Qualitative Probabilities for Image Interpretation

Allan Jepson* Richard Mann†
*Department of Computer Science, University of Toronto, Toronto M5S 1A4 CANADA
†NEC Research Institute, Inc., 4 Independence Way, Princeton, NJ 08540 USA

Abstract

Two basic problems in image interpretation are: a) determining which interpretations are the most plausible amoungst many possibilities; and b) controlling the search for plausible interpretations. We address these issues using a Bayesian approach, with the plausibility ordering and search pruning based on the posterior probabilities of interpretations. However, due to the need for detailed quantitative prior probabilities and the need to evaluate complex integrals over various conditional distributions, a full Bayesian approach is currently impractical except in tightly constrained domains. To circumvent these difficulties we introduce the notion of qualitative probabilistic analysis. In particular, given spatial and contrast resolution parameters, we consider only the asymptotic order of the posterior probability for any interpretation as these resolutions are made finer. We introduce this approach for a simple card-world domain, and present computational results for real blocks-world images.

1 Introduction

A fundamental problem in image understanding is to choose among many possible consistent interpretations for an image. For example, consider the image segments shown in Fig. 1a. A human observer might infer that the corresponding scene is probably made up of three objects, namely a triangular card, a quadrilateral card, and a stick. Moreover, it is plausible that there was a flaw in the image edge extraction process, which caused "drop-out" segments in the image data for the stick and the quadrilateral. This interpretation is depicted in Fig. 1b. However, note that there are many other possible interpretations within such a "card-world" domain, such as the two depicted in Figs. 1c,d.

The standard explanation for why the interpretation in Fig. 1b is preferred is that it is non-accidental [1, 10, 11]. That is, any other interpretation involves a careful (i.e. 'accidental') alignment of either the objects within the scene, or the viewer with respect to the scene, or both. In contrast, the scene model in the naturally selected interpretation involves only generic processes, namely a triangular card has been placed in front of a convex quadrilateral and a stick. Similarly, the imaging process for this interpretation is also non-accidental, involving only two relatively common flaws (i.e. drop-out segments) in the feature extraction process.

However, as described in [7, 8], technical difficulties arise in the attempt to develop a computational formalism based on non-accidental properties. A central contribution of this paper is that we formalize this type of reasoning, extend it, and ground it in Bayesian analysis through the use of 'qualitative probabilistic reasoning'.

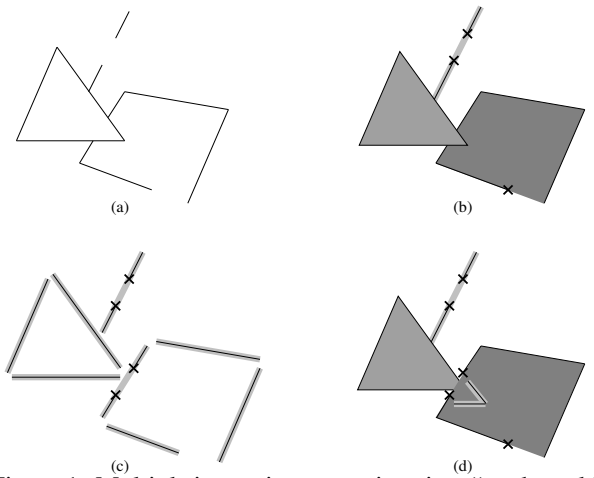


Figure 1: Multiple image interpretations in a "card-world" domain consisting of sticks and convex, opaque polygons. *Legend:* Image segments are shown as thin black lines. Thick grey lines depict sticks. Shaded grey regions depict (opaque) polygonal cards. Crosses depict breakpoints in sticks and/or polygon edges. (a) Input image consisting of image segments. (b) The preferred interpretation consisting of a triangle in front of a quadrilateral and a stick. Drop-out segments arise in the image of the stick and the quadrilateral due to an imperfect image line-finder. (c) An alternative interpretation consisting of eight sticks, two of which have drop-out segments. (d) A third interpretation, this time with the triangle behind the quadrilateral. To explain the image in (a) two additional sticks and one more drop-out must be included in the interpretation, as compared to that in (b).

Our work differs from related work in perceptual grouping [3, 5, 14, 15], and to work in object recognition [2, 4, 6, 10], in its adherence to a Bayesian formulation.

2 Qualitative Probabilities

It is useful to first present a brief intuitive overview of qualitative probabilistic analysis and it's application to the interpretation images. To do this we begin by defining an image interpretation, which consists of models for both the scene and the imaging process (see Sec. 2.1). Given this specification, we then discuss how qualitative probabilities can be used to: 1) provide a computationally simple plausibility ordering of the possible interpretations (see Sec. 2.3), and 2) prune the search for the most plausible ones (see Sec. 2.4). Finally, in subsequent sections we describe an

implementation of this theory and provide some computational results on real images.

