Non-Gibbsian Markov random field models for contextual labelling of structured scenes

Daniel Heesch and Maria Petrou
Imperial College London
Communications and Signal Processing Group
Department of Electrical and Electronic Engineering
London SW7 2AZ, UK
{daniel.heesch,maria.petrou}@imperial.ac.uk

Abstract

In this paper we propose a non-Gibbsian Markov random field to model the spatial and topological relationships between objects in structured scenes. The field is formulated in terms of conditional probabilities learned from a set of training images. A locally consistent labelling of new scenes is achieved by relaxing the Markov random field directly using these conditional probabilities. We evaluate our model on a varied collection of several hundred hand-segmented images of buildings.

1 Introduction

Recent years have seen notable improvements in the performance of object classifiers. Greater robustness against occlusion and intraclass variability has been achieved by describing objects by a large number of local and largely view-invariant features (e.g. [15, 5, 18, 14]). For single classes efficient classification methods such as boosting allow recognition to be in real-time (e.g. [17]). Some of these models have the additional benefit of biological plausibility. The hierarchical feed-forward architecture of [13] aims to mimic the ventral stream of visual information processing and is able to predict with great accuracy whether or not an object is present in a scene.

It seems, however, that in order to be able to scale to the several thousands of categories humans discriminate without effort, appearance based object classification needs to be complemented by techniques that utilise contextual information. Context may be described as any dependency between the object to be recognised and everything else in the scene, be these other objects or the scene as a whole. Experimental evidence suggests that humans do exploit both types of dependency during object recognition. It is well established, for example, that the nature of a scene can be recognised based on low spatial frequency information [11]. Recent neuro-imaging studies support the view that low spatial frequencies are processed in the cortex at a very early stage during visual recognition [2], suggesting that perception involves top-down facilitation. Much like the gist of a scene, the spatial relationships between objects can be determined without high frequency information. Bar and Aminoff in [1] establish early activation of cortical "context networks" that appear to store spatial relationships, pointing to a key role of spatial context as an early facilitator during object recognition.

Our goal is to learn these spatial and topological relationships from the data and to utilise this information in a Markov random field (MRF) model to achieve a consistent labelling of new scenes. The MRF is defined not over a pixel array but the set of regions that correspond to objects. From training data we learn the probability distribution over labels for a region, given the objects in its local neighbourhood. These supply the conditional probabilities that define the MRF and are used during an iterative relaxation scheme to find a probable realisation given the structural relationships observed in a new scene.

Unlike the MRFs hitherto used in computer vision, the MRFs we use here are non-Gibbsian, i.e. they cannot be expressed in terms of cliques and a global cost function. This is because the interactions between units are directional and non-symmetric (A influences B differently from how B influences A). Such MRFs are characteristic of natural complex systems and they may be used to model, for example, the interaction between neurons in the human brain, population dynamics or company interactions. Complex systems subject to such unit interactions tend to oscillate between different states rather than converge to a single state [9]. In the case of human perception, the human brain is then somehow able to select from the possible interpretations the most appropriate one. In this paper we use a relaxation method appropriate for producing the states of such an MRF and a criterion that allows us to select the right state.

We validate our approach on a set of about 250 photographs of buildings that were manually segmented and labelled. This domain is particularly interesting as it exhibits sufficiently tight structural constraints to benefit from our approach, and a fair amount of structural variability to challenge it.

This paper is structured as follows. Section 2 presents related work. Section 3 introduces the non-Gibbsian model. Section 4 details how it is used to label new scenes. Section 5 describes a series of experiments to validate our approach. Section 6 concludes the paper.

2 Related work

We here consider related works that are concerned with modelling peer-to-peer, rather than hierarchical, dependencies. A natural choice for probabilistic modelling of local dependencies are Markov random fields [8], defined either on a segmentation of the image as in [10, 4] or on a rectangular grid as in [7, 6, 14]. The authors in [6] and [14] define a conditional random field over individual pixels. In [14], contextual information is incorporated by using the joint boosting algorithm [16] for learning potential functions and by employing a novel feature that captures local dependencies in appearance. Neither work explicitly considers spatial relationships, although in [6] the absolute position of a site is included in the potential function.

In [4], it is assumed that training images are associated with a bag of words with no explicit mapping between regions and terms. This renders the learning task more difficult but makes it easier to get hold of large amounts of training data. The MRF is specified through single and pair-wise clique potential functions learned from the data. To make the estimation problem tractable, potential functions are symmetric with respect to their arguments (labels of adjacent image regions). The model does not capture asymmetric dependencies, nor does it take into account spatial relationships.

In [10], an MRF is defined over image regions by specifying the clique functions for all types of single and pair-wise cliques. The potential functions are taken to be a weighted sum of m basis functions whose parameters are set manually.

Our objectives are similar to those in [4] and [10]. Unlike those two, however, we allow neighbouring blobs to influence each other differently depending on their relative spatial position. The asymmetry thereby introduced forbids the definition of cliques and thus the formulation of the MRF in terms of a Gibbs distribution. Our model consists of conditional probabilities that are learned directly from the data using structural information as can be obtained from the low spatial frequency content of an image.

3 The model

3.1 Non-Gibbsian MRF

Let $S = \{1,\ldots,N\}$ index a set of regions in an image. We assume that each region is associated with a random variable f_i which takes its value from a discrete set of class labels. The field $F = \{f_i : i \in S\}$ is assumed to be Markovian in the sense that the probabilistic dependencies among f_i are restricted to spatial neighbourhoods $\mathcal{N}_i$, that is,

$$P(f_i|f_{S-i},R) = P(f_i|f_{\mathcal{N}_i},R_i), \tag{1}$$

where R denotes the matrix of pair-wise spatial relationships between regions, and R_i the row pertaining to region i. We assume, therefore, that the conditional dependencies depend not only on the identity of the neighbouring regions but also on their relative spatial relationships with the ith region. This is an important component of our model as it allows us to capture the non-isotropic nature of many scenes. For convenience, we refer to a particular observation pair $(f_{\mathcal{N}_i},R_i)$ as the *neighbourhood configuration* or simply *configuration*, and to the ith region associated with it as the *focal region*.

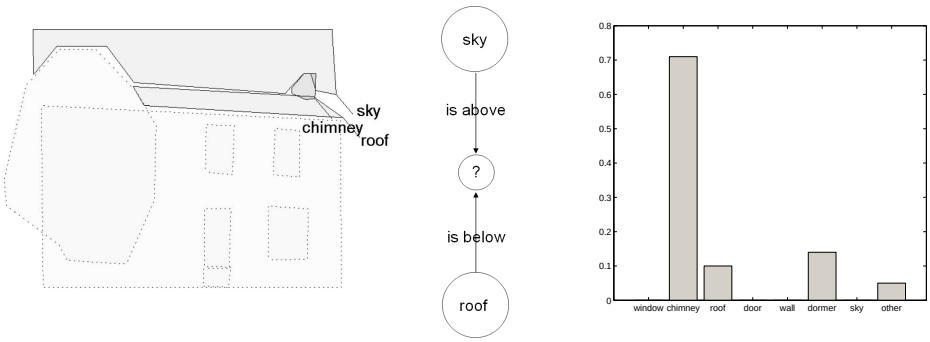

Figure 1: A particular configuration associated with a chimney (left), a schematic representation of the configuration $(f_{\mathcal{N}_i},R_i)$ (middle) and the conditional probability distribution over all labels associated with that configuration, $P(f_i|f_{\mathcal{N}_i},R_i)$, as obtained from training images (right). The distribution tells us that a region below sky and above a roof is a chimney (71%) but may also be a dormer (14%) or another roof (10%).

3.2 Neighbourhoods

Since we need to learn the conditional distributions from a relatively small training set, we limit the neighbourhood to at most six regions: the neighbour above, below, to the left and to the right of region i, as well as the region containing and being contained by region i. The

neighbourhood relation is reciprocal and two regions are neighbours if they are separated by no more than a certain distance threshold. The distance between two regions $A, B \subset \mathbb{R}^2$ is computed as

$$d(A,B) = \sum_{i \in \{x,y\}} \min_{a \in A, b \in B} |a_i - b_i|, \tag{2}$$

where a_x represents the x coordinate of point a. Other choices of a distance function are of course conceivable. This particular one has the effect that two regions need not be the same to have a zero distance but may be (i) overlapping, (ii) exactly adjacent or (iii) contained in one another. For example, a wall that surrounds a number of windows has a zero distance from each of them. If regions are non-overlapping, the distance along each direction is given by the smallest Euclidean distance between any two points of the two regions. This has the advantage that the distance between two regions is not affected by their respective sizes (as would be the case under many metrics such as the Hausdorff metric). For a distance cutoff of 0, the neighbourhood consists of all regions whose bounding boxes overlap with or touch the focal region. Were the regions regularly arranged like pixels, the resulting neighbourhood would be the familiar 8-pixel neighbourhood. The optimal distance cutoff is learned through cross-validation. Figure 2 depicts the distribution over configuration sizes for the optimal zero cutoff. The right figure illustrates how the configurations become larger as the distance cutoff increases.

Given a distance threshold, the conditional probability distributions (eq. 1) are learned by noting for each region i observed in a set of training images its corresponding configuration $(f_{\mathcal{N}_i}, R_i)$. The results can conveniently be stored in the form of a hashtable with the key being a particular configuration and the value being the conditional probabilities over labels for the focal region. Given a region with known neighbourhood configuration, we can thus rapidly obtain a probability distribution over labels at the focal region. To ensure that the joint distribution of the MRF is nowhere zero, we add a small positive value to each zero-valued conditional probability and subsequently normalise.

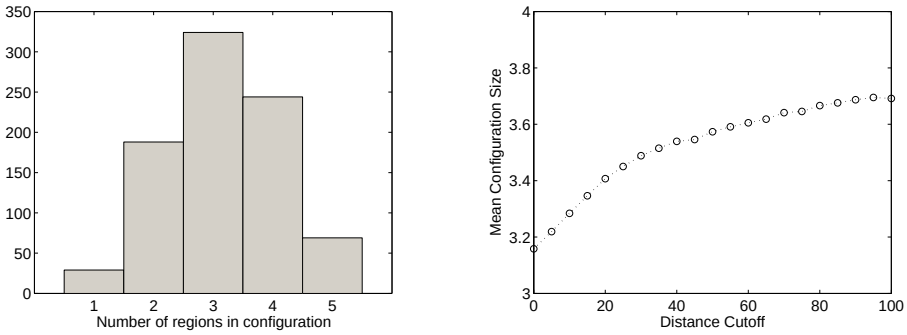

Figure 2: Frequency distribution of different configuration sizes for a distance cutoff of zero (left). As we increase the distance threshold, the configurations become larger (right).

4 Labelling of new scenes

This section details how to obtain probable realisations of the MRF given a new scene. We make the assumption that scenes have been segmented into regions where each region corresponds to an object to be recognised. How these regions are obtained in the first place

is a problem in its own right and outside the focus of this work. We shall simply take it for granted that an appropriate segmentation has been achieved.

4.1 Global Gibbsian versus local non-Gibbsian relaxation

A standard technique to find a probable realisation of an MRF is simulated annealing which allows a stochastic label update at a site to be retained with a certain probability P_r even if the new realisation of the field is less probable. By letting P_r converge to zero, the field eventually settles at a maximum of the joint probability distribution. In other words, simulated annealing strives to find solutions that are globally maximally consistent.

Because of the impossibility to define cliques, our non-Gibbsian field is formulated purely in terms of local, conditional probability distributions (Equation 1). We aim to find labellings that are locally consistent by repeatedly sampling from these conditional distributions.

4.2 Graph colouring

In order to iteratively update regions based on the current labelling of their neighbourhood, we partition the set of regions into a set of codings. The idea of a coding was first introduced by Besag [3] in the context of the iterated conditional mode algorithm for MRF parameter estimation. A coding is equivalent to the concept of a vertex colouring of a graph, that is, it constitutes a partitioning of the set of vertices (= regions) so that no two adjacent vertices (= neighbouring regions) belong to the same partition. Because of the assumption of Markovianity, the likelihood over vertices of the same colour reduces to a simple product of the respective conditional probabilities. We employ a greedy strategy to achieve a vertex colouring, in which vertices are visited in order of decreasing vertex degree (i.e. number of neighbours). Each vertex is assigned the first possible colour from a list of colours. One example of a colouring is given in Figure 3. The wall has the largest number of neighbours and is correspondingly assigned the first colour (‘1’).

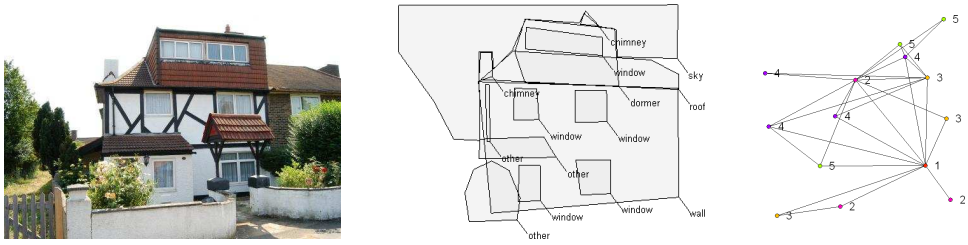

Figure 3: Original image (left). Hand-segmented and hand-labelled training image (middle). Vertex colouring of the neighbourhood graph (right): vertices with the same number have non-overlapping neighbourhoods.

4.3 Choosing a solution

Regions are updated within each coding by retrieving and sampling from the probability distribution corresponding to that region’s current neighbourhood configuration. If the configuration has not been seen before, because it was not observed in the training set, the new label is drawn from a uniform distribution. This scheme on its own is not guaranteed to

converge and indeed it seems to have no tendency to do so. Following each update, we compute for each coding $\mathscr{C}_j$

$$P(f_{\mathscr{C}_j}|R) = \prod_{i \in \mathscr{C}_j} P(f_i|f_{\mathscr{N}_i}, R)$$

Our estimate of the overall probability of the data is obtained by averaging over $P(f_{\mathscr{C}_j}|R)$. Because the codings are generally of different size, the arithmetic average sometimes used for regular MRF is unsuitable. Instead, we estimate the joint probability as

$$P(f_1, \ldots, f_N) \approx \frac{1}{N} \sum_j |\mathscr{C}_j| \left[\prod_{i \in \mathscr{C}_j} P(f_i|f_{\mathscr{N}_i}, R) \right]^{\frac{1}{|\mathscr{C}_j|}}. \quad (3)$$

Let p be the ratio between the estimated joint probability after and before the update. We accept the change with probability 1 if $p > 1$ and with probability $p^{\frac{1}{T}}$ otherwise. T is the temperature parameter whose value decreases exponentially with time. Figure 4 shows an example of how the value given by eq. 3 increases over successive iterations. One iteration here involves the update of the labels of all regions.

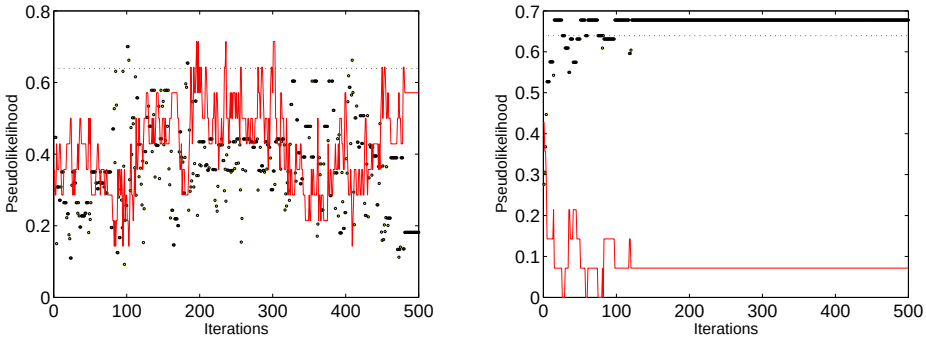

Figure 4: Dynamics of stochastic updating process with and without maximisation of the pseudolikelihood. The dotted line marks the pseudolikelihood associated with the true labelling. The continuous line shows the proportion of misclassified regions. In both diagrams, regions are updated based on the conditional probabilities. For the left diagram, a new labelling is always accepted, for the right diagram, a labelling is accepted when it improves the current optimum or when it is worse by no more than a value that decreases with time.

5 Experiments

For our experiments, we collected 253 images of buildings from the World Wide Web. Each image was manually segmented into regions that correspond to parts of the building or parts of the environment such as sky or vegetation. Each region was labelled by hand using an annotation tool similar to LabelMe. The complete dataset contains nearly 6,000 regions covering a dozen of classes.[1]

[1]The images along with the annotation and segmentation information is available at http://www.commsp.ee.ic.ac.uk/~dheesch/ngmrf/data/

We allow for the following seven labels (with respective frequencies): ‘window’ (0.507), ‘chimney’ (0.054),‘roof’ (0.053), ‘door’ (0.087), ‘wall’ (0.089) , ‘dormer’ (0.015), ‘sky’ (0.055), ‘other’ (0.14). The ‘other’ label aggregates all remaining structures that were annotated (e.g. ’pipes’ and ’balcony’). We report performance of different algorithms in terms of classification accuracy, i.e. the proportion of regions that have been labelled correctly. To estimate how the algorithm will be able to predict data that it was not trained on, we use the leave-one-out method of cross-validation, i.e. we remove one image from the set at a time to be our test image and train on the remaining 252 images.

5.1 Comparison with other methods

We compare our non-Gibbsian MRF model with two other classification models, a non-contextual Bayes classifier and an alternative contextual model that uses probabilistic relaxation to find a locally consistent labelling.

5.1.1 Non-contextual Bayes classifier

As a non-contextual benchmark we implemented a Parzen classifier that classifies regions based on the posterior probabilities given measurements of a number of low-level features from the region. We use a set of three features that can easily be obtained from the low-frequency content of a scene: the mean intensity, the normalised area of the region and its vertical position. For each feature, the posterior probabilities over classes is given by Bayes rule with the class-conditional densities being approximated using a Parzen window with a Gaussian kernel function centred on a set of class exemplars E_c

$$p(x|c) \propto \sum_{x_i : i \in E_c} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|x - x_i|}{2\sigma^2} \right), \tag{4}$$

where σ is learned through cross-validation. We assume each feature to be conditionally independent given the class, and thus compute the overall class probability density as a product of feature-specific posteriors.

5.1.2 Probabilistic relaxation

The second comparison is with an alternative contextual labelling technique known as probabilistic relaxation [12]. The contextual information consists of the conditional probabilities of a label, given that another label is found in a particular relative position to the first. In each iteration of the relaxation process, the label probabilities are updated based on the probabilities at the previous time step, modulated by the support a particular label f_i receives from neighbouring labels,

$$P^{(n+1)}(f_i = c) = \frac{P^{(n)}(f_i = c) Q_i(c)}{\sum_{\mu \in \mathscr{L}} P^{(n)}(f_i = \mu) Q_i(\mu)} \tag{5}$$

with support function

$$Q_i(c) = \sum_{j : f_j \in \mathscr{N}_i} \sum_{\nu \in \mathscr{L}} P(f_i = c | f_j = \nu, r_{ij}) P(f_j = \nu). \tag{6}$$

Here $\mathscr{L}$ denotes the label set. The compatibilities are learned from the data in a similar way as are the conditional distributions for neighbourhood configurations in the MRF model. Note that unlike the MRF model, which allows configurations to comprise up to six regions, this particular formulation of probabilistic relaxation is limited to binary dependencies. This makes statistical learning of dependencies easier but comes at the expense of limited modelling power.

5.1.3 Results

Table 1 shows the results for probabilistic relaxation and our NG-MRF when using the output of the Parzen classifier to initialise the labelling. In order to assess the variability in performance, we have opted for a leave-one-out strategy. The results are the average over 253 images with more than 5,000 regions.

The best results are obtained by the non-Gibbsian MRF, followed closely by the non-contextual classifier. It is noteworthy that this particular version of probabilistic relaxation, instead of improving the results of the non-contextual Parzen method, makes them worse.

Regions	Unique cfgs	Prior	Parzen	PR	NG-MRF
5,682	0.904	0.521 (0.0006)	0.690 (0.125)	0.568 (0.134)	0.729 (0.124)

Table 1: Performance comparison for different classification methods. Prior: each region is given the same, most frequently occuring label; Parzen: non-contextual classification; PR: probabilistic relaxation; NG-MRF: non-Gibbsian Markov random field. Performance is measured in terms of the proportion of regions classified correctly (standard deviation in brackets). The second column gives the proportion of unique configurations in the test set for which a conditional distribution has been learned from the training images.

Table 1 does not show how performance varies between different classes. As the confusion matrix in Table 2 indicates, by far the greatest accuracy is achieved for windows. That many other classes are misclassified as windows may be attributed to the strong prior on the 'window' class that influences the result through the non-contextual Parzen initialisation. Note that doors in particular are frequently mistaken for windