PARALLEL STRUCTURE RECOGNITION WITH UNCERTAINTY: COUPLED SEGMENTATION AND MATCHING

Paul R. Cooper

Technical Report #5 • September 1990

Parallel Structure Recognition with Uncertainty: Coupled Segmentation and Matching

Paul R. Cooper

September 1990

The Institute for the Learning Sciences
Northwestern University
Evanston, IL 60201

This work was supported in part by the Department of Defense, specifically by the Defense Advanced Research Projects Agency, the Air Force Office of Scientific Research, and the Office of Naval Research. The Institute for the Learning Sciences was established in 1989 with the support of Andersen Consulting, part of The Arthur Andersen Worldwide Organization. The Institute receives additional funding from Ameritech (an Institute Partner) and IBM.

Abstract

This paper addresses the problem of recognizing structurally composed objects from uncertain image-derived evidence. The solution is a pair of cooperating networks that simultaneously segment and recognize the objects in the scene. The segmentation problem is posed as a problem of labelling a graph that represents object parts and their relationships at a high level of abstraction. Recognition is achieved by computing partwise correspondence between object parts in the scene and object model parts in a model base. The same labelling scheme is used for the recognition network. A coupled Markov Random Field provides a single unified formal framework for both labelling computations.

In the segmentation network, evidence from the image expressed as likelihoods on the labels is combined with clique potentials representing both qualitative a priori constraints and domain dependent knowledge. Clique potentials in the recognition network represent constraints that enforce pairwise consistency in the matching of parts, and coupling constraints between the networks ensure that the segmentation and recognition decisions are in agreement. The domain problem is the recognition of Tinkertoy objects. Implementation experiments show the framework can interpret ambiguous scenes with occlusion, accidental alignment, and noise.

1 Introduction

Noise and the projection of a 3D world into two dimensions mean that image data can provide only uncertain information for use in higher level visual processing, such as recognition. Visual recognition therefore requires inference and decision making.

In fact, at least two related decisions are required. First, the *true structure* of the world must be inferred. That is, a decision about the physical characteristics of the world must be made, based on the uncertain input information and prior knowledge. Objects must be segmented and their physical parameters must be determined. Signal must be separated from noise, and inferences about what caused the signal must be made. This decision is a measurement decision.

A second decision, the recognition decision, is required to determine the *identity* of the objects in the world. A match must be computed – the models that best explain the facts of the physical world must be selected. Recognition is achieved only when the measured world is explained in terms of what was known before. But with uncertain information, this also requires a decision.

It has long been recognized that these problems are best addressed together. This paper describes a framework whereby this can be accomplished directly – a single network is used to pose and make both decisions simultaneously, in a unified and completely coupled way.

Clearly, the decisions are intimately related. Without some decision about the world's physical characteristics, object identities can never be determined. So the recognition decision depends to some extent upon the measurement or structure inference decision. But the structure inference decision can also depend upon the recognition decision. Determining the physical parameters of the world, especially in the presence of noise, is by itself often an under-determined ill-posed problem. Prior or high-level knowledge in many forms can play a role in resolving this uncertainty. In particular, partial information about the recognition decision, such as the sighting of characteristic identifying features, can provide evidence that helps resolve the uncertainty in the measurement task. Other prior knowledge, such as smoothness assumptions imbedded in low-level vision operations, or situation context, can also play a role.

This paper presents a massively parallel network that computes these two decisions together. The domain problem that is addressed is the recognition of Tinkertoy objects. Both the structure inference decision and the recognition decision for Tinkertoy objects are posed as labelling problems. Each of the two labelling problems is itself represented as a Markov Random Field (MRF), and the two fields (representing the two labelling problems) are connected together to form a single coupled Markov Random Field. Prior knowledge about both decisions is represented within the network as weights between the variables, and uncertain evidence is represented as likelihoods at the labels. Finally, an estimation algorithm on the network is used to infer both the true scene structure and an object-model match simultaneously.

The paper first gives some background context and a detailed description of the network. Following this, some implementation experiments are described. The first two experiments involve only the segmentation subnet, and demonstrate how the combination of prior knowledge and evidence in a high-level representation can overcome local ambiguity arising from occlusion and accidental alignment. The last experiment shows how coupling the recognition and segmentation processes together can yield a correct interpretation of a scene even when the local evidence favors the incorrect interpretation.

2 Background

2.1 Recognition from Structure and Structure Inference

The recognition of Tinkertoy objects is the goal of The Tinkertoy Project [Cooper, 1989]. Previous work [Cooper and Swain, 1989; Swain and Cooper, 1988; Cooper, 1988; Cooper and Hollbach, 1987; Cooper and Swain, 1988] has addressed issues in the parallel recognition of objects from structure, assuming that discrete essentially error-free descriptions of composed objects can be obtained from images. Recognition from structure addresses the role of parts and the spatial relations between them in the recognition process [Witkin and Tenenbaum, 1983; Lowe, 1985; Pentland, 1987; Pentland, 86; Biederman, 1985; Hoffman and Richards, 1986]. Because their identity is defined primarily by the spatial relationships between simple parts, Tinkertoys provide a convenient domain task for examining recognition from structure.

When perfect information assumptions are *not* made, developing a principled solution to the problem of recognition from structure is considerably more difficult. One traditional approach is to assume that the descriptions produced by early segmentation processes may contain errors. Structure recognition then requires solving the inexact matching problem for structures. Even if a reasonable definition for distance or "best match" can be defined, heuristics of some kind must be adopted to circumvent the combinatorial nature of the resulting problem [Shapiro and Haralick, 1981; Eshera and Fu, 1986].

Inferring the true structure of the physical world subsumes both the problems of segmentation and reconstruction. Evidential approaches to these problems have been proposed before [Feldman and Yakimovsky, 1974; Chou, 1988; Sher, 1987], but usually with a much lower-level visual representation. Regularization methods for early vision [Poggio *et al.*, 1985] exploit prior knowledge and can be framed as network computation, but these methods do not exploit information about the recognition computation and generally yield lower-level representations as well.

Relaxation and constraint satisfaction have often been used as the basis for the recognition of objects described discretely [Mackworth, 1977; Hinton, 1977; Kitchen and Rosenfeld, 1979; Hummel and Zucker, 1983]. Probabilistic frameworks for recognition have also been used in the past [Binford *et al.*, 1987; Bolles, 1977] and high-level information has been exploited in model-based vision [Brooks, 1986]. Most closely related to this work is the work of the Yale Stickville project [Mjolsness *et al.*, 1988; Utans *et al.*, 1989]. In that work, a neural network architecture is described that recognizes composed objects by the optimization of an objective function encoding many of the same constraints used here. But no coupled recognition and segmentation process was proposed.

2.2 Markov Random Fields in Vision

Markov Random Fields (MRFs) have been used as the basis of an evidential approach to many computer vision tasks in recent years [Geman and Geman, 1984; Marroquin, 1985; Cross and Jain, 1983; Chou, 1988]. Most of this work has addressed very low-level representations and processes, and has used MRFs that are essentially rectangular arrays. In some cases coupled MRFs were used, as in Chou [1988], who showed how the segmentation and reconstruction processes could be coupled.

The theory of Markov Random Fields extends beyond simple arrays, and can be applied to arbitrarily structured graphs like the one used later to build a high-level structure representation. Some of the relevant aspects of MRF theory and its application to labelling problems are now very briefly reviewed [Kindermann and Snell, 1980].

Consider a set $\mathbf{X}$ of discrete-valued random variables X. Associate with the random variables an undirected graph G defined as a set S of sites (or vertices) and a neighborhood system (or set of edges) E. The random variables of the field are indexed by the graph vertices as X_s. Variables are neighbors in the MRF when the associated vertices are adjacent in the graph. In the formulation of a labelling problem as an MRF, the variables in the labelling problem are the random variables of the MRF.

The value ω_s of a random variable may be any member l_i of the state space set L. Because of the application of the field to the labelling problem, the event elements of the set L will be called labels. An assignment of values to all the variables in the field is called a configuration, and is denoted ω.

We are interested in the probability distributions P over the random field $\mathbf{X}$. Markov Random Fields have a locality property:

$$P(X_s = \omega_s | X_r = \omega_r, r \in S, r \neq s) = P(X_s = \omega_s | X_r = \omega_r, r \in N_s) \qquad (1)$$

that says roughly that the state of site is dependent only upon the state of its neighbors (N_s). MRFs can also be characterized in terms of an energy function U with a

Gibb's distribution:

$$P(\omega) = \frac{e^{-U(\omega)/T}}{Z} \tag{2}$$

where T is the temperature, and Z is a normalizing constant.

If we are interested only in the prior distribution $P(\omega)$, the energy function U is defined as:

$$U(\omega) = \sum_{c \in C} V_c(\omega) \tag{3}$$

where C is the set of cliques defined by the neighborhood graph G, and the V_c are the clique potentials.

Specifying the clique potentials V_c provides a convenient way to specify the global joint prior probability distribution P. The clique potentials can be conveniently viewed as weights in a connectionist network. They provide a mechanism to express soft constraints between labels at related variables. Unary clique potentials in effect express first order priors, while binary clique potentials express the constraints between pairs of variables in the field.

Suppose we are instead interested in the distribution $P(\omega|O)$ on the field after an observation O. An observation constitutes a combination of spatially distinct observations at each local site. The evidence from an observation at a site is denoted $P(O_s|\omega_s)$ and is called a likelihood. Assuming likelihoods are local and spatially distinct, it is reasonable to assume that they are conditionally independent. Then, with Bayes' Rule we can derive:

$$U(\omega|O) = \sum_{c \in C} V_c(\omega) - \sum_{s \in S} \log P(O_s|\omega_s) \tag{4}$$

To summarize, the MRF represents a labelling problem. Evidence about the hypotheses is expressed as label likelihoods, and prior knowledge is expressed in terms of the clique potentials, generalized weights that express soft constraints between spatially related variables.

Inference on the MRF network can be framed in terms of the energy function. For example, the maximum a posteriori probability can be computed by finding the minimum of the non-convex energy function U. Needless to say, this is a non-trivial problem and not the focus of this work. In the experiments which follow, a deterministic approximation algorithm called Highest Confidence First (HCF [Chou, 1988]) is used to find a good minimum of the energy function. This minimum corresponds to a particular selection of labels for each variable.

3 Network Description

The design of a coupled MRF network for the segmentation and recognition of structured objects will now be given. The constraints on the design arose from the network's two major functions: representing with uncertainty possible Tinkertoy scenes, and matching such scenes to object models. An additional design constraint was minimizing the network's complexity so that it could actually be realized.

The network description is given in five parts. First, definitions are provided for the variables, labels and connections that specify the MRF. Then a description of the network weights, prior knowledge expressed as clique potentials, is given. Finally, the form of the evidence from the problem instances is described.

3.1 Variables

By far the most important part of the problem is selecting an appropriate set of variables. The separate task of each subnet in the coupled MRF suggests that different sets of variables be used in each part. However, the task of coupling the decisions together is simplified if an analogous set of variables is used in each subnet.

For the recognition subnet, past work involving recognition with discrete perfect data provides one starting point [Cooper and Hollbach, 1987]. A convenient parallel formulation of the structure matching problem as a labelling problem, based on the unit/value principle [Barlow, 1972; Feldman and Ballard, 1982; Ballard, 1984], is to consider all possible object-model *part correspondences* simultaneously. In this scheme, the variables are defined to be the object parts, and the possible labels are the potentially corresponding model parts. This architecture, a cross-product matching table, is common in many connectionist designs [Mjolsness *et al.*, 1988; Goddard, 1988]. It remains to select part-variables that adequately encode the richness of the domain. The most obvious choice for parts is the two physical Tinkertoy parts, the rod and junction disk, but it is difficult to encode junction geometry with this primitive part selection. Junction geometry is encoded implicitly if the *slots* on the disk (discrete junction connection points) are chosen as logical "parts" rather than the disks themselves.

But recognition is only half the problem. Simply representing with uncertainty the range of possible Tinkertoy scenes is a difficult problem. As can be seen in Figure 1, even images of very simple Tinkertoy scenes contain significant uncertainty. Part parameters such as rod length are clearly difficult to determine reliably. But self-occlusion and noise make the geometry of the junctions difficult to determine, and this might have a profound effect on the task of determining the identity of the object. In short, in a real image of even a very simple part-and-junction world, the existence, identity and parameterization of the parts might all be uncertain.

The crucial uncertainty is uncertainty about structure: what is attached to what. In an arbitrary Tinkertoy scene, any two slots not on the same disk could be connected with a rod. A convenient representation for these possibilities is to define a set of variables called *virtual rod variables*, one variable for each possible slot/slot connection. Whether or not a particular variable will represent an actual rod in the scene will depend on the scene and the image evidence.

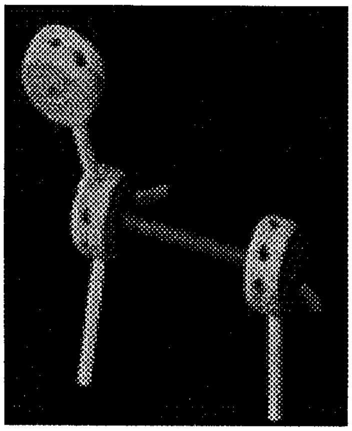

Figure 1: Real Tinkertoy Image

To summarize, the complete MRF will have two analogously defined sets of variables, one set for the subnet doing recognition, and one set for the subnet doing structure inference. Variables will also come in two types: *slot variables* (called simply *slots*) and *virtual rod variables* (or *vrods*).