2.1 Interpretations

An interpretation of a set of image features involves a scene model and an imaging model which together account for the given set of features. In general, we take card-world scene models to consist of such a 2D layered arrangement of sticks and convex polygonal opaque cards. Since the objects are opaque, the depth layering dictates the visibility of each point on any object. For example, the interpretation depicted in Fig. 1b involves a scene model consisting of a stick, a triangular card, and a quadrilateral card. The triangular card is in front of both the stick and the quadrilateral. For objects which do not intersect in the image, such as the stick and the quadrilateral, the depth relation is undefined (see [15] for a more general 2D layered scene model).

Also, to keep things simple in this introductory example, we take the imaging process to be almost veridical. We assume that the only error is that subsegments (or all) of a visible scene edge may be missed by the image line-finder. The imaging model, which relates the scene model to the observed image features, must specify the occurrence and the endpoints of each of these 'drop-outs'. For example, the interpretation in Fig. 1b has an imaging model which specifies one drop-out on the stick and another on the quadrilatural (the X's mark all the variable or 'free' endpoints of the drop-outs, namely those which are not coincident with the endpoint of a visible scene edge).

Together, the scene and imaging models given above account for all of the observed image segments in Fig. 1a. Moreover, they do not predict the presence of any additional visible, but unobserved, edges. We define an *image interpretation* to be any such pair of scene and imaging models which, in this manner, account precisely for the image data. In general, the condition that an interpretation must account for all of the image data, and no more, provides a consistency relation between interpretations and sets of image data (cf. [12]).

2.2 Qualitative Priors

In order to select plausible interpretations we consider a form of Bayesian analysis. Such an analysis requires prior probability distributions for both the scene and imaging models. Instead of using precise quantitative specifications, the qualitative probabilistic approach considers simple equivalence classes of priors (see [8]). In this paper we define the equivalence classes in terms of a spatial resolution parameter, ϵ, and an image contrast resolution parameter, δ. Here $\epsilon > 0$ corresponds to the spatial resolution in the image of each coordinate for an observed point, or the normal position of a detected line. Similarly, $\delta > 0$ corresponds to the variation in image contrast needed to detect an image edge. Smaller values of ϵ and δ represent finer resolutions.

The prior $p(C_n)$ for any particular convex n-gon C_n is taken to be a function of the n vertex positions $\{\vec{x}_i\}_{i=1}^n$. The critical condition on this prior is that it is both bounded from above, and bounded away from zero from below. It then follows that the occurrence of any given n-gon, up to a resolution of ϵ for each of the $2n$ degrees of freedom in C_n, has prior probability of order $\Theta(\epsilon^{2n})$.[1] This order estimate follows from the volume of a ball of radius ϵ in the $2n$-dimensional parameter space for C_n.

For the current paper we take the shape and position of any object to be independent of the shape and position of other objects. For example, let $M_b(\vec{\alpha})$ denote the scene model depicted in Fig. 1b. Here the 'b' refers to panel b in Fig. 1, and $\vec{\alpha}$ is an eighteen dimensional vector of parameters which specify the 2D position of the vertices of the triangle and the quadrilateral, as well as the two endpoints of the stick. Let $M_b(A_0(\epsilon))$ denote the set of scene models which have parameters within the ball, $A_0(\epsilon)$, centered on $\vec{\alpha}$ with radius ϵ. Then the prior for $M_b(A_0(\epsilon))$ is simply the product of the prior for generating the triangle (which is $\Theta(\epsilon^6)$), the quadrilateral ($\Theta(\epsilon^8)$), and the stick ($\Theta(\epsilon^4)$). Note that, since the arrangement in depth layers only involves binary choices of which object of an overlapping pair is in front, there is no contribution to the order of the prior for arranging these objects in depth. As a result, we find $p(M_b(A_0(\epsilon))) = \Theta(\epsilon^{18})$.

The prior for the imaging model or, rather, when given a particular image, the *likelihood* of the imaging model, is defined in a similar way. Recall that the only variable property of the current imaging model is the presence and placement of the drop-outs. The occurrence of a drop-out is taken to appear with probability proportional to δ, where δ represents the resolution in image contrast. Moreover, the imaging model needs to account for the endpoints of the drop-out, which are determined to a spatial resolution of ϵ along the corresponding image segment. In particular, the drop-out in the interior of the stick in Fig. 1a requires two additional spatial parameters, and therefore has a likelihood of $\Theta(\epsilon^2\delta)$. The drop-out on the quadrilateral extends from an interior point on the scene edge right up to the edge's endpoint, so we require only one additional parameter in the imaging model. Thus its likelihood is $\Theta(\epsilon\delta)$. Combining these, we find the likelihood of generating the image I in Fig. 1a from the scene model $M_b(A_0(\epsilon))$, is $p(I|M_b(A_0(\epsilon)) = \Theta(\epsilon^3\delta^2)$.

