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

‡Currently at: Dept. of Computer Science, Univ. of Waterloo, Waterloo, Ont. N2L 3G1 CANADA

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 blocks-world images.

1 Introduction

Two fundamental problems in image understanding are: a) choosing a plausible interpretation from many possible consistent interpretations for an image; and b) controlling the search for plausible interpretations. 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 were flaws in the image edge extraction process which caused “drop-outs” in the image data for both 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 nonaccidental [2, 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 nonaccidental, involving only two relatively common flaws (i.e. drop-out segments) in the feature extraction process.

A central contribution of this paper is that we formalize this type of nonaccidental reasoning, extend it, and ground it in Bayesian analysis through the use of ‘qualitative probabilistic reasoning’. This provides a rigorous probabilistic framework for integrating information from nonaccidental features. We present computational results on blocks-world images. These results demonstrate that our approach forms a convenient basis for reasoning about images of scenes with multiple objects and occlusion. The implementation also indicates how this analysis can be used to significantly reduce the search complexity by pruning implausible search paths.

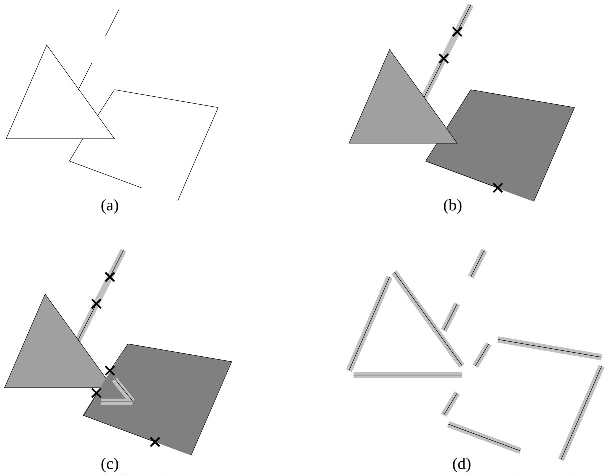
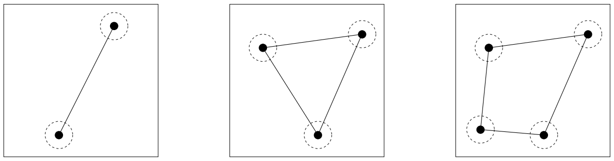


Figure 1: Multiple image interpretations in a “card-world” domain consisting of sticks and convex, opaque polygons. *Legend:* Image segments are shown as thin black lines. Thick grey lines depict sticks. Shaded grey regions depict (opaque) polygonal cards. Crosses depict breakpoints in the image segments for sticks and 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) A less preferred interpretation with the triangle behind the quadrilateral. To explain the image two additional sticks and one more drop-out must be included in the interpretation, as compared to that in (b). (d) A trivial interpretation obtained by explaining every image segment with a stick.

Our approach differs from current work in perceptual grouping [4, 6, 14, 16], image interpretation [1, 9], and object recognition [3, 5, 7, 10] in its use of the qualitative probabilistic framework. In particular, most work in perceptual grouping and object recognition has focussed on finding salient groups of image features based on



(a) (b) (c)

Figure 2: The resolution of endpoints and vertices for a stick (a), a triangle (b), and a quadrilateral (c).

some quantitative measures such as the fraction of model edges covered, energy functions which enforce consistency among features, and so on. Other researchers have presented Bayesian or minimum description length formulations for object recognition and scene interpretation. In contrast to both of these approaches, qualitative probabilities capture the structure available in nonaccidental features, but abstract away most of the information required for the quantitative prior distributions and likelihood functions.

Our approach is similar in spirit to ϵ-semantics [12] in knowledge representation, and to other approaches using defaults [4], but differs in that here the quantitative tools of probability theory are applied to weakly specified priors. In particular, the interaction of various defaults is cleanly and completely specified with our current approach.