Formally, we specify the entire Markov Random Field $\mathbf{X}$ consisting of 2 sub-fields, $\mathbf{X}_{struc}$ and $\mathbf{X}_{recog}$. Each sub-field has two types of variables, denoted $\mathbf{X}^{slot}$ and $\mathbf{X}^{vrod}$. The variables are most conveniently described in terms of an arbitrary but fixed size set of disks, and an arbitrary slot labelling on the disk. Thus we define the sets:

$$\begin{aligned} \text{DISKS} &= \{0, 1, 2, \ldots \text{NumDisks}\} && (5) \\ \text{SLOTS} &= \{0, 1, 2, \ldots \text{NumSlotsPerDisk}\} && (6) \end{aligned}$$

In the following definitions, the subnet types are defined as $\{\text{struc}, \text{recog}\}$. Entities associated with the recognition subnet are often denoted with a primed notation. The slots are basically defined as an ordered pair describing the disk and slot number on the disk. The virtual rods are defined in terms of which slots they connect.

$$\begin{aligned} \mathbf{X}^{slot}_{struc} &= \{(\text{struc}, n, m) | n \in \text{DISKS}, m \in \text{SLOTS}\} && (7) \\ \mathbf{X}^{slot}_{recog} &= \{(\text{recog}, n', m') | n' \in \text{DISKS}, m' \in \text{SLOTS}\} && (8) \\ \mathbf{X}^{vrod}_{struc} &= \{(\text{struc}, a, b) | a, b \in \mathbf{X}^{slot}_{struc}\} && (9) \\ \mathbf{X}^{vrod}_{recog} &= \{(\text{recog}, a', b') | a', b' \in \mathbf{X}^{slot}_{struc}\} && (10) \end{aligned}$$

3.2 Labels

The set of labels attached to each variable determines the decision being made in the labelling problem. In this network there are two labelling problems, each encoded as a subnet in the MRF. Corresponding to each subnet and decision is a different set of variables.

Structure Inference Subnet		Matching Subnet	
Variables	Labels	Variables	Labels
slots	doesn't exist exists and empty exists and full	slots′	doesn't matching anything matches model slot X matches model slot Y, etc.
virtual rods	doesn't exist exists and has length L1 exists and has length L2 exists and has length L3	virtual rods′	doesn't match anything matches model rod A matches model rod B, etc.

Table 1: Definition of Variables and Labels

For the structure inference decision, the labels correspond to the hypothetical physical parameters the parts and their composition can actually have in the world. Thus, the labels for the *vrod* variables are the possible lengths of the rods, and the labels for the *slot* variables describe whether the slot is *empty* or *filled* (i.e. has a rod plugged into it). An important part of the label sets is provision for the possibility that the parts do not exist. This is necessary to account for the possibility that what seems to be a part is in fact an artifact of signal noise.

Formally, we can define the label sets for the variables $\mathbf{X}_{struc}$ as follows:

$$L^{slot}_{struc} = \{\text{doesn't exist}, \text{exists} \wedge \text{empty}, \text{exists} \wedge \text{full}\} \tag{11}$$

$$L^{vrod}_{struc} = \{\text{doesn't exist}, \text{exists} \wedge \text{length L1}, \text{exists} \wedge \text{length L2}, \text{exists} \wedge \text{length L3}\} \tag{12}$$

In the cross-product matching array, the labels of the $\mathbf{X}_{recog}$ variables are the potentially matching parts of the same type from the model base. The label set is augmented by the possibility that no corresponding model part can be found to match an object part. If we call the sets of model parts ModelSlots and ModelRods respectively, we can define the sets of labels as:

$$L^{slot}_{recog} = \{l | l \in \{\text{doesn't match anything}\} \cup \text{ModelSlots}\} \tag{13}$$

$$L^{vrod}_{recog} = \{l | l \in \{\text{doesn't match anything}\} \cup \text{ModelRods}\} \tag{14}$$

The interpretation of a variable in the set $\mathbf{X}_{recog}$ having a particular label is that the *object* part represented by the *variable* is matched to the *model* part represented by the *label.*

The definitions of the variables and labels for both halves of the field are summarized in Table 1.

3.3 Connections: The MRF Graph

Connections between associated variables are required to represent the constraints that can occur between entities in the problem. As in any connectionist design, it is

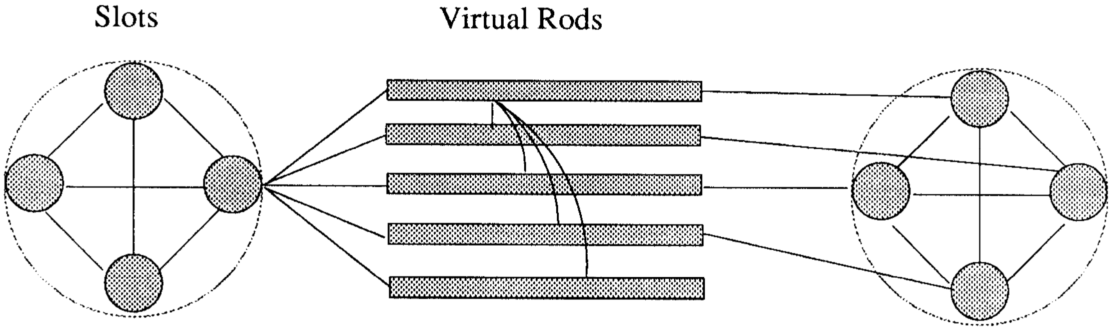

Figure 2: Fragment of MRF Graph. The shaded objects are the MRF sites (slots and rods). The solid lines represent edges in the MRF graph. The virtual rods show all the connections from one slot to slots on one other disk, as well as the dangling rod possibility. The set of virtual rods is a clique; only some of the connections are shown.

establishing these connections that encodes the essence of the computation. In the case of the MRF, the connections define the MRF graph, and represent adjacencies between related variables in the field.

The topology of the overall network consists of the same network replicated twice, once for each labelling problem, and a set of coupling connections. Within each subnet there exist the following edges, defined in this case for the structure inference subnet.

Each slot on a disk is connected to every other slot on the disk.

$$E_{struc}^{slot/slot} = \{ < (\text{struc}, n, m), (\text{struc}, n, p) > | n \in \text{DISKS}, m, p \in \text{SLOTS}, m \neq p \} \quad (15)$$

Each slot is connected to a set of rods, each of which can connect it to any other slot.

$$\begin{aligned} E_{struc}^{slot/vrod} &= \{ < (\text{struc}, n, m), (\text{struc}, a, b) > | n \in \text{DISKS}, m \in \text{SLOTS}, & (16) \\ & \quad a, b \in \mathbf{X}_{struc}^{slot}, a = (\text{struc}, n, m) \} & (17) \end{aligned}$$

Finally, each set of rods attached to a particular slot forms a completely connected subgraph.

$$E_{struc}^{vrod/vrod} = \{ < (\text{struc}, a, b), (\text{struc}, a, c) > | a, b, c \in \mathbf{X}_{struc}^{slot}, b \neq c \} \quad (18)$$

A sketch of a fragment of the graph defined by these connections is given in Figure 2. Edges in the subnet devoted to the recognition labelling problem are defined exactly analogously.

The overall structure of the net is loosely suggested in Figure 3. The coupling connections, which turn out to primarily be connections between the corresponding variables in the two subfields, are not shown, but are implied by the alignment of

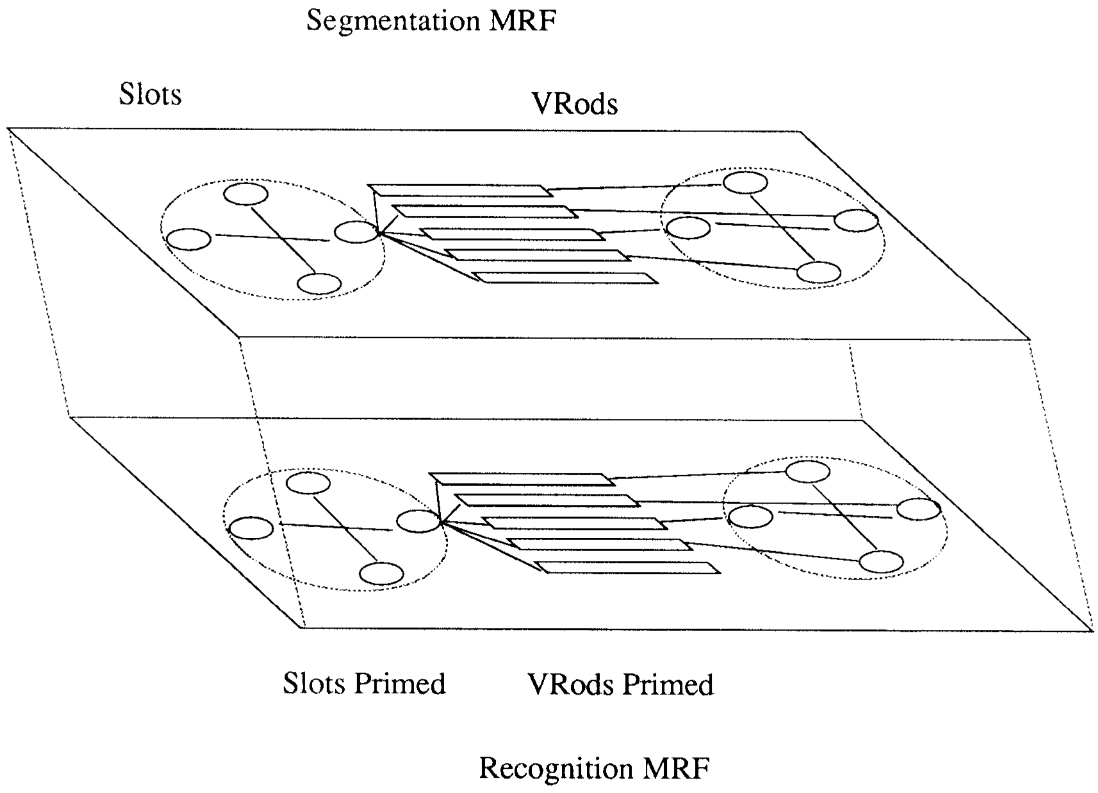

Figure 3: Recognition and Segmentation MRFs are Structurally Identical

the two subnets. The complete set of coupling connections is given by the following equations:

$$E^{slot/slot'}_{coupled} = \{< (\text{struc}, n, m), (\text{recog}, n, p) > | n \in \text{DISKS}, m, p \in \text{SLOTS}\} \quad (19)$$

$$E^{vrod/vrod'}_{coupled} = \{< (\text{struc}, a, b), (\text{recog}, a, b) > | a, b \in \mathbf{X}^{slot}_{struc}\} \quad (20)$$

The constraints implemented along the connections as clique potentials are described momentarily. The non-homogeneous, non-isotropic nature of the graph structure reflects the application of MRF theory to the represention of high-level structure and recognition, and differs greatly from traditional image-based array MRF applications. The label sets and their interpretations also differ substantially from previous MRF applications, which mainly used binary on/off label sets.

3.4 Network Weights: Prior Knowledge or Clique Potentials

To complete the definition of the MRF, it is necessary to provide data for the field. The evidence and prior knowledge must be specified.

In any inference problem involving perception, there are only two sources of information: sensor evidence for this problem instance, and what was known before. In probabilistic frameworks including MRFs, previous knowledge is expressed as priors. The joint prior distribution on an entire MRF is expressed through the clique potentials, or network weights. Domain dependent Tinkertoy knowledge, both qualitative and quantitative, is represented by clique potentials in the Tinkertoy MRF.

Most of the constraints in the field reflect qualitative facts about Tinkertoys, and can be considered "hard" constraints. It is convenient to describe three different kinds of constraints, corresponding to the three kinds of connections in the network: within the structure inference subnet, within the recognition subnet, and between the two subnets.

Constraints on Structure Inference

Qualitative Constraints The variables and labels in the structure inference subnet represent possible partwise physical interpretations of the scene. Potentially, such local partwise interpretations could be physically unrealizable when considered together. A set of qualitative constraints derived from the nature of Tinkertoys ensure that the partwise interpretations are physically consistent over all the parts. Table 2 lists the 2-cliques that enforce this consistency.

Consider, for example, two slots on a disk. It is locally possible that one slot be labelled as *doesn't exist*, and the other as *exists*. But this situation is physically

Clique		Potential
Slot	Slot	
doesn't exist	doesn't exist	consistent
doesn't exist	exists, full	inconsistent
doesn't exist	exists, empty	inconsistent
exists, empty	exists, empty	frequency dependent
exists, empty	exists, full	frequency dependent
exists, full	exists, full	frequency dependent
Vrod	Vrod	
doesn't exist	doesn't exist	consistent
doesn't exist	exists, $L1 \mid L2 \mid L3$	consistent
exists, $L1 \mid L2 \mid L3$	exists, $L1 \mid L2 \mid L3$	inconsistent
Slot	Vrod	
doesn't exist	doesn't exist	consistent
doesn't exist	exists, $L1 \mid L2 \mid L3$	inconsistent
exists, empty	doesn't exist	consistent
exists, empty	exists, $L1 \mid L2 \mid L3$	inconsistent
exists, full	doesn't exist	inconsistent
exists, full	exists, $L1 \mid L2 \mid L3$	good

Table 2: 2-Clique Potentials for the Structure Inference Subnet. Constraints enforcing consistent structure inference

impossible, a fact known a priori. Furthermore, slots are only connected to other slots on the same disk, so a 2-clique potential value which discourages the formation of this pairwise interpretation is built into the network. This constraint is reflected in lines two and three of the table. It would also be physically inconsistent to have two rods plugged into the same slot. A vrod/vrod clique potential expresses this fact. Another inconsistent pairwise interpretation would be if a slot were labelled *empty*, but an adjacent virtual rod variable (representing a rod plugged into that slot) were labelled *exists*.

Quantitative Prior Knowledge While many of the constraints encoded as potentials in the MRF represent hard facts known a priori, clique potentials can also express softer constraints where it is appropriate – quantitative prior knowledge. Clique potentials can represent the frequency with which local properties occurred in past problem instances. For example, unary clique potentials at the labels indicating rod length can be thought of as encoding first order statistics about the lengths of rods in previous problem instances. Salient second and higher order features (such as junction geometry at a disk) might also be represented in the network. The slot/slot 2-clique potentials might represent, for example, the fact that junctions with two rods 90 degrees apart occur frequently in Tinkertoy problems, while junctions with two rods at 180 degrees occur rarely. In this way, statistics based on a domain of previous problem instances can influence perceptual inference in the current problem instance. One might think of the domain dependent prior knowledge as "smoothing" the current evidence to a solution during the inference process on the network. Some of the power this possibility offers is explored in the first implementation experiment.

It should also be possible to learn the clique potential values by measuring the frequencies of local properties over a domain of problem instances. While converting frequency data to potentials may in general be difficult [Pearl, 1988], approximations may yield reasonable results [Swain, 1989]. Thus, clique potentials offer a principled mechanism for exploiting certain kinds of learnable domain dependent knowledge.

Constraints on Recognition

The variables and labels in the recognition subnet represent possible partwise object/model matches. Again, such local partwise computations could be globally inconsistent. The constraint in this case enforces consistency between matches of pairs of parts. In other words, if two object parts are connected in the object, the two model parts they respectively match must also be connected in the model. This is a common constraint in network-based formulations of graph matching [Mjolsness *et al.*, 1988]. For example, all the slots at an object disk must be matched against all the slots on a single model disk so that they can align.

Clique		Potential
Slot′	Slot′	
matches some model slot X	matches a consistent slot Y	consistent
matches some model slot X	matches an inconsistent slot Y	inconsistent
matches some model slot X	doesn't match anything	inconsistent
Vrod′	Vrod′	
matches some model rod	doesn't match anything	consistent
matches some model rod	matches any model rod	inconsistent
Slot′	Vrod′	
matches some model slot A	matches a model rod X connected to A	consistent
matches some model slot B	matches a model rod Y not connected to B	inconsistent

Table 3: 2-Clique Potentials for the Recognition Subnet. Constraints enforcing consistent partwise object/model matching

Note that these constraints are dependent upon the structure of the candidate model, and are thus derived from the model base. A summary of the the 2-clique potentials implementing this constraint is given in table 3.

Coupling Constraints

The coupling constraints between the subnets are, of course, crucial to the entire coupled approach. These constraints reinforce interpretations of the physical world that are consistent with explaining the world in terms of the model. Consider, for example, the possibility that object rod "A" matches model rod "2". Now the length of model rod "2" is known. If the model rod "2" were length $L1$, there would be a potential weight encouraging the physical interpretation of rod "A" as length $L1$. These coupling constraints are also hard qualitative constraints derived directly from the model, and are summarized in Table 4.

Clique Potential Values

The tables above specify clique potentials in a qualitative and somewhat vague way. In actuality, specific numbers were used to express the qualitative constraints. The numbers themselves are uninteresting, however, for at least two reasons. First, comparable constraints are implemented by comparable numbers in the network. Inconsistencies were all treated similarly. Second and more importantly, in practice once the appropriate order of magnitude was established for the representation of the clique parameters, variations in the values of the potentials had little effect on network performance.

Clique		Potential
Slot	Slot′	
exists filled	matches a filled model slot	consistent
exists filled	matches an empty model slot	inconsistent
exists empty	matches a filled model slot	inconsistent
exists empty	matches an empty model slot	consistent
Vrod	Vrod′	
exists, length L1	matches a model rod length L1	consistent
exists, length L1	matches a model rod not length L1	inconsistent
exists, length L2	matches a model rod length L2	consistent
exists, length L2	matches a model rod not length L2	inconsistent
exists, length L3	matches a model rod length L3	consistent
exists, length L3	matches a model rod not length L3	inconsistent

Table 4: Coupling 2-Clique Potentials

3.5 Evidence from the Problem Instance: Likelihoods

To complete the network definition, the image-derivable evidence constituting the particular problem instance must also be defined. The evidence from the image is organized as likelihoods at the labels in the segmentation subnet. For example, the likelihood that a particular rod has length $L1$ might be 0.7, and the likelihood that the rod has length $L2$ might be 0.1. While the theory of likelihood generation is well-established for low-level vision [Sher, 1987], generating fully justified probabilistic likelihoods at a high level of abstraction requires a complete probabilistic treatment of intermediate vision. For the experiments, the existence of a likelihood generation operator that actually gathers evidence from the image was assumed, and evidence was synthesized artificially. Generally the evidence was constructed from qualitative criteria, such as "very certain", "almost no evidence for this", etc.

The experiments deliberately probed a wide range of input conditions including worst cases, reflecting possible data that could arise from real images. Furthermore, the network behavior was extremely robust to variations in the exact values for the evidential input, reducing the significance of the fact that the data were synthetic.

A schematic overview of the entire system, showing the sources of prior information and evidence and the relationship between the subnets, is given in Figure 4.

3.6 Complexity and Correctness

With network algorithms, space complexity is relevant. The basic parameter of the network is the size of the scene that is representable, in terms of the number of available slots n. $O(n^2)$ virtual rods are required to represent all possible slot/slot

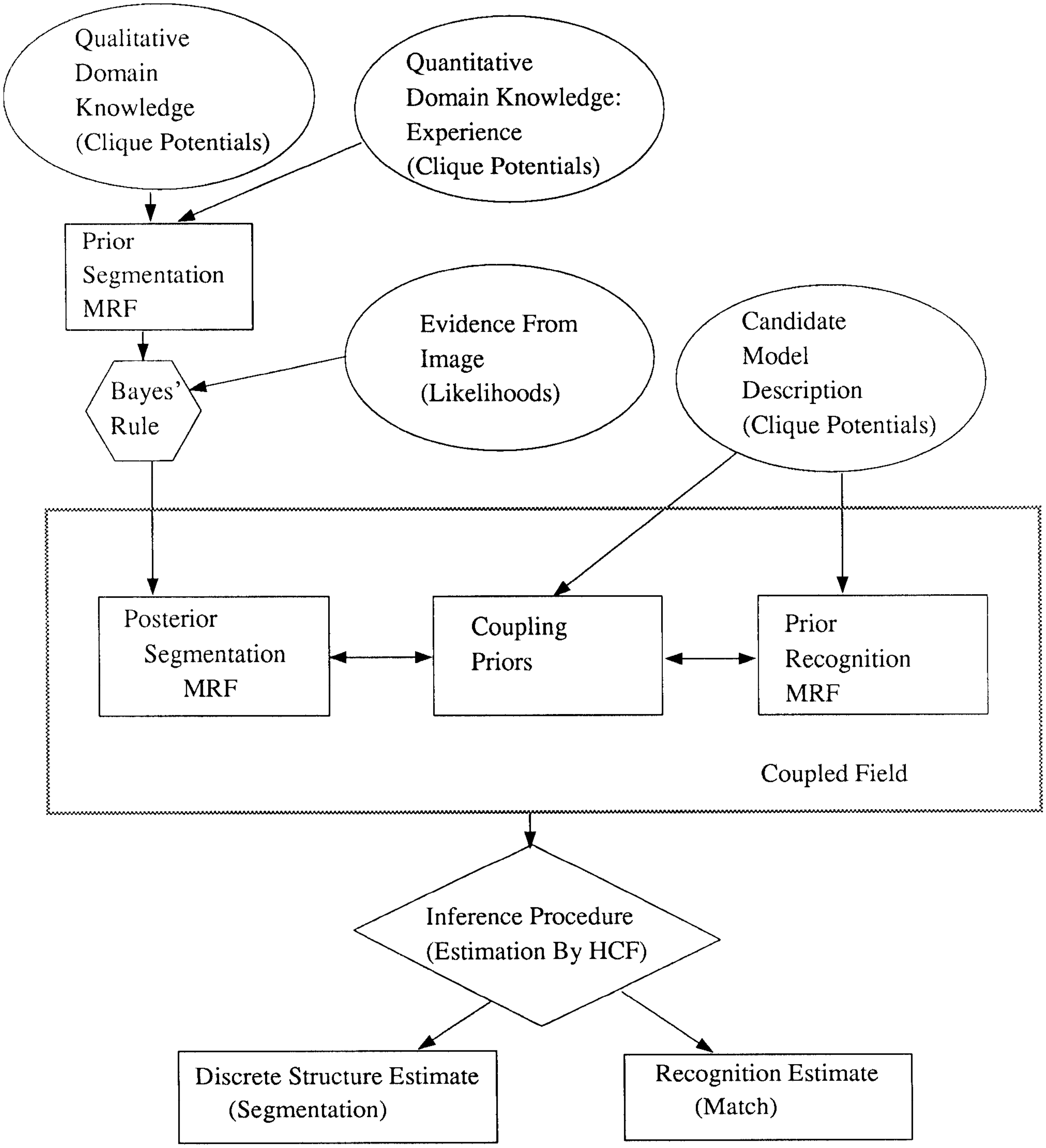
Qualitative Domain Knowledge (Clique Potentials)
Quantitative Domain Knowledge: Experience (Clique Potentials)
Prior Segmentation MRF
Evidence From Image (Likelihoods)
Candidate Model Description (Clique Potentials)
Bayes' Rule
Posterior Segmentation MRF
Coupling Priors
Prior Recognition MRF
Coupled Field
Inference Procedure (Estimation By HCF)
Discrete Structure Estimate (Segmentation)
Recognition Estimate (Match)

Figure 4: System Overview

connections, and connecting the virtual rods together ultimately requires $O(n^3)$ 2-cliques. Some other cubic terms arise from the representation of order 3 cliques, but the $O(n^3)$ factor dominates the space complexity.

With simpler networks, it is sometimes possible to prove correctness of a network algorithm and to determine time complexities [Cooper, 1988]. But the problem addressed here is complex enough to require the minimization of a non-convex energy function. Therefore, as is well known and expected with any hard problem, it is difficult to make substantive general statements about correctness and time complexity. In general, performance depends upon both the algorithm used to do the minimization, and upon the character of the energy space itself, which is defined both by the architecture of the network and by the data of a problem instance. Minimization algorithms were not a focus of this research. HCF [Chou, 1988], the minimization algorithm that was used, is guaranteed to converge, has exponential worst case time complexity but excellent performance in practice, and has been observed in many experiments in different domains to find good minima in the energy space. Other algorithms, such as simulated annealing [Kirkpatrick *et al.*, 1983] or continuation methods [Blake and Zisserman, 1987] could also have been used.

Of more interest here was the definition of a well-behaved energy space. The network design, incorporating a choice of variables and constraints, is a major determiner of the quality of the energy space. In the experiments in the next section, the energy function appeared well-behaved, good minima were detected, and network performance was robust with respect to variation in both the potentials and the evidence. Other network architectures addressing essentially the same problem have reported difficulties with energy spaces that are not so well-behaved [Utans *et al.*, 1989].

4 Implementation Experiments

This section describes results obtained with an implementation of the MRF network described above. The implementation was built with the Rochester Connectionist Simulator [Goddard *et al.*, 1988].

4.1 Structure Inference Only: Accidental Alignment

The first two experiments were designed to explore and demonstrate the representational power of the framework, and to explore tradeoffs between the evidence and priors in the process. These experiments were thus implemented with only the structure inference or segmentation subnet of the complete MRF network.

