BMVC
2002

A Qualitative, Multi-scale Grammar For Image Description and Analysis

Derek Magee
School of Computing, University of Leeds, UK
drm@comp.leeds.ac.uk

Abstract

A qualitative image description grammar with automatic image fitting and object modelling algorithms is presented. The grammar is based on assigning a square sub-region of an image one of a finite number of qualitative labels, based on the occurrence of object boundaries within this region and how these intersect the region boundary. In the general case there is an infinite number of such labels, however the use of a multi-scale approach allows a finite (small) number of labels at each scale. This makes the problem tractable within a constraint satisfaction type framework. Constraints are put on neighboring labels based on the premise that all object boundaries are continuous, having no ending within an image. A minimum description length (MDL) approach is suggested for description hypothesis selection (based on colour histograms) and methods for (constraint based) hypothesis generation/adaption and (Hidden Markov Model based) a-priori shape modelling are presented.

1 Introduction

The problem of image and image sequence interpretation is often divided into two tasks; i) extracting visual information, and ii) reasoning about the scene. Extracting visual information is often performed using a model based approach. Models are, in general, quantitative descriptions of object or scene characteristics. These descriptions are often encapsulated as statistical distributions (e.g. the point distribution model (PDM) [14]). Increasingly within the spatial reasoning community qualitative representations are being used (see [3] for an overview) for their simplicity (and thus power) when used for formal reasoning. When dealing with real scenes these qualitative descriptions are formed from quantitative descriptions produced by an (often imperfect) low level system. This can lead to errors in the qualitative description, for example; if two close objects are being tracked using PDM's the qualitative property 'disjoint' formed from this may actually be more related to the noise of the tracking process than the actual relation.

In this paper a qualitative grammar for representing visual information (in terms of object boundaries) is presented. Qualitative Relations such as joint/disjoint etc. may be directly extracted from this description. Qualitative descriptions can, to a certain extent, encapsulate some of the observational uncertainty often modelled using statistical techniques in machine vision systems. However, we use statistical techniques alongside

BMVC 2002 doi:10.5244/C.16.27

qualitative representations in order to fully deal with uncertainty. This can be a powerful combination, as demonstrated by Fernyhough *et al.* [6].

A minimum description length approach [12] is taken to hypothesis selection. Potential hypotheses are generated based on a set of constraints for valid hypotheses using a guided constraint satisfaction programming (tree search) technique. This is applied at multiple scales from course to fine, with constraints being propagated from one level to the next. Tree search constraint satisfaction methods are an alternative to stochastic search procedures (such as particle filters [8]) in many machine vision applications. Cucchiara *et al.* [4] use such a search technique for model based image analysis, and even have a qualitative element to their model. Further back Barrow and Popplestone [2] use a similar technique to match 'image description graphs' and 'object model graphs' for object recognition (a theme taken up by several other researchers since). Waltz (described in [1]) applies constraint satisfaction to line labeling problems.

A priori shape information may be encapsulated into the grammar by the use of hidden Markov models that model a (variable length) sequence of edge descriptors in a statistical manner.

2 A Qualitative Grammar for Image Description

The purpose of the grammar presented is to describe an image as a set of logically consistent object boundaries. To achieve this the image is divided up into square regions. The aim of the grammar is to provide a qualitative description of each region based on the object boundaries and their intersections with the square region boundaries. Given an arbitrarily complex image there is an infinite number of such descriptive labels. If the constraint that only a single object boundary may intersect a side of a particular square region boundary is introduced this number becomes finite as illustrated in figure 1. It should be

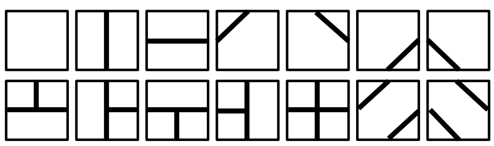

Figure 1: A Qualitative Grammar

noted that these labels are qualitative, based on which square region edges that the object boundaries intersect. Figure 2 illustrates several examples which may be described by the second state.