2 Qualitative Probabilities

To begin, consider the prior probability for the occurrence of a single line segment in an image, as depicted in Fig. 2a. Let $p(L(\vec{x}_1, \vec{x}_2))$ denote the prior probability density for any particular line segment, $L(\vec{x}_1, \vec{x}_2)$, having endpoint positions $\vec{x}_1$ and $\vec{x}_2$. Instead of selecting a particular quantitative prior, $p(L)$, here we consider a wide equivalence class of such prior densities. The critical condition we impose on $p(L(\vec{x}_1, \vec{x}_2))$ is that

$$0 < d_0 \leq p(L(\vec{x}_1, \vec{x}_2)) \leq d_1 \qquad (1)$$

for some constants d_0 and d_1. That is, the prior probability density is bounded both from above and from below, away from zero.

To make use of this weak prior information, we consider an asymptotic analysis as a spatial resolution parameter becomes increasingly finer. In particular, suppose that the endpoints of a line segment can be resolved to within a radius of r pixels, and that the whole image is L pixels in either direction. Let $\epsilon = r/L$ denote the spatial resolution parameter. Consider the prior probability that one endpoint, say $\vec{x}_1$, of the line segment occurs within some disk of radius ϵ (see Fig. 2a). From equation (1) it follows that the prior probability of such an event is of order $\Theta(\epsilon^2)$ as $\epsilon \rightarrow 0$.[1] Similarly the prior probability of observing both endpoints to be within a radius of ϵ of the predetermined points $\vec{x}_1$ and $\vec{x}_2$ is $\Theta(\epsilon^4)$. Both of these asymptotic results follow directly from equation (1) by integrating the density over the set of possible endpoint positions for a given resolution parameter ϵ.

Clearly such an asymptotic analysis can be extended to more general objects. For example, consider the prior density for any particular convex n-gon C_n, which can be taken to be a function of the n vertex positions $\{\vec{x}_i\}_{i=1}^{n}$. The critical condition on this prior density is again that it is both bounded above, and bounded away from zero from below. A similar asymptotic analysis now shows 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 $p(C_n) = \Theta(\epsilon^{2n})$. In particular, the prior for the triangle depicted in Fig. 2b is of order $\Theta(\epsilon^6)$, while the quadrilateral in Fig. 2c is of order $\Theta(\epsilon^8)$.

We also need to consider the arrangements of several objects within the scene. For card-world we take scene models to consist of 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 taken to be undefined (see [15] for a more general 2D layered scene model).

Qualitative priors can be derived for such scene models consisting of multiple objects. 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, the scene model depicted in Fig. 1b is a particular layered 2D arrangement of a triangle, a quadrilateral, and a stick, with the position of each endpoint and vertex resolved to within a disk of radius ϵ. Let $M_b(\vec{\alpha})$ denote this scene model, where the 'b' refers to panel b in Fig. 1, and the parameter vector $\vec{\alpha}$ denotes the 18 parameters needed to specify the locations of the endpoints and vertices of the three objects. By the independence assumption, the prior for $M_b(\vec{\alpha})$ is then 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 the prior probability for the scene model depicted in Fig. 1b is $p(M_b) = \Theta(\epsilon^{18})$.

2.1 Posterior Probabilities

An interpretation of a set of image features, say I, involves a scene model and an imaging model which together account for I. Given the image data I, we wish to compute the posterior probability of any particular scene model,

[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 $\epsilon \rightarrow 0$.

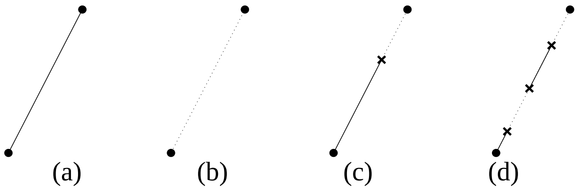


Figure 3: Possible image data for a scene model consisting of one stick. Here dotted lines denote line-finder drop-outs, and X's mark interior breakpoints. (a) Scene model. (b) Missing edge, $\Theta(\delta)$. (c) Missing end segment, $\Theta(\delta\epsilon)$. (d) Missing interior and end segments, $\Theta(\delta^2\epsilon^3)$

$\mathcal{M}$. According to Bayes theorem, the posterior probability of $\mathcal{M}$ satisfies,

$$p(\mathcal{M}|I) = \frac{p(I|\mathcal{M})p(\mathcal{M})}{p(I)}. \qquad (2)$$

Here $p(I|\mathcal{M})$ is the likelihood of observing the data I given the scene model $\mathcal{M}$, and $p(\mathcal{M})$ is the prior for $\mathcal{M}$. We refer to their product as the *unnormalized posterior*.

