The Importance of Reasoning about Occlusions during Hypothesis Verification in Object Recognition

Charlie Rothwell

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

The Importance of Reasoning about Occlusions during Hypothesis Verification in Object Recognition

Charlie ROTHWELL

N° 2673

Octobre 1995

PROGRAMME 4

Rapport de recherche

ISSN 0249-6399

The Importance of Reasoning about Occlusions during Hypothesis Verification in Object Recognition

Charlie ROTHWELL*

Programme 4 — Robotique, image et vision
Projet Robotvis

Rapport de recherche n° 2673 — Octobre 1995 — 27 pages

Abstract: In this paper we study the limitations of current verification strategies in object recognition and suggest how they may be enhanced. On the whole *object topology* is exploited little during verification. In practice, understanding the connectivity relationships between features in the image, or on the object, can lead to significantly more accurate evaluations of recognition hypotheses.

Usually adjacent features on an object should be party to mutual visibility constraints. This is to say that when a model is hypothesized in a scene, two features which are adjacent within the model of an object should either both be visible in an image of the object, or both occluded. If not, then we can broadly say that an occlusion event should exist between the two features. In the case that this event fails to be measurable, we can start to infer that the model hypothesis is incorrect. Similar reasoning can be used to exploit image topology and the uniqueness of sets of model-image correspondences. Generally, such a line of thinking departs from traditional approaches in which topological interactions between features are not exploited fully.

Testing out our algorithms for topology and occlusion analysis has involved the implementation of a complete object recognition system. The system we have built measures planar algebraic invariants in real images and uses these to index into a model base. The result of the indexing step is a list of hypotheses. These hypotheses are evaluated both using traditional verification algorithms, and also using our more detailed methods.

Key-words: object recognition, junction detection, occlusion reasoning

(Résumé : tsvp)

*Funded by an HCM grant from the European Community. Email: crothwel@sophia.inria.fr

Unité de recherche INRIA Sophia-Antipolis
2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex (France)
Téléphone : (33) 93 65 77 77 – Télécopie : (33) 93 65 77 65

De l'importance de l'analyse des occultations pour la vérification d'hypothèses en reconnaissance d'objets

Résumé : Dans ce rapport, nous étudions les limitations des stratégies actuelles de vérification en reconnaissance d'objets et suggérons comment celles-ci pourraient être améliorées. Globalement, on peut dire que les *propriétés topologiques des objets* sont trop peu exploitées durant l'étape de vérification. En pratique, comprendre les relations de connectivité liant les primitives de l'image ou du modèle peut mener à un meilleur traitement des hypothèses.

En général, les primitives adjacentes d'un objet sont liées par des contraintes de visibilité mutuelle. Cela signifie que si l'on suppose que l'on observe un objet (hypothèse) dans une scène, deux primitives adjacentes dans le modèle doivent être toutes deux soit visibles soit occultées. Si cela n'est pas le cas, on peut dire qu'un évènement d'occultation doit exister entre les deux primitives. Dans le cas où cet évènement ne peut être mesuré, on peut faire la supposition que l'hypothèse était fausse. Un raisonnement similaire peut-être utilisé pour exploiter la topologie de l'image et l'unicité des appariements entre primitives images et modèles. Cette méthode se distingue de la plupart des approches traditionnelles par une meilleure exploitation des interactions topologiques entre les primitives.

L'évaluation des performances de nos algorithmes pour l'analyse de la topologie et des occultations a necéssité l'implémentation d'un système complet de reconnaissance d'objets. Celui-ci mesure des invariants algébriques d'objets plans à partir d'images réelles et utilise ceux-ci pour indexer une base de données de modèles. Le résultat de l'étape d'indexation est une liste d'hypothèses. La qualité de celles-ci est alors évaluée en utilisant à la fois les algorithmes de vérification traditionnels et nos méthodes plus complètes.

Mots-clé : reconnaissance d'objets, détection de jonctions, analyse des occultations

Contents

1 Introduction

The process which tends to be used most to differentiate between different hypotheses in object recognition is verification. However, verification is only the final step in a long chain of visual processing events which takes the user from grey level images to the verified hypotheses which suggest object identities and poses in images. In this paper we concentrate on object recognition from single images. Recognition systems which treat single images tend to start off by extracting features from the image, then they do feature grouping and model selection (frequently called indexing), and penultimately they include a correspondence stage which pairs together model and image features. In mature systems a verification step is then included which finally determines the degree of correctness of the model-image correspondences.

In this paper we study how different verification strategies can affect the quality of the results returned by a recognition system. Overall, we will show that an adequate understanding of model and image *topology* can lead to enhanced verification strategies, and thus ultimately towards better system performances. We use the term topology to represent the connectivity relationships between features. Locally the topological representations we use convey whether two model features are adjacent and should be seen as such in an image. More globally we can ask whether a continuous chain of features should be present between any pair of features, and so define notions of global connectivity.

Topology allows us to make neighbourhood based inferences. The observance of a particular model feature in an image would most likely indicate that all features adjacent to the first feature on the model should also be visible. The failure to find the adjacent features in the image would thus indicate either of two things: perhaps an occlusion is present and so the features are hidden; or conversely that the hypothesis is wrong and should either be discarded or at least have its importance diminished. Discrimination between these two types of incident is achieved via topology and occlusion analysis. There are a number of variants on this type of reasoning which we pursue throughout this paper.

1.1 Conflation of correspondence and verification

Object recognition strategies have formed a fundamental part of computer vision research. The large number of different design strategies have thus been recorded extensively in the literature as a whole, and more particularly have been the subject of a number of review books and articles. One such review is the relatively recent book by Grimson [10]. This book is actually more than just a summary of the literature, but is in fact an in-depth study of the capabilities of a number of different approaches to the correspondence stage of recognition. From reading this book it quickly becomes apparent that the major differences between the various recognition systems lies in their approach to tackling the correspondence problem. Grimson provides a breakdown of recognition into three general phases:

- *Selection:* what subset of the data corresponds to the object?
- *Indexing:* which object model corresponds to the data subset?

- *Correspondence:* which individual model features correspond to each data feature?

This decomposition is partly the key to the simplicity of many verification algorithms. All too frequently a system implementation possesses no explicit verification stage but rather embeds the final steps of reasoning about a recognition hypothesis within the correspondence phase. It is this conflation of correspondence and verification which means that verification methods have to use the same reasoning mechanisms as those employed for correspondence. Thus they seldom have a marked effect on the resultant recognition hypotheses.

Some of the many systems which develop an emphasis for correspondence are those created by Ayache and Faugeras [1], Pollard, *et al.* [21], Grimson and Lozano-Pérez [11], Bolles and Horaud [2], Faugeras and Hebert [9], Thompson and Mundy [26], Huttenlocher and Ullman [18], Lamdan and Wolfson [19], Stein and Medioni [25], and Califano and Mohan [5]. These examples can be loosely classed into the different categories of interpretation trees, alignment, and geometric hashing. Of course this list is somewhat modest and excludes many other examples of recognition systems. However, it is largely representative in the range of algorithms and the differences in performances which are available. Anyway, the specifics of the correspondence algorithm are not intended to be the focus of this paper, we only mention these algorithms to highlight the fact that verification is often little more than an in-depth correspondence analysis, rather than an independent process.

1.2 The need for hypothesis verification

Commonly the final conclusion of a recognition algorithm is that a set of model features match a set of image features. The same set of correspondences imply a geometric mapping from the model to the image (or to a three-dimensional reconstruction of the scene for 3D sensing techniques). Thirdly, there is also usually a measure of the number of matching features as a percentage of the whole. The acceptance of a correspondence hypothesis is therefore often based on the following three criteria:

1. That a suitably large number of matching model and image primitives should be found. These primitives might be line segments, corners, certain types of surface patch, etc.
2. That the pose of the object in the scene should be realistic. A conceivable bad pose would be that the hypothesized object lies somewhere below a known solid ground-plane.
3. That a suitably large proportion (say two-thirds) of *all* of the model data can be explained by observed image features. For three-dimensional objects, a complete model description might be a set of small surface patches derived via a triangulation of a CAD-type model, or similarly a set of planar curves for a two-dimensional object.

By way of example, we can see the presence of the first and last of these processes in the two-dimensional HYPER system of Ayache and Faugeras [1]. In this system an interpretation tree is used to match a *sufficiently large* set of coplanar line-segments from the model to the image. Normally three or four correspondences might be enough to compute the transformation between the model and image. The matching process continues until a certain percentage of the model boundary has been explained by image features, or potentially until the system realizes that enough of the model cannot be recovered due to occlusion or some other reason (in which case the hypothesis can be rejected).

Good examples of the use of the third criteria can be found in the verification phase of the system of Thompson and Mundy [26] which relies on the computation of the distance transform with Chamfer filters (Borgefors [4]). Another instance is in the matching algorithm of Huttenlocher, *et al.* [17], and their use of the Hausdorff distance.