2.3 Preferred Interpretations

These computed orders for the scene and imaging models can be used to determine preferences between the various possible interpretations. To see this recall that Bayes theorem states that, given image data I, the posterior prob-

[1] Throughout this paper we use $\Theta(\epsilon^k)$ to denote the sharp order estimate, that is, $f(\epsilon) = \Theta(\epsilon^k)$ if and only if there exists constants $K_1,\ K_2 > 0$ such that $K_1\epsilon^k \leq |f(\epsilon)| \leq K_2\epsilon^k$, as $\epsilon \to 0$.

ability for any set of interpretations, say $\mathcal{M}$, is

$$p(\mathcal{M}|I) = \frac{p(I|\mathcal{M})p(\mathcal{M})}{p(I)}. \quad (1)$$

We refer to the numerator here as the *unnormalized posterior*. For $\mathcal{M} = M_b(A_0(\epsilon))$ it follows from the results of the previous section that the unnormalized posterior is of order $\Theta(\epsilon^{18})\Theta(\epsilon^3\delta^2) = \Theta(\epsilon^{21}\delta^2)$. Similarly, for the interpretations depicted in Figs. 1c,d, we find the unnormalized posteriors are of orders $\Theta(\epsilon^{36}\delta^2)$ and $\Theta(\epsilon^{31}\delta^3)$, respectively. Equation (1) then ensures that the posterior probabilities for these three interpretations are just the common factor $1/p(I)$ times these unnormalized posteriors. Thus, for sufficiently fine resolutions ϵ and δ, the interpretation $M_b(A_0(\epsilon))$ is much more probable than the other two.

In order to compare two unnormalized posteriors we may require information about the relative sizes of δ and ϵ. For example, to compare the two interpretations depicted in Fig. 1c,d, with the unnormalized posteriors given above, we need to compare $\Theta(\epsilon^5)$ and $\Theta(\delta)$ (here we have removed a common factor of $\epsilon^{31}\delta^2$). A simple but general form for a such a comparison policy is to suppose that e_- an e_+ are two constant exponents such that $e_- \leq e_+$ and

$$\epsilon = O(\delta^{e_-}), \quad \delta^{e_+} = O(\epsilon), \quad (2)$$

as $\epsilon, \delta \to 0$. For example, $(e_-, e_+) = (1, 10)$ ensures that $\epsilon^5 = o(\delta)$, and hence the interpretation in Fig. 1d would be preferred over that in Fig. 1c. In general, however, such a comparison policy leads to a partial ordering of interpretations, not a total ordering (eg. the terms ϵ and δ^2 cannot be ordered when $(e_-, e_+) = (1, 10)$).

In the computational examples we find that the selection of the preferred interpretations are not sensitive to the precise choice of e_- and e_+, and for convenience there we set $e_+ = e_- = \infty$. This gives a simple total ordering on the interpretations, with an unnormalized posterior of order $\Theta(\epsilon^{e_1}\delta^{d_1})$ preferred over one with order $\Theta(\epsilon^{e_2}\delta^{d_2})$ if and only if either $e_1 < e_2$, or $e_1 = e_2$ and $d_1 < d_2$.

Thus we find that a qualitative probabilistic analysis leads to an intuitive preference order on the interpretations in Fig. 1, with the interpretation in (b) preferred over (d), and both of these are preferred over (c). One of the primary research issues we wish to address is whether the qualitative probabilistic approach can similarly capture the most preferred interpretations of real images.

2.4 Hypothesis Generation and Search

A second important test for qualitative probabilistic analysis is whether it can form the foundation for effective search heuristics. Even in the simple card-world domain the search space grows exponentially with the number of edges, so a brute-force search is impractical.

Here we show how qualitative probabilities can be used to determine the plausibility of partial interpretations,

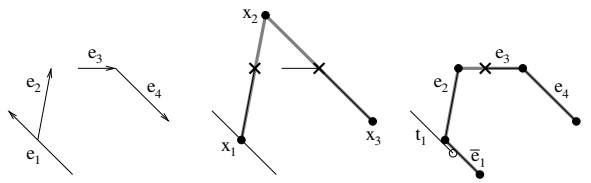


Figure 2: (a) Image edges. (b) Part of an example polygon in the set $C_n(\vec{x}_1, \vec{x}_2, \vec{x}_3)$. (c) Part of a polygon covering e_2, e_3, e_4 and the subsegment $\bar{e}_1$ of e_1. The remaining 'tail' t_1 has a free endpoint denoted by the circle on $\bar{e}_1$.

which can then be used to prune the search. For example, consider the image data in Fig. 2a along with the hypothesis depicted in Fig. 2b. In particular, the hypothesis is that the two image edges e_2 and e_4 are covered by consecutive edges of some convex n-gon for some $n \geq 3$. Let $C_n(\vec{x}_1, \vec{x}_2, \vec{x}_3)$ denote the set of all convex n-gons which have the points $\vec{x}_1$, $\vec{x}_2$, and $\vec{x}_3$ as three consecutive vertices. These vertices $\vec{x}_i$ are as shown in Fig. 2b. The edges e_2 and e_4 are directed so that the interior of the n-gon is to their right side. Since the three specified vertices are given to a spatial resolution of ϵ, the prior probability for the hypothesis $C_n(\vec{x}_1, \vec{x}_2, \vec{x}_3)$ is $\Theta(\epsilon^6)$.

