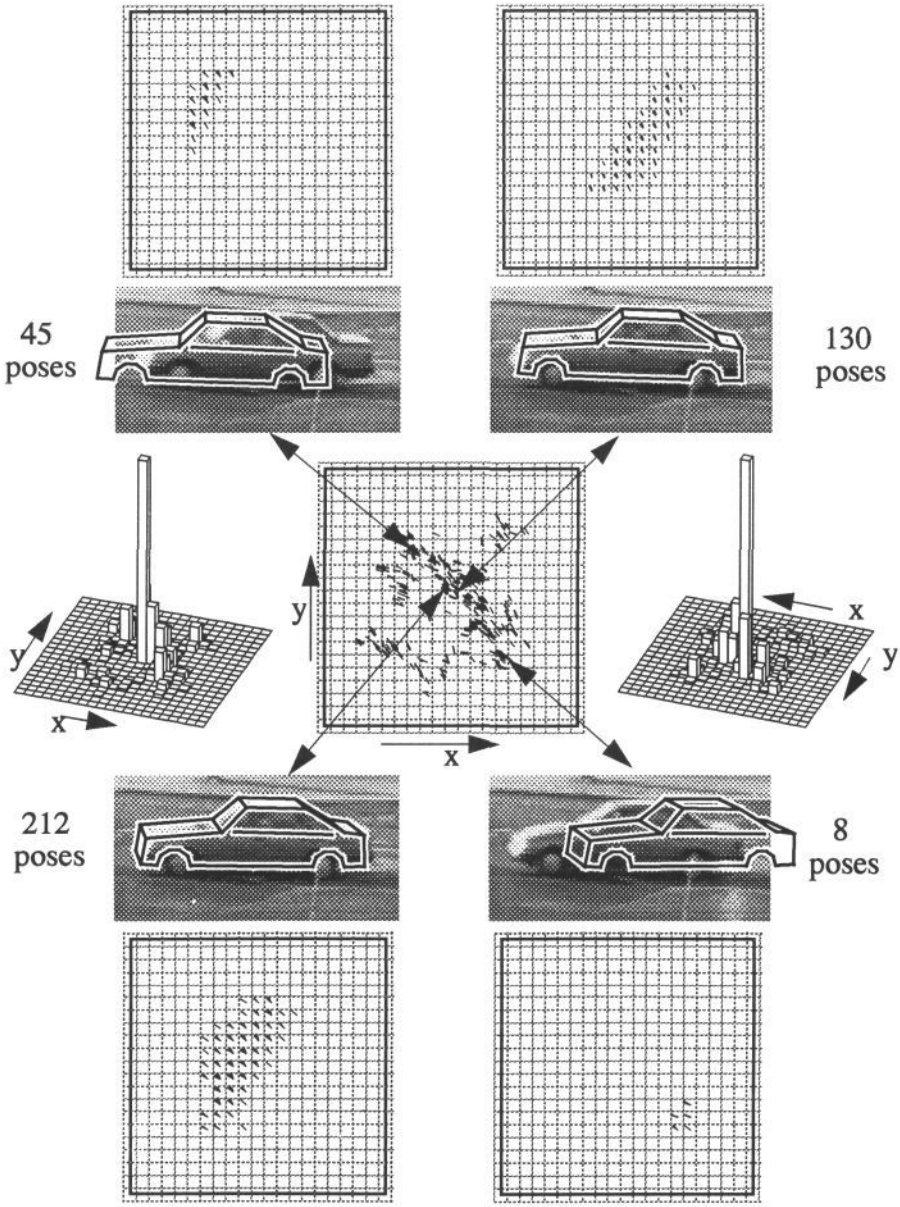

Figure 5 Details of convergence test for saloon car of Figure 8, showing selected attractors with equivalence classes (see text).

