Learning Enhanced 3D Models
for Vehicle Tracking

J.M. Ferryman, A. D. Worrall and S. J. Maybank
Computational Vision Group, Department of Computer Science
The University of Reading RG6 6AY, UK
J.M.Ferryman@reading.ac.uk

Abstract

This paper presents an enhanced hypothesis verification strategy for 3D object recognition. A new learning methodology is presented which integrates the traditional dichotomic object-centred and appearance-based representations in computer vision giving improved hypothesis verification under iconic matching. The "appearance" of a 3D object is learnt using an eigenspace representation obtained as it is tracked through a scene. The feature representation implicitly models the background and the objects observed enabling the segmentation of the objects from the background. The method is shown to enhance model-based tracking, particularly in the presence of clutter and occlusion, and to provide a basis for identification. The unified approach is discussed in the context of the traffic surveillance domain. The approach is demonstrated on real-world image sequences and compared to previous (edge-based) iconic evaluation techniques.

1 Introduction

The aim of this work is to extend previous research on hypothesis verification, and to improve the accuracy and robustness of pose refinement for object tracking. In recent years, top-down hypothesis verification has received relatively little attention in the vision literature. Notable exceptions are [2, 4, 10, 11]. The particular application domain for this work is vehicle tracking [2, 4, 5, 11]. This domain poses significant problems for object recognition: vehicles, for example, exhibit considerable within-class and between-class appearance variations.

Traditionally, edge-based techniques have been employed for iconic evaluation but edge models, *i*) exploit only a small part of the image structure, and *ii*) are ambiguous. Moreover, previous exemplar-based learning schemes with partial likelihoods have performed poorly in highly-variable natural scenes. Furthermore, in conventional model-based tracking, the hypothesised model is verified independently in each frame of the sequence. Therefore, it cannot accumulate knowledge of the object's appearance over time. Such problems can be alleviated to some extent by extending the iconic evaluator to be context sensitive. This can be done by including learnt information where "features" are described by an "appearance-based model". The proposed approach to iconic matching is to adopt a unified geometric/appearance-based approach based on learning.

BMVC 1998 doi:10.5244/C.12.87

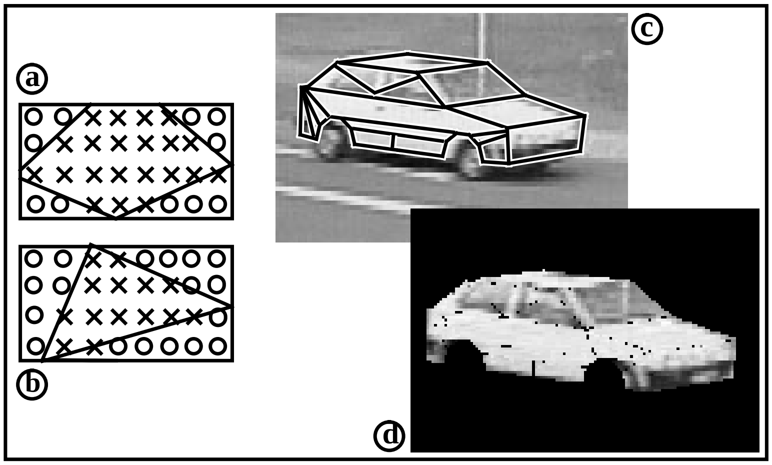

Figure 1: Appearance model. (a) and (b) illustrate sampling of the appearance model representation for two vehicle surfaces, (c) shows the model features projected under a hypothesised pose, and (d) the appearance model reconstruction using 20 eigenvectors. In this particular example the facet tesselation does not include the wheels. In practice, these features are used.

2 The approach

The approach taken is to exploit 3D geometric knowledge to segment specific image regions corresponding to a variety of feature types. In this work, we concentrate on features corresponding to vehicle surfaces. An important observation is that the modelled features are more specific to individual vehicles and therefore best used during verification.