The likelihood term depends on the probability of the imaging model, which relates the scene model $\mathcal{M}$ to the observed image data I. 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 various subsegments (or all) of a visible scene edge may be missed by the image line-finder (see Fig. 3). The imaging model, then, must specify the occurrence and the endpoints of each of these 'drop-outs'.

The likelihood can now be defined in a similar way to the prior probability for a scene model. The occurrence of a drop-out is taken to appear with probability proportional to δ, where δ represents the resolution in image contrast necessary for the line-finder to detect an image edge. Moreover, the imaging model needs to account for the spatial positions of the endpoints of the drop-outs, which are determined to a spatial resolution of ϵ along the corresponding image segments.

In particular, consider the various ways that the single stick in Fig. 3a might be imaged. The likelihood of missing the stick entirely, as in Fig. 3b, is $\Theta(\delta)$, which depends only on the image contrast resolution δ. Similarly, a drop-out at one end of the stick (see Fig. 3c), requires one additional spatial parameter, and therefore has a likelihood of $\Theta(\epsilon\delta)$. (The other end of the drop-out in this case is dictated by the end of the stick, and is attributed to the scene model, not the imaging model). For Fig. 3d we require three drop-out endpoint parameters and there are two segments at which there is a loss of contrast. We take these separate drop-out segments to be independent, and therefore the likelihood is taken to be $\Theta(\epsilon^3\delta^2)$.

We can combine these likelihood computations with the previous asymptotic results for the priors to obtain expressions for the unnormalized posterior probability of various interpretations. For example, one interpretation for Fig. 1a is given by the scene model $M_b(\vec{\alpha})$, discussed above, together with an imaging model which accounts for the drop-outs on the quadrilateral and the stick. These drop-outs have likelihoods of $\Theta(\epsilon\delta)$ and $\Theta(\epsilon^2\delta)$, respectively. Therefore we find the likelihood of generating the image I in Fig. 1a from the scene model $M_b(\vec{\alpha})$, is $p(I|M_b(\alpha)) = \Theta(\epsilon^3\delta^2)$. Since the prior for the scene model, $p(M_b(\vec{\alpha}))$, was shown above to be $\Theta(\epsilon^{18})$, the unnormalized posterior for this interpretation is then $\Theta(\epsilon^{18})\Theta(\epsilon^3\delta^2) = \Theta(\epsilon^{21}\delta^2)$.

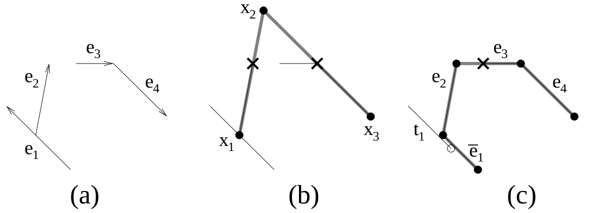


Figure 4: (a) Image edges. (b) Part of an example polygon in the set $C_n(\vec{x}_1, \vec{x}_2, \vec{x}_3)$. (c) Part of a polygon covering e_2, e_3, e_4 and the subsegment $\bar{e}_1$ of e_1. The remaining 'tail' t_1 has a free endpoint denoted by the circle on $\bar{e}_1$.

Similarly, for the interpretations depicted in Figs. 1c,d, we find the unnormalized posteriors are of orders $\Theta(\epsilon^{31}\delta^3)$ and $\Theta(\epsilon^{40})$, respectively. Equation (2) then ensures that the posterior probabilities for these three interpretations are just the common factor $1/p(I)$ times these unnormalized posteriors.