The basic structure of the first experiment is given in Figure 5. The scene shows a 2D Tinkertoy scene with 2 objects, a man and his dog, accidentally aligned so that the

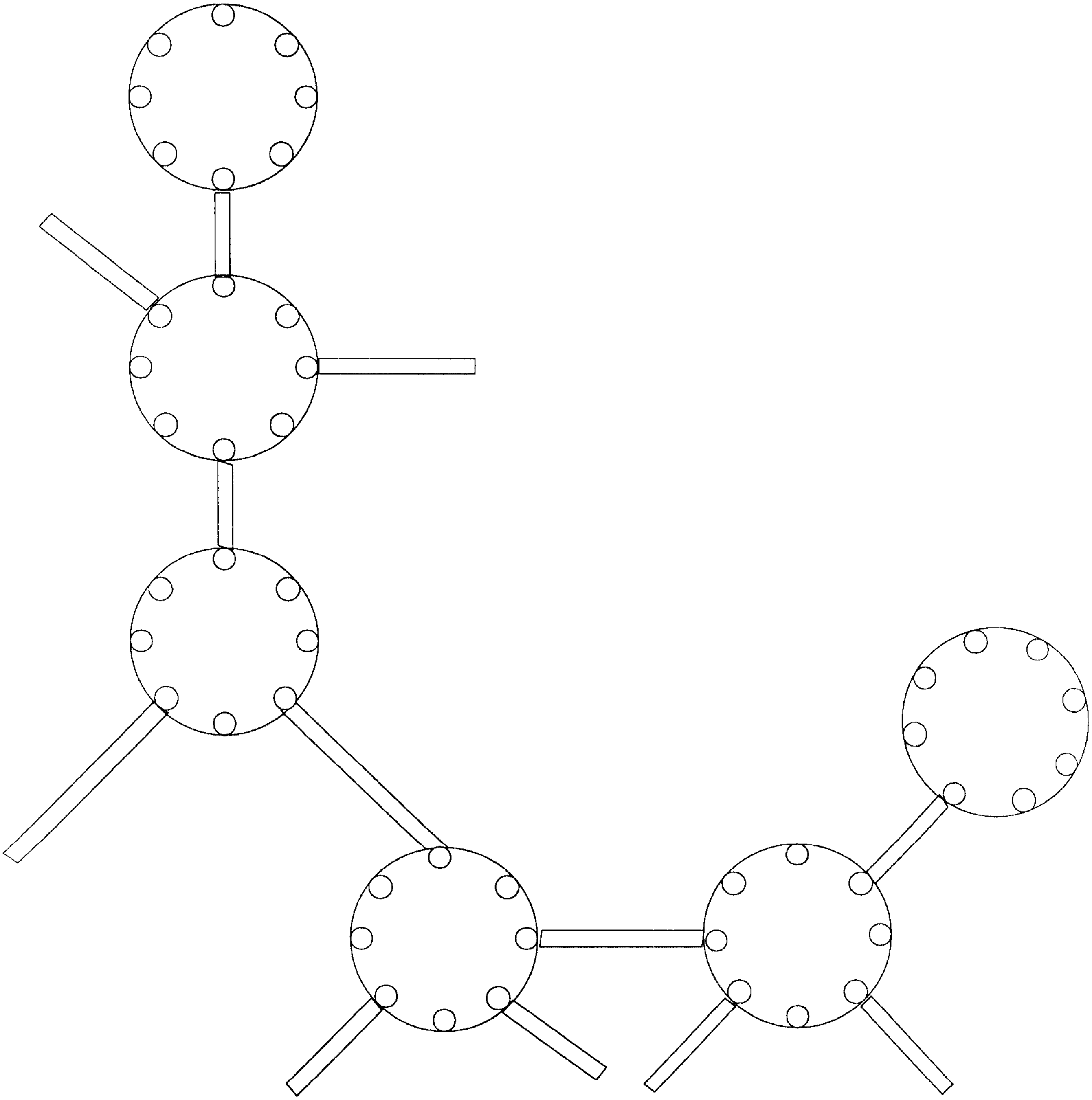

Figure 5: A Man and His Dog: schematic of image showing accidental alignment

hypothesis that there is only one object is reasonable. (Alternative interpretations that have been suggested include a caterpillar and a molecule with 6 atoms.)

The scene represents a non-trivial problem of representation and inference. Some interesting features of the experiment are as follows. First, it demonstrates a scene in which a segmentation ambiguity is present, and shows how the evidence in such a scene might occur and be represented. Local labelling ambiguity is present and simple to represent — for example, different rod lengths have different likelihoods. The experiment also contains non-trivial structural uncertainty — is it one object or two? Second, it demonstrates the way both priors and evidence combine to yield a decision. In this experiment, the evidence about the major segmentation decision is (by design) inconclusive. The priors must therefore provide the information necessary to achieve an interpretation. This represents one possible balance that can exist between the evidence and the priors. In some cases with ambiguous evidence, prior knowledge alone is inadequate, and interpretation mistakes are made. Third, it demonstrates the power of the inference procedure in resolving such an ambiguous decision problem. The sequential trace of the inference process is particularly impressive in this regard, because it involves a local decision-reversal. In this case, the global energy of the later decision is better than the first decision.

The input evidence for the problem instance is presented graphically in Figure 6. In the figure, the likelihoods are shown for each label by bar graphs located near the spatial hypothesis they describe. The lines are scaled to represent likelihoods between zero and one.

The evidence surrounding the connection of the two objects (at point A in the figure) is very ambiguous. The 'connected' and 'disconnected' hypotheses both have very similar evidence. Note that the hypothetical likelihood generator was fairly confident about the length of the rod, just not about whether it connected the two slots or not. (Both hypotheses have about the same likelihoods at each of their labels.) This is an example of how true structural ambiguity is represented in the net. Note also that the evidence at the slot hypothesis (B in the figure) is completely ambiguous. In effect, because of the accidental alignment, the likelihood generator would find evidence for the *full* label. On the other hand, a reasonable likelihood generator would probably have knowledge about slot-rod junctions, and would thus know that lack of perpendicularity at the junction is evidence for the *empty* label.

In Figures 7 through 11 the progress of the inference process on the experiment is shown. Figure 7 shows the first few label commitments that were made: these are the MRF sites with the least ambiguous evidence.

Of particular interest is Figure 8 when HCF has incorrectly labelled the scene as one object. Once the slot (the man's 'hip' joint, designated A in the figure) is labelled *full*, excitation energy exists in favor of some rod being attached. Then the 'connecting' virtual rod (B) is chosen because its local evidence is slightly stronger than that of the competing 'dangling' virtual rod (C.) When the connecting virtual