Since there are two drop-outs each with a free endpoint (see Fig. 2b), the likelihood of generating e_2 and e_4 as a subset of the image data is $\Theta(\epsilon^2\delta^2)$, Thus the unnormalized posterior for $C_n(\vec{x}_1, \vec{x}_2, \vec{x}_3)$ is of order $\Theta(\epsilon^8\delta^2)$, for all $n \geq 3$.[2]

To determine the level of evidence for such a hypothesis, we use the odds of this hypothesis in comparison to one in which the same subset of image data is explained only by sticks. These odds are given by the ratio of posterior probabilities for the two hypotheses which, by Bayes rule (1), is just the ratio of the two unnormalized posteriors. Thus the odds for $C_n(\vec{x}_1, \vec{x}_2, \vec{x}_3)$ are $\Theta(\epsilon^8\delta^2)/\Theta(\epsilon^8) = \Theta(\delta^2)$ (the denominator here is the unnormalized posterior for the hypothesis that e_2 and e_4 arise from two sticks). We see that the odds actually favor the two sticks hypothesis, for δ sufficiently small. Similar calculations show that the odds for (e_2, e_3) being covered by consecutive edges of an n-gon are $\Theta(\epsilon^{-1}\delta)$, and the odds for edges (e_3, e_4) are $\Theta(\epsilon^{-2})$.

Finally, consider the hypothesis, say $H_v(e_1, e_2)$, that the pair of edges (e_1, e_2) arise from consecutive edges of an n-gon, such as the one depicted in Fig. 2c. This implies an under-segmentation error has occured to form image edge e_1, and some other (independent) process must explain the 'tail' of edge e_1 extending outside of this polygon. Let $\bar{e}_1$ denote the subsegment of e_1 covered by the n-gon, and t_1 the tail segment, where t_1 can overlap $\bar{e}_1$, as long as its right endpoint (the open circle in Fig. 2c) is within $\bar{e}_1$.

The unnormalized posterior for $H_v(e_1, e_2)$ is then the product of the unnormalized posterior for the 'V' formed

[2] Similar interpretations which have drop-outs on both ends of edges e_2 and/or e_4 are possible, but give smaller unnormalized posteriors.

by $(\bar{e}_1, e_2)$, namely $\Theta(\epsilon^6)$, and the unnormalized posterior for the tail t_1. In the worst case, this latter term could be $\Theta(\epsilon^3)$, corresponding to the hypothesis that t_1 arose from a stick with an uncertain right endpoint. Alternatively, in the best case, the tail could be explained by a part of some other polygon. For a perfectly imaged n-gon the unnormalized posterior is $\Theta(\epsilon^{2n})$, and therefore the prorated value for each of the n edges is $\Theta(\epsilon^2)$. Using this best case, the hypothesis $H_v(e_1, e_2)$ has an unnormalized posterior of order $\Theta(\epsilon^6)\Theta(\epsilon^2) = \Theta(\epsilon^8)$. And the odds for $H_v(e_1, e_2)$ are simply $O(1)$, in other words, there is no asymptotic evidence in favor of this hypothesis.

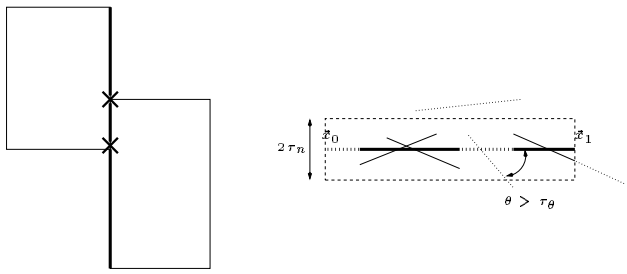
A detailed discussion of search using information provided by qualitative probabilities is beyond the scope of the current paper. Here we use a simple search heuristic based on the log of the odds for each hypothesis. These odds are compared to the maximal odds that can be obtained for any hypothesis in which m edges of an n-gon are partially covered by image edges (i.e. at least a subsegment of each of these scene edges is accounted for by an image edge, see Fig. 3b). For convex n-gons with $m < n$, we find that the maximal odds are $\Theta(\epsilon^{2m+2})/\Theta(\epsilon^{4m}) = \Theta(\epsilon^{-2m+2})$, as attained for convex open chains of m edges. We define $\tau(m)$ to be the log of these maximal odds, namely $\tau(m) = (2m-2)|\log(\epsilon)|$ for $m < n$.