2.2 Preferred Interpretations

In order to compare two unnormalized posteriors we may require information about the relative sizes of δ and ϵ. A simple form for this is to suppose that

$$\epsilon = \Theta(\delta^q), \qquad (3)$$

as $\epsilon \to 0$ for some constant $q \geq 0$. In the computational examples we find that the selection of the preferred interpretations are not sensitive to the precise choice of q. For convenience here we treat $\epsilon \ll \delta$ (i.e. $q = \infty$). This corresponds to the assumption that missing features are much more likely to occur than accidental alignments. As a result, an unnormalized posterior of order $\Theta(\epsilon^{n_1}\delta^{m_1})$ is preferred over one with order $\Theta(\epsilon^{n_2}\delta^{m_2})$ if and only if either $n_1 < m_2$, or $n_1 = n_2$ and $m_1 < m_2$.

This provides an intuitively plausible ordering for the three interpretations depicted in Fig. 1. In particular the posterior distribution for the interpretations depicted in Fig. 1b (with an unnormalized posterior of $\Theta(\epsilon^{21}\delta^2)$) is asymptotically much larger than that of Fig. 1c ($\Theta(\epsilon^{31}\delta^3)$), which in turn is much larger than that of Fig. 1d ($\Theta(\epsilon^{40})$). In fact, this same ordering remains valid so long as $q > 1/3$ in (3).

3 Hypothesis Generation and Search

It is critical that our qualitative probabilistic analysis 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.

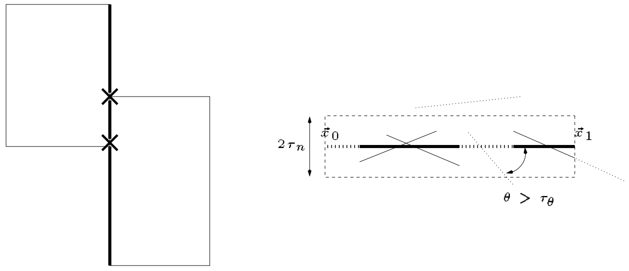
Here we show how qualitative probabilities can be used to determine the plausibility of partial interpretations, which can then be used to prune the search. For example, consider the image data in Fig. 4a along with the hypothesis depicted in Fig. 4b. In particular, the hypothesis is that the two image 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 (as shown in Fig. 4b). 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 $\Theta(\epsilon^6)$.

Moreover, since there are two drop-outs each with a free endpoint (see Fig. 4b), 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 (2), 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, for sufficiently small δ, the odds actually favor the two sticks hypothesis. 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. 4c. This implies an under-segmentation error has occurred 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. 4c) 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 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) (b)

Figure 5: (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.

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. 4b). 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. 4a 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. 6.1 we discuss the results of applying this pruning heuristic to image data.

4 Detailed Imaging Model

To apply our system to real images it is critical that the imaging model accounts for typical imperfections in the

[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.

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; and 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. 5). The last two types of imperfections are discussed below.

4.1 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. 5a). 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. 5a).

4.2 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. 5b). 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 Fig. 5b). In addition, we specify a minimum length for any drop-outs and false-targets.

5 Application to Blocks-World Scenes

We demonstrate our approach on blocks-world scenes (see Fig. 7). 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. 6). 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.

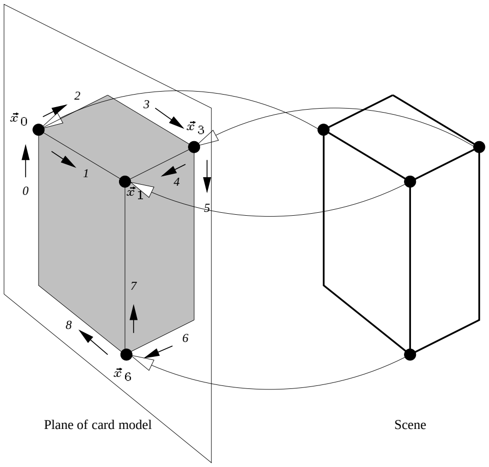


Figure 6: The card-world representation of a block is the orthographic projection of a block, depicted here in grey. The arrows on the edges within the card indicate the orientation convention used.

An orthographic view of a general 3D block is fully specified by four points in the image plane, as shown in Fig. 6. 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. 7c 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. 7c.