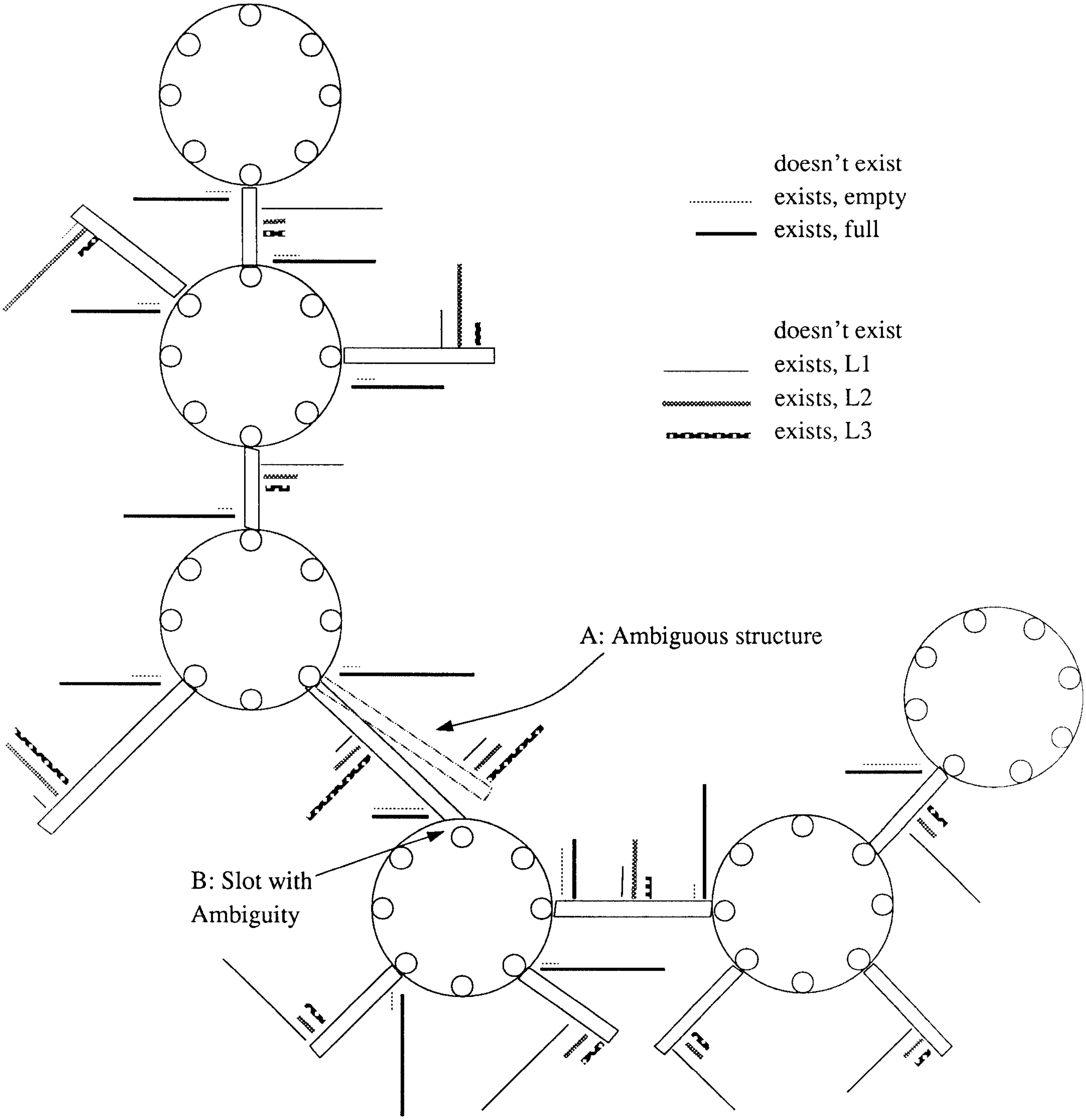

Figure 6: Input Evidence from the Image: bar graphs of label likelihoods

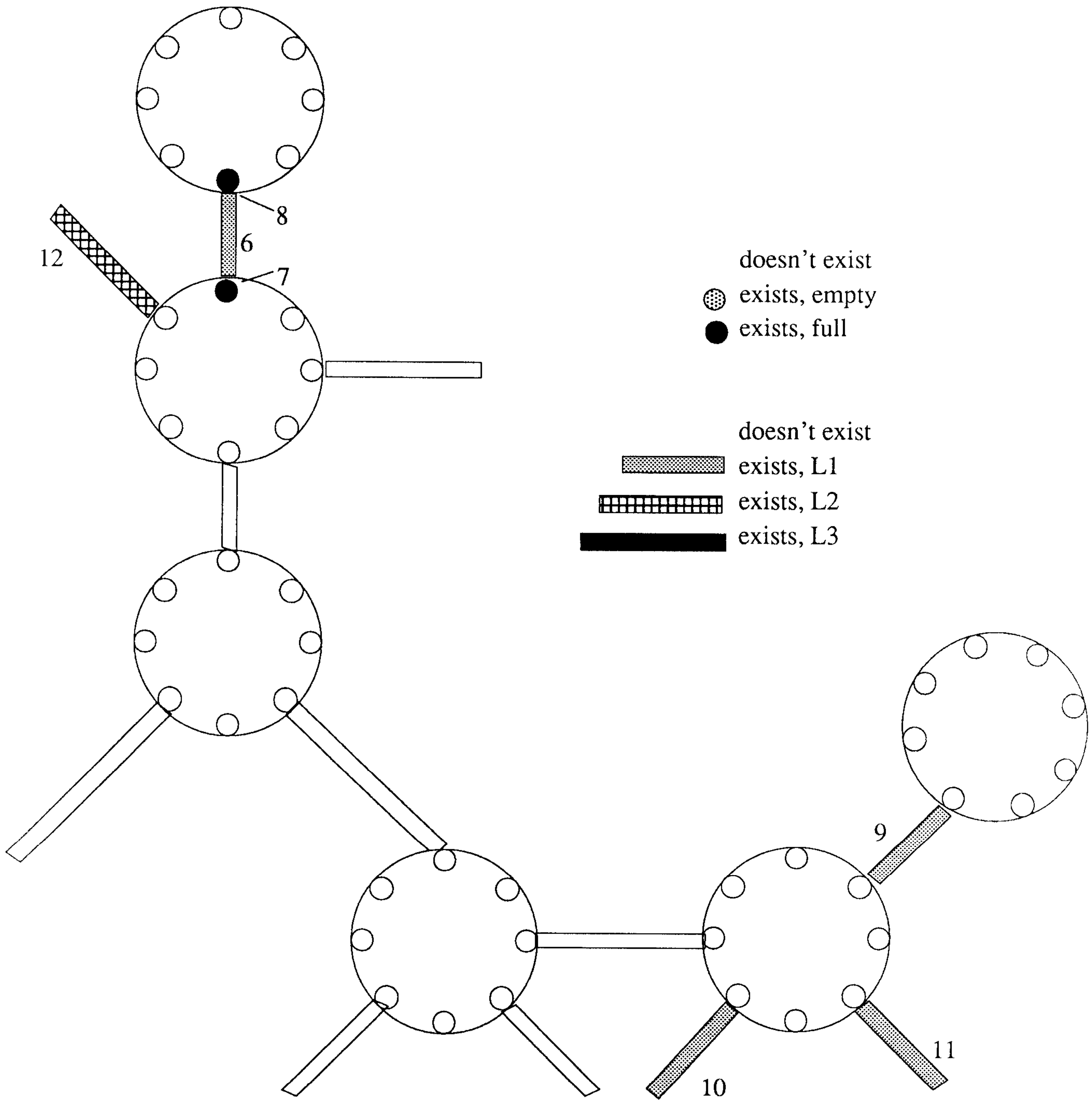

Figure 7: Segmentation Inference on MRF: First Few Committed States

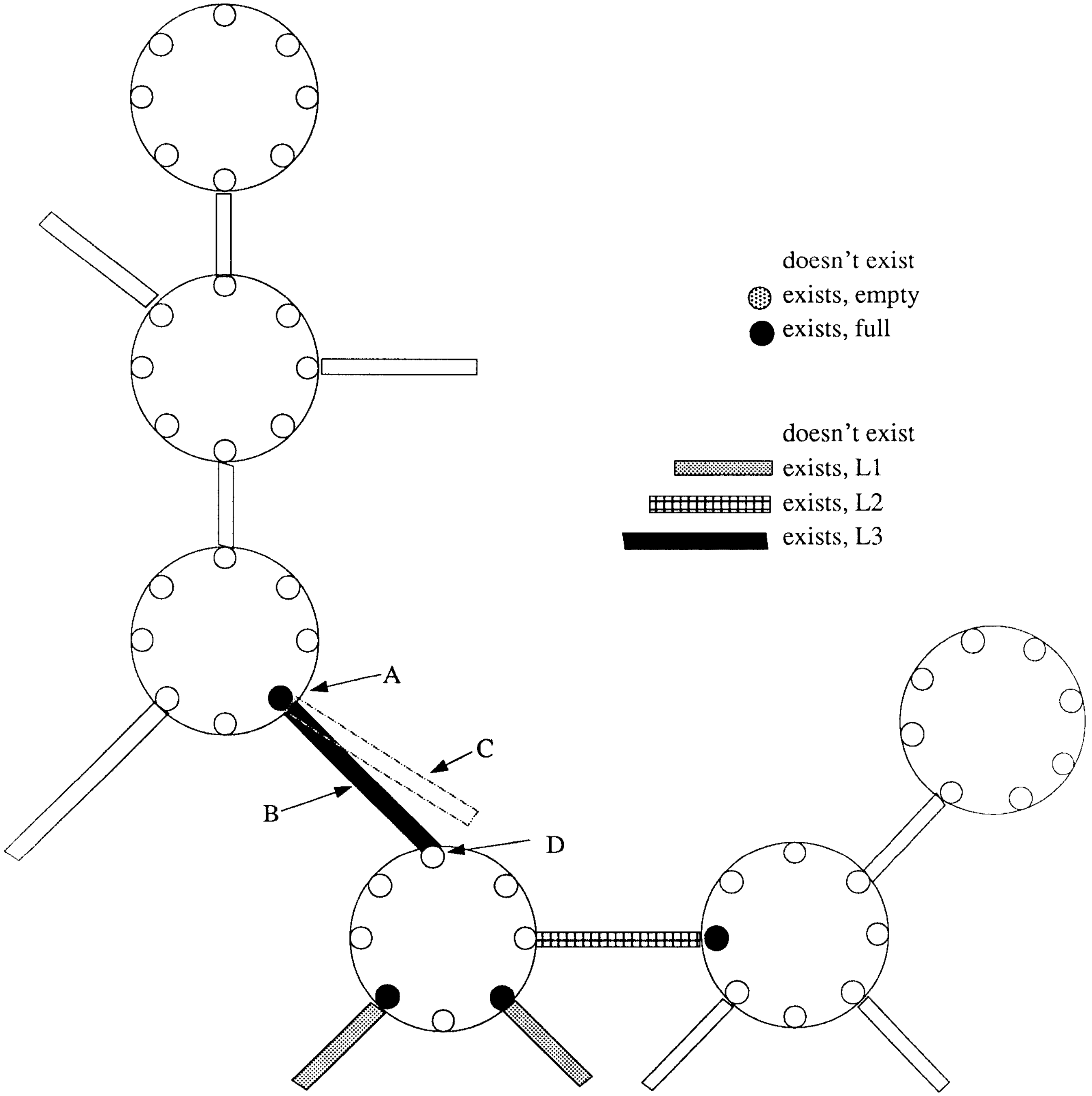

Figure 8: Segmentation Inference on MRF: Incorrect vrod hypothesis "B" is chosen over correct hypothesis "C", providing excitation energy for the "full" label at "D"

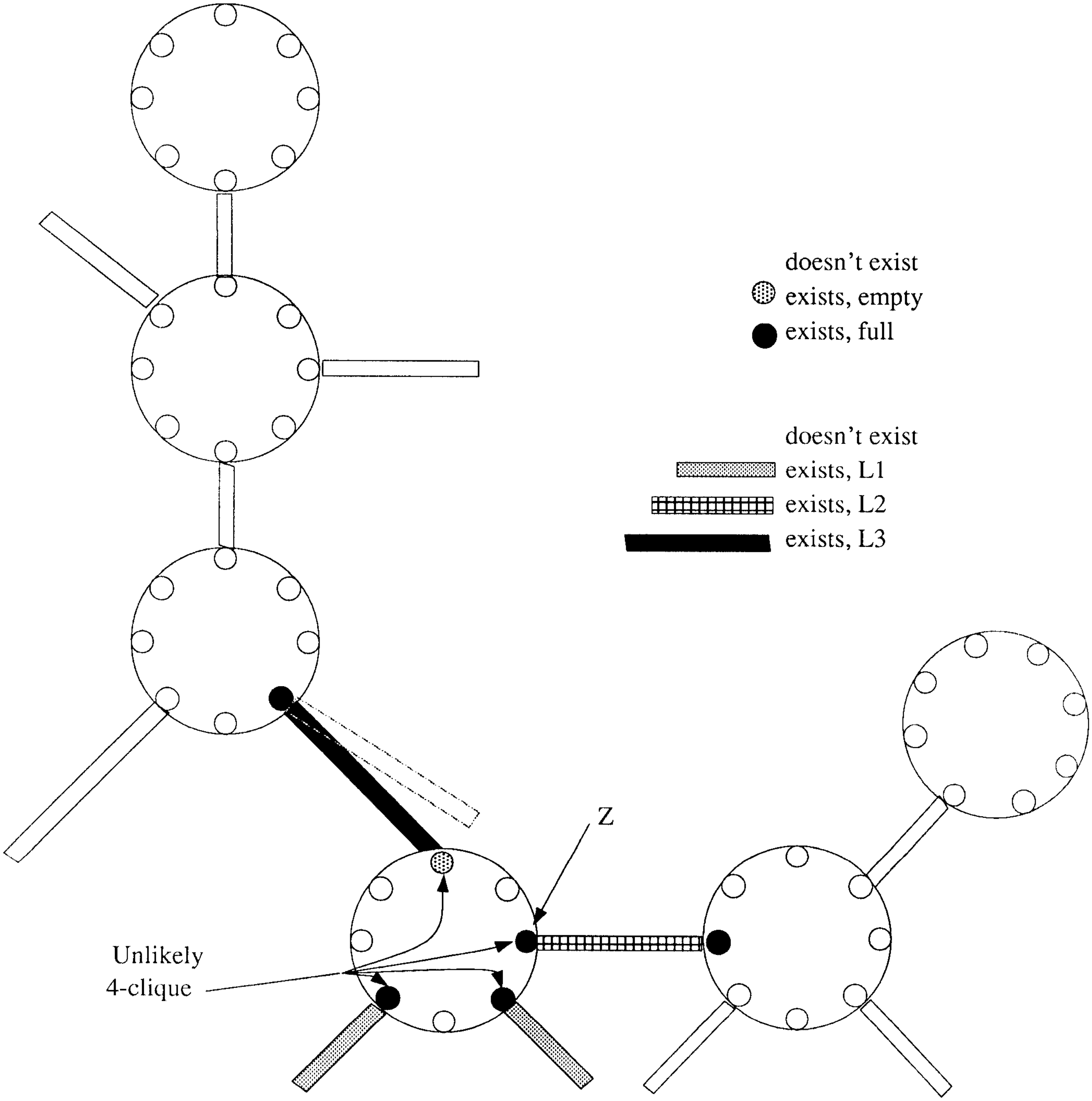

Figure 9: Segmentation Inference on MRF: "full" decision at "Z" makes inhibitory 4-clique relevant

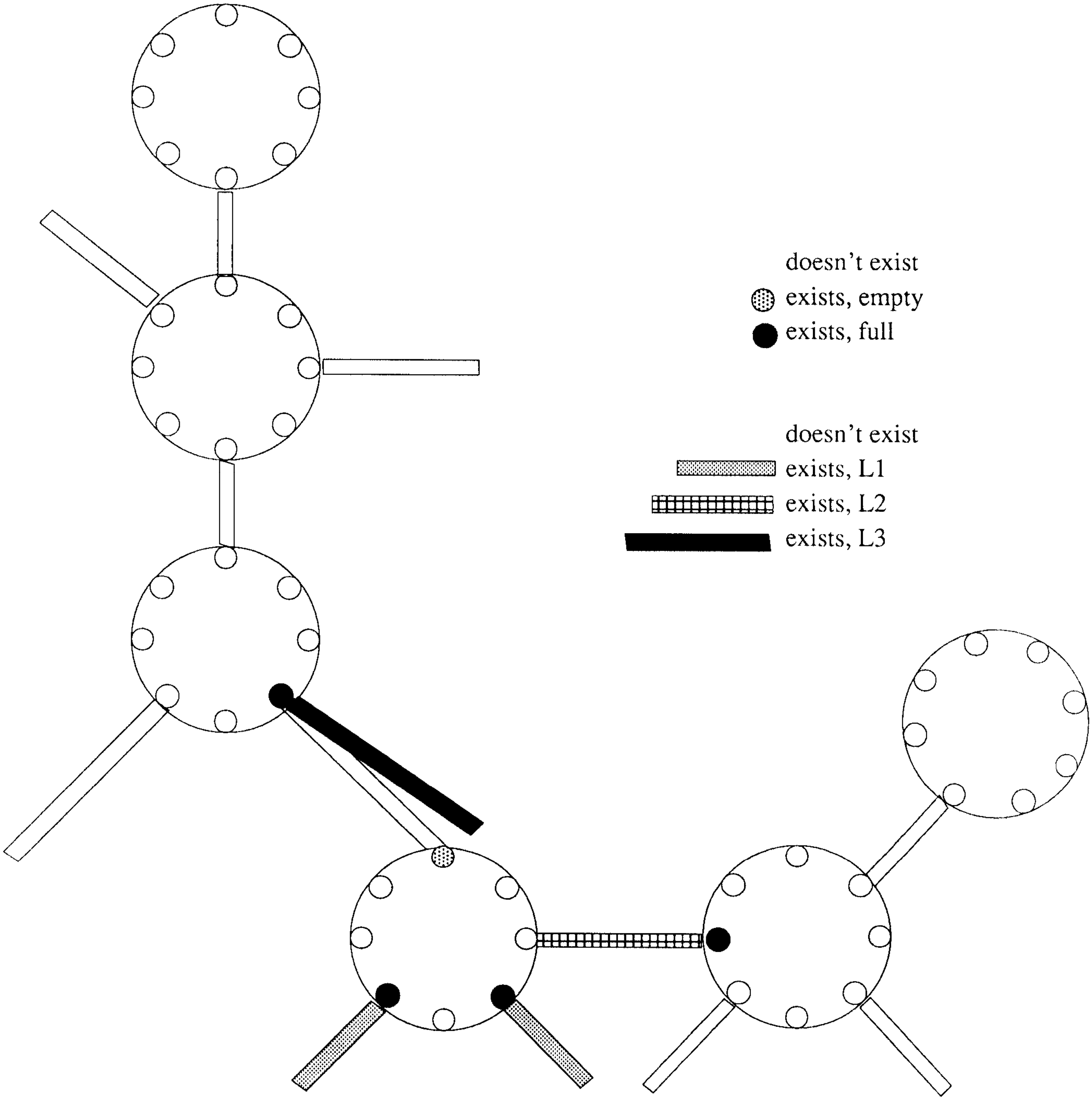

Figure 10: Segmentation Inference on MRF: inconsistent "empty" slot with connecting vrod causes change of decision

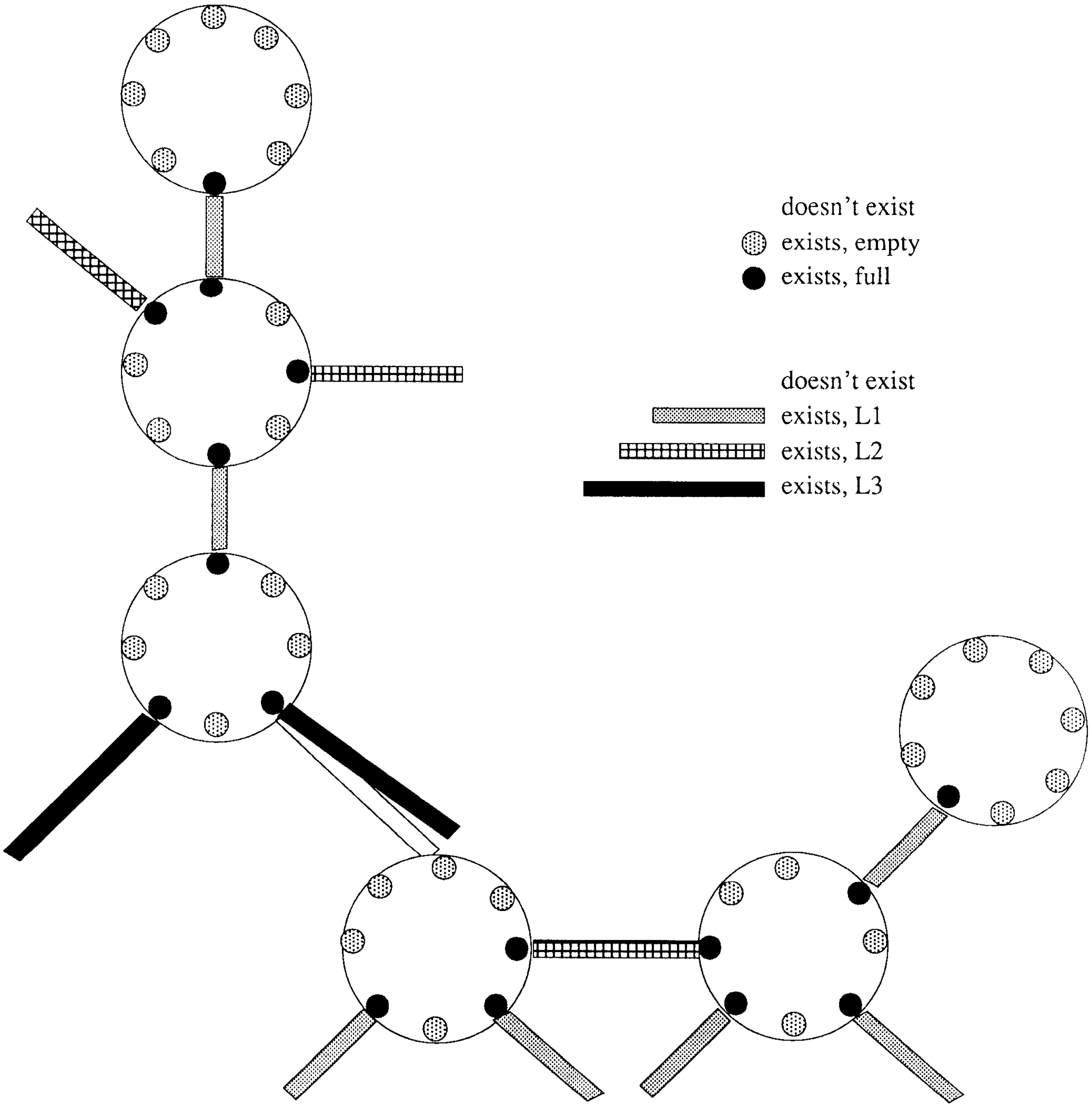

Figure 11: Final Segmentation

rod is labelled as existing, the associated unlabelled slot D gets energy encouraging the *full* label.