The 'plausible garden path' search heuristic involves pruning any hypothesis which consists of an n-gon with m partially covered scene edges and has log odds smaller than $\rho\tau(m)$ (for simplicity we ignore terms in δ). Here we use $\rho = 0.5$. From the odds computations above, we find that the only unpruned hypotheses involving pairs of edges from Fig. 2a are the ones which cover (e_2, e_3) and (e_3, e_4) with consecutive edges. Similarly, the only 3- and 4-edge plausible convex groups are (e_2, e_3, e_4), (e_1, e_3, e_4), or (e_1, e_2, e_3, e_4), which are intuitively reasonable. In Sec. 5.1 we discuss the results of applying this heuristic to real image data.

3 Detailed Imaging Model

To apply our system to real images it is critical that the imaging model accounts typical imperfections in the feature extraction process. Here the only features we consider are image segments, as extracted by a typical line-finder. Therefore we need to model the different types of imperfections in the line-finder results, namely: 1) drop-outs; 2) false-targets; 3) over- and under-segmentation errors; 4) limited resolution; 5) errors in the position and orientation estimates.

The first three of these types of imperfections can be dealt with using the approach described in Sec. 2.1. In particular, drop-out segments are explicitly accounted for by the imaging model within an interpretation. We use sticks in the scene model to represent false targets, which arise from unmodeled scene structure such as shadows, texture, or image noise. Over and under segmentation errors are dealt with by allowing several image segments to account for a single scene edge, and also several visible scene edges to account for a single image segment (see Fig. 3). The last two types of imperfections are discussed below.



(a) (b)

Figure 3: (a) Multiple scene edges (thin lines) can map to single image segment (thick line), as in abutment. Scene edges are merged whenever they project to nearly collinear image segments. We also predict the location of potential breakpoints (denoted "X") whenever there is a change in the surface on either side of an edge. (b) One visible scene edge, $[\vec{x}_0, \vec{x}_1]$ (thick line), can account for multiple image segments (thin lines). We use a fixed tolerance for perpendicular distance (τ_n) and angular error (τ_θ) between the observed image segments and the ideal edge. Unexplained parts of the scene edge ("drop outs") are shown as thick dotted lines. A scene edge may account for only a subsegment of an image edge, as for the rightmost image segment.

3.1 Visible and Resolvable Edges

Visible scene edges are defined to be the set of visible points on the edges of the polygonal cards or sticks. (A point is visible if and only if that point is not within any other object which is deemed to be in front.) However, for a scene model in which two objects nearly abut, there can be two visible edges which are nearly colinear in the image (see Fig. 3a). The image line-finder will be unable to resolve such a pair if the perpendicular distance between them is too small.

To account for this limit in resolution we merge nearby visible scene edges into single visible edges. In addition, in order to predict the location of potential defects (such as drop-outs) in the observed image segments, we partition visible edges into collections of subedges that share common object boundaries (see Fig. 3a).

3.2 Position and Orientation Errors

The imaging model is conveniently described in terms of covering relationships between an image segment and a visible scene edge (see Fig. 3b). A visible scene edge can "cover" all or part of an image segment, and thereby provide an appropriate explanation for how that image segment (or part thereof) could have arisen. Conversely, an image segment can cover all or part of a visible scene edge, and thus provide data supporting the hypothesis of that scene edge's existence. We use a fixed perpendicular and angular tolerance to model the error in image segments (see

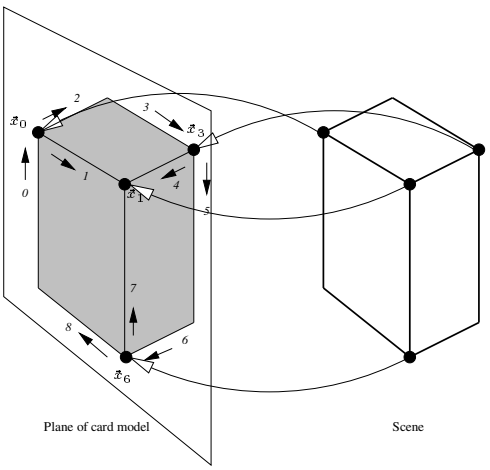


Figure 4: The card-world representation of a block is the orthographic projection of a block, depicted here in grey. The arrows on the edges within the card indicate the orientation convention used.

Fig. 3b). In addition, we specify a minimum length for any drop-outs and false-targets.

4 Application to Blocks-World Scenes

We demonstrate our approach on blocks-world scenes (see Fig. 5). For simplicity we restrict the viewing conditions such that the image of any individual block is well-approximated by an orthographic view. The key to modeling blocks-world using layered arrangements of 2D cards is to depict the orthographic view of some 3D rectangular block on the cards themselves (see Fig. 4). Each of the visible edges of the corresponding 3D block is represented by an edge on this card, and the interior region of the card is considered to be opaque. The length of the sides can be varied, along with the particular viewpoint depicted on the card.

An orthographic view of a general 3D block is fully specified by four points in the image plane, as shown in Fig. 4. The coordinates of the three arrow junctions ($\vec{x}_0$, $\vec{x}_3$, and $\vec{x}_6$) and that of the Y-junction ($\vec{x}_1$) uniquely determine a block. Roberts [13] showed that a necessary and sufficient condition for such a parameterized 2D model to correspond to some orthographic view of a rectangular block is that each of the three interior edges (labelled 1, 4, and 7) must form an obtuse or right angle with both of the other two interior edges.

We use a qualitative prior for the scene models and a qualitative likelihood for the imaging models. Since the orthographic projection of a rectangular block is described by eight parameters, we take any particular block (up to a spatial resolution of ϵ) to have a prior probability of $\Theta(\epsilon^8)$. The likelihood of drop-outs are determined exactly as in Sec. 2.1. For example, the interpretation depicted Fig. 5c for image E involves 4 blocks, 2 sticks, and 10 drop-outs (four of which correspond to entirely missed edges) with a total of 7 free endpoints (marked by X's in the figure). The unnormalized posterior is then $\Theta(\epsilon^{4*8}\epsilon^{2*4}\epsilon^7\delta^{10}) = \Theta(\epsilon^{47}\delta^{10})$, as given by the 'score' in Fig. 5c.

5 Experiments

For our experiments we consider image data from real images of simple blocks-world scenes. We used the line-finder in the *Khoros 1.2* package to generate the image data in Fig. 5a. Note that even without the original images, human observers have no difficulty interpreting the scenes from this edge data.

5.1 Individual Blocks

We first implemented a brute-force search similar to the IT search in [4], except our imaging model is less restrictive (allowing subsegments of image edges to be matched to scene edges) and our scene model has more (i.e. 8) parameters. This search finds all maximal subsets of the image data that are consistent with the orthographic image of one block. In agreement with [4] this search proved to be impractical, taking 4 to 6 hours on a Pentium 133MHz processor. For the simple image data sets in Fig. 5a several thousand different individual block hypotheses were found. The majority of these were partially-instantiated blocks, that is, blocks for which the parameters are not completely specified by the covering set of image edges.

The odds for each block hypothesis, versus the hypothesis that the same subset of image edges come from sticks (see Sec. 2.4), was used as a plausibility measure to sort the list of blocks. The true blocks typically appeared in the top one percent of this sorted list. The only case this failed was for the leftmost block in Image 4. Here the line-finder failed to detect enough edges to fully instantiate the block, and the true block was not found. Instead the search algorithm grouped edges from `Block1` with the edges from this leftmost block (see Fig. 5b).

An order of magnitude speed-up in the search was obtained by using the plausible garden path heuristic discussed in Sec. 2.4. That is, we pruned any block hypothesis whose log odds fell below half the maximum possible log odds for any block hypothesis with the same number of partially or completely covered edges. Another order of magnitude speed-up was obtained by first finding plausible local block fragments, and then grouping these fragments using the same odds measure to prune implausible hypotheses.[3] The resulting search ran in under a minute on each of the examples in Fig. 5. The resulting search found every true block, except the bottom left block in Image 4 (missed by the full search) and the bottom left block in Image 7 (for which an isolated edge exhibiting a tail needs to be grouped with the edges forming the lower portion of the block).

5.2 Complete Interpretations

[3] For details of both of these pruned search algorithms see [9].

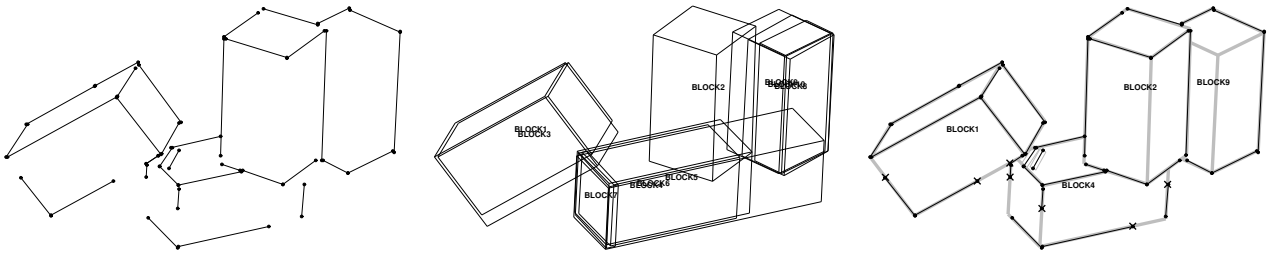


Image "E": 34 Segments | 10 Blocks | Score:$\epsilon^{47}\delta^{10}$ ($\epsilon^{50}\delta^{10}$)

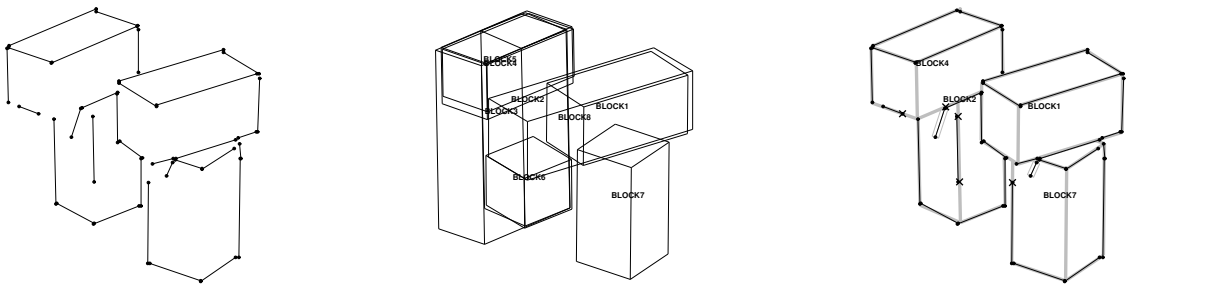


Image "7": 32 Segments | 8 Blocks | Score:$\epsilon^{45}\delta^{8}$ ($\epsilon^{50}\delta^{11}$)

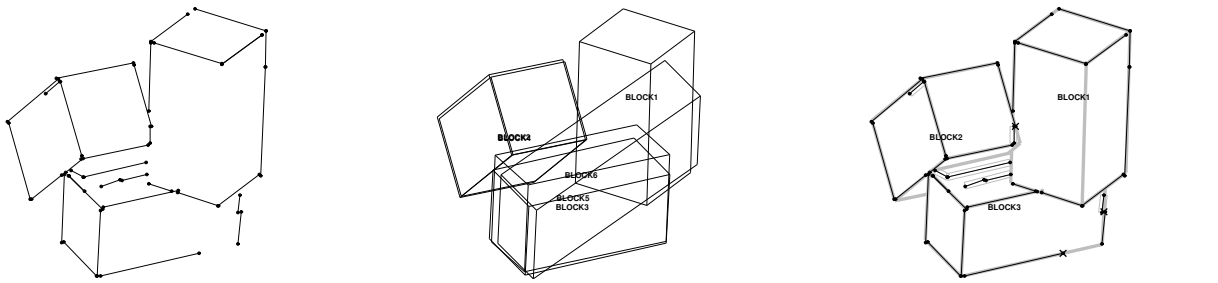


Image "D": 34 Segments | 6 Blocks | Score:$\epsilon^{51}\delta^{8}$ ($\epsilon^{53}\delta^{7}$)

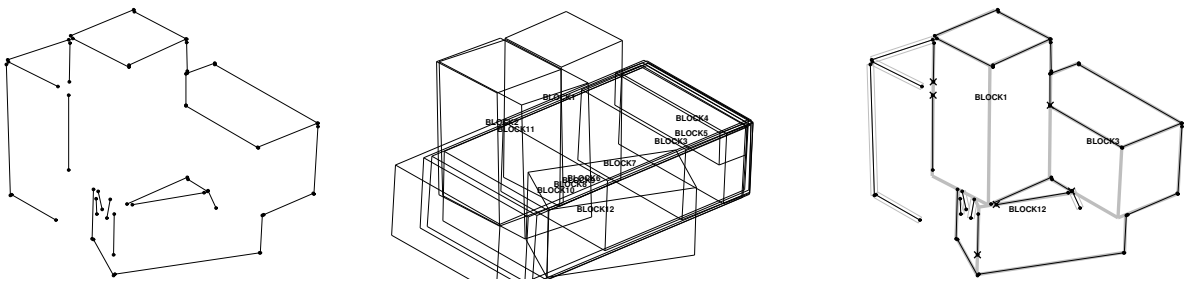


Image "4": 30 Segments | 12 Blocks | Score:$\epsilon^{62}\delta^{8}$ ($\epsilon^{65}\delta^{9}$)

(a) (b) (c)

Figure 5: Experimental results for several blocks-world images. (a) Line segments found by line finder. (b) All individual fully-instantiated blocks found. For clarity, we have shown the individual blocks with a slight random "jitter". (c) Preferred interpretation. *Legend:* Image segments are shown as thin black lines. Object edges are shown as thick grey lines. Breaks are denoted by crosses. Sticks are shown as thin grey boxes outlining one or more image segments. The unnormalized posterior (i.e. score) is shown below each interpretation. The score for the second-best interpretation is shown in brackets.

Image	Num. Interp.	σ_o	Nodes Visited	Unique Nodes	Time (sec.)
E	24	12	2915	1167	528
7	6	12	737	341	114
D	21	12	483	335	99
4	16	12	935	751	244

Table 1: Performance of the search for interpretations for the image data in Fig. 5a. The timing was done on a Pentium II 300MHz processor.

Here we consider the full 2D layered interpretations for the blocks-world image data in Fig. 5a. To simplify the implementation, we restrict our consideration to *fully instantiated* blocks. This restriction means the overlap between pairs of blocks is fully specified, which significantly simplifies the algorithm for determining the feasible depth layerings. A second simplification is that we restrict any sticks to be in front of all the blocks. The search space, then, consists of all subsets of the fully instantiated blocks (see Fig. 5b) arranged in all possible layered depth relations, plus zero or more sticks in the foreground.

To find plausible interpretations, we perform a depth-first search, beginning with the interpretation consisting of no blocks (i.e. all sticks), and adding blocks one at a time with the remaining uncovered image edges always modeled by sticks. Each new block is inserted in all possible depth positions within the existing layering. For each layering, the unnormalized posterior of the interpretation is computed. A branch is pruned if the unnormalized posterior is not increased with the addition of the new block by at least a factor of $\Theta(\epsilon^{-\sigma_o})$. Here we used $\sigma_o = 12$.

The run-times for this algorithm are reported in Table 5.2. The most preferred interpretation the algorithm found is displayed in Fig. 5c, along with the unnormalized posteriors for this interpretation and that of the second most preferred interpretation (given in parentheses). The algorithm arrived at a plausible solution in all cases but Image 7, where it's input data was missing the appropriate hypothesis for the bottom-left block. Also, for Image D, the algorithm placed the top-left block in front of the top-right block, because noise in the edge detector has resulted in a small "tail" from `Block2` that must be explained by a stick when `Block2` is placed behind `Block1`. The latter is actually the correct interpretation, and is the second most preferred.

6 Conclusion

Our results indicate that the qualitative probabilistic analysis provides a natural preference ordering on interpretations for simple card-world and blocks-world scenes. Moreover the analysis motivated the choice of search heuristics which proved to be quite effective.

The same style of analysis can be applied to other domains, such as model-based object recognition [2, 4, 6, 10], curve and surface grouping [3, 14, 15], and simple motion interpretation (as discussed in [8]).

For a given problem domain we advocate the following approach. First select scene and imaging models, along with appropriate qualitative priors. Given a comparison policy, such as (2), this setup then fully specifies: 1) the search space for interpretations; 2) the qualitative unnormalized posterior probabilities for both full and partial interpretations; and 3) a preference ordering on the interpretations. The computational vision problem for the given domain can then be formulated as finding a (or all) maximally preferred interpretation(s) within this search space. A general strategy for searching such a space would complete this package, and is an important area for future research.

Note that for imaging systems with finer resolution parameters (i.e. finer image resolution and/or a better feature extraction process) one can expect that the odds used in the search will become more extreme. Thus the various decisions the search algorithm needs to make will be more clear cut and, if the representation is appropriate, we can expect the search to become easier for the same scenes. In passing, we note that the resolution of the images used for Fig. 5 was about the equivalent of the resolution the human fovea achieves on your thumbnail held at arms length (i.e. 2 degrees of arc). In light of that fact, our computational results may seem more comparable to the performance of the human visual system.

References

[1] T. Binford. Inferring surfaces from images. *AIJ*, 17:205–244, 1981.

[2] S. Dickinson, A. Pentland and A. Rosenfeld. From volumes to views: An approach to 3-D object recognition. *CVGIP*, 55(2):130–154, 1992.

[3] J. Feldman. Regularity-based perceptual grouping. *Computational Intelligence*, 13(4):582–623, 1997.

[4] W.E.L. Grimson and T. Lozano-Perez. Localizing overlapping parts by searching the interpretation tree. *IEEE PAMI* 9(4), 1997, pp.469-482.

[5] D. Jacobs. Robust and efficient detection of convex groups. *CVPR-93*, pp.770–771, 1993.

[6] D. Jacobs. Matching 3-D models to 2-D images. *IJCV*, (21)1/2:123–153 ,1997.

[7] A.D. Jepson. Comparing stories. In D. Knill and W. Richards, eds., *Perception as Bayesian Inference*, pp. 478–488. Cambridge University Press, 1996.

[8] A.D. Jepson, W. Richards, and D. Knill. Modal structure and reliable inference. In *Perception as Bayesian Inference*, pp. 63–92.

[9] A.D. Jepson and R. Mann. Qualitative probabilities for image interpretation. *In preparation*.

[10] D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, 1985.

[11] K. Nakayama and S. Shimojo. Experiencing and perceiving visual surfaces. *Science*, 257:1357–1363, 1992.

[12] R. Reiter and A. Mackworth. A logical framework for depiction and image interpretation *Artificial Intelligence*, 41:125–155, 1989.

[13] L.G. Roberts. Machine perception of three-dimensional solids. TR 315, Lincoln Lab, MIT, May 1963.

[14] E. Saund. Perceptual organization of occluding contours of opaque surfaces. CVPR-98 Workshop on Perceptual Organization, Santa Barbara, CA.

[15] L. Williams and A. Hanson. Perceptual completion of occluded surfaces. *CVIU*, 64(1):1–20, 1996.