Figure 6 shows the results using the active method (left) and the passive method (right). It can be seen that both methods agree well in continuous tracking (dotted lines), with some slight indication of better consistency for the active model. However, the Passive model recovers pose accurately over longer gaps of time (and therefore distance). Note how the continuous lines follow the dotted lines further to the right in the right hand graphs than the left. According to these data, the passive model seems able to recover

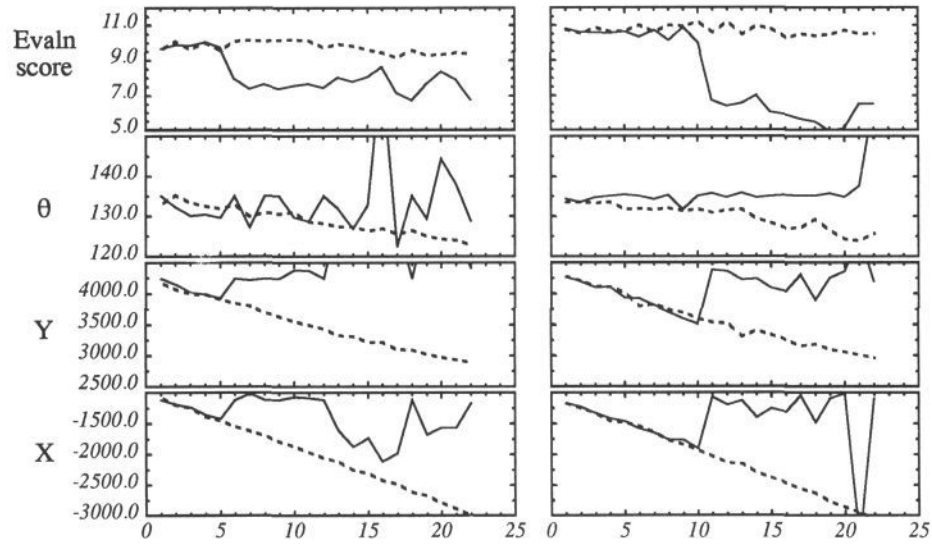

Figure 6 Tracking performance of active (left) and passive (right) pose refinement. All values recovered as functions of the number of fields from start. Dotted lines show results from field-by-field tracking. Continuous line shows pose recovered in field n in one step from field 0

from time gaps of approximately twice as long as the active model, indicating that a successful tracking system would only need to analyse the image sequence one half as frequently.

On the other hand, the numbers of iterations required by the two systems are very different. The passive model typically required 60-80 image evaluations, and this was fairly independent of the time gap. The active model succeeded after as few as 5-10 iterations for time gaps of 1-3 fields, but required far more as the gap increased; beyond a gap of 5 fields the active model failed completely.

This experiment bears out our earlier assertion that the active model is fast and accurate, but only if the starting pose is very close to the true pose; the passive model is better able to recover from larger errors, but is less precise and far slower. An optimal strategy might require the judicious fusion of the two methods.

7 Conclusion

We have described a perspective inversion approach towards pose refinement using an "active" model which offers improved performance over existing methods. The method is effective both in recovering the pose of a rigid model, and in recovering the structure of the deformable model described in [3].

The new approach requires far fewer iterations of a top-down search for image evidence than is needed using our previous "passive" model, and it appears to be more smoothly convergent. However, the range of initial pose errors from which convergence is likely is less - it is more prone to being caught up on a local maximum. An experimental investigation of the relative performance of the two methods is reported, giving some idea of the trade-off between the costs and benefits of the two methods.

8 Appendix

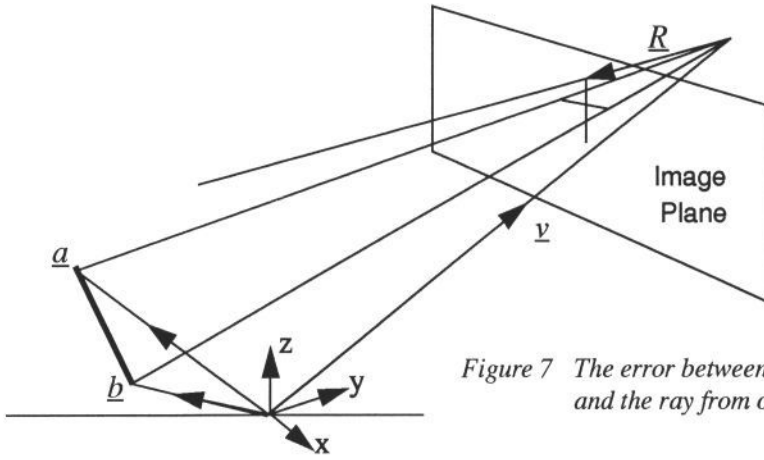

Figure 7 The error between a model line and the ray from one normal

8.1 Pose recovery

Consider a model at some initial pose on the ground plane. It is assumed that this is close to the "true" position, which is displaced from it by small translations t_x and t_y and rotation θ about the z axis. The nodal point of the camera lies at $\underline{v}$ in the initial model coordinate frame

By searching in the image normal to the projected model lines, we recover points of significant image gradient (see section 2), which are deemed to be due to the model line. Let one such point define a ray $\underline{R}$, and the corresponding model line be defined by its endpoints $\underline{a}$ and $\underline{b}$ (see Figure 7).