The appearance of a moving vehicle is learnt during the motion. The aim here is to construct and refine (i.e. learn) the appearance of a vehicle and employ this representation to constrain the matching in subsequent tracking (sections 3 and 4 below). The model of the vehicle is projected into the image with hidden line removal. An appearance model (see Section 2.1) is constructed for each model feature representing a set of 3D points on the vehicle surface (this stage is performed offline). Each vehicle surface is treated as an independent feature. This approach allows *i*) a feature to be sampled (with hidden points removed) under full perspective projection, and *ii*) an equal number of samples to be obtained from each feature. The feature samples are used to learn the representation (see Section 2.2).

2.1 Appearance model

An appearance model is constructed for each model surface. For each surface on the vehicle a set of ($n \approx 500$) points are constructed on the surface. Currently this is done by using a frontal parallel view of the surface and raster scanning the boundary box of the surface in such a way as to generate the desired number of points. This technique is general and

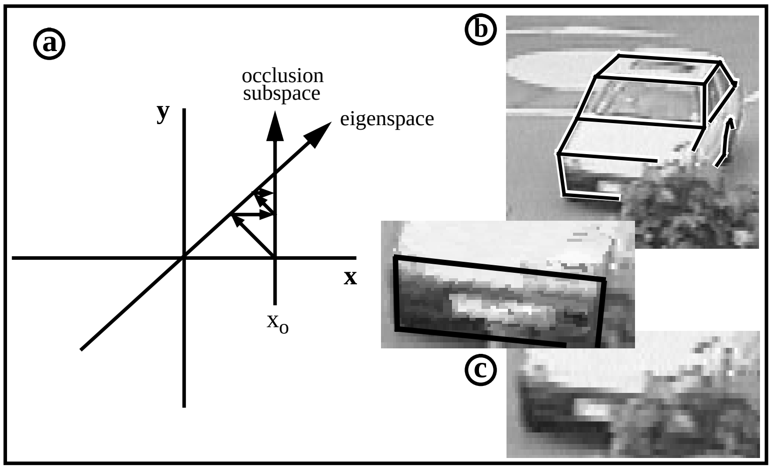

Figure 2: Occlusion handling. (a) shows a simple 2D case; (b) shows a projected model overlaid onto an image vehicle; and (c) shows the best reconstruction for the vehicle's occluded "front" feature using 20 eigenvectors.

does not impose topological relationships between the points[1]. This step is performed offline. During verification, a model is projected into the image under a hypothesised pose. A check is then performed to determine whether the points are clipped by either the vehicle or scene model. The set of 3D points is then projected into the image. Figure 1(b) illustrates the appearance model representation for two example features, Figure 1(c) shows a vehicle model projected into the image and Figure 1(d) the same imaged vehicle reconstructed using the appearance-based model for all visible features.

2.2 Learning the representation

In order to learn the global eigenspace representation for the background and vehicles sample features are taken for both. The method employed[2] is to sample the background image using Monte Carlo techniques, and to sample instances of vehicles with correctly fitted models. This stage is performed offline. Each feature is treated as a vector of dimension n by placing the grey-level image samples (with bilinear interpolation) in order. The set of k normalised training feature vectors $\{r_1, r_2, ..., r_k\}$ are represented by a lower-dimensional subspace based on principal component analysis (PCA). The eigenvectors (U) and eigenvalues (Σ) are efficiently computed using SVD. Let D (of dimension $n \times k$) be the matrix of feature vectors, $C = DD^T$, λ are the eigenvalues of C, $\lambda \propto s_i^2$ where s_i are the singular values of D. The rank of C is less than or equal to k so at most k eigenvectors are computed. The ordered eigenvectors are complete, orthogonal and form a basis that spans the k-dimensional training space. The input features are approximated

[1] Another technique using triangular facets which does impose a topological constraint is being considered.