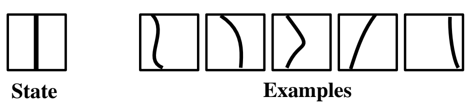

Figure 2: Quantitative Examples Described by the Second State

Objects where multiple object boundaries intersect a side of a particular square region boundary cannot be represented **at that resolution**, however extending the grammar to multiple scales allows a compact description of scenes with arbitrary topology. This is achieved by repeatedly dividing square regions into four equal squares. The resolution required to represent particular objects within a scene is application specific and, for many high level scene/object analysis tasks, may be reasonably low. Figure 3 illustrates a scene and it's corresponding multi-scale description.

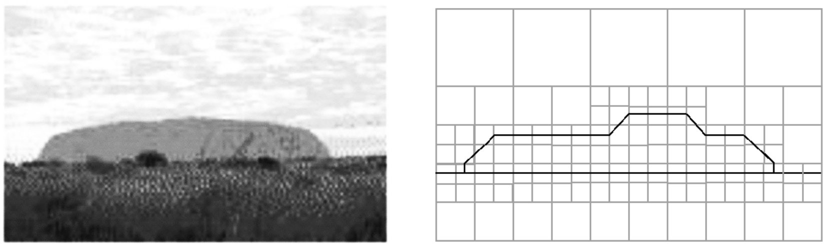

Figure 3: Describing a View of Ayres Rock

The observant reader will notice there are no possible labels in which an object boundary exactly intersects a corner of a square region. This is for two reasons; The first is a philosophical objection. When describing a real life scene/object there is always uncertainty (based on observation noise, quantisation errors etc.) and object boundary configurations cannot be exactly determined (hence the need for qualitative and statistical methods). States where object boundaries pass exactly through a square region corner effectively define an exact location for a point on the boundary. This is against the philosophy of the grammar. The second reason is more practical, based on the fact that more states complicates hypothesis selection and makes this less computationally efficient. These extra states are not necessary as existing states describe lines which are infinitely close to these corners.

2.1 Rule based constraints on 'Valid' Descriptions

As stated previously valid descriptions contain no broken object boundaries (i.e. object boundaries do not terminate within an image, although they may do at the image edge). This constraint is enforced as a set of rules for adjacent square region descriptions. Figure 4.a illustrates a valid pair of neighbors, whereas Figure 4.b illustrates an invalid pair.

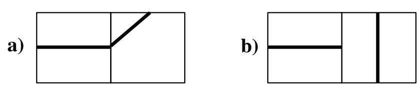

Figure 4: a) Valid pair of adjacent region labels, b) Invalid pair of adjacent region labels

Rules for invalid/valid neighbors are given in Appendix A. In addition to these rules about boundaries objects must be taken into consideration, for example figure 5 shows a description that is valid based on the object boundary rules but invalid as a description of an image. The description in figure 5 is invalid because two sub-regions of the top right square region relate to the same object, but are divided by an object boundary. We define a set of region equivalence rules to detect such occurrences by labeling sub-regions with object labels. These rules define which sub-region pairs must relate to the same object, which sub-region pairs cannot relate to the same object and which sub-region pairs for which no constraints exist. These rules are described in Appendix B. Section 4 gives details of a scheme to apply these rules within a hypothesis generation scheme.