6 Experiments

For our experiments we consider image data from images of simple blocks-world scenes. We used the line-finder in the *Khoros 1.2* package to generate the image data in Fig. 7a. Note that, even for human observers given this

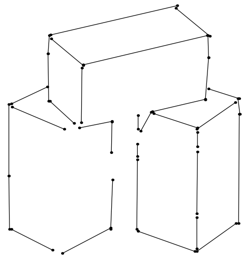
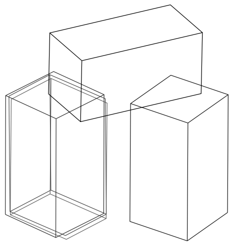
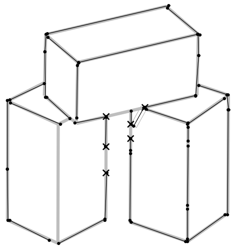


Image "9": 34 Segments



Individual blocks: 5



Score:$\epsilon^{34}\delta^{5}$ (next best $\epsilon^{42}\delta^{7}$)
Search (secs): 6, 10, 17, 20
Nodes: 29, 44, 69, 81

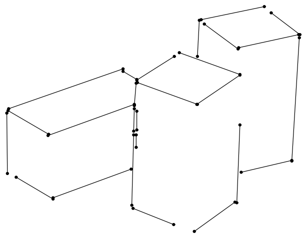
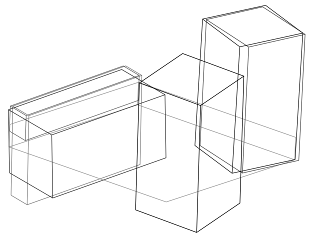
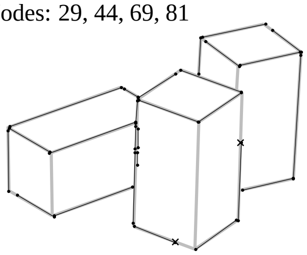


Image "5": 26 Segments



Individual blocks: 7



Score $\epsilon^{26}\delta^{5}$ (next best $\epsilon^{35}\delta^{7}$)
Search (secs): 12, 17, 31, 57
Nodes: 53, 67, 115, 188

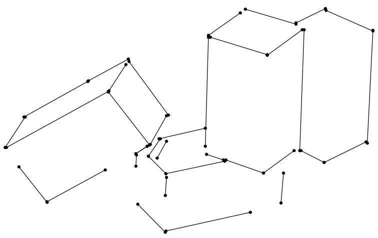
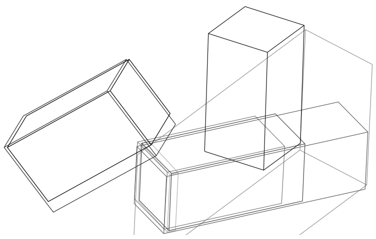
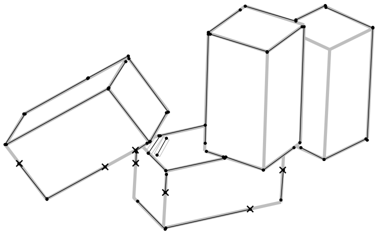


Image "E": 34 Segments



Individual blocks: 12
(Best 8 shown)



Score $\epsilon^{47}\delta^{10}$ (1 other within ϵ^{2})
Search (secs): 56, 72, 192, 257
Nodes: 150, 210, 493, 671

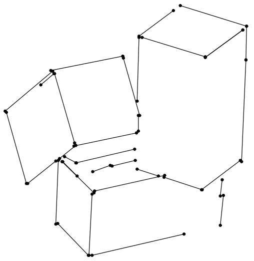
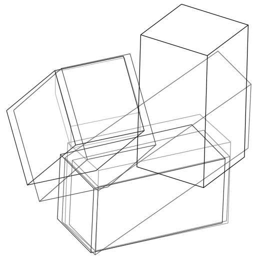
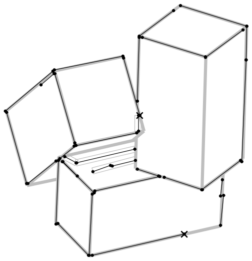


Image "D": 34 Segments



Individual blocks: 11
(Best 8 shown)