[2] This is performed by projecting the vehicle model and sampling the appearance model for each visible feature.

by the first w eigenvectors with $w \ll k$. In practice, $n = 500$ grey-levels[3] and $w = 20$ eigenvectors has shown to provide good reconstruction of input training features with minimal residual error. An important property of the eigenspace representation which is exploited in this work is that the closer the projections of two input features are in eigenspace, the more highly correlated the original vectors. The distance in eigenspace is an approximation of image cross-correlation. A global eigenspace is chosen so as to reconstruct both the background and vehicles (i.e. it is able to discriminate between them). The appearance model we use is a point in the 20 dimensional eigenspace for each of the model features.

2.3 Subspace updating

The eigenspace can reconstruct features present in the training set but the dimensionality of the current basis feature space may not be sufficient to encode the observed feature. This can be detected by looking at the reconstruction error. There are two cases in which the underlying feature representation may need to be updated: i) when the appearance of a previously unseen vehicle feature cannot be well approximated by the eigenspace, and ii) a vehicle approaches the camera and features can be resolved in detail which may not have been well represented by the training set. It is computationally impractical to re-compute the SVD from scratch the complete set of training input features. Fortunately, there has recently been research into fast and stable updating algorithms for SVD. In this work we adopt the adaptive method of Manjunath *et al.* [7] for SVD updating. For a new feature A_{i+1} the new SVD is computed as $[U_i \Sigma_i V_i^T A_{i+1}] = U^{'} \Sigma^{'} V^{'T}$ where U_i and V_i are matrices whose columns are the first w left- and right- eigenvectors respectively, and $\sum_i$ the corresponding matrix of eigenvalues computed after obtaining i measurements of the feature. Full details are given in [7]. For the experiments reported in this paper the dimensionality of the eigenspace is kept constant. In practice, this is done by reducing the dimensionality of the space after each SVD update back to the original size.[4]. The global eigenspace is updated on a per feature basis using a mean reconstruction error criterion. In practice, highly-textured features (e.g. the front of the vehicle) require the representation to be updated more often than other, more homogeneous features. The advantage of updating on a per feature basis is that the feature space is only updated as necessary to maintain the discriminatory ability of the feature and thus the robustness of feature tracking. The same eigenspace is used for all features on one vehicle. In general, a single eigenspace update is only required at the start of tracking which eliminates the requirement to frequently change the underlying representation.

2.4 Occlusion handling

An advantage in adopting the eigenspace representation is that it can be used to reconstruct the best approximation of a feature when there is missing data. This occurs when there is occlusion. The missing data can be treated as free variables and the observed data as fixed. The free variables are changed in order to minimise the distance between

[3] At present we only consider grey-level intensity information but the approach is easily extended to include further contexual information (e.g. colour).

[4] the alternative approach would be to increase the dimensionality and map the observed data into the new space.

```
1. Compute projection coefficients ($proj_j$) for feature vector $r_i$ by
   taking dot product of $r_i$ with $U_j, 1 \leq i \leq w, 1 \leq j \leq n$.
2. REPEAT
   for all points $j = 1..n = 500$ in the feature do
       if the point $j$ is occluded then
           (a) $reconstruction = \sum_{i=1}^{w=20} proj_i U_i$
           (b) $sum$ += $(reconstruction - vec_j)^2$
           (c) $vec_j = reconstruction$
       endif
   endfor
   Compute projection coefficients ($proj_j$) for feature (as 1 above).
UNTIL (sum $\leq$ thd)
```

Figure 3: Algorithm for feature reconstruction under occlusion.

the n dimensional feature vector and its approximation in the w dimensional eigenspace. The problem then becomes one of determining the minimum distance between each unobserved point in the occlusion subspace (OS) and the eigenspace (ES). Figure 2(a) illustrates a simple 2D case. Here the eigenspace is the diagonal line ($y = x$) and the y value is unobserved. The optimal value of y can be obtained by projecting from the initial point ($x_0, 0$) onto the line $y = x$ and then back onto the line ($x = x_0$). This projection is repeated until the minimum distance is found. In the example the two spaces intersect but in general this is not the case. Figure 2(b) illustrates an example of occlusion and (c) the best reconstruction for the "front" feature. Figure 3 illustrates the complete algorithm for feature reconstruction under occlusion. Note that the quality of the reconstruction depends upon the ability of the eigenspace to reconstruct the feature appearance as discussed in Section 2.3.

