3D Grouping by Viewpoint Consistency Ascent

L. Du, G.D. Sullivan & K. D. Baker
Intelligent Systems Group
Department of Computer Science
University of Reading, RG6 2AY, UK
L. Du @READING. AC. UK.

Abstract

The viewpoint consistency constraint (VCC) provides a powerful way to discover extended feature groups and to test hypotheses in object recognition. Lowe's incremental method fails in complex scenes, and an exhaustive tree search (eg Grimson &Lozano-Perez) is too expensive. We present a state space approach in which transitions are made which monotonically ascend a measure of viewpoint consistency

1 Introduction

Model based vision usually relies on a hypothesis-test-refine cycle to recover an accurate estimate of an object's pose [1, 2, 3, 10]. An initial pose estimate, typically based on the detection of an object-specific cue feature, allows a search to be made for additional evidence which supports the cue. This in turn provides a better estimate of the pose, based on the extended set of feature correspondences. We call this process 3D grouping using the viewpoint consistency constraint (VCC).

Lowe [10] has reported an incremental approach to the search for extended feature sets, successively aggregating features which meet simple acceptance criteria. We have found that this algorithm frequently fails in images containing clutter.

An alternative approach is to use the initial pose to identify all plausible extended features, and to search among the combinations for sets of mutually consistent features. This problem can be represented as an Interpretation Tree, as used by Grimson & Lozano-Perez, but in matching 3D objects to 2D data the evaluation of a node of the tree requires a (multi-feature) viewpoint inversion, which cannot be effectively pre-complied into (pairwise) look-up tables.

We present a state space representation of the problem and a search algorithm which we call Viewpoint Consistency Ascent (VCA). Each state has a value (defined by a measure of the viewpoint consistency), and transitions take place in a two stage process between states which differs by a single feature match, according to the steepest ascent of the value.

2 VCC concepts

A set of 2D features, each matched to specific object features of a 3D model, is called a *3D clique*. The process of deriving a 3D clique is called *3D grouping* and the number of matches in a clique is its *cardinality*.

2.1 Measurement of viewpoint consistency

A 3D clique of cardinality at least 3 usually enables the use of perspective inversion to determine the pose. In turn this allows the generation of an instance of the model which

BMVC 1991 doi:10.5244/C.5.7

can be projected onto the image as a 2D template. The *viewpoint consistency of a 3D clique* is a measure of the accuracy with which features in this template coincide with the data features in the clique.

The agreement between a model feature and an image feature can be measured in the image domain by three discrepancies: difference of orientation m_θ (in degrees), perpendicular distance m_d (in pixels) and the length of any unmatched segment of image feature m_e (in pixels), as illustrated in Figure 1. [NB our definition of the unmatched segment of the image feature tolerates truncated image features but disfavours extended edges likely to rise from coincidental alignment.]

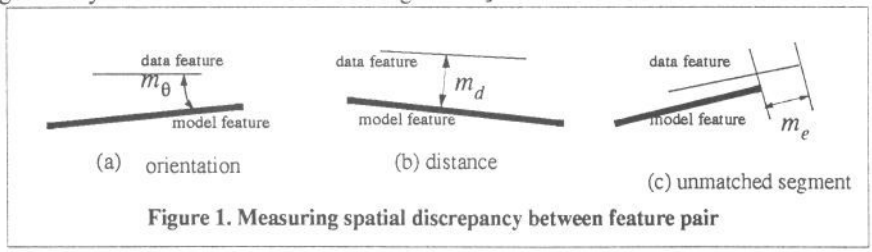


Figure 1. Measuring spatial discrepancy between feature pair

It is important to note that we use the viewpoint consistency constraint to the full and that a clique is always taken as a whole. We use two measures for VCC.

(1) As a graded measure, we define the *inconsistency of a 3D clique* as:

$$V=\text{MAX}\ \{(w_\theta m_{\theta i} + w_d m_{di} + w_e m_{ei}) \mid \forall \text{ matches (i) in the clique}\}$$

The terms w_θ, w_d and w_e are weights for the different discrepancy measures, empirically chosen so that the three terms have roughly equal impact on the graded measure when they are maximal, given the size of the area of interest in the image.

(2) To define a binary acceptance criterion we use a threshold (τ), A clique is acceptable if

$$(m_{\theta i} \le \tau_\theta w_\theta)\ and\ (m_{di} \le \tau_d w_d)\ and\ (m_{ei} \le \tau_e w_d)\quad (\forall i)\,.$$

2.2 3D grouping

The starting point for 3D grouping is a set of *candidate features* which are established on the basis of the initial pose estimate. This initial pose may have been derived from an initial hypothesis in a single frame [13], or from extrapolation of the previous frame in a tracking problem [14]. To account for possible inaccuracy with the initial pose, we accept all features meeting the acceptance criterion with a lax value of τ.

We thereby have a set of n_d data features $D = \{d_i | i = 1, \ldots, n_d\}$, and a set of n_m model features $M = \{m_i | i = 1, \ldots, n_m\}$. Note that n_m is the number of model features having at least one candidate d_i and is usually considerably lower than the total number of features on the model. Each model feature (m_i) is associated with a subset of D, $D_i = \{d_{ij} | j = 1, \ldots, n_i\}$ (see Figure 2(c)).

An exhaustive exploration of all possible sets of pairings $\{m_i, d_{ij}\}$ ($m_i \in M$, $d_{ij} \in D_i$) involves $\prod_{i=1}^{n_m} (n_i + 1)$ cases (1 is added to each n_i to account for the case of a null match). In practice the problem is slightly reduced since in order to allow perspective

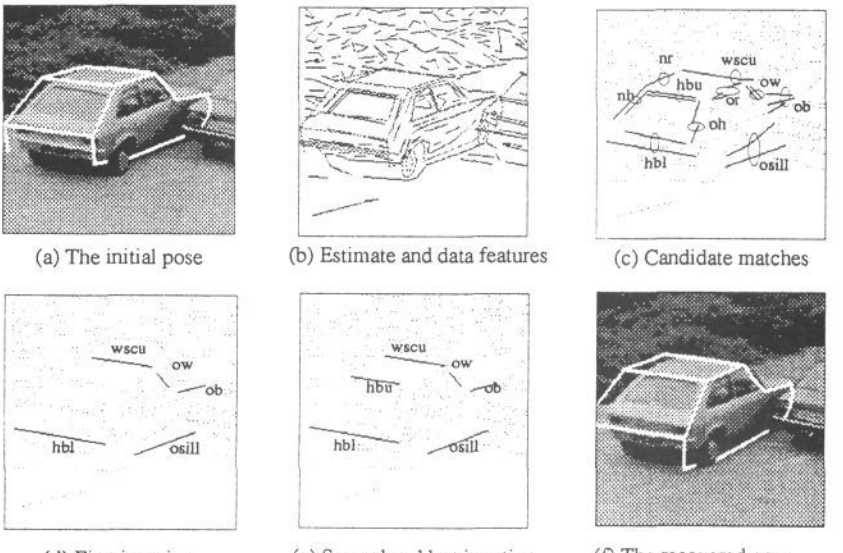
inversion we require that each clique contains at least 3 matches between object and image features.

The evaluation of consistency (including pose inversion) constitutes the main computational burden for 3D grouping, so the *computational cost* of a strategy can be estimated by the number of VCC evaluations. In our experiments we assess performance with respect to subjectively defined correct matches obtained by visual examination of the image features. We define the *rate of misses* (rm) and *rate of false-matches* (rf) by means of this ground truth: rm is the proportion of visible object features failing to be matched by the clique; rf is the proportion of false matches in the resulting clique. The *perfect clique* with rm=rf=0 is expensive to find. It also usually contains substantial redundant information, since our purpose is to rule out cases of accidental conspiracy and to obtain sufficient correct evidence to determine the object position. We call a clique of low rm and rf=0 a *desirable clique.* While there is a unique perfect clique, there exist multiple desirable cliques. We pursue any one of these desirable cliques by allowing a few misses.

3 Previous methods for exploiting the VCC

3.1 Lowe's incremental model matching [10]

Given a 3D grouping problem as defined above, Lowe's incremental model matching method works iteratively. At each iteration, a probability of conspiracy is given to each candidate image feature according to its spatial relation to the estimated model feature. For each subset D_i containing more than one image feature, an ambiguity is calculated according to the maximal and minimal probabilities in this subset. Candidates of low conspiracy probability (using a threshold) and having low ambiguity (using a second threshold) are accepted. Before the next iteration, the pose estimate is refined using the newly added matches. The process finishes when no image feature passes those thresholds.



(a) The initial pose (b) Estimate and data features (c) Candidate matches

(d) First iteration (e) Second and last iteration (f) The recovered pose

Figure 2. Incremental approach to 3D grouping fails to identify correct matches

A major advantage of an incremental approach is that no backtracking is needed so that it is computationally cheap. However applied to images of complex objects in cluttered scenes it often fails to converge correctly (see Figure 2). The method was applied to 4 similar images. The rate of false-matches range from 0.3 to 0.5, which indicates very low reliability.

The failure is not coincidental. Firstly, there is little reason to believe that a data feature is more likely to be correct than another simply because it is slightly closer to the originally estimated position. Secondly, incrementally aggregating individual interpretations of data features may reduce the global viewpoint consistency. Mismatches from previous iterations cannot be corrected nor can further mismatches be prevented. In Figure 2, the algorithm is fatally affected by the initial error in mismatching the ground shadow with the sill.

3.2 Alternative model-based approaches

The failure of Lowe's algorithm is typical of other systems matching 3D models to 2D edges. ACRONYM [3] was a very ambitious attempt to accommodate a wide range of objects and had a sophisticated constraint management system. However it has been criticised for its inability to apply the viewpoint consistency constraint accurately which limited its robustness[10].

Both Goad's method and Bray's re-implementation and extension use "shape-only" local constraints specifying legitimate ranges (thresholds) of angle, direction and distance between feature pairs. Their main advantage is that hypothesis-verification cycles are not needed even for single frame. However, "shape-only" constraints cannot discriminate against clutter and introduce errors by tessellation of the viewsphere. Bray tested his re-implementation by adding random lines and shortening the data features to simulate noise effects. This may appear distracting to humans but it bears little resemblance to structured clutter which often introduces strong ambiguity. Furthermore no indication was given, in either report about the impact of changing thresholds on robustness. To ensure global consistency, Bray suggested a final stage of model testing using Lowe's method. Therefore it is natural to anticipate the same type of failure as Lowe's.

4 The VCC used to prune an interpretation tree

4.1 3D Grouping as a constrained search of an interpretation tree

Following Grimson and Lozano-Perez' 2D-2D [5] and 3D-3D [6] systems, the 3D grouping problem can be represented as a constrained search of an Interpretation Tree (IT), which can be searched in a depth-first manner.

The VCC provides an "acceptance" criterion to justify pruning of the tree. Whenever inconsistency is detected a whole subtree is pruned and back-tracking occurs. Any consistent group, at least as large as 3, is recorded. When the search finishes the largest consistent cliques are assessed and the best is taken as the resultant interpretation.

4.2 Experimental findings

The IT is combinatorial, but the pruning operation removes subtrees and greatly reduces the number of nodes encountered. However, the need for null matches and the fact that the size of a clique must be at least 3 to invoke the VCC causes a substantial minimum computational cost. For a typical problem, with $n_m = 10$ and the average size of $D_i = 3$, the cliques of cardinality 3 number $C_3^{10}3^3 = 3240$.

We have experimented with different acceptance criteria by changing the threshhold τ. It has proved difficult to select a threshold which works successfully for our test images. In any case, with a null match allowed, pruning is largely ineffective [12], so the IT approach is very costly.

5 The VCC as a best-first heuristic

5.1 State space formulation of 3D grouping

We may regard the search of different 3D cliques as a state space problem, and use the scalar measurement of view consistency V as a heuristic to choose state transitions. For a given problem, there are as many states as the number of possible cliques. We only consider transitions between two states differing by a single match, either by changing a match or by dropping a match (see Figure 3).

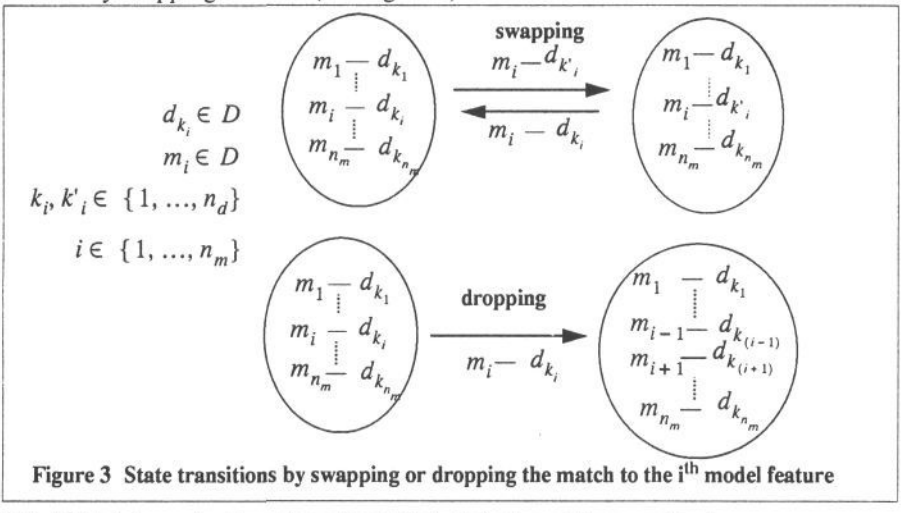


Figure 3 State transitions by swapping or dropping the match to the i[th] model feature

5.2 Initial state and a two stage best-first state transitions method

Within the state space representation, the 3D grouping process becomes a sequence of state transitions, with performance determined by the initial state, the transition steps and the termination condition. At each transition we improve the VCC monotonically - hence the algorithm is called Viewpoint Consistency Ascent (VCA).

We establish the initial clique by using the unique maximal clique containing on the longest data feature from each subset D_i. The initial clique may be formally represented as:

$$C_0 = \{ (m_k, d_{kl}) \,|\, length(d_{kl}) \geq length(d_{kj}) \, (\forall j) \, ; k = 1, \ldots, n_m \}$$

This seems a good choice of starting state: in our experiments we have found that 60~75% of the matches in this initial clique are correct. All the potentially matched model features are included, so the cardinality of this initial state is n_m.

The remaining features comprise a set of "spares" D_s, in which there are n_m subsets, $D_{si} = \{ d_{ij} \,|\, j = 1, \ldots, n_i, d_{ij} \neq d_{l_i} \}$ some of which may be empty. (The association between each candidate and its model feature is thereby preserved).

State transitions are made in two stages, both using a local best-first heuristic.

(1) Swapping stage

State transitions in this stage attempt to move towards a goal state by exchanging single data features from D_s with their counterparts in the clique. The aim of this stage is to minimise misses.

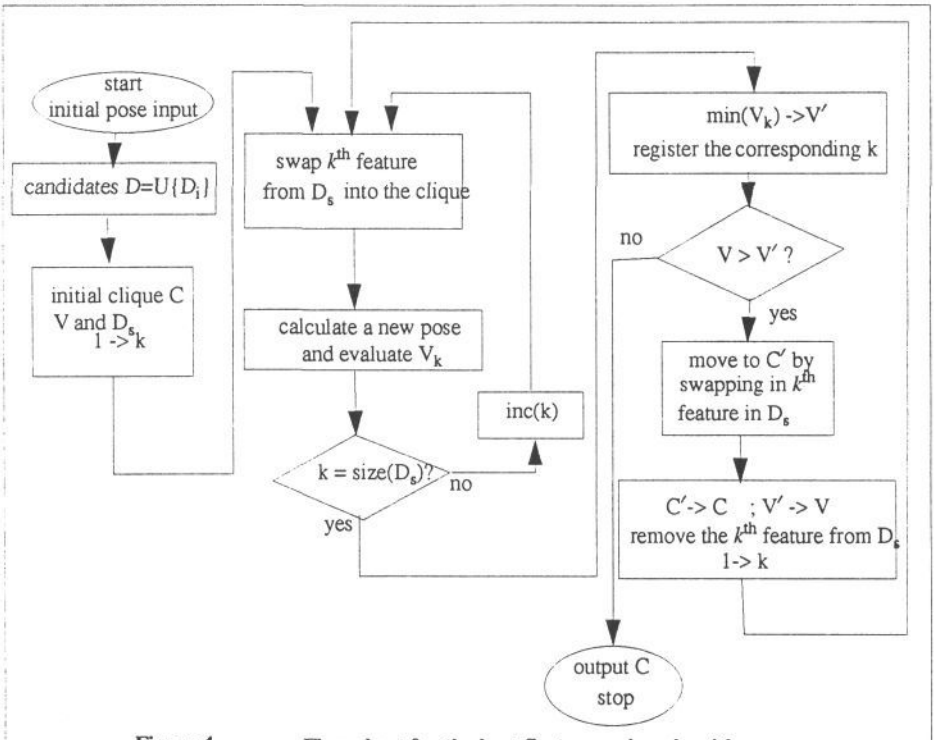


Figure 4 **Flow chart for the best-first swapping algorithm**

The algorithm for the swapping stage is illustrated in Figure 4. At any state, we examine all possible single swaps, compute all consequent new poses and their VCC scores. We take the swap with the best consistency improvement as the state transition. A series of transitions bring a monotonic improvement in consistency scores, which stops when further consistency improvement cannot be obtained.

(2) Dropping stage

At the end of swapping stage, the system will generally have reached a state much closer to a desirable clique containing fewer false matches than the initial clique C. The second stage seeks to drop any remaining false matches.

The algorithm for the dropping stage is illustrated in Figure 5. It bears close resemblance to the first stage in the sense of monotonic consistency improvement and examination of all possible single drops before taking a transition. This stage iteratively drops single matches from the current clique. It stops when no change brings about improvement in consistency or the cardinality of the clique reduces to 3.

(3) Incorporation of hysteresis

The simple monotonic improvement test in Figure 4 and 5 (V>V′) leads to a slow convergence. Therefore hysteresis is introduced by means of a threshold (V>V′-Thd) to prevent unprofitable sequences of small improvements.

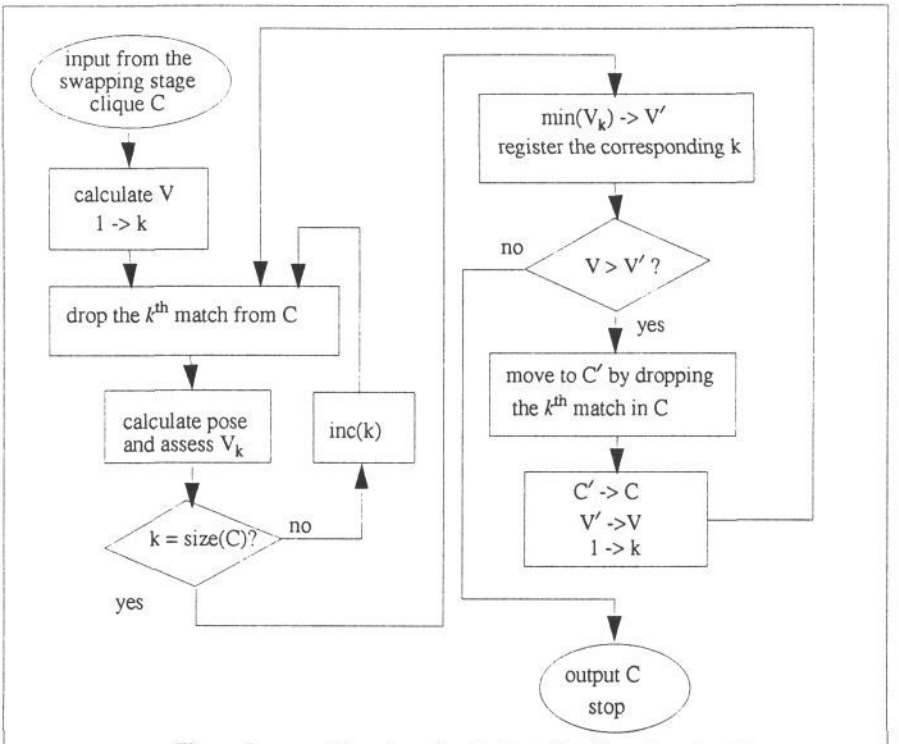


Figure 5 **Flow chart for the best-first dropping algorithm**

5.3 Performance

The worst case cost in the swapping stage (without replacement of the rejected d_i in D_s) is $\sum i\,(i = 1, \ldots, n_d - n_m)$ However the process is very unlikely to run into the worst case because the majority of matches in the initial clique are likely to be correct. The worst case cost for the dropping stage is $n_m^2/2$. Again this is very unlikely in practice because only a small portion of incorrect matches left over from the previous stage are incorrect and need to be dropped. The worst case cost for the VCA method, which is a combination of two stages, is therefore $n_d^2 - 2n_d n_m$, but the actual cost is likely to be much lower.

The various stages of the algorithm are illustrated in Figure 6. We see that the initial ground shadow error (which upset Lowe's algorithm, Fig 2) is effectively overruled by the other evidence. Illustrations of the results of the method in three other cases are shown in Figure 7. Each column represents one case. Images and the initial pose are shown at the top; the result of Lowe's incremental algorithm is in the middle; the result of the VCA method is at the bottom. [In case1, the problem in the middle figure is mainly due to a mismatch of the side of the hatchback. In case 2 attention should be paid to the rear of the car. In case 3, significant error is to be found at the rear and the sill.]

The main costs of different algorithms are shown in Table 1. The VCA provide better performance than Lowe's incremental method at a cost far below that of searching an interpretation tree.

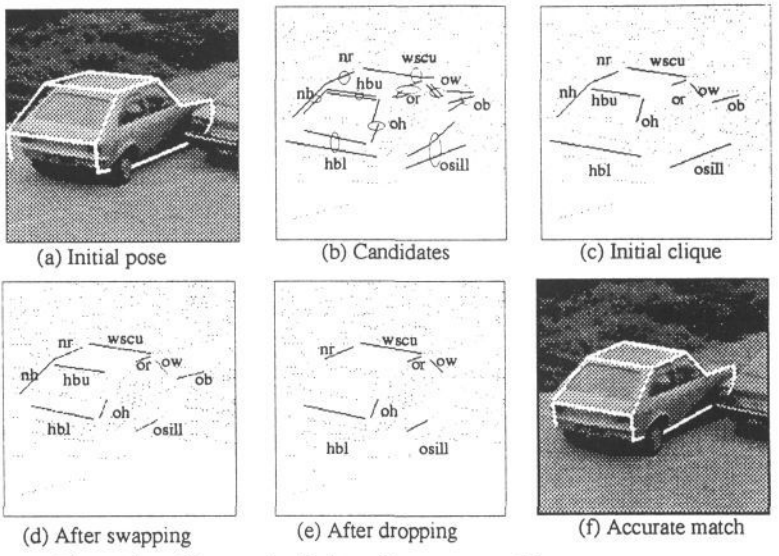


Figure 6 3D grouping by best-first state transition

6 Conclusions

The significance of this study is two-fold. First we have presented a best-first state transition approach to 3D grouping which has improved reliability, and a worst casecomplexity of $O(n^2)$. Secondly, we believe that the study carries important implications for 3D-2D model matching in general. When images are good and contain little structured clutter, model matching can be made very efficient by using methods such as Lowe's which treat the VCC inexactly. But in domains such as that studied here, the image is more complex and new methods have to be adopted to impose the VCC more stringently.

7 References:

[1] Bodington, Sullivan & Baker, "Experiments on the use of the ATMS to label features for object recognition", Computer Vision-ECCV'90, Spring-Verlag, 1990.

[2] Bray, A., "Recognising and Tracking Polyhedral Objects", Ph.D Thesis, Sussex University , UK, 1991.

[3] Brooks, A. B., "Model Based Computer Vision", UMI Research Press, 1984.

[4] Bolles, R., et al, "3DPO:A Three-Dimensional Part Orientation System", Int. J. of Robotics Research.

[5] Grimson, et al, "The combinatorics of object recognition in cluttered environments using constrained search", ICCV'88

[6] Grimson, et al,"Model-based Recognition and Localization from Sparse Range or Tactile Data", Int. J. of Robotics Research.

[7] Chen, et al, "A Robot Vision System for Recognising 3D Objects in Lower-order Polynomial Time", IEEE PAMI, No 6, 1989.

[8] Goad, C., "Special Purpose Automatic Programming for 3D model-based vision", Proceeding of the ARPA image understanding Workshop, Arlington, Virginia, 1983.

[9] Kim, W. and Kak, C., "3D Object Recognition Using Bipartite Matching Embedded in Discrete Relaxation", IEEE PAMI, No 4, 1991.

[10] Lowe, D., "The viewpoint consistency constraint", International Journal of Computer Vision, 1987

[11] Lowe, D., "Fitting Parameterised 3D Models to Images", IEEE PAMI, No 5, 1991

[12] Grimson, "The combinatorics of heuristic search termination for object recognition in cluttered environments", MIT, AI Lab Memo 1111.

[13] Rydz, A., et al, "Model based Vision Using a Planar Representation of the Viewsphere", Alvey Vision Conference'88, Manchester, 1988.

[14] Worrall, A., et al, "Model Based Tracking", BMVC'91, Glasgow, 1991

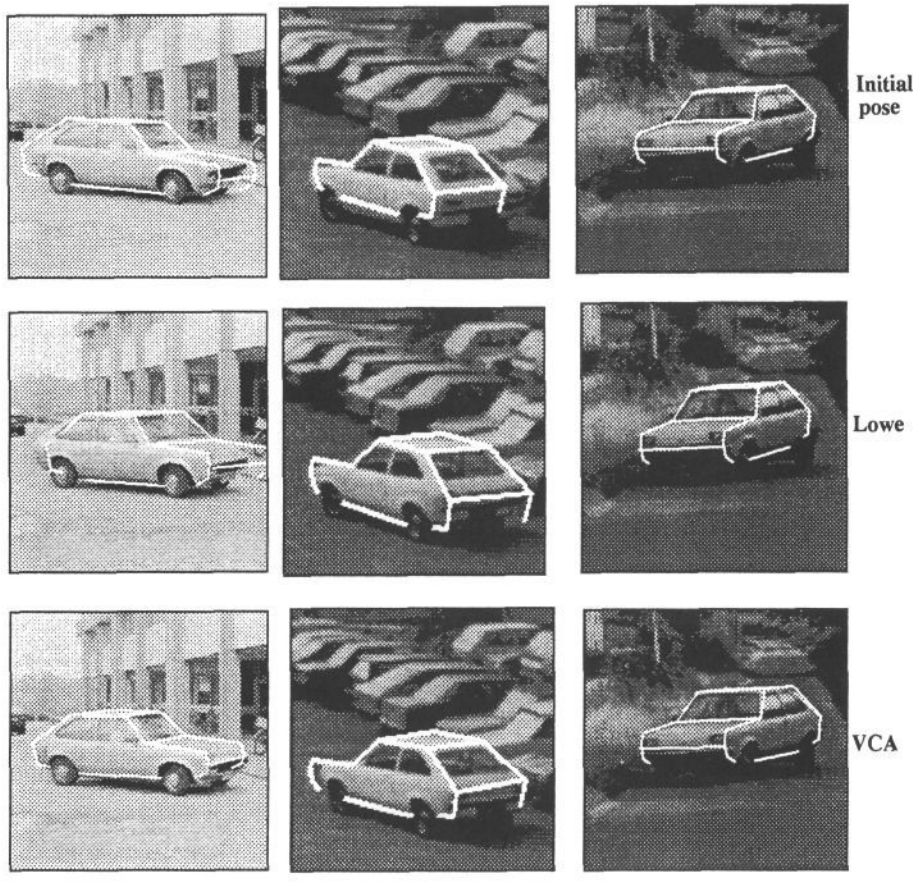


Figure 7 Comparison of Lowe's incremental aldorithm and VCA

	Lowe	VCA	Cliques of cardinality = 3
case 1	3	119	2068
case 2	2	63	1248
case 3	5	215	8652

Table 1 Comarison of the costs in VCC computation of different methods