Score:$\epsilon^{42}\delta^{7}$ (1 other within ϵ^{2})
Search (secs): 20, 37, 67, 142
Nodes: 80, 143, 241, 475

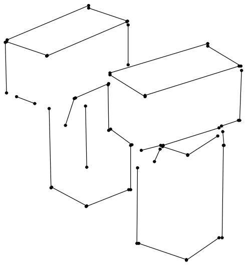
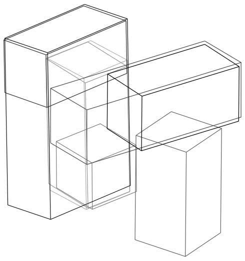
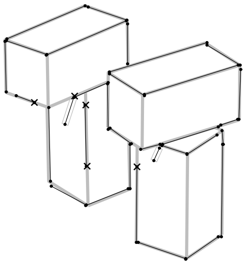


Image "7": 32 Segments



Individual blocks: 10
(Best 8 shown)



Score:$\epsilon^{45}\delta^{8}$ (1 other within ϵ^{2})
Search (secs): 21*, 66, 132, 228
Nodes: 72*, 182, 351, 578

Continued from previous page.

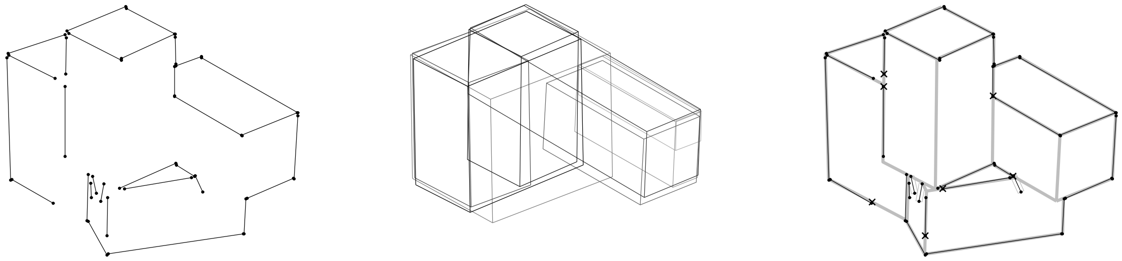


Image “4”: 30 Segments

Individual blocks: 18
(Best 8 shown)

Score:$\epsilon^{51}\delta^{9}$ (7+ others within ϵ^{2})
Search (secs): 65*, 160*, 419, 518
Nodes: 221*, 479*, 1247, 1602

(a) (b) (c)

Figure 7: Experimental results for several blocks-world images. (a) Line-finder results. (b) The fully-instantiated blocks with the highest individual odds. For clarity, the individual blocks are shown with a slight random “jitter”, and lighter greys indicate blocks with lower odds. (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 preferred interpretation. The ambiguity of the preferred interpretation is indicated by either the score of the next best interpretation found, or by the number of other interpretations found with scores within ϵ^2 of this best one. The execution times (on a Pentium II 300MHz processor) and the number of unique interpretations visited for the band search algorithm are reported for $N = 1, 2, 4, 8$, respectively (here * denotes the search at this bandwidth failed to find the preferred interpretation). Note the preferred interpretations for images E and 4 involve individual blocks that are ranked 11^{th} and 17^{th}, respectively, and are therefore not displayed in column (b).

data, there are some minor ambiguities in the interpretations (eg. the bottom-left block in image “7”, and the leftmost block in image “4” are not completely resolved).

6.1 Individual Blocks

We first implemented a brute-force search similar to the IT search in [5]. The significant differences are that 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 [5] this search proved to be impractical, taking 4 to 6 hours on a Pentium 133MHz processor. For the simple image data sets in Fig. 7a 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. 3), 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 cases this failed was for the left block in image “4” and the bottom-left block in “7”. Here the line-finder failed to detect enough edges to fully instantiate the block, and therefore the true block could not be found. Instead, in both of these cases the search algorithm grouped additional edges together with those from the true block (see Fig. 7b).

An order of magnitude speed-up in the search was obtained by using the plausible garden path heuristic discussed in Sec. 3. 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 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. The resulting search ran in under a minute on each of the examples in Fig. 7. The resulting search found every true block except for the two blocks missed by the full search.