3 Appearance matching

In this section we describe our approach to hypothesis verification given the object representation introduced in earlier sections. For each model feature we have a point in the 20D eigenspace which acts as a prototype for the feature. The problem we now address is how to match between this set of points in the eigenspace which represents the appearance model and the set of points used for the image reconstruction in the same eigenspace.

3.1 Evaluation function

The approach is to adopt a probabilistic framework. The underlying assumption is that projections of the same observed feature over several images can be modelled by a Gaus-

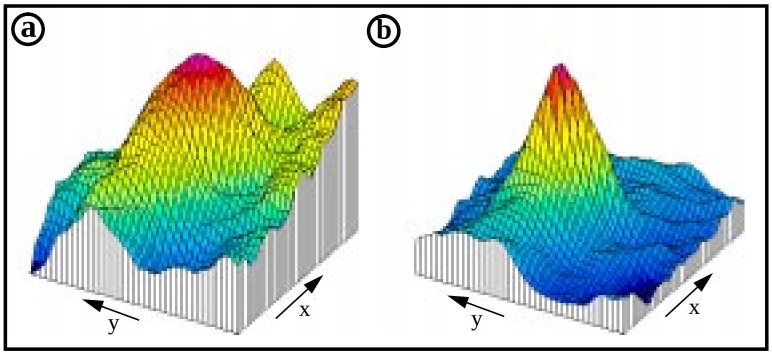

Figure 4: Evaluation surfaces in pose XY space for an edge-based evaluator (a) and the new appearance-based evaluator (b) for the same imaged vehicle and model projection.

sian density, and its response is given as

$$d(x) = exp^{\frac{-(x-\bar{x})C(x-\bar{x})}{2}} \tag{1}$$

where x is the 20D feature vector, $\bar{x}$ is the Gaussian mean and C is the feature covariance matrix. For initialisation, C is set to the covariance of the global eigenspace.

During tracking, the mean and covariance of the features in parameter space are updated based on a Gaussian sampling in pose space centred on the final recovered pose obtained using the simplex optimisation scheme. The mean and covariance is updated using a weighting technique. This can be viewed as an ellipsoid of uncertainty in feature location that collapses as the pose-parameter space mapping approximation improves. In practice, it is not necessary to update the estimate of the distribution for every frame but only when a significant change in vehicle orientation or depth occurs. The overall evaluation score (goodness-of-fit) for a projected model is the normalised sum of Gaussian responses and is weighted by the number of observations of the feature (i.e. number of frames), the projected feature surface area and the amount of occlusion

$$eval = \sum_{features} weight \cdot d(x) \tag{2}$$

The evaluation can be viewed as the minimisation of the distance of each feature from its prototype (i.e. Gaussian mean) in the eigenspace. The "smoothness" of the evaluation surface is affected by i) the completeness of the representation and ii) the data dimensionality. Figure 4 illustrates the evaluation surface for an edge-based evaluator (a) and the new appearance-based evaluator (b) for the same imaged vehicle and model projection. The vehicle was tracked through an image sequence with the vehicle pose

Edge-based	Appearance-based
Cheaper	Expensive
More stable at near-camera	More sensitive to view catastrophies
More sensitive to geometry	Less sensitive to geometry
More prone to aliasing (multimodal)	Unimodal evaluation surface
Less robust to occlusion	More robust to occlusion
Cannot track far objects (poor resolution)	Track distant objects reliably

Table 1: Comparison of edge-based and appearance-based evaluation methods for vehicle tracking.

refined at each frame to learn the sppearance. The model was then displaced from the final recovered pose by fixed amounts ranging from -1m to 1m along both the x and y axes and from -25 to 25 degrees of rotation around the vertical axis. The edge-based surface is multimodal with significant potential for a deterministic downhill optimisation method (e.g. simplex) to get trapped in local maxima. The appearance-based evaluation surface, however, is unimodal thus enabling faster, more reliable determination of the maximum aposteriori (MAP) estimate of the correct vehicle pose.