The slot D does not get labelled immediately. Instead, later, the slot designated Z in Figure 9 gets labelled as *full*. Once the slot Z is *full*, if slot D were full as well, this would commit 4 slots on that disk as *full*. But the set of priors used for the experiment states that this particular configuration of 4 rods at a disk is unlikely, as marked in Figure 9. This 4-clique prior inhibits the simultaneous labelling of all 4 of the slots as *full*. The inhibition energy is sufficient to commit the fourth slot, slot D, to the state *empty*, as shown in Figure 9. Of course, an empty slot is incompatible with the vrod B hypothesizing connection, so that vrod gets relabelled as *not existing*, and the alternative rod C is relabelled as *existing and L3*.

Other parts of the input evidence are comparatively ambiguous as well. The man's right leg, for example, has very uncertain evidence about the correct length. (An explanatory assumption for the evidence might be that that region of the image was noisy, so the likelihood generator had difficulty discriminating lengths.) This uncertainty, purely local, is easily resolved by the network.

Eventually, the whole scene is correctly labelled, as shown in Figure 11.

4.2 Structure Inference Only: Occlusion

A more representative experiment is now given. This experiment demonstrates the kind of uncertain information that could arise in a typical occlusion, and how it might be resolved. Compared to the extremely unlikely accidental alignment that was correctly interpreted in the last experiment, this occlusion is a much simpler problem for the system to handle. In contrast, most vision systems make no provision for occlusion at all.

The scene and the input evidence can be seen in Figure 12. Note the evidence about a variety of structural hypotheses surrounding the location of the occlusion. In order for there to be any possibility at all of a mistaken interpretation, it is once again necessary to contrive an accidental alignment of the occluded rod and two occluding slots. (Otherwise, the likelihood generators would not have ambiguous local information, and the correct global interpretation would follow simply from the evidence.)

Let's examine the evidence relevant to the occlusion. At a high-level of analysis, there are two main possible hypotheses: the true hypothesis (dog occluding giraffe) with a single occluded rod connecting the giraffe's head to its shoulders, and the mistaken hypothesis that the giraffe and dog are awkwardly connected. The latter hypothesis requires two short rod's connecting the giraffe's head and shoulders to the dog, respectively. Note that the local evidence about the vrods representing the hypotheses is ambiguous; they have exactly the same likelihoods. Consider also the evidence about the slots. On the dog's head, the two slots aligned with the occluding

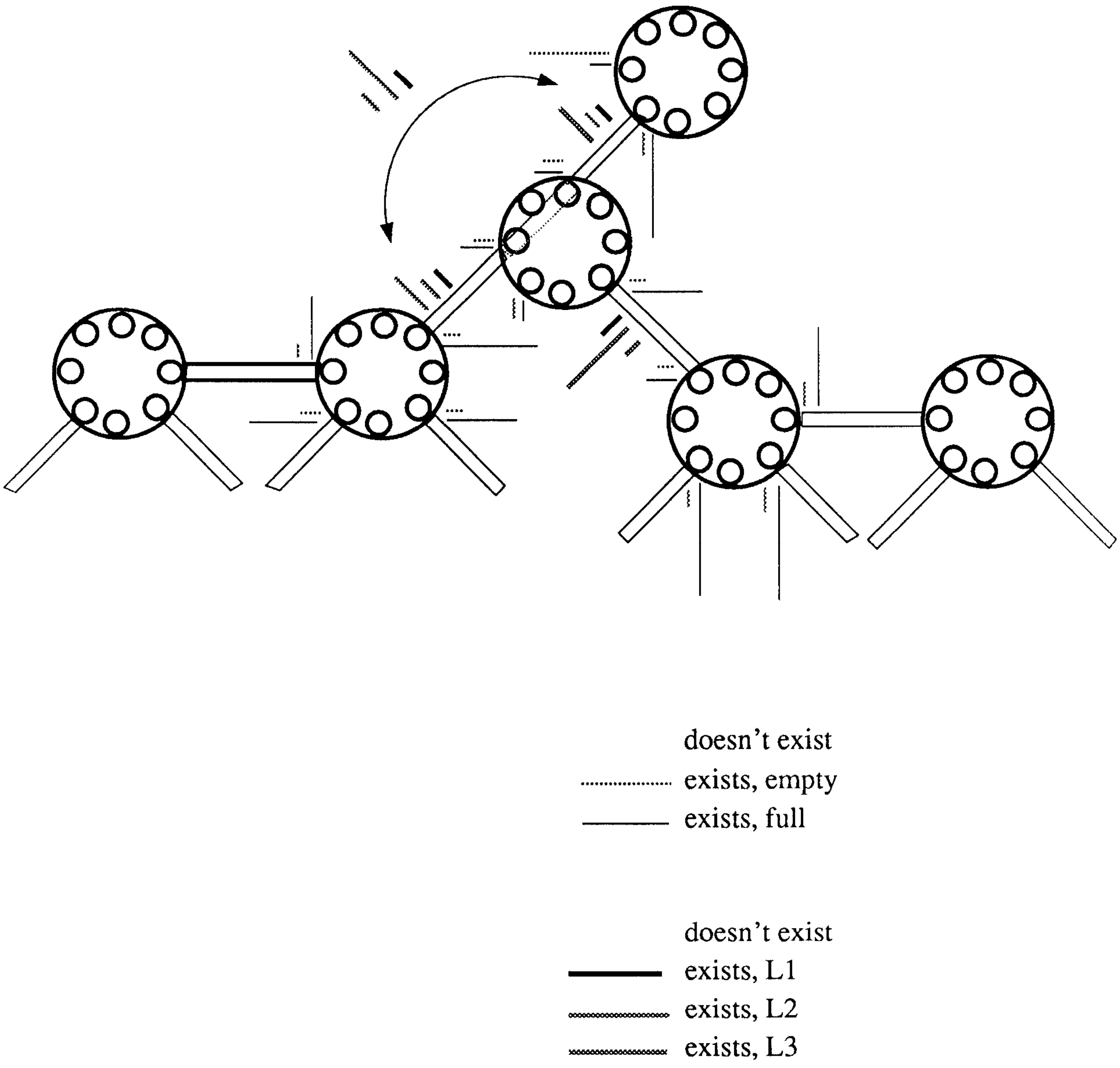

Figure 12: Dog Occluding Giraffe, Input Evidence: label likelihoods

rod show better evidence for *full* than *empty*. The local evidence is thus actually ranked in inverse order to the truth.

For this problem, making a segmentation decision based on the usual simple criteria will obviously not suffice. In particular, a threshold will yield the *wrong* answer, because of the incorrect ranking of the evidence at the slots.

As can be seen in Figure 13, the Tinkertoy MRF eventually achieves the right interpretation, correctly segmenting the two overlapping objects. In this problem, the statistically derived priors reflecting the frequency of junction-pattern occurrence play little role. Instead, constraints propagating to the competing rod hypotheses at the occlusion allow the correct interpretation. Consider the "giraffe neck" hypothesis versus its competitors. The evidence at the slot at both ends of the neck is strongly in favor of the slots being labelled *exists and full*. As a result, HCF commits those states early. At this point, the connecting "neck" virtual rod is receiving excitation energy from *both* slots. The competing (incorrect) hypotheses are each compatible at *one end only*. As a result, the "neck" hypothesis commits to existence, winning the vrod WTA competition, and forcing the competitors to turn off. In short, the correct global interpretation is more compatible with the evidence, and thus has a lower energy.

4.3 Coupled Recognition and Segmentation

The final experiment involves the same scene as the first experiment, but without a set of priors that assist in correctly interpreting the scene. Instead, the coupling of the recognition and segmentation processes will yield a correct interpretation.

The input evidence for the experiment is the same as previously. Now however, there are additional matching and coupling constraints, as suggested by the presence of the candidate model in Figure 14.

In Figure 15 through Figure 17 the progress of the inference process on the experiment is shown. Figure 15 shows some of the first label commitments that were made. The segmentation labelling decisions are again shown by the shading of the parts in the scene. The recognition or matching decisions are shown by the lines connecting the model parts to object parts. In the model there is only one rod of length $L2$. As a result, unique labellings can be easily found that map each rod in the image of length $L2$ to the dog's trunk axis.

In Figure 16 the inference procedure has once again incorrectly labelled the scene as one object, for the same reason. But local partwise matching decisions are being made in parallel as well. The correspondence labelling has matched the dog's shoulders correctly. Eventually (Figure 17) the correspondence of the dog's trunk axis propagates to match of the trunk-hip slot ('A' in the figure.) As a result of this, all the slots on the dog's hip disk become matched. At this point, because of the correct recognition, a conflict exists. *In the model*, the crucial slot ('B') is empty. But

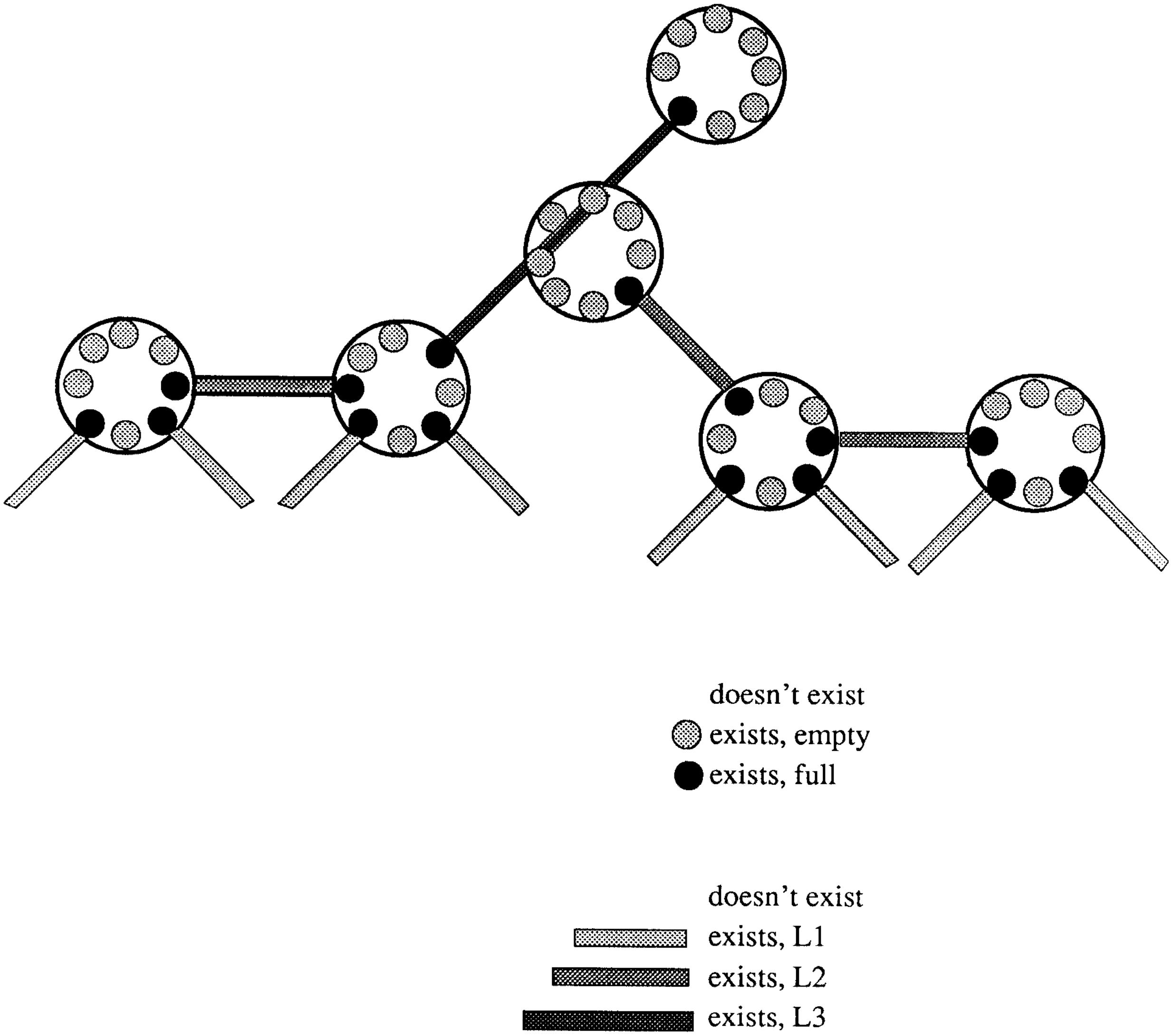

Figure 13: Dog and Giraffe: Final Segmentation Labelling

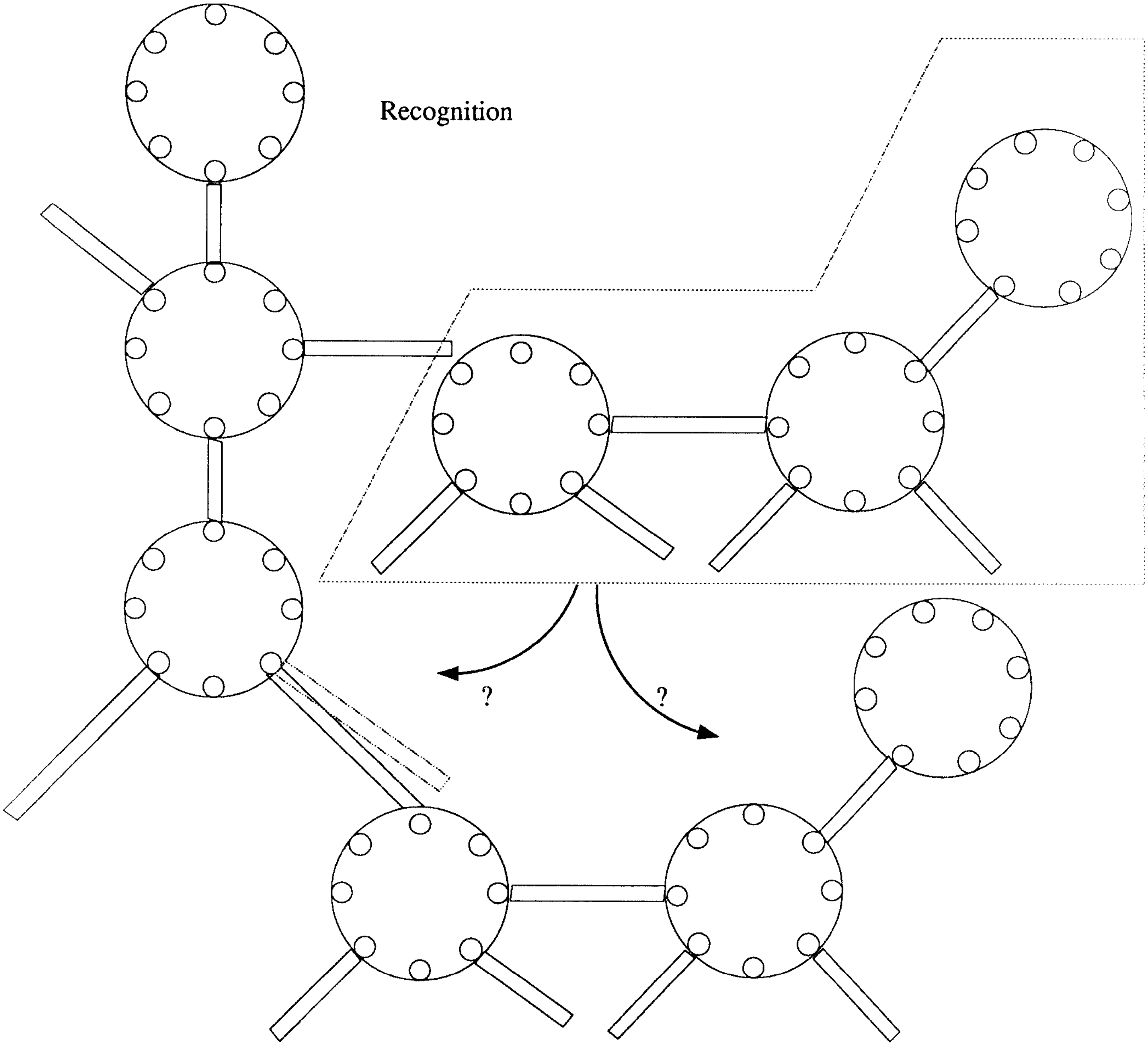

Figure 14: Recognition Problem: Can a dog be recognized from the image evidence?

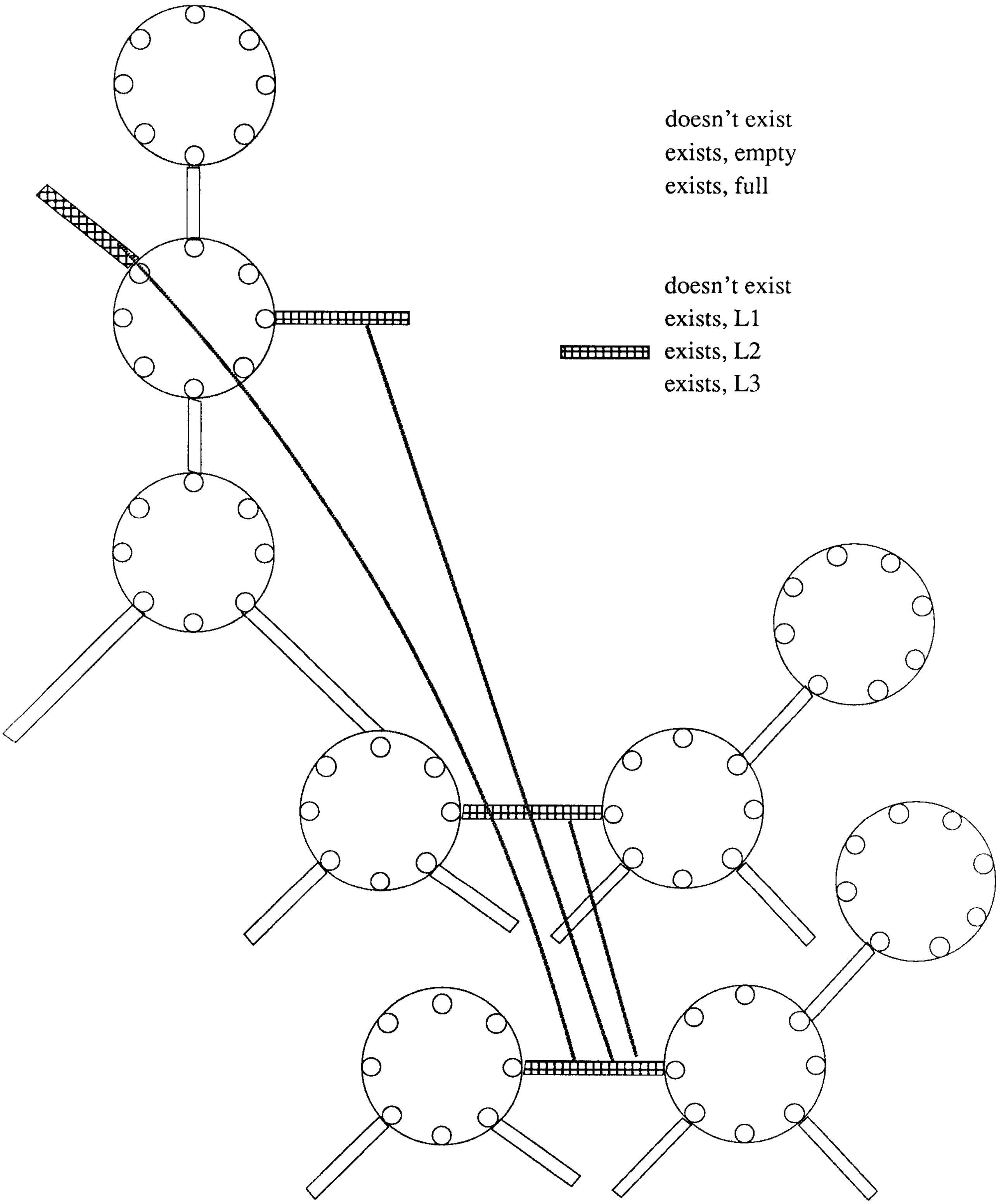

Figure 15: Coupled Segmentation and Recognition Experiment, 1: some initial segmentation and partwise recognition inferences

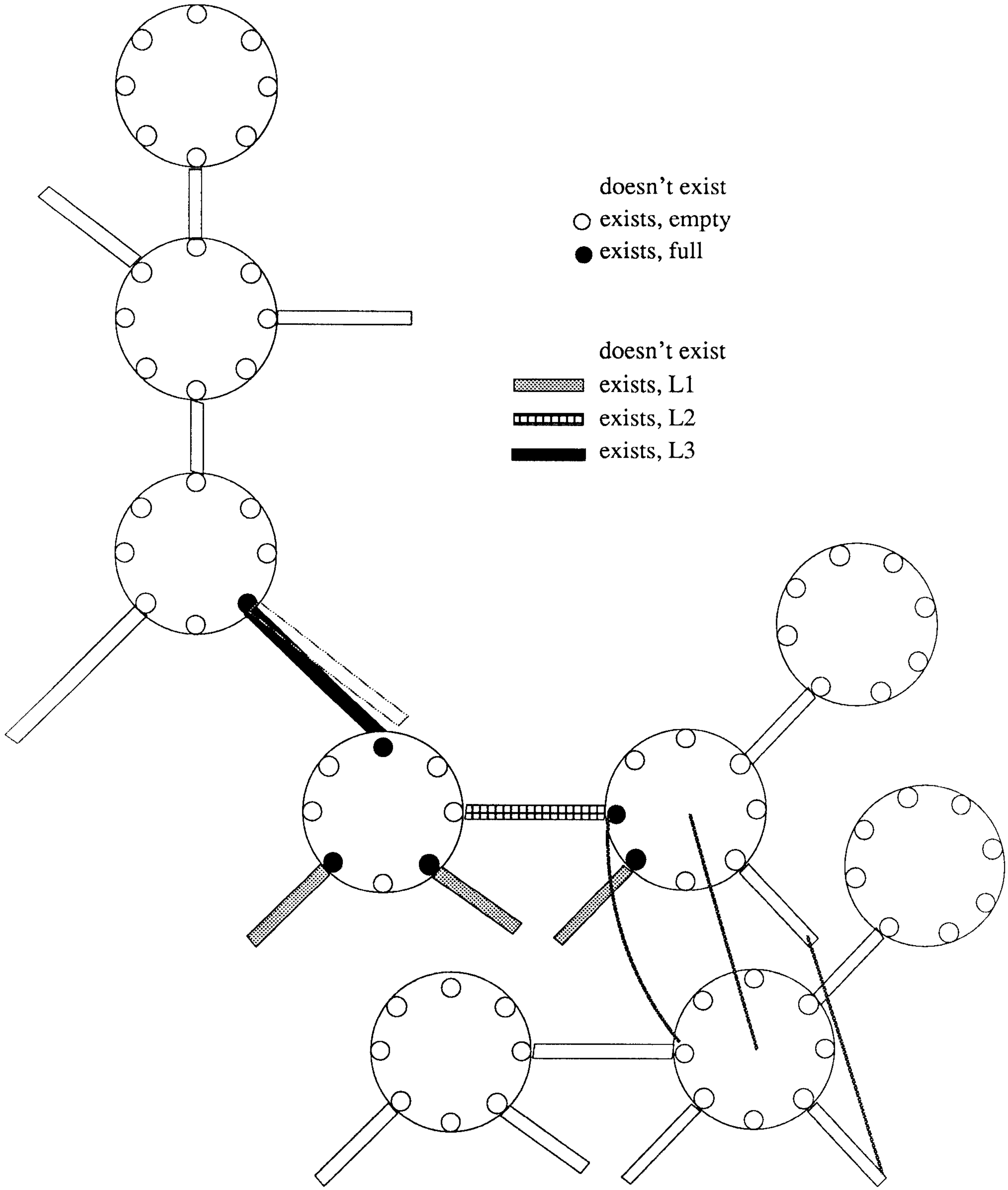

Figure 16: Coupled Experiment, 2: incorrect segmentation decision, and some further correspondence inferences

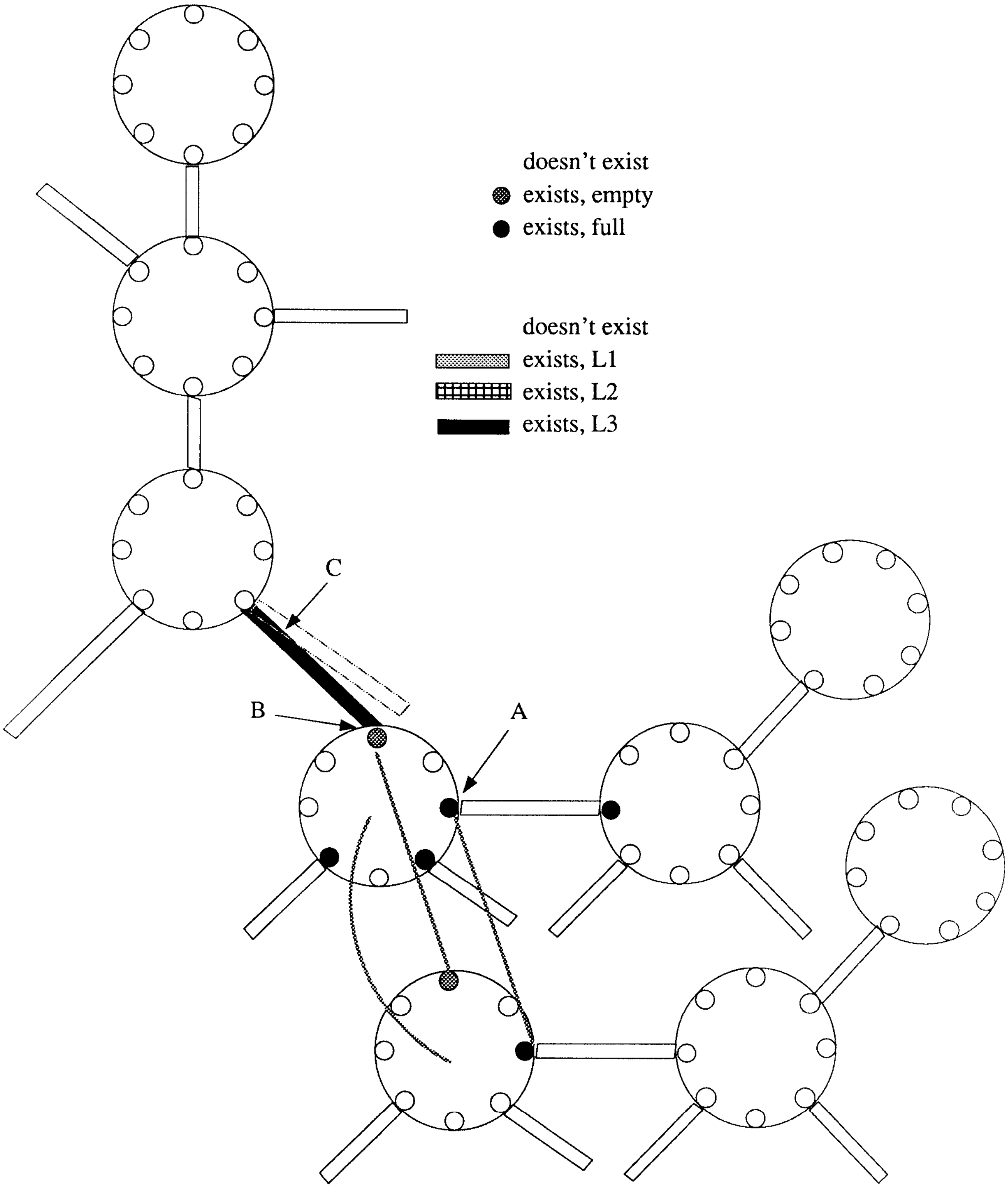

Figure 17: Coupled Experiment, 3: match has propagated to slot “A”, causing match of all slots at that disk, including “B”. The labelling at vrod “C” suggests slot “B” should be *full*, but the correspondence with the model dictates that “B” is labelled *empty*

the segmentation labelling of the connecting man's leg ('C') is reinforcing the *full* label at B. The coupling from the recognition field labels slot B as *empty*, and an inconsistency between the labelling of B and C exists. Ultimately, the segmentation labelling is reversed and the objects are correctly segmented.

Simultaneous, truly coupled segmentation and recognition from structure has been accomplished for a non-trivial scene. It is worth emphasizing that in this and other experiments the computation was remarkably insensitive to variations in the exact values of either the evidence or the prior clique weights. This reflects the fact that an appropriate selection of variables and labels defines an architecture with a well-behaved energy function. Other choices may not yield equally robust behavior.

4.4 Multiple Models

The network can also process more complex recognition problems, such as those with multiple models, and multiple instantiations of a single model. The parallel framework for partwise matching suggests this possibility, but the coupled interaction of the recognition and segmentation processes pose difficulties. Simulation experiments have shown that in many cases, the network does successfully process multiple models in parallel [Konopka, 1990]. (These results will be reported more completely in the future.)

Multiple instantiations of a single model can be recognized in parallel. In one experiment, multiple instantiations of a "dog" model were recognized, even though the scene required segmenting ambiguous evidence into the two objects. The variable/label representation is designed so that each object part in the scene is assigned a label corresponding to a model part. More than one variable may have the same label, so more that one instance of a model part may occur in a scene.

Different models can also be matched in parallel to the object parts in the scene. For example, models of both the "dog" and the "man" can be matched simultaneously against the scene of section 4.3. For multiple different models to successfully match, a recognizable feature must exist on each to differentiate it from the others. Furthermore, the structure inference must infer the feature correctly prior to the recognition of the object with the feature.

5 Conclusion

This paper has described a coupled network that solves the recognition problem from uncertain information by inferring the solution to both the segmentation problem and the matching problem simultaneously. Within a probabilistic network framework, both the evidence and relevant prior constraints can interact to yield good global answers to both problems, even when either problem on its own is underdetermined, and even when the local evidence is ambiguous or favors the incorrect interpretation. Visual inference decisions can be computed that would be very difficult to achieve successfully with traditional vision system architectures.

The work also demonstrates a novel application of Markov Random Fields in a non-homogeneous, non-isotropic, high-level application. MRFs are convenient for the representation of labelling problems, and are particularly convenient for the expression of the arbitrary spatial relationships that arise in the representation of spatially complex objects. The coupling of two MRFs, each one addressing a different inference problem, was extremely important to achieving a solution. The basic coupled framework should also be extendible for richer scene domains, so scene parameters other than structure can be simultaneously computed. Finally, the role of clique potentials has been viewed as the representation of constraints, both qualitative "hard" *a priori* truths and "soft" frequency-related constraints that should be learnable.

References

[Ballard, 1984] D.H. Ballard, "Parameter Networks," *Artificial Intelligence*, 2(1):235–267, 1984.

[Barlow, 1972] H. B. Barlow, "Single units and sensation: A neuron doctrine for perceptual psychology?," *Perception*, 1:371–392, 1972.

[Biederman, 1985] I. Biederman, "Human image understanding: recent research and a theory," *Computer Vision, Graphics and Image Processing*, 32(1):29–73, 1985.

[Binford *et al.*, 1987] Thomas O. Binford, Tod S. Levitt, and Wallace B. Mann, "Bayesian Inference in Model-Based Machine Vision," In *Proceedings, 3rd Workshop on Uncertainty in AI*, pages 86–97, 1987.

[Blake and Zisserman, 1987] A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press, 1987.

[Bolles, 1977] R.C. Bolles, "Verification Vision for Programmable Assembly," In *Proceedings: IJCAI-77*, pages 569–575, 1977.

[Brooks, 1986] R. Brooks, "Model Based 3-D Interpretation of 2-D images," In Alex P. Pentland, editor, *From Pixels to Predicates*, pages 299–321. Ablex Publishing Corporation, 1986.

[Chou, 1988] Paul B. Chou, "The Theory and Practice of Bayesian Image Labeling," Technical Report TR 258, Department of Computer Science, University of Rochester, August 1988.

[Cooper, 1988] Paul R. Cooper, "Structure Recognition by Connectionist Relaxation: Formal Analysis," In *Proceedings: Conference of the Canadian Society for Computational Studies of Intelligence, CSCSI-88*, Edmonton, Alberta, June 1988.

[Cooper, 1989] Paul R. Cooper, "Parallel Object Recognition from Structure (The Tinkertoy Project)," Technical Report 301 (Ph.D. Thesis), Dept. of Computer Science, University of Rochester, July 1989.

[Cooper and Hollbach, 1987] Paul R. Cooper and Susan C. Hollbach, "Parallel Recognition of Objects Comprised of Pure Structure," In *Proceedings of the DARPA Image Understanding Workshop*, pages 381–391, February 1987.

[Cooper and Swain, 1988] Paul R. Cooper and Michael J. Swain, "Parallelism and Domain Dependence in Constraint Satisfaction," Technical Report TR 255, Department of Computer Science, University of Rochester, December 1988, Submitted for publication.

[Cooper and Swain, 1989] Paul R. Cooper and Michael J. Swain, "Domain Dependence in Parallel Constraint Satisfaction," In *Proceedings IJCAI-89: International Joint Conference on Artificial Intelligence*, August 1989.

[Cross and Jain, 1983] G.R. Cross and A.K. Jain, "Markov Random Field Texture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5:25–39, 1983.

[Eshera and Fu, 1986] M. A. Eshera and King-Sun Fu, "An Image Understanding System Using Attributed Symbolic Representation and Inexact Graph-Matching," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, PAMI-8(5), 1986.

[Feldman and Ballard, 1982] J. A. Feldman and D. H. Ballard, "Connectionist Models and Their Properties," *Cognitive Science*, 6:205–254, 1982.

[Feldman and Yakimovsky, 1974] J. A. Feldman and Yoram Yakimovsky, "Decision Theory and Artificial Intelligence: I. A Semantics-Based Region Analyzer," *Artificial Intelligence*, 5:349–371, 1974.

[Geman and Geman, 1984] Stuart Geman and Donald Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *PAMI*, 6(6):721–741, November 1984.

[Goddard, 1988] N.H. Goddard, "Representing and Recognizing Event Sequences," In *Proceedings 1988 Connectionist Summer School*, 1988.

[Goddard *et al.*, 1988] Nigel H. Goddard, Kenton J. Lynne, and Toby Mintz, "Rochester Connectionist Simulator," Technical Report TR233, University of Rochester, March 1988.

[Hinton, 1977] Geoffrey E. Hinton, *Relaxation and Its Role in Vision*, PhD thesis, University of Edinburgh, 1977.

[Hoffman and Richards, 1986] D. Hoffman and W. Richards, "Parts of Recognition," In Alex P. Pentland, editor, *From Pixels to Predicates*, pages 268–294. Ablex Publishing Corporation, 1986.

[Hummel and Zucker, 1983] Robert A. Hummel and Steven W. Zucker, "On the Foundations of Relaxation Labeling Processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5:267–287, 1983.

[Kindermann and Snell, 1980] Ross Kindermann and J. Laurie Snell, *Markov Random Fields and their Applications*, American Mathematical Society, 1980.

[Kirkpatrick *et al.*, 1983] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, "Optimization by Simulated Annealing," *Science*, 220:671–680, 1983.

[Kitchen and Rosenfeld, 1979] Les Kitchen and Azriel Rosenfeld, "Discrete Relaxation for Matching Relational Structures," *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-9:869–874, 1979.

[Konopka, 1990] R. Konopka, "Parallel Recognition of Multiple Objects from Structure," M.S. project report, 1990.

[Lowe, 1985] David Lowe, *Perceptual Organization and Visual Recognition*, Kluwer Academic Publishers, 1985.

[Mackworth, 1977] Alan K. Mackworth, "Consistency in Networks of Relations," *Artificial Intelligence*, 8:99–118, 1977.

[Marroquin, 1985] Jose Luis Marroquin, "Probabilistic Solution of Inverse Problems," Technical report, MIT Artificial Intelligence Laboratory, September, 1985.

[Mjolsness *et al.*, 1988] Eric Mjolsness, Gene Gindi, and P. Anandan, "Optimization in Model Matching and Perceptual Organization: A First Look," Technical Report YALEU/DCS/RR-634, Department of Computer Science, Yale University, June 1988.

[Pearl, 1988] Judea Pearl, *Probabalistic Reasoning in Intelligent Systems*, Morgan Kaufman, 1988.

[Pentland, 1987] A. Pentland, "Recognition by Parts," In *Proceedings, ICCV87: First International Conference on Computer Vision*, June 1987.

[Pentland, 86] Alex P. Pentland, "Parts: Structured Descriptions of Shape," In *Proceedings AAAI-86, American Association for Artificial Intelligence*, pages 695–701, 86.

[Poggio *et al.*, 1985] Tomaso Poggio, Vincent Torre, and Christof Koch, "Computational Vision and Regularization Theory," *Nature*, 317:314–319, September 26, 1985.

[Shapiro and Haralick, 1981] Linda G. Shapiro and Robert M. Haralick, "Structural Descriptions and Inexact Matching," *IEEE-PAMI*, 3(5), 1981.

[Sher, 1987] David B. Sher, "A Probabilistic Approach to Low-Level Vision," Technical Report 232, Department of Computer Science, University of Rochester, October 1987.

[Swain, 1989] M. J. Swain, "Estimating MRF Clique Parameters from Frequency Measurements," unpublished, 1989.

[Swain and Cooper, 1988] Michael J. Swain and Paul R. Cooper, "Parallel Hardware for Constraint Satisfaction," In *Proceedings AAAI-88, the American Association for Artificial Intelligence Conference*, St. Paul, Minn., August 1988.

[Utans *et al.*, 1989] Joachim Utans, Gene Gindi, Eric Mjolsness, and P. Anandan, "Neural Networks for Object Recognition within Compositional Hierarchies: Initial Experiments," Technical Report 8903, Yale University, February 1989.

[Witkin and Tenenbaum, 1983] Andrew P. Witkin and Jay M. Tenenbaum, "On the Role of Structure in Vision," In Jacob Beck, editor, *Human and Machine Vision*, pages 481–543. Academic Press, 1983.