6.2 Complete Interpretations

Here we consider the full 2D layered interpretations for the blocks-world image data in Fig. 7a. 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. 7b) arranged in all possible layered depth relations, plus zero or more sticks in the foreground.

To find plausible interpretations, we perform a greedy band-search. This is an iterative process which maintains a “band” of the best interpretations found so far. The ordering is prescribed by the power of ϵ in the unnormalized posterior; the term in δ is ignored. During the search the band is pruned by deleting all interpretations whose unnormalized posterior is asymptotically smaller than that of the N^{th} ranked interpretation. (Note that, in the case of ties for the N^{th} position, the band can contain more than N interpretations.) We refer to N as the search bandwidth.

The search proceeds by iteratively updating the interpretations in the band. Initially the band is set to contain only the trivial interpretation consisting of all sticks. To update the band at each iteration, new candidate interpretations are generated by adding one block to each interpretation in the band. These additional blocks are inserted at each feasible depth. Since the addition of a block could occlude much of an existing block, or even several blocks, single blocks are then greedily deleted from the candidate interpretations so long as the scores are increased by doing so. The union of current band and this resulting set of candidate interpretations is pruned at the N^{th} ranked score to form the band for the next iteration of the search. This process continues until the band does not change from one iteration to the next. Note that just one block can be added to any interpretation within the band during each iteration, and thus we again require a "garden path" to the preferred solution.

The results of this algorithm are presented in Fig. 7c, where the most preferred interpretations are displayed along with their unnormalized posteriors. The run-times for bandwidths $N = 1, 2, 4,$ and 8 are observed to grow roughly linearly with N, as do the the number of unique interpretations visited. The search algorithm arrived at the most preferred solution in all cases except for Image 7 with bandwidth 1, and for Image 4 with bandwidths 1 and 2. Note that images 4 and 7 exhibit various accidental alignments which allow for edges from different blocks to be incorrectly grouped into larger blocks. These larger blocks are selected early in the search process and, for small bandwidths, cause it to be misled.

The ambiguity in the interpretation of the images is also represented in Fig. 7c. Images 5 and 9 are found to be strongly unambiguous, with the second best unnormalized posteriors 8 or 9 orders of ϵ smaller than the preferred one. For image E two interpretations differing only in a depth reversal of the rightmost block were found with the same maximal score. For Image D the ambiguity is between which of the top two blocks occludes a small part of the other. Finally, the ambiguities for images 7 and 4 arise because the image data is insufficient to resolve one of the blocks in each case, and several plausible choices exist. All these ambiguities seem natural given the image data.

Note that for imaging systems with finer resolution parameters (i.e. a 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 comparison, note that the resolution of the images used for Fig. 7 was about the equivalent of the resolution the human fovea achieves on your thumbnail held at arms length (i.e. 2 degrees of arc).

7 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 effective search heuristics.

The same style of analysis can be applied to other domains, such as model-based object recognition [3, 5, 7, 10], curve and surface grouping [4, 14, 15], and simple motion interpretation [8]. These are important areas for further study.

An open question concerns how our approach based on qualitative probabilities performs compared to quantitative approaches for scene interpretation (eg., [1, 3, 7, 9]), and further if we can exploit quantitative probability information (such as the relative frequency of various types of objects) to obtain a stronger ordering on our interpretations and/or better search heuristics.

Acknowledgements

This research was supported by NSERC Canada. We would like to thank David Jacobs, Jacob Feldman, and Chakra Chennubhotla for helpful comments on an earlier version of this paper.

References

[1] E. Adelson and A. Pentland, The perception of shading and reflectance. In D. Knill and W. Richards, eds., *Perception as Bayesian Inference*, Cambridge University Press, 1996.

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

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

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

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

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

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

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

[9] Y. LeClerc, Constructing Simple Stable Descriptions for Image Partitioning. *IJCV*, (3):73–102, 1989.

[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] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Pub., 1988.

[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.

[16] L. Williams and K. Thornber, A Comparison of Measures for Detecting Natural Shapes in Cluttered Backgrounds *ECCV-98*, p.432, 1998.