3.2 Relation to previous work

The appearance-based paradigm in computer vision has recently been successfully employed for 3D object recognition by Nayar and Murase [9] amongst others. The main problems with their approach are obtaining adequate segmentation of an object of interest from the background, and taking account of occlusion. Furthermore, it is difficult to see how an appearance model *alone* can be used as an object representational scheme. It seems impossible to acquire a full appearance model for all vehicle poses which could occur under perspective projection even with the ground-plane constraint (GPC). Mundy *et al.* [8] performed an experimental comparison of appearance and geometric approaches. The authors, in line with our own opinion, consider that object representation schemes that complement each other are fertile ground for new research. More specifically, a model should be more than geometry alone and therefore this suggests the combination of the two representations. The most closely related work to ours is Black *et al.* [1] on EigenTracking and Cootes *et al.* [3] on Active Shape Models (ASM's). In [1] rigid and articulated objects are tracked using a view-based representation. The main limitation is that all processing is performed in the image-plane with no "notion" of 3D. This is also the case for the morphable model approach of Jones *et al.* [6]. In the PDM approach of [3] each model point is associated with a model of expected image evidence: grey-level models generated offline from sets of training examples. It is noted that the main limitations of this approach are in application domains (e.g. outdoor scenes) where the grey-level appearance can change significantly and unpredictably. Our approach learns the object appearance online based upon experience.