The three simple tests are quite likely to provoke some rather specific problems. One major concern is expressed in Fig. 1. In the figure we consider the apparently simple task of trying to match either of two objects in a model base (a rectangle and a square) to an image of the rectangle occluded by a third object. The image has been produced merely via a plane Euclidean distortion (more normally we have to cope with far more varied transformations in computer vision applications). Referring to the figure, if the our matching thresholds are too generous then we will be able to find a near perfect match in the image for the features of the square and yet only be able to match a proportion of those for the rectangle. One would therefore be led to believe that the hypothesis for the square is more likely than that for the rectangle. Note that this image situation may be considered to be *generic* as it does not involve any significant accidental alignments (occlusions of this nature are forever present in the images we have studied). However, it is worth noting that there is actually enough information in the image to make is obvious that the square hypothesis is incorrect.

The additional information involves the nature of the topological structure of the models in the model library: features are joined by well-defined junction types and not related purely by geometric measures (such as angles and distances). Previously little emphasis has been placed on model topology. For the purpose of this paper we represent topology at the lowest level by *vertices* between which span *edges*, and chains of edges which are called *one-chains*. Closed one-chains form cycles which are called *faces*. Adjacency of low-level image features is then tested by reasoning about whether they come from the same edge, or two edges in the same one-chain, etc.

Along the two occlusion events in Fig. 1 we observe two 'T' junctions (in the terminology of Waltz [27]). In contrast, had the square hypothesis been correct we would have expected to find 'L' junctions in the same area of the image. The lack of the 'L' junctions immediately suggests that the square hypothesis is incorrect. Moreover, the observation of 'T' junctions in general implies occlusion and so provides an immediate explanation of why the rectangle model failed to receive image support. It thus strengthens the presence of a rectangle hypothesis and suggests that basing verification scores purely on the degree of model-image alignment is inadequate.

The need for this type of reasoning was first remarked on a long time ago. For instance, Guzman [12] showed how junction analysis would not only improve recognition performance, but could be used as a basis of in-depth scene reasoning. However, more recent approaches to recognition have paid little heed to explaining the unexplained in images, and are usually

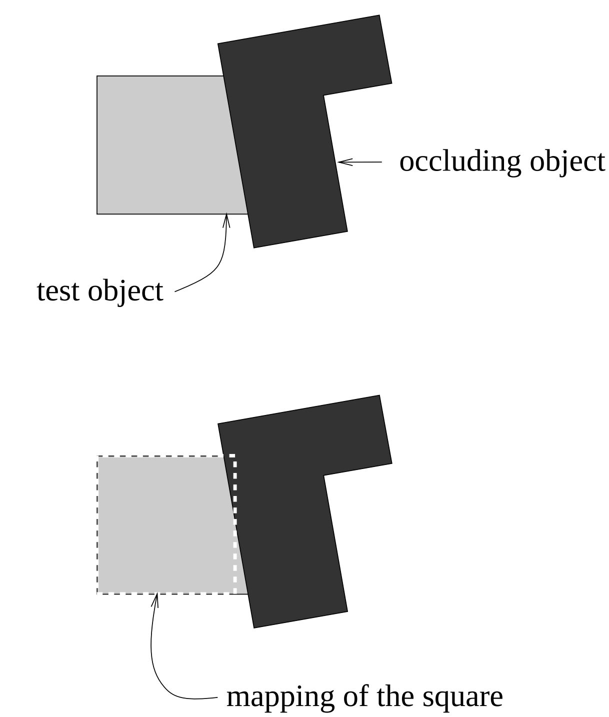


Figure 1: *If our recognition system has a model base of a rectangle and a square, we will find that the square is more likely to score higher for the occluded test object than the rectangle which is really the correct match.*

content to interpret only positional evidence. Once topological information has been introduced one can begin to include in-depth reasoning about negative evidence of hypotheses rather than working purely with positive evidence.

1.3 The requirement for an implemented recognition system

Our work is based on a re-implementation of the LEWIS recognition system of Rothwell [24]. It derives object hypotheses via invariant-indexing based on planar algebraic invariants. We summarize the processing steps later in Section 2. We are not really interested in the way that recognition hypotheses are derived, and so we do not go into detail about how the system works (in fact we could equally have used any of a number of other different recognition systems). The importance of working with an actual system is that recognition hypotheses are recovered from images with realistic measures of confidence and error. Thus we can examine the limits of the verification procedures more thoroughly.

LEWIS behaves in a way typical of recognition systems in that it does not derive perfect recognition results. Depending on the choice of a set of parameters, one can vary the number of false positives and false negatives returned. However, due to the desire to recognize occluded objects in fairly complex scenes one is unlikely to produce only correct hypotheses. Therefore, it is sufficient to say that although a correspondence stage manages to rule out most of the false hypotheses, there are still likely to be incorrect interpretations left after the completion of the processing of the image. In [24] it is reported that roughly as many false positives were recorded as correct evaluations for moderately complex scenes with occlusions and a relatively large model base.

This level of failure could be construed as being somewhat worrying, though we believe that the depth of scene analysis provided by LEWIS was insufficient and that far more can be done. Principally, hypotheses were accepted if a sufficient amount of positive evidence for model features was found in an image. In practice this criterion was satisfied by the recovery of fifty percent image support for a hypothesis. Certain other hypotheses could be ruled out through pose considerations (the hypotheses would indicate that the objects would need to be too distant in the scene to be observed clearly in the image, or similarly have too high a slant in the world for reliable feature extraction to be performed). However, at no stage were the following considerations evaluated:

- The presence of negative evidence must be explained. Primarily we should be able to locate occlusion events which justify the lack of measurement of a scene-model match. Failure to find occlusion events reduces the likelihood of a hypothesis being correct.

- Some notion of object topology must be preserved, unless there is positive evidence of occlusion. Therefore two image features should not be marked as coming from adjacent features on an object's boundary unless they are either connected in the image, or unless there exists an occlusion event between them.

- Under generic viewing conditions there must be uniqueness of solution. A single feature cannot belong to more than one object. Multiple partitioning of image data was previously allowed due to a lack of confidence in the existent verification procedures. With improved verification we can perhaps move towards deriving absolute conclusions about scene hypotheses.

This last point requires development. The nature of the results given by different recognition systems varies dramatically depending on the application. Often the recognition problem is posed (perhaps implicitly) as the task of finding a specific object in a scene, then once the object has been found processing is terminated. This problem is significantly different from (and simpler than) that of attempting to identify all objects in a scene which might correspond to any of a number of objects in a model base. In this situation a single mistake can lead to catastrophic failures in interpretation because any decision is considered as final and affects all subsequent processing. It is perhaps therefore wiser to be conservative and to allow multiple interpretations so that no truly correct hypothesis is discarded. However, most applications might be expected to provide unique and accurate scene interpretations. We are therefore interested in moving towards the notion of single feature interpretations whilst still working in relatively unstructured and unknown environments.

1.3.1 Outline

Prior to discussing a number of new verification methods we discuss our implementation of the LEWIS recognition system in Section 2, and provide some examples of the system working in Section 3. Then, the different verification methods are discussed in Section 4 along with examples which demonstrate how the methods work in images.

2 The recognition architecture

We now briefly describe the main aspects of our implementation of the LEWIS planar object recognition system described in [24]. Overall our system is very close to the original implementation, and so full explanations are omitted. The reader is advised to refer to [24] for complete details, though in many ways the precise working of the system is not so important, but rather the fact that it represents an implemented system which derives recognition hypotheses. The main functional difference between our new implementation and the original version lies in our more developed verification algorithms. The mechanisms and effects of these algorithms are the main contribution of this paper.

The application of the system lies in the recognition of planar algebraic feature sets (namely lines and ellipses), and so all important model aspects are restricted to being two-dimensional. However, the model descriptions, which are invariant measures, are projective and so the positions and orientations of the objects to be recognized in three-dimensional space can be almost arbitrary. The recognition system is principally built around a pipeline architecture which computes indexes from scene features that form model-image feature hypotheses. The main steps are briefly:

1. **Edge detection:** edges are detected in the image using our implementation of the Canny [6] edge detector. Full details of the filter are given in [23]. The main difference between our edge detector and previous versions is that much better image topology is recovered from the scene (using adaptive thresholding). As is standard, edgel chains are extracted from the edge image with sub-pixel accuracy.

2. **Feature fitting:** lines and ellipses are fitted to the edgel chains. This is done using incremental algorithms based on orthogonal regression for straight line segments and a modification of the Bookstein [3] algorithm for ellipses.