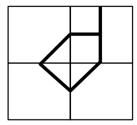

Figure 5: Invalid Description which does not break label adjacency rules

3 MDL Hypothesis Selection Using Colour Histograms

The purpose of the grammar presented in the previous section is the analysis of real scenes and objects via digitised images (or image streams). This requires some way of ranking how well a particular description describes a given scene or object. The Minimum Description Length (MDL) [12] approach is taken based on colour histograms constructed from pixels in the square regions (or in sub-regions of these). This approach is based on premise that the best description of a system is the the one that can be coded with the fewest bits. The total information content in a representation may be written as:

$$I_{Total} = I_{Model} + I_{Params} + I_{Residual} \tag{1}$$

Where I_{Model} is the information required to code the model, I_{Params} the information required to code the model parameters which describe the particular instance of the model and $I_{Residual}$ is the information required to code the residual (the difference between the model and the reality). Shannon [13] states that, given an optimal coding scheme, the number of bits required to code a probabilistic model/system is given as the entropy of the system:

$$Entropy = \sum p(n) \log_2 p(n) \tag{2}$$

Entropy is a convenient measure of information as it is simple to compute and independent of coding method. For these reasons this forms the basis of our model selection criteria. This criteria is based on the average information required to code a pixel given the grammar, a description based on this grammar and a set of colour histograms (one for each sub-region). It should be noted that we have not found a way of directly evaluating the qualitative description and, as such, it must be converted into a quantitative description to build the colour histograms for evaluation. Figure 6 gives the quantitative sub-divisions used for each qualitative state.

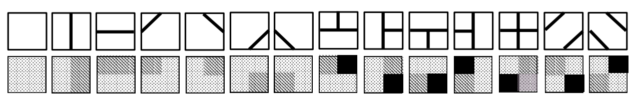

Figure 6: Quantitative Region Sub-division Used in Hypothesis Selection

The qualitative subdivisions given in figure 6 were chosen as they are easy and efficient to work with. Ideally an ‘optimal’ qualitative subdivision would be defined, based on image information / MDL criteria, however this would be computationally expensive and the current approach works reasonably well.

When calculating per-pixel ‘mean information’ it is necessary to consider the alternative models (sub-region colour histograms) for each region and the qualitative region label given by the grammar. The model information (I_{Model} in equation 1) may be divided into two parts i) the information required to code the model(s), and ii) the information required to code which model is selected. We ignore i) as this is a per-region rather than per-pixel information and is as such negligible in comparison with per-pixel information.

The second part is easily coded using the entropy formula in equation 2. The probabilities ($p(n)$) in this equation are calculated in proportion to the relative size (in pixels) of the sub-regions e.g. For the first state (no region boundary) the model information is zero ($1 \times \log_2 1 = 0$), and for the second it is 1 ($0.5 \times \log_2 0.5 + 0.5 \times \log_2 0.5 = 1$).

The parameter information (I_{Param} in equation 1) is based on the colour histogram distribution ($p_h(n)$) and the relative size of the sub-regions ($p_s(m)$) represented as:

$$\sum_{m=1}^{M} p_s(m) \sum_{n=1}^{N} p_h(n) \log_2 p_h(n) \tag{3}$$

Where N is the number of bins in the histograms and M is the number of sub-regions. The residual information ($I_{Residual}$ in equation 1) is divided into two parts; i) Quantisation error (for correct colour histogram bin classifications), and ii) mis-classification error. Quantisation error may be estimated by estimating the distribution of colour values represented by a particular quantisation. Davies *et al.* [5] approximate this with a Gaussian distribution, although in our case a limited range uniform distribution may be more appropriate. In practice we ignore this information as, for histograms with a fixed number of bins, it is constant. The second part of $I_{Residual}$, histogram bin mis-classification error, is zero in a perfect coding system. However, this residual is a useful measure of histogram compactness and, as such, we assume an imperfect coding scheme that codes the histogram parameters (i.e. which bin is assigned to a given pixel) stochastically from the corresponding region histogram distribution. (This would be the case for a coding scheme with access to the region histograms only and no access to the underlying pixel values.) As such the residual information is calculated by forming histograms of errors in the three colour dimensions (red, green, blue) as:

$$p_r(e) = \sum_{r(n)-r(m)=e} p_h(n) \times p_h(m) \tag{4}$$

$$p_g(e) = \sum_{g(n)-g(m)=e} p_h(n) \times p_h(m) \tag{5}$$

$$p_b(e) = \sum_{b(n)-b(m)=e} p_h(n) \times p_h(m) \tag{6}$$

Where $r(n)$, $g(n)$ and $b(n)$ are the mean (center) values of the red, blue and green components for colour histogram bin n. Residual information ($I_{Residual}$) is calculated as the sum of the entropies of these three error histograms using equation 2.

4 Hypothesis Generation and Adaptation

The simplest approach to hypothesis generation is to generate random label sets or search every combination of labels and to apply the rules described in section 2.1 to determine validity. These approaches are far from computationally efficient. Our approach is to only generate hypotheses that are valid. We take a constraint satisfaction programming (CSP) approach to hypothesis generation [15], in particular we use a tree search with value ordering and dynamic variable ordering. In plain English this means we build a tree of partial hypotheses allocating a label to a particular square region at each level of the tree. Figure 7 illustrates this for a simple image with two regions and two possible labellings (A+B).

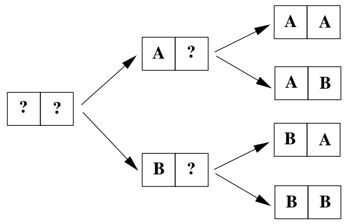

Figure 7: Hypothesis Generation By Tree Building

If at any point during the tree building a partial hypothesis becomes invalid with respect to the rules described in section 2.1 branching from this 'node' ceases. The order in which the nodes are branched is based on the relative description length of the partial hypothesis (see section 3) with respect to the zero hypothesis (in every square region has no boundary intersecting it) as:

$$I_{Rel} = \sum_{Label(x,y) \neq ?} I_{Total}(x,y) - I_z(x,y) \tag{7}$$

The node with the lowest relative information I_{Rel} is selected as the node to branch next (the 'value ordering' approach). The region to be assigned a value at the branch is selected as the region with the fewest remaining valid alternatives for that node (the 'dynamic variable ordering' approach). The actual hypotheses are assigned in arbitrary order. In some circumstances it is possible to generate a complete set of hypotheses in a short timescale, however this is not always possible. The use of the value ordering approach allows the search to be terminated after a fixed number of complete hypotheses have been generated (or a fixed time period has elapsed, or minimum information is reached). The subset of valid hypotheses produced are generally the lowest in terms of relative information and a substantial reduction in computational cost may be achieved using this approach.

4.1 The Use of Attention Mechanisms

To reduce the size of the search problem Attention Mechanisms may be used. In dynamic scenes motion may be used as a cue, however static images (which is our initial focus) require an alternative. Our solution is to examine absolute entropy. Regions where information (or some other measure such as motion) is higher than a threshold are considered as potentially containing edges. All other regions are set to the zero state (no edge boundaries intersecting). In the analysis of dynamic scenes it is possible to propagate states from the previous time-step based on motion information (i.e. if there is no motion the state is unchanged).

4.2 Extending the Search to Multiple Resolutions

At low resolution (e.g. blocks of 32×32 pixels in a 256x256 image) it is possible to compute the complete set of valid hypotheses in a short timescale (often $<$ 1 sec), and a useful subset of these at approaching video frame rates. For higher resolutions the problem size increases approximately with the order of the number of labels in the grammar and the computational cost increases accordingly. One solution to this is to perform the search at multiple resolutions. Information from one resolution may be propagated to the next highest resolution, imposing initial constraints at the higher resolution to reduce the problem size. This is achieved using the best hypothesis from the lower resolution problem (in terms of MDL criteria) and applying two sets of constraints. Firstly four square regions at the higher resolution must contain an object boundary if the corresponding region at the lower resolution does and must not if it doesn't. The second set of constraints relate to the possible labels at the higher resolution given the lower resolution label. A set of rules (given in Appendix C) constrain these possible labels. This multi-scale approach can reduce the problem size by many orders of magnitude.

4.3 Hypothesis Adaptation

The MDL hypothesis selection method described in section 3 is imperfect (as is almost any hypothesis selection method for any model for real data). This can lead to errors being propagated from one resolution to the next. A solution to this is to combine the (constrained) global search strategy with a local adaptation scheme in which hypotheses are refined locally. This is analogous to a 'local search' strategy such as used in Kass' Snakes algorithm [10]. The approach we take is to refine small 3×3 region blocks using the relative information calculated as in equation 7 as a mechanism to select which set of regions to refine. If this relative information is positive for the center region of the group it is a candidate for adaptation. The adaptation process simply involves setting the state of these 9 regions unknown and running the constraint satisfaction algorithm described at the beginning of section 4. This is a highly efficient procedure as the total number of regions is small (9) and constraints on valid labels are imposed by surrounding labels.

5 Incorporating A-priori Shape Statistics

As with other methods that include no prior shape information (such as snakes [10]) the model/grammar fitting process can be disturbed by noise and clutter. A sensible approach in such circumstances is to include prior information about the shape of possible objects to be described. Such techniques are also useful for object identification. Cootes and Taylors Point Distribution Model [14] does this for a parameterised shape model. Our model contains a variable number of discrete parameters and, as such, their modelling technique is not suitable. Our approach is to borrow a technique used for modelling variable length time sequences, the Hidden Markov Model (HMM) [11]. One formulation of this, the Cyclic Hidden Markov Model (CHMM) [9], is particularly applicable to the modelling of object boundaries (closed or open). The transition graph for this is shown in figure 8.

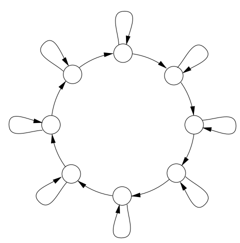

Figure 8: Architecture of the 'Cyclic Hidden Markov Model'

Each state (node) of the CHMM represents a probability distribution of some characteristic sequentially along the boundary. Initially we tried to model the state labels along the boundary, however this approach was not rotationally (or scale) invariant. Our latest approach is to model the absolute distance from the centroid of an external object boundary to the center of each square region along this boundary, using a continuous (single 1D Gaussian) distribution per state. If this distance is normalised (e.g. by division by the maximum or mean distance) this representation becomes scale invariant in addition to rotation and translation invariant. The Baum Welch optimisation algorithm [11] is used to optimise the CHMM based on a training set. Currently we have no automatic way of determining the number of hidden states required, however methods are suggested in [9].

6 Results and Evaluation

The grammar fitting process with no *a-priori* information included, as described in sections 2-4 was applied to images from a standard image database of objects [7] . Results for 32x32 pixel block resolution are given in figure 9.

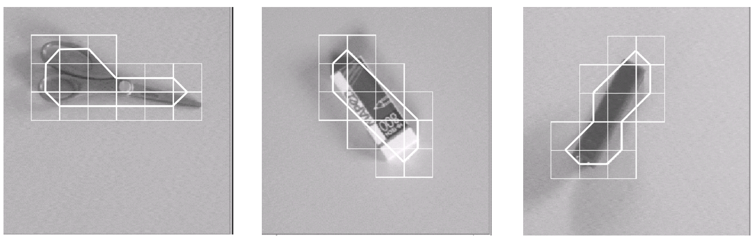

Figure 9: Results for the Fitting Algorithm with no *a priori* models

The algorithm, as described, works well in uncluttered scenes at low resolution. At higher resolutions, and in the presence of clutter, the model fits to the strongest region boundaries, regardless of shape or topology. This is a problem exhibited by other methods that include no *a priori* information such as snakes [10]. Section 5 describes the first step to including *a priori* information in this scheme, model building. To evaluate this shape modelling scheme sets of models were built of four of the objects in the database [7]; An Eraser, A pair of Scissors, a Stapler and a Set Square (Triangle). Between 5 and 10 examples of each item were used for this evaluation viewed approximately from above and at several different rotations. The resolution used was 8x8 pixel squares. When a model was compared to the object it was trained on 'leave one out' tests were used to ensure no test objects were used to train the model used in that evaluation. The Viterbi algorithm (see [11]) was used to compute mean log likelihood for 5 examples of each object given each model. The results are given in figure 10 over all tests.

Data / Model	Eraser	Scissors	Stapler	Triangle
Eraser	-0.577	-1.873	-0.802	-2.655
Scissors	-5.582	-1.400	-3.834	-2.841
Stapler	-0.838	-1.680	-0.691	-3.463
Triangle	-7.760	-2.096	-6.242	-1.260

Figure 10: Evaluation of HMM Classification Accuracy (mean of mean log likelihoods)

The higher (less negative) the mean log likelihood the better the fit. Using this as a classifier 19 out of 20 object examples were classified correctly with the only incorrect classification classifying the Stapler as almost equally likely to be the Eraser or the Stapler (-0.772 vs. -0.789 mean log likelihood). These are fairly similar objects in shape and it is perhaps encouraging that this has been picked out.

7 Future Work and Discussion

In this paper we have outlined a multi-scale, region based grammar for the qualitative description of images/scenes. This is based on sub-dividing the image into square regions and assigning a qualitative label to each. A constraint satisfaction type search procedure has been proposed for hypothesis generation, and it has been shown how the minimum description length principle may be applied to hypothesis selection. We have also shown how *a priori* statistical shape information may be encapsulated within this grammar by use of Cyclic Hidden Markov Models.

We have only just begun to explore the possibilities of the scene description grammar presented. In particular the integration of the *a priori* model into the fitting process is still to be done. Another work in progress is an improved attention mechanism. The attention mechanism is crucial to the efficiency of the hypothesis generation search and the current approach is not well suited to cluttered scenes. We also wish to apply this technique to dynamic scenes / image sequences.

References

[1] D. Ballard and C. Brown. *Computer Vision*. Prentice-Hall, 1982. (Ch. 9.5.3).

[2] H. Barrow and R. Popplestone. Relational descriptions in picture processing. *Machine Intelligence*, 6:377–396, 1971.

[3] A. G. Cohn and S. M. Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1-2):1–29, 2001.

[4] R. Cucchiara, E. Lamma, P. Melolo, and M. Milano. An interactive constraint-based system for selective attention in visual search. In *Proc. Int. Symposium on Methologies for Intelligent Systems*, 1997.

[5] R.H. Davies, T.F. Cootes, C.J. Twining, and C.J. Taylor. An information theoretic approach to statistical shape modelling. In *Proc. British Machine Vision Conference*, pages 3–12, 2001.

[6] J. Fernyhough, A. Cohn, and D. Hogg. Constructing qualitative event models automatically from video input. *Image and Vision Computing*, 18:81–103, 2000.

[7] D. Hall. View sphere database. http://www-prima.inrialpes.fr/perso.php?name=Hall.

[8] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29:5–28, 1998.

[9] D. Magee and R. Boyle. Detecting lameness in livestock using resampling condensation and multi-stream cyclic hidden markov models. In *Proc. British Machine Vision Conference*, pages 332–341, 2000.

[10] M.Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *First International Conference on Computer Vision*, pages 259–268, 1989.

[11] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[12] J. Rissanen. Modelling the shortest data description. *Automatica*, 14:465–471, 1978.

[13] C.E. Shannon. A mathematical theory of communication. *The Bell System Techical Journal*, 27:379–423, 1948.

[14] T.F.Cootes, C.J. Taylor, D.H. Cooper, and J.Graham. Training models of shape from sets of examples. In *Proc. British Machine Vision Conference*, pages 9–18, 1992.

[15] E. Tsang. *Foundations of Constraint Satisfaction*. Academic Press, 1993.

Appendix A: Region Label Adjacency rules

These rules consist of two types; horizontal adjacency rules (for regions to the left/right) and vertical adjacency rules (for regions to the top/bottom). In the following tables 0 represents invalid and 1 represents valid. The rows/columns relate to the states presented in figure 1 in the order presented there.

	Left													
	1	1	0	1	0	0	1	0	0	0	1	0	0	0
	1	1	0	1	0	0	1	0	0	0	1	0	0	0
	0	0	1	0	1	1	0	1	1	1	0	1	1	1
	0	0	1	0	1	1	0	1	1	1	0	1	1	1
	1	1	0	1	0	0	1	0	0	0	1	0	0	0
	1	1	0	1	0	0	1	0	0	0	1	0	0	0
Right	0	0	1	0	1	1	0	1	1	1	0	1	1	1
	0	0	1	0	1	1	0	1	1	1	0	1	1	1
	1	1	0	1	0	0	1	0	0	0	1	0	0	0
	0	0	1	0	1	1	0	1	1	1	0	1	1	1
	0	0	1	0	1	1	0	1	1	1	0	1	1	1
	0	0	1	0	1	1	0	1	1	1	0	1	1	1
	0	0	1	0	1	1	0	1	1	1	0	1	1	1
	0	0	1	0	1	1	0	1	1	1	0	1	1	1

	Above													
	1	0	1	1	1	0	0	1	0	0	0	0	0	0
	0	1	0	0	0	1	1	0	1	1	1	1	1	1
	1	0	1	1	1	0	0	1	0	0	0	0	0	0
	0	1	0	0	0	1	1	0	1	1	1	1	1	1
	0	1	0	0	0	1	1	0	1	1	1	1	1	1
	1	0	1	1	1	0	0	1	0	0	0	0	0	0
Below	1	0	1	1	1	0	0	1	0	0	0	0	0	0
	0	1	0	0	0	1	1	0	1	1	1	1	1	1
	0	1	0	0	0	1	1	0	1	1	1	1	1	1
	1	0	1	1	1	0	0	1	0	0	0	0	0	0
	0	1	0	0	0	1	1	0	1	1	1	1	1	1
	0	1	0	0	0	1	1	0	1	1	1	1	1	1
	0	1	0	0	0	1	1	0	1	1	1	1	1	1
	0	1	0	0	0	1	1	0	1	1	1	1	1	1

Figure 11: Horizontal and Vertical Region Label Adjacency Rules

It should be noted that efficient application of these constraints may be performed by storing possible labels and rules as a bitstream (with one bit per label) and using the bitwise AND operators to impose constraints.

Appendix B: Sub-Region Object Equivalence Rules

Due to space constraints we cannot present the complete set of sub-region equivalence rules, however figure 12 shows the rules for regions lying to the right of a region given the second label given in figure 1 as an example.

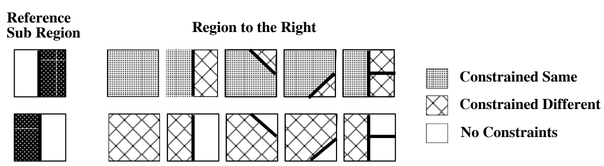

Figure 12: Sub-region Object Equivalence Rules for Regions to right of 2^{nd} Label Region

Appendix C: Multi-resolution Propagation Rules

Due to space constraints we cannot present the complete set of multi-resolution propagation rules, however figure 13 presents the possible states when the region label is the second label in figure 1 as an example.

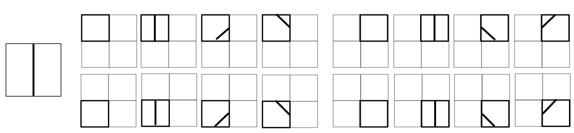

Figure 13: Multi-resolution Propagation Rules for Second Label Region

The reasoning behind these rules is that no description at the higher resolution can be formed that is qualitatively different from that at the lower resolution (except the zero hypothesis, which is dealt with separately).