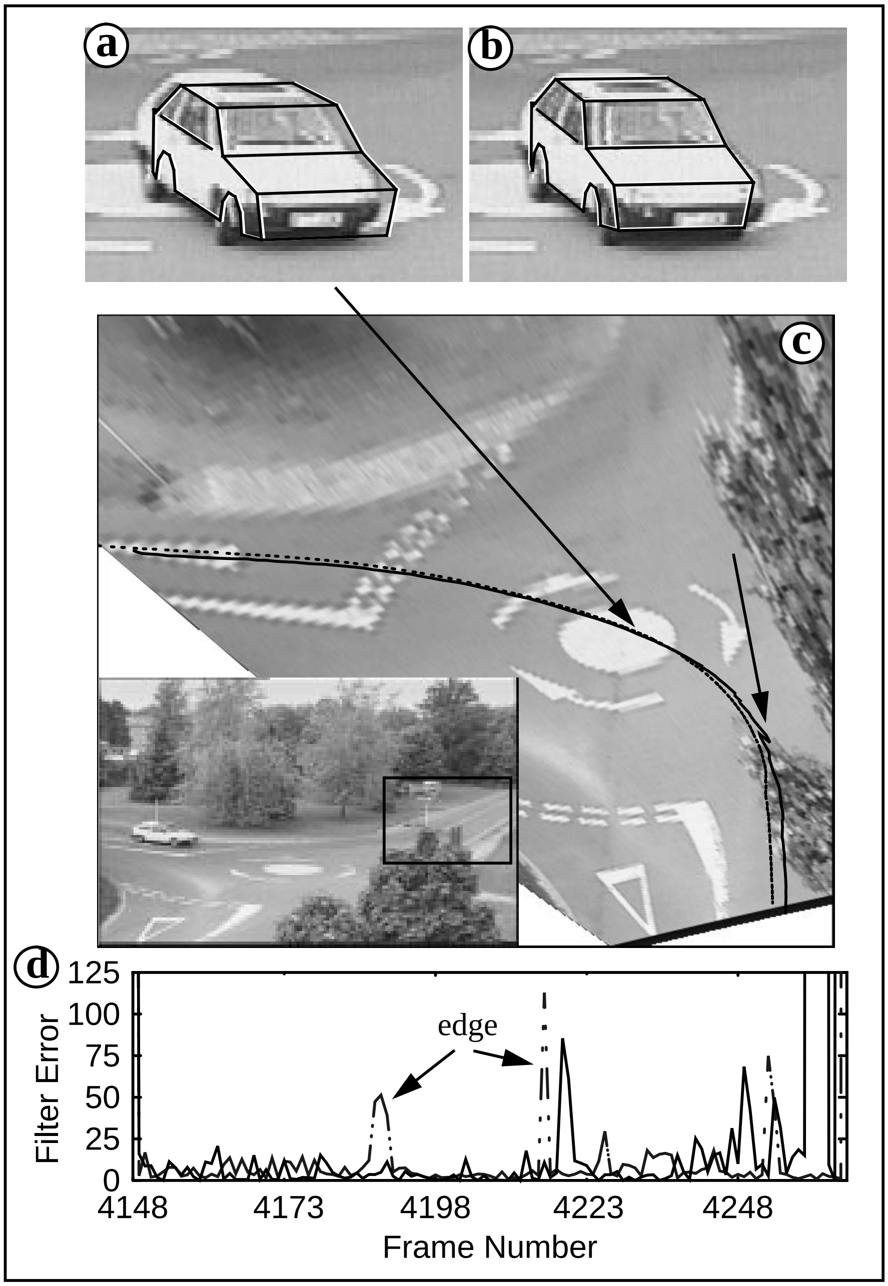

Figure 5: Examples of tracking using edge-based iconic matching (solid line in (c)) and the appearance-based model evaluation scheme (broken line in (c)). (a) illustrates a poorly recovered vehicle pose using the edge-based method while (b) indicates the pose recovered using the appearance-based approach for the same frame. The graph (d) illustrates the Mahalanobis distance between the recovered pose and IEKF prediction for the edge-based method (broken line) and appearance-based approach (solid line).

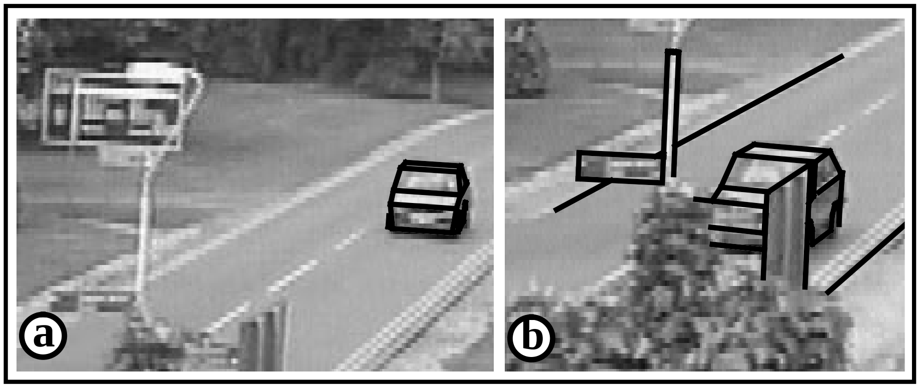

Figure 6: Advantages of appearance-based tracking for (a) far-camera and (b) with significant (40%) occlusion. The edge-based tracker fails for both cases.

4 Tracking

A number of experiments have been performed on different image sequences obtained using a static camera. For each sequence, model pose initialisation was performed by eye and either an edge-based evaluator [11] or the appearance-based evaluator used to refine and track the vehicle through the sequence using an Iterated Extended Kalman Filter (IEKF). Figures 5 and 6 illustrate two particular sequences of interest. In Figure 5, the white hatchback enters the scene from the left and navigates a roundabout approaching the camera. The vehicle undergoes significant change in depth and scale, and it is partially occluded by the tree (modelled by it's silhouette) near the camera. The results of tracking are illustrated in Figure 5 (the lower graph compares the Mahalanobis distance between the recovered pose and filter prediction) which shows that the edge-based tracker (dashed trajectory in upper figure) is much more affected by image clutter than the appearance-based tracker (solid trajectory in upper figure). Note also that the edge-based tracker is distracted by the occlusion as indicated in Figure 5(c).