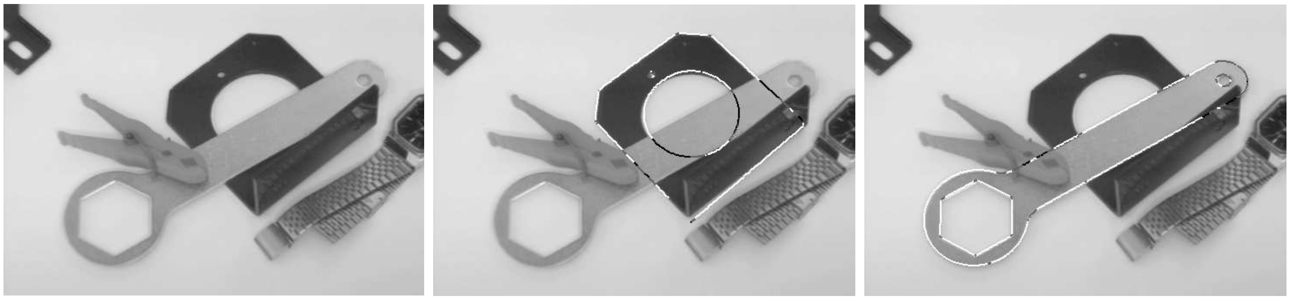
3. **Grouping:** lines are *grouped* by connectivity into feature groups and conics by proximity. The goal of the grouping process is to derive sets of invariant feature groups of the following three forms: groups of five lines; three lines and a conic; and pairs of conics. Each of these feature groups possess a number of projective invariants which are suitable invariant descriptions for the sets of plane algebraic features.

4. **Indexing:** the invariants for each of the feature groups are used to *index* into indexes spaces. The index spaces (one for each type of invariant) are represented as hash tables. Should the invariants for an image feature group match those for a model, then a hypothesis is constructed which expresses the model-image *correspondence* formally.

5. **Formation of extended hypotheses:** The invariant feature groups provide only *local* descriptions of objects, and do not encompass all of the features of a model. The distribution of such feature groups around the boundary of an object frequently leads to the formation of a number of hypotheses for a single object in a scene. Compatible hypotheses from different feature groups should be merged together prior to verification to form *extended hypotheses*.

6. **Verification:** the model-to-image transformation is computed using the corresponding model and image features in the extended hypotheses. A projective transformation must be found which maps all of the model algebraic features sufficiently closely to the image features (lines and ellipses), otherwise the hypothesis is rejected out-of-hand. The entire set of model features can then be projected into the image and compared to the image data. The model features are represented by edgel sets recovered from an acquisition image of the model object. An extended hypothesis is accepted if more than fifty percent of the model features are found to project to within five pixels of image edge data of the right orientation (the orientations must differ by no more than fifteen degrees).

Lewis also included additional stages of verification based on pose computation. We have not yet included them in our implementation as it appears that a more complete analysis of the quality of the hypotheses can be achieved prior to resorting to pose-based computations.



a **b** **c**

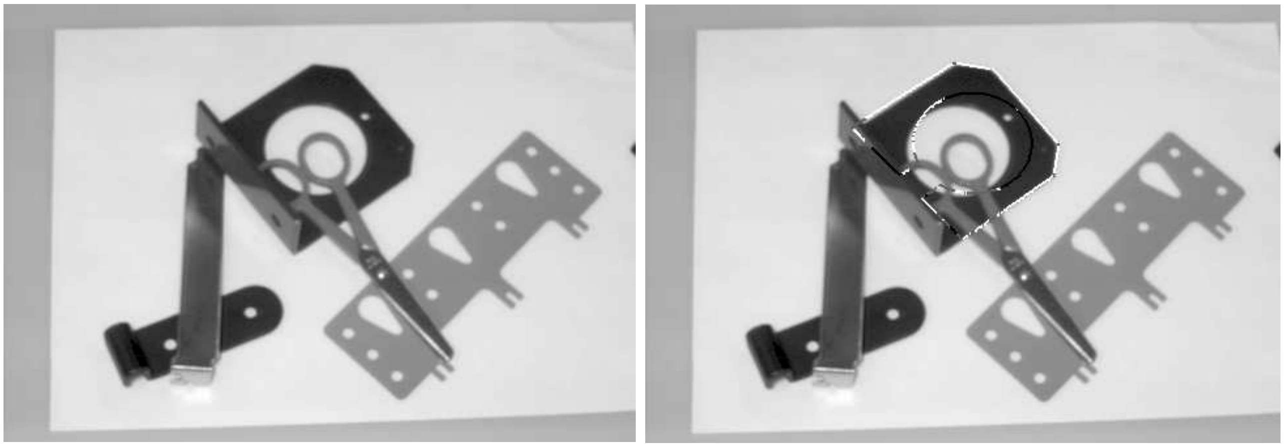
Figure 2: *The bracket is recognized despite being occluded by the spanner. Completion of the original verification process is demonstrated in (b) by the back-projection of the bracket model (which is a set of edge data) into the scene. The white data from the model represents all of the model points which found matching image edge data within five pixels and having a matching orientation difference of less than fifteen degrees. In all* 70.5% *of the bracket model finds support in the image. The grey parts of the outline represent model features which find image edgels within five pixels of their projections, but which have an incorrect orientation. The black segments of the model outline are those which find no matching image data. Neither of these last two classes traditionally count towards the verification score. Recognition of a spanner is shown (c) with a* 77.8% *level of image edge support for its model.*

3 Recognition examples

We now discuss some recognition examples to see how the system performs and to provide an idea of the sort of environments in which it works. The images are courtesy of [24], though the processing has been done in our re-implemented system. Using the same the examples provides a direct comparison of the effectiveness of our reasoning against the tests reported for the original implementation. It thus demonstrates where the verification analysis breaks down and suggests where we can look to improve performance.

The first recognition example is shown in Fig. 2 where both a bracket and a spanner are recognized in the presence of mutual occlusion and clutter. The verification scores used involve estimating the proportion of projected model data which is supported by image edge data of the right orientation (within fifteen degrees) and which is sufficiently close (no more than five pixels away). The bracket received 70.5% image support and the spanner 77.8%; both of scores exceed fifty percent and are thus accepted as correct hypotheses (the requirement for visibility of fifty percent of the object outline is the same as the original verification criterion).

As we shall see, just marking model features as good or bad is an insufficient match criterion. In Fig. 2 we have also shown which model edgels lie within five pixels of image edge data but fail the orientation test (marked in grey), and also those edgels which fail



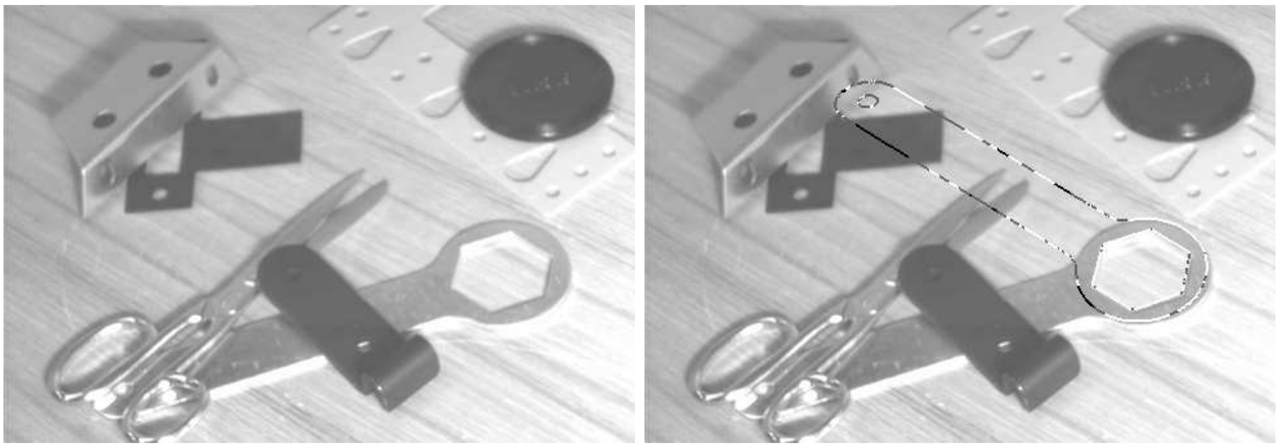
a **b**

Figure 3: *A model from the model base which has a five-line configuration similar to that of the bracket is hypothesized in the image as corresponding to a view of the bracket (on the strength of the single five-line invariant matching). Back projection of the model recovers* 55.0% *level image support, which is sufficient under the verification criteria to be marked as accepted.*

both the distance and orientation tests (in black). We shall see that it is those edgels which lie in the second category (black) that provide the first clues of a potential breakdown in the construction of a hypothesis, and hence image regions around these edgels need to be examined in more detail than has been done previously.

A very characteristic example of the failure of recognition is shown in Fig. 3. Here a model has been incorrectly hypothesized via indexing and is given further support by the fact that an apparently correct model-to-image transformation can be computed. A 55.0% level of image support is found via edge matching and so the hypothesis is marked as accepted. Obviously there is an error as a similar object from the model base contains a local feature group with similarity to that of the bracket, though using the original verification measures it cannot be ruled out.

A different type of failure mechanism is shown in Fig. 4. This is where a false positive is created due to the presence of spurious scene features which result from the existence of a linear texture. (Coincidentally the correct model was chosen from the model base, but its pose was incorrectly calculated. We also managed to create a hypothesis with the correct pose, but as we do not use any higher-level processing at this stage, we are unable to arbitrate between the correct and the false hypotheses.) The problem arises because unconnected image features provide support for the model hypothesis in areas away from the feature group which initialized the index. In this case 55.2% of the model edges found



a **b**

Figure 4: *In this case an invariant configuration causes the estimation of the incorrect pose for an object. This is due to the symmetry of the model (the correct pose is also recovered by another hypothesis). Under verification, and due to the presence of significant linear texture, the poor-pose hypothesis is accepted with* 55.2% *image support.*

matches in the scene even though only a few of them actually projected onto the features used for indexing.

4 Enhanced verification methods

The examples in Figs 3 and 4 highlight two areas where verification fails due to the insufficiency of current reasoning methods. The fundamental problem is that the approaches used up until the present time treat each individual model element separately, and so make no use of the concept that a model is really an integrated topological structure. Advances can be made quite rapidly when we start to diffuse the local acceptance or rejection of a hypothesis around an object boundary. In short, in considering the acceptance of any one model element, we must analyse the outcome of the verification procedure for its neighbouring elements.

The two driving notions in verification thus become *topology* and the understanding of *occlusion*. The topology of the image features which are assigned correspondences to model features must match the topology of the model features exactly. Exception to this can only be permitted due to occlusion, or when we are faced with the time-old problem of extracting reliable segmentations from images. We thus need to develop algorithms for analyzing the topology (connectivity) of features and for estimating the presence of occlusions.

The existence of recognition hypotheses does however help to resolve one of the hardest problems which is faced in image analysis. The extraction of low-level features such as occlusion events is very difficult without *a priori* information (which is normally the case). However, a hypothesis suggests where each of a specific type of the low level features that we need to examine is located, and so we are able to employ what is really just an extreme model-based approach for their evaluation. Consequently, we can use top-down processing to provide low-level information for topology reasoning, or we can simply re-use bottom-up information that we recovered previously. In fact, both types of data are used during detailed verification.

Incidentally, the exact way in which a hypothesis is considered is very important with relation to the algorithms that can be used in verification. We believe very firmly that a hypothesis should be *presumed to be true until proved otherwise to be false* rather than a more aggressive approach which tries to maintain as few hypothesis as possible at any one time and so prunes associations quickly. Proceeding along the more conservative route ensures that any hypothesis corresponding to a correct interpretation is unlikely to become discarded. In contrast, the second approach is very good in situations where a single object might be in a scene, and where that object has little similarity to other objects in the model base, but it can often be forced into making incorrect decisions.

4.1 Occlusion reasoning

The first step towards better verification involves reasoning about the presence and nature of occlusion events. In cluttered scenes we are very seldom able to recover image support for the entire boundary of the projection of a model hypothesized through indexing. Often the objects are occluded in scenes, or perhaps they may cause partial self-occlusions if they actually have their own three-dimensional structure. The key issue is that the projection of the model into the image indicates where occlusions appear to be arising (due to the loss of image support) and so we should be able to find independent evidence for the occlusions. If we cannot, then it is likely that the original recognition hypothesis is incorrect and so we might do better by considering a different interpretation.

Occlusion events are typically marked by the presence of 'T' junctions in the image edge structure. The understanding of the nature of different types of junction was originally studied in the line-labeling programs of Guzman [12], Clowes [7], Huffman [16], Mackworth [20] and Waltz [27]. Principally an object feature which undergoes occlusion should be cut and terminated by a locally straight transverse line segment. Thus when occlusions are suggested by the sudden loss of image support for the projection of the model we should try and look for 'T' junctions.

In fact we have to be a little more careful in deciding where occlusions might be. Just following along projected model features until we first fail to find a match leads to a significant over estimation of occlusions. This is because small sets of projected model edgels frequently fail to find correctly oriented image edgels even though edges of roughly the right type of structure can be found very close to the projected model. Normally this is due to the edge orientations computed by the edge detector becoming erroneous near corners and junctions (and hence towards the ends of object features).

We were thus led towards using a three-level classification for the projected model data, that is: strongly matching features; those with close image support but where the edgels are of the wrong orientation; and thirdly those features which find no image support nearby. The third category indicates very markedly the presence of an occlusion, though the locations of transitions into this class tend to be some distance from the real occlusion events (often about five pixels). However, the transition into the second class (when caused by a real occlusion) is usually located within a pixel of the actual occlusion event, and so we need to employ the accuracy of the transitions into the second class along with the robustness of the membership of the third class.

The hypothesizing of occlusions is consequently derived from a two step process:

1. Find all of the projected model edgels which have no image support. These form connected sub-sets. Go to the boundaries of these sets (to where image support is first totally lost).

2. If these boundaries are adjacent to projected model edgels which have full image support (distance and orientation), we are potentially at an occlusion event; therefore store the current location. Otherwise, track along the projected model curve until we first find a model edgel with full support; mark this location as the position of a possible occlusion.

Once the sites of potential occlusions have been hypothesized we may evaluated their likelihoods in two different ways. The first makes direct use of the edge detector information which was computed as the initial step of image processing. Initially we recover all of the junctions in the original edge image which have an order greater or equal to three. Junctions of order three are those where three edgel chain curves meet at a single point, and may be classed as either 'Y', *arrow*, or 'T' junctions. As we are only interested in the latter class, there is no need to draw a distinction between the arrow and 'Y' junctions. We declare the presence of a 'T' junction when the angle between any pair of the edges meeting at a junction are within twenty degrees of 180 degrees. The large tolerance reflects the fact that edge contours are frequently displaced by significant amounts near junctions, and it also allows for occlusion by curved objects. If the 'T' junction lies sufficiently close (a small number of pixels, such as five) to where the recognition hypothesis deduced that there should be an occlusion event, then we can add confidence to the hypothesis.

We also accept higher order junctions such as when four edge contours meet at a point. Although these events are particularly rare, they *might* reasonably be interpreted as occlusion events so long a similar angle constraint to that used above is satisfied.

This first level of processing does not resolve all of the occlusion events we find in an image. This is because most edge detectors are notoriously poor at recovering meaningful connectivity at junctions. Although the edge detector we use [23] provides better topology than a filter such as the Canny [6], it still fails to recover all of the 'T' junctions in the image. We therefore resort to a second test which examines the overall structure of the image intensity data near to the location of a potential occlusion. This is done by parametric model fitting to junctions such as suggested in the approaches of Hueckel [15], Rohr [22], or Deriche and Blaszka [8]. In fact we use the algorithm of [8] in our tests.

The type of parametric model fitted assumes that the surface is composed of a number of constant intensity plateaux which meet at the junction and are separated by straight edges. Each plateau represents an image region and smoothing is accounted for between the image regions through an approximate parametrization of Gaussian smoothing. The algorithm of [8] fits such a model over a specific window size at a given location in the image (where we suspect that there is an occlusion), and returns a number of different parameters which represent the interpretation of the intensity surface. The key measures which are returned are a fitting cost, the grey level values of the plateaux, and the angles at which the edges come into the junction. We can then estimate whether the junction is a real 'T' junction by looking at the angles between the edges and by making sure that the plateaux have sufficiently different grey levels (they should be spaced out by at least ten grey levels; we are using 256 grey level images).

4.1.1 Examples - edge contour junctions

At this stage it is educational to analyse some examples which demonstrate how the two methods work on the images which are of interest. Given this knowledge we can then proceed to see how the different 'T' junction reasoning processes can be used to enhance or detract from the different object recognition hypotheses as they are produced by the indexing system.

First, in Fig. 5 we show how a set of edgel data from the edge detector hypothesizes the presence of 'T' junctions near to where occlusion events should be found on the strength of a specific recognition hypothesis. This is an example of positive support for an object hypothesis, with the measurement of the low-level junction description enhancing the acceptance of a hypothesis. In Fig. 6 we demonstrate an example of negative evidence which renders a hypothesis unlikely (or perhaps may even cause it to be rejected outright). The projected model curves are shown initially following a contour in the image, but at one stage (due to an incorrect model being hypothesized), the projected model curve parts quite clearly from the image curve. The edge detector fails to find any junctions near the point of departure and so one can start to doubt whether the original hypothesis is correct.

This type of reasoning appears very attractive on the weight of the two examples given. However, most edge detectors are notoriously poor at recovering correct edgel contour connectivity near junctions. This means that failure to record the presence of a 'T' junction can be taken as only partial evidence against a hypothesis, and should not be used too strongly. Conversely, if we are able to find a 'T' junction where predicted by a recognition hypothesis then we can add considerable weight to the hypothesis. The progress we make in marking 'T' junctions is apparent as some of the projected edge data from the model of the hypothesis originally failed to gain image support and was treated neutrally, now the lack of image support for those edgels can now be treated as *positive support*.

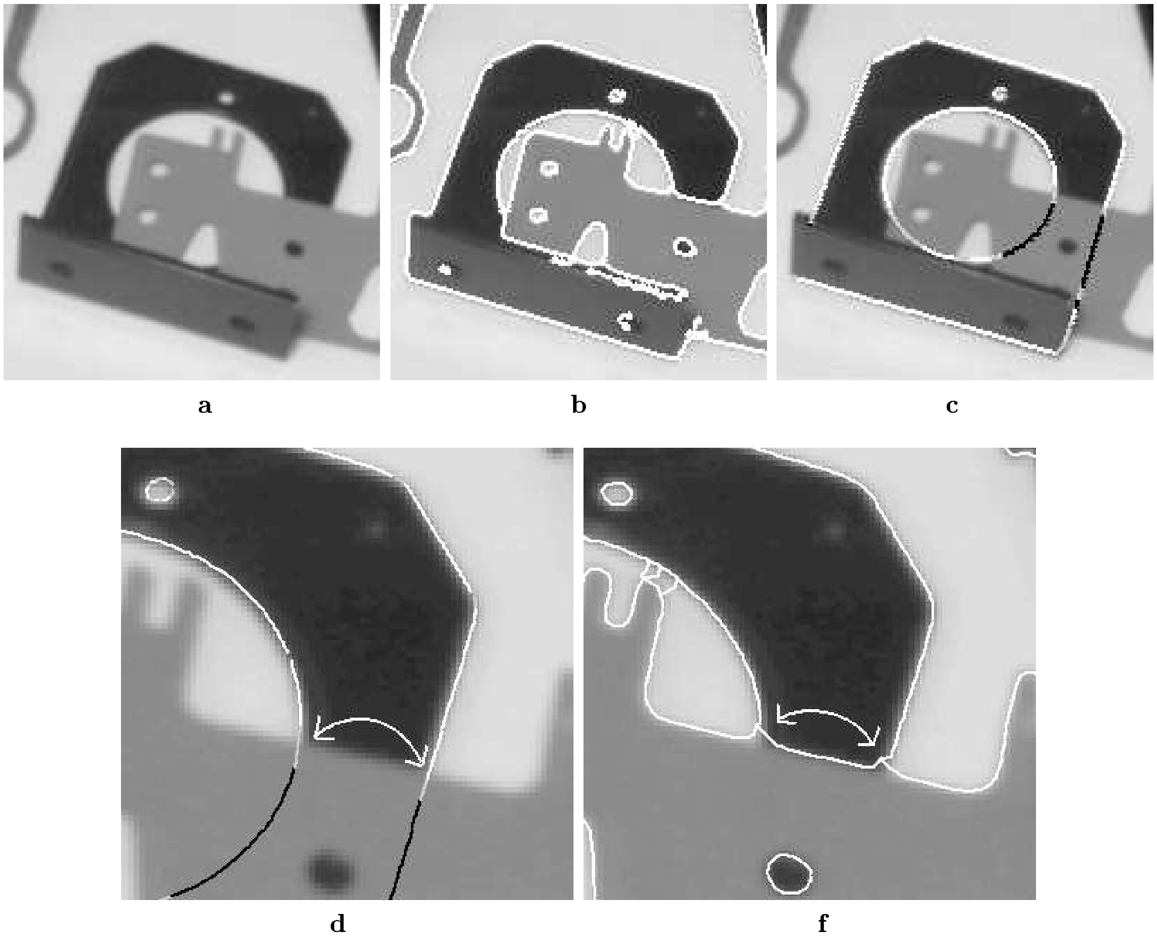


Figure 5: *In (a) we show part of an image which includes an object from the model base. The edges found by the edge detector are shown in (b), and the (correct) projection of the matched model in (c). The recognition hypothesis suggests the presence of occlusions at the transition zones from white to grey of the projected model; this is shown in more detail by the arrows in (d) for two of the occlusion regions. Near to these occlusion points are triple junctions in the edge description which have the form of a 'T'. This is shown in (f) which is a close-up of (b), and so the hypothesis appears to be correct.*

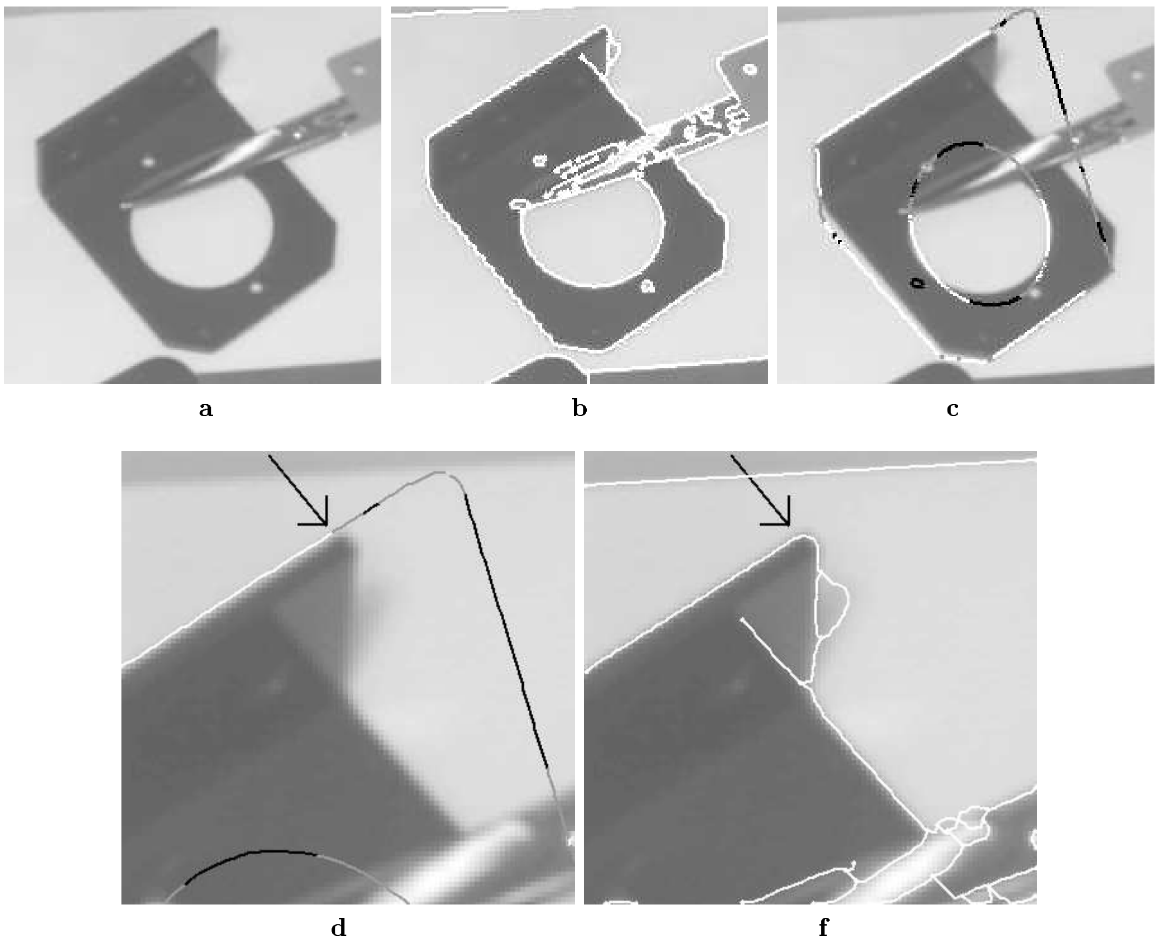
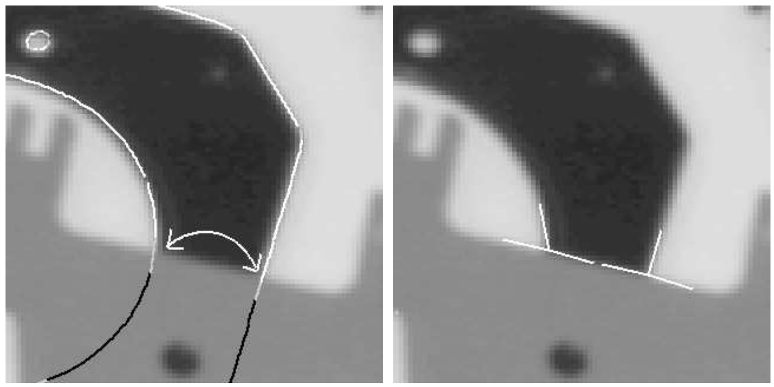
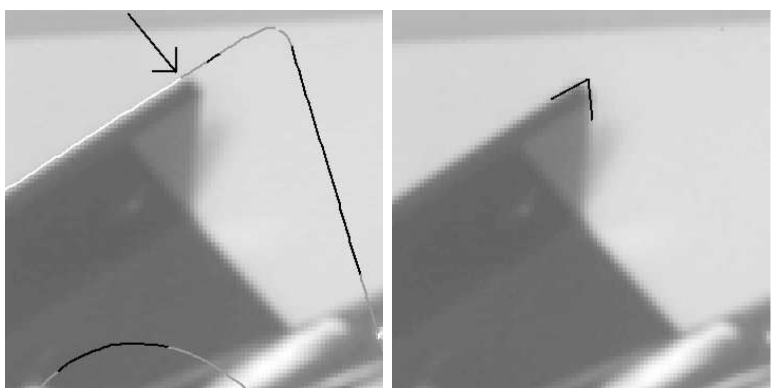


Figure 6: *For the image section in (a) we have computed the edgels shown in (b). From these and a subsequent fitting process we compute invariants to form the incorrect hypothesis shown in (c). The features used to compute the hypothesis are not shown. The indexing has incorrectly hypothesized a match to a model which finds significant image support. Near the white-grey transition of the projected outline in (d), which is a close-up of (c), we would hope to find a junction in the edge description, (that is if the hypothesis were correct). However, as shown in (f), the failure of the edge detector to record a junction in the right place suggests that the hypothesis is false.*



a **b**

Figure 7: *In (a) the close-up of the correct hypothesis shown in Fig. 5d is depicted, and the output of the Deriche-Blaszka junction detector is shown in (b). The left-most junction possesses an angle of 179.3 degrees between the straight part of the 'T', and the right-hand junction 175.8 degrees. Both of these are sufficiently close to 180 degrees, and so can be accepted as 'T' junctions. What is more, the differences between the closest grey levels for both junctions are just over 55 grey levels, and so the different plateaux are clearly distinct.*



a **b**

Figure 8: *The Deriche-Blaszka filter also fails to find a suitable 'T' junction near the potential occlusion event already discussed in Fig. 6. Instead it finds an 'L' junction which is shown in (b). Such a feature has no relation to an occlusion event, and so the original hypothesis is likely to be false.*

4.1.2 Examples - parametric junction models

The susceptibility of the approach based of finding edge contour junctions to poor segmentation actually led us to examine the effectiveness of parametric junction model fitting. In

Figs 7 and 8 we show similar examples to those for the edge junction reasoning, but this time the presence of 'T' junctions is evaluated using the Deriche-Blaszka operator [8]. In the first case a pair of suitable 'T' junctions are found and shown superimposed in the figure, and in the second case no such fit could be found and so the hypothesis is marked as being unlikely. Thus we see that parametric model fitting can enhance the understanding of recognition hypotheses when directed by prior topological reasoning.

Again, these two examples of the use of the parametric junction model approach do not describe the whole truth. Over repeated trials with a considerable number of images we have found that the Deriche-Blaszka model does not actually represent the image intensity surface correctly. Certainly the constant grey level model is correct for image areas where the global intensity gradient is small (though not the local intensity gradients which represent edges), where the objects are very much polyhedral or polygonal, and where the overall complexity of the scene is low. The examples we have shown so far all satisfy these image conditions. More typically the differences in grey levels between the different regions of the image are very slight near junctions, and the grey levels are better represented by sloping surfaces rather than constant grey levels. Thus the simple parametric model fitting approach fails. One could of course try to use a more sophisticated junction model which allows the grey levels in a region to vary linearly, or by some other model. However, it is likely that such a model would have too many parameters, so the process of fitting instances to the image surface would be likely to be unstable and so not to yield reliable or accurate results. We would certainly be interested in employing such a parametric fitting model having a stable performance, but to our knowledge no such feature detector exists. Nevertheless, as development of the system continues we shall investigate a number of other features detectors such as those of Heitger and von der Heydt [14] and Forstner [13].

Consequently, it is reasonable to consider the use of junction-model fitting to be similar to that of the edge contour junctions: accept a 'T' junction if it is hypothesized in the image, but failure to record an instance does not mean that such a feature does not really exist. Overall, we have found that the Deriche-Blaszka model provides much better confidence of the presence of a 'T' junction than the edge based approach for controlled images, but the fewer assumptions made during edge detection about the shape of the intensity surface means that an edge detector often outperforms the more complex approach in marginal cases. One such example of this is shown in Fig. 9, were the Deriche-Blaszka model fails completely to find a suitable fit for a triple-junction model, but the edge detector (when using exactly the same parameters as for recognition) succeeds in finding a junction of the right nature.

4.1.3 Summary

So far we have demonstrated how occlusion events can be hypothesized by studying the model topology information contained in hypotheses produced by a typical recognition system. Hypothesized occlusions can be evaluated in a number of different ways. We have demonstrated two such methods: the first using bottom-up information recovered from the original output of an edge filter; and the second derives top-down data resulting from the

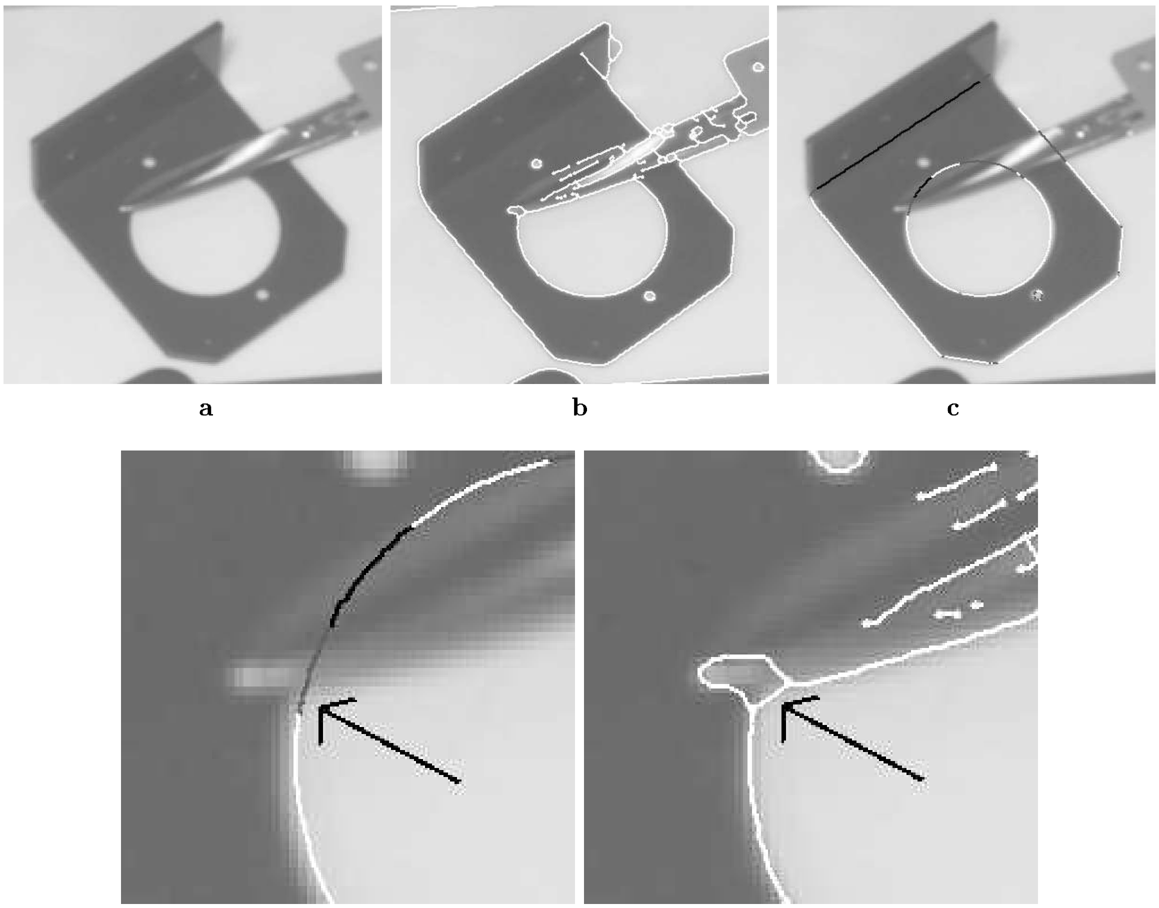
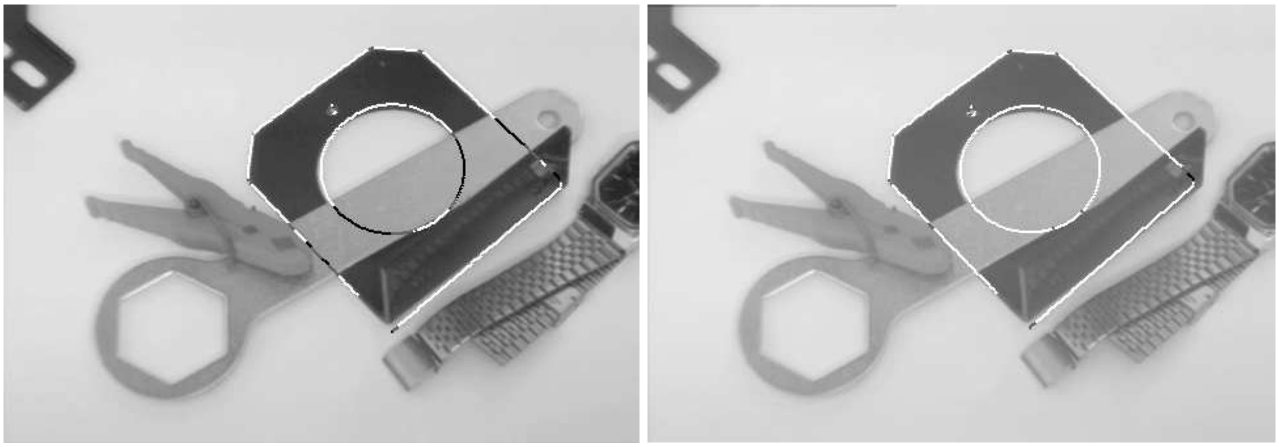


Figure 9: *Again we consider the image in Fig. 6, but this time we are interested in a correct hypothesis. The edge detector output is given in (b) and the projection of the model is in (c). When we consider the edges around a real occlusion event as shown in (f) we find a triple-junction which bears resemblance to a 'T' junction. However, due to the similarity in the grey values around this part of the image, and also as a result of specularities on the occluding object, the Deriche-Blaszka filter fails to find fit a suitable junction model.*



a **b**

Figure 10: *The original level of verification for the bracket produced only a* 70.5% *score for the hypothesis shown in (b). However, after the prediction and verification of the various 'T' junctions bounding the invisible part of the hypothesized object, we increase the verification score* 93.6%. *Edgels which were previously unmatched, but are now marked as positively occluded, are depicted in white in (b). The overall verification score is incremented by the number of edgels in the positively identified occluded regions.*

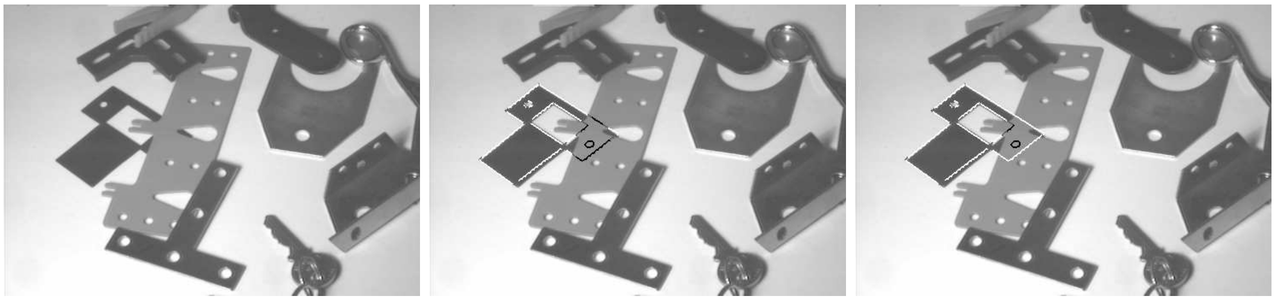
application of parametric junction model fitting. For certain cases the second method is more accurate and more reliable, but in general it relies on making incorrect assumptions about the shape of the intensity surface.

Nevertheless we can employ both methods with caution. Whenever either approach suggests the presence of a 'T' junction, we can be be fairly confident that it is right, however they both frequently reject junctions which do actually correspond to occlusion events.

4.2 How to update hypotheses

The two methods described above for evaluating the presence of 'T' junctions allow us to update the scores given to hypotheses during the back projection stage of verification. Whenever we find an occluded region which is terminated at one end by a verified 'T' junction, it is marked as making a positive contribution to the hypothesis. This is to say that the score of *visible model edgels* used to compute the overall verification score should be incremented by the number of edgels in the occluded region.

Consequently we can transform a hypothesis such as that shown in Fig. 10 (which is the same hypothesis as shown in Fig. 2b) from a 70.5% recognition score to a re-evaluated score of 93.6%, and hence have little doubt that the hypothesis is correct. Ideally we would hope that the new score would tends towards 100%, but there are always short sections of



a **b** **c**

Figure 11: *The original verification score for the object in (b) was* 70.7%. *After occlusion reasoning the confidence level rises to* 83.6% *which is a clear indication that the hypothesis is correct.*

projected model curve which project near to image features with the wrong orientation (and so are not marked as being caused by occlusion). Taking these into account would make the hypothesis in Fig. 10 take on a score of little less than 100% (in fact 98.2% when we ignore small section of less than five pixels in length). The dominance of a good hypothesis such as this one significantly enhances our understanding of the scene. From a number of experiments we are able to conclude that a final matching score of over 90% leaves little doubt as to the identity of an object (this score includes occlusion reasoning and the removal of short unmatched segments of less than five pixels in length).

A second example is shown in Fig. 11 where the score of an occluded object is increased from 70.7% to 83.6% after occlusion reasoning (and 91.3% after removal of short unmatched chains). Such a high recognition score usually only occurs when the hypothesis really is correct, and so we may have a significant degree of confidence in this hypothesis.

4.3 Correctness of image topology

So far in this paper the effects we have been interested in have been dominated by model topology and cause-and-effect reasoning for connected parts of the model to be either occluded or visible. It is quite understandable that we can make reciprocal considerations with regard to image topology, or more properly between the consistency of both of the model and image descriptions. In an ideal world (where we would of course cease to be frustrated by problems in segmentation), the image topology should exactly match the projection of the model topology.

Even with our current segmentation abilities we can develop simple tests which show some robustness to errors in the extraction of the geometric interpretation of the image, and which indicate whether the grounds for believing a hypothesis should be reduced. Note that this approach fits in with the *true until proved false* philosophy in that we try not to discredit hypotheses totally until the evidence becomes insurmountable. For instance,

a **b**

Figure 12: *The poorly oriented hypothesis in Fig. 4 required support from a large number of unconnected image features caused by an underlying texture. These result in an image topology (many disconnected edge curves) which is inconsistent with the model topology (a few long curves). As there are no indications of occlusion events at the end of each image curve, we infer that the correspondences are incorrect. Subsequently we can doubt the hypothesis.*

we can start off by seeing which sets of image features have been given a correspondence with a single model feature. Due to the trivial connectivity constraints which exist on a lone model feature one would also expect the image features to be connected. There are a number of different reasons for the images features becoming distinct, though still remaining topologically linked either directly, or by being in the same one-chain. Perhaps the segmentation and fitting procedures separated the features at the geometric level by attributing each one to different algebraic objects, or perhaps the edge detector erroneously placed a junction between them which provides connectivity with other features. However, in both of these cases the image topology is consistent with a single model feature, and so we need not doubt the integrity of a particular hypothesis.

Conversely, out attention should be drawn to instances in which connectivity has been lost. Of course there might be a perfectly reasonable explanation such as the presence of an occlusion event, but if not we should add further doubt to the interpretation. Thus, our way of reasoning is again led back to the detection of occlusion events and hence 'T' junctions.

We show an example of the success of this line of reasoning in Fig. 12. Here model features have been projected into the image and have found sufficient image support along their lengths. However, the support has actually been provided by sets of unconnected image features. We thus test for the presence of occlusion events at the ends of the image features, and if they are not found, we mark the hypothesis as being un-reasonable.

4.4 Uniqueness of description

By this stage of the proceedings the hypotheses have undergone a fairly detailed level of topological analysis. Those which have been attributed near perfect scores are very likely to be correct, whilst those with poorer verification tallies may either be erroneous, or might just be suffering due to difficulties in segmentation or occlusion event detection.

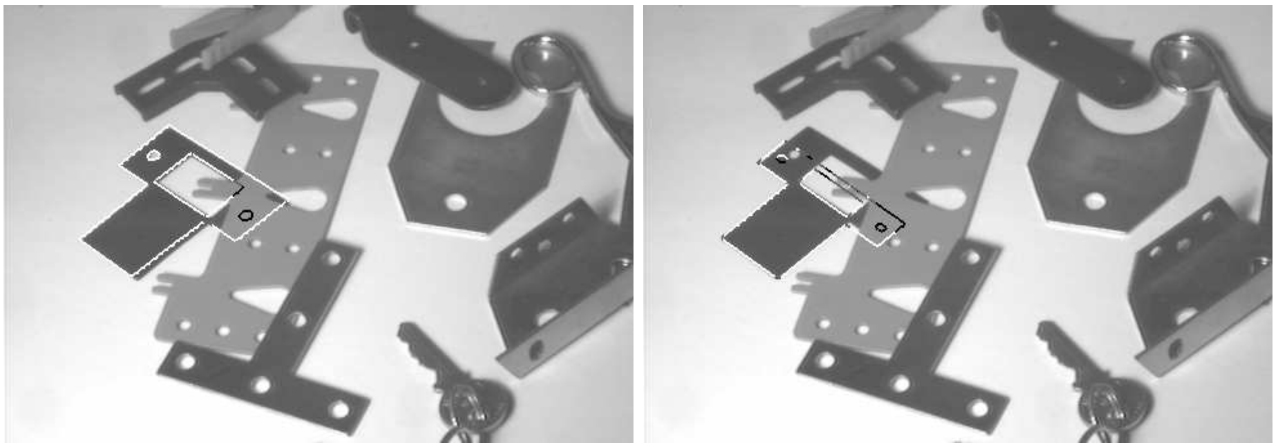
We now turn to the basic fact that a single feature in an image is almost certainly caused only by a single scene feature. Therefore, a consequence of recognition should be that the correspondence between model and image features is at most one-to-one. Should two hypotheses match a single image feature we can be sure that *at least one* of the hypotheses is incorrect, and so should try and eliminate the least likely. This process is risky should the confidence levels in the hypotheses by poorly defined, but as our abilities at verification improve, we can start to attribute error measures with a reasonable degree of accuracy.