The normal to the plane containing the model line and the centre of projection is given by $\underline{n} = (\underline{v} - \underline{a}) \wedge (\underline{v} - \underline{b})$. The endpoints of the model line $\underline{a}$ and $\underline{b}$ depend on the local transformation parameters, and we can define $f(t_x, t_y, \theta) = \underline{R} \cdot \underline{n}$.

If the ray $\underline{R}$ intersects the model line then it lies in the plane, and $f(t_x, t_y, \theta) = 0$. The left hand expression can be Taylor expanded in the normal way about (0,0,0) and the resulting linear equation solved for t_x, t_y and θ.

We seek the simultaneous solution of $f = \underline{n} \cdot \underline{R} = 0$ for all significant rays found for all visible model lines. This over-determined set of equations may be solved by using the linear least squares method.

8.2 Recovery of structure parameters

The function f may also be expressed in terms of the reduced PCA space, and minimised in the same way to determine the best-fitting structure parameters. In this case the transformation of model point $\underline{a}$ is written as $\underline{a} + S^a \underline{\sigma}$, with

$$\underline{\sigma} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \cdots \\ \sigma_6 \end{bmatrix} \qquad \boldsymbol{S}^a = \begin{bmatrix} S_1{}^a{}_x & S_2{}^a{}_x & & S_6{}^a{}_x \\ S_1{}^a{}_y & S_2{}^a{}_y & \cdots & S_6{}^a{}_y \\ S_1{}^a{}_z & S_2{}^a{}_z & & S_6{}^a{}_z \end{bmatrix}$$

where $\underline{\sigma}$ gives the coordinates in PCA space of the model instance, and the element $S_n{}^a{}_x$ of the matrix $\boldsymbol{S}^a$ gives the x displacement from the mean of model vector $\underline{a}$ under a unit shift of the n^{th} PCA coordinate.

The partial derivatives of $\underline{a}$ with respect to the structure parameter σ_i are given by the appropriate column of the structure matrix $\boldsymbol{S}^a$

$$\frac{\partial}{\partial \sigma_i}\underline{a} = \left[S_i{}^a{}_x \; S_i{}^a{}_y \; S_i{}^a{}_z \right]^T$$

We can therefore compute the partial derivatives of the normal vector $\underline{n}$ with respect to the structure parameters, and express the function f as a Taylor expansion of the structure parameters, to solve as before.

References

[1] Lowe, D. G. *Perceptual Organisation and Visual Recognition*, Kluwer Academic Publications, 1985.

[2] Press, W. H. *et al. Numerical Recipes*, CUP 1986.

[3] Ferryman, J. M., Worrall, A. D., Sullivan, G. D. and Baker, K. D. A generic deformable model for vehicle recognition. (Submitted to BMVC95).

[4] Sullivan, G. D. Visual Traffic Understanding using the Ground-plane Constraint, Proc Int Conf on Signal Processing, Cyprus, 1993.

[5] Sullivan, G. D. Model-based Vision for Traffic Scenes using the Ground-plane Constraint, *Real-time Computer Vision*, Eds: C Brown and D Terzopoulos, CUP, 1995.

[6] Sullivan, G. D. Visual interpretation of known objects in constrained scenes, Phil. Trans. R.Soc. Lon., B, 337, pp 361-370, 1992.

[7] Tan, T. N. Sullivan, G. D. and Baker, K. D. Fast Vehicle Localisation and Recognition Without Line Extraction and Matching, Proc. 5th British Machine Vision Conference, 13-16 September, University of York, York, pp 85-94, 1994.

[8] Tan, T. N. Sullivan, G. D. and Baker, K. D. Recognising Objects on the Ground-Plane, Proc. 4th British Machine Vision Conference, 1993.

[9] Worrall, A. D., Baker, K. D. and Sullivan, G. D. Model-based perspective inversion, Image and Vision Computing Journal, 7(1), pp 17-23, 1989.

[10] Worrall, A.D., Sullivan, G. D. and Baker, K. D. Advances in Model-based Traffic Vision, Proc. 4th British Machine Vision Conference, pp 559-568, 1993.