Figure 6 illustrates two instances in which edge-based tracking fails (far-camera and significant occlusion) but appearance-based tracking succeeds. A comparison of the pros and cons of edge-based and appearance-based tracking is given in Table 1. An experiment was also performed to compare the performance of integrated edge-based and appearance-based tracking for pose refinement. However, it is not clear at the moment how the evidence from independent trackers should be weighted and therefore the control problem remains an open research issue.

4.1 Extension to deformable models

The methodology discussed applied so far assumes fixed rigid models. The methodology presented easily extends to linear deformable PCA models [5]. The PCA model is constructed so that the surfaces remain planar under a change in the PCA parameters. This means that there is a plane-to-plane homography for each surface. This homography maps between the surface at the mean PCA parameters and the same surface at the current PCA parameters.

5 Conclusions and future work

This paper has introduced a new methodology which integrates the traditional dichotomic object-centred and appearance-based representations, leading to improved hypothesis verification using iconic matching. In particular we have demonstrated the successful application of appearance-based techniques to vehicle tracking, resulting in more reliable model-based tracking particularly with respect to occlusion.

Future work aims to investigate whether maintaining a fixed topology of the points across vehicle surfaces improves the ability of the system to perform reconstruction. At present, the eigenspace has to allow for differences in the toplogy. A fixed topology may allow a better representation of surface information. Additionally, we intend to investigate surface lighting/reflectance models for normalising the input data prior to the eigenspace analysis. Future work will also consider active pose refinement and initialisation issues in the appearance-based paradigm.

References

[1] M. J. Black and A. D. Jepson, EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation, *Int. Journal of Computer Vision,* Vol. 26, No. 1, pp 63-84, 1998.

[2] K. Brisdon, Evaluation and Verification of Model Instances, *Proceedings Alvey Vision Conference,* Cambridge, pp 33-37, 1987.

[3] E. C. Di Mauro, T. F. Cootes, C. J. Taylor and A. Lanitis, Active Shape Model Search using Pairwise Geometric Histograms, *Proceedings (eds. R. Fisher and E. Trucco) British Machine Vision Conference,* 1, pp 353-362, 1996.

[4] L. Du, G. D. Sullivan and K. D. Baker, On Evidence Assessment for Model-Based Recognition, *Proceedings (eds. D. Hogg and R. Boyle) Alvey Vision Conference,* 1, pp 149-186, 1992.

[5] J. M. Ferryman, A. D. Worrall, G. D. Sullivan and K. D. Baker, A Generic Deformable Model for Vehicle Recognition, *Proceedings 6th British Machine Vision Conference,* pp 127-136, 1995.

[6] M. Jones and T. Poggio, Multidimensional Morphable Models, *Proceedings 6th Int. Conf. Computer Vision,* Bombay, India, pp 683-688, 1998.

[7] B. S. Manjunath, S. Chandrasekaran and Y. F. Wang, An Eigenspace Update Algorithm for Image Analysis, *Proceedings IEEE Int. Symposium on Computer Vision,* Coral Gables, Florida, pp 551-556, 1995.

[8] J. Mundy et al, Experimental Comparison of Appearance and Geometric Model Based Recognition, *Lecture Notes in Computer Science,* 1144, pp 247-269, 1996.

[9] H. Murase and S. K. Nayar, Visual Learning and Recognition of 3D Objects from Appearance, *Proceedings Int. Journal of Computer Vision,* 14, pp 5-24, 1995.

[10] C. Rothwell, The Importance of Reasoning about Occlusions during Hypothesis Verification in Object Recognition, *Technical Report No. 2673,* INRIA Sophia-Antipolis, October 1995.

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