Although we have not yet been able to perform a complete and deep study, it appears that by the time that the occlusion analysis has been performed we can start to mark hypotheses more clearly for acceptance or rejection. We therefore proceed by accepting the single hypothesis which has gained the highest recognition score. Then, all of the image features which have been given a correspondence to any of the features in this model are marked as being explained. All other hypotheses which have correspondences with these image features are marked as being inconsistent, and rejected. We then take the next best hypothesis, and proceed by examining its image features. In short, this process ensures a uniqueness of description of the image features. An example of this type of reasoning is given in Fig. 13 where we are able to rule out the hypothesis in (b) because it shares scene features with the very highly scored hypothesis in (a).

Currently this process seems to work quite well, though it is possible that the introduction of such a severe pruning step at this point is premature. It might be that more intermediate steps should be introduced prior to the enforcement of the uniqueness of description. Some of these other possible phases are discussed in Section 5, though until we have tested the system more we cannot be sure about which directions to follow next.

5 Conclusions

In this paper we have demonstrated how reasoning about model and image topology enhances our verification abilities in object recognition. A typical recognition system such as that of Thompson and Mundy [26] computes the final match score for any hypothesis by determining whether a set of independent model features finds support in an image. Rothwell [24] demonstrated that such a strategy does not produce conclusive recognition results. We have found that diffusing verification information around connected components of a model means that a lack of image support can actually be turned into *positive evidence*. This in turn means that verification thresholds can perhaps be raised from 50% up to somewhere in excess of a 90% level of image support.



a b

Figure 13: *In Fig. 11 we were able to find two hypotheses which matched to common scene features. After all of the topological processing the hypothesis in (a) scored* 91.3% *and that in (b)* 68.2%. *Any score over* 90% *provides very strong confidence, and so we can eliminate any other hypotheses which match to the same image features.*

The development of our verification algorithm involves reasoning about the discrepancies between the model and image topologies. The differences are used to hypothesize where occlusion events ('T' junctions) should lie in the image. The presence of such events strengthens a recognition hypothesis, and the lack of one suggests that a hypothesis might be false. The detection of the junctions is done via both edge detector output [23], and the Deriche-Blaszka feature detector [8]. Neither of these detectors function perfectly, though when either hypothesizes the presence of a 'T' junction we can be relatively sure that the claim is true. Notably, the ease of use of the Deriche-Blaszka detector is enhanced through the use of top-down processing gained from the use of recognition hypotheses.

Of course our results are not entirely complete. Whilst working within single images, we need to analyse further the effects of other feature detectors. There are a large number of other filters which require evaluation in either of the domains of bottom-up or top-down processing. We also need to test the algorithms on a more varied range of objects of which a three-dimensional model base is the ultimate goal.

On a different level, we have only made use of the boundary information contained within the models. Such geometric primitives obviously provide very easy access to object descriptions. However, a full verification scheme should include analysis about surface properties such as texture, and perhaps even colour. Certainly with the aid of top-down segmentation based on the recognition hypotheses one would be able to test out other object properties in conjunction with the more geometric aspects.

Acknowledgments

I would like to acknowledge the input of a number of people who have influenced this work. The re-implementation of the LEWIS recognition system has been done using the TargetJr software originally developed at the General Electric CRD Center, and as such I would like to thank Bill Hoffman from GE for resolving a number of software related problems. The original LEWIS implementation evolved out of discussions with Joe Mundy, David Forsyth and Andrew Zisserman. Thierry Blaszka provided his implementation of the Deriche-Blaszka parametric junction detector which is discussed in this paper. I would also like to thank Theo Papadopoulo and Nour-Eddine Deriche for other discussions.

This research was funded by a Human Capital and Mobility grant from the European Community.

References

[1] N. Ayache and O.D. Faugeras. HYPER: A New Approach for the Recognition and Positioning of Two-Dimensional Objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(1):44–54, January 1986.

[2] R. C. Bolles and Radu Horaud. 3DPO: A Tree-dimensional Part Orientation System. *International Journal of Robotics Research*, 5(3):3–26, 1986.

[3] F. Bookstein. Fitting conic sections to scattered data. *Computer Vision, Graphics, and Image Processing*, 9:56–71, 1979.

[4] G. Borgefors. Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6):849–865, November 1988.

[5] Andrea Califano and Rakesh Mohan. Systematic design of indexing strategies for object recognition. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 709–710, New-York, NY, June 1993. IEEE Computer Society, IEEE.

[6] J. F. Canny. Finding edges and lines in images. Technical Report AI-TR-720, Massachusets Institute of Technology, Artificial Intelligence Laboratory, June 1983.

[7] M. B. Clowes. On seeing things. *Artificial Intelligence Journal*, 2:79–116, 1971.

[8] R. Deriche and T. Blaszka. Recovering and characterizing image features using an efficient model based approach. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 530–535, New-York, June 1993. IEEE Computer Society, IEEE.

[9] O.D. Faugeras and M. Hébert. The representation, recognition, and locating of 3d shapes from range data. *The International Journal of Robotics Research*, 5:27–52, 1986.

[10] W. Eric L. Grimson. *Object Recognition by Computer: The Role of Geometric Constraints*. The MIT Press, Cambridge, Massachusetts, London, England, 1990.

[11] W.Eric L. Grimson and T. Lozano-Pérez. Localizing overlapping parts by searching the interpretation tree. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):469–482, 1987.

[12] A. Guzman. Decomposition of a visual scene into three-dimensional bodies. In *FJCC*, volume 33, pages 291–304, 1968.

[13] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision*, volume 1. Addison-Wesley, 1992.

[14] Friedrich Heitger and Rüdiger von der Heydt. A computational model of neural contour processing: Figure-ground segregation and illusory contours. In *Proceedings of the 4th Proc. International Conference on Computer Vision*, pages 32–40, Berlin, Germany, May 1993. IEEE Computer Society Press.

[15] M.H. Hueckel. A local visual operator which recognizes edges and lines. *Journal of the ACM*, 20(4):634–647, 1973.

[16] D. A. Huffman. Impossible objects as nonsense sentences. In B. Meltzer and D. Michie, editors, *Machine Intelligence*, volume 6, chapter ? Edinburgh University Press, 1971.

[17] Daniel P. Huttenlocher, Jae J. Noh, and William J. Rucklidge. Tracking non-rigid objects in complex scenes. In *Proceedings of the 4th Proc. International Conference on Computer Vision*, pages 93–101, Berlin, Germany, May 1993. IEEE Computer Society Press.

[18] Daniel P. Huttenlocher and Shimon Ullman. Recognizing Solid Objects by Alignment with an Image. *The International Journal of Computer Vision*, 5(2):195–212, November 1990.

[19] Y. Lamdan and H.J. Wolfson. Geometric Hashing: A General and Efficient Model-Based Recognition Scheme. In *Proceedings of the 2nd International Conference on Computer Vision, Tampa, Florida*, pages 238–249, December 1988.

[20] A. K. Mackworth. Interpreting pictures of polyhedral scenes. *Artificial Intelligence Journal*, 4:99–118, 1973.

[21] S.B. Pollard, J. Porrill, J.E.W. Mayhew, and J.P. Frisby. Matching geometrical descriptions in three-space. *Image and Vision Computing*, 5(2):73–78, 1987.

[22] K. Rohr. Recognizing corners by fitting parametric models. *The International Journal of Computer Vision*, 9(3):213–230, 1992.

[23] C. Rothwell, J. Mundy, W. Hoffman, and V.-D. Nguyen. Driving vision by topology. In *IEEE Computer Vision Symposium*, 1995. to appear.

[24] Charles Rothwell. *Object recognition through invariant indexing.* Oxford University Science Publications. Oxford University Press, February 1995.

[25] F. Stein and G. Medioni. Structural Indexing: Efficient 3-D Object Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):125–145, February 1992.

[26] D.W. Thompson and J.L. Mundy. Three-dimensional model matching from an unconstrained viewpoint. In *Proceedings of the International Conference on Robotics and Automation, Raleigh, NC*, pages 208–220, 1987.

[27] D. Waltz. Understanding line drawings of scenes with shadows. In Patrick H. Winston, editor, *The Psychology of Computer Vision*, pages 19–91. McGraw-Hill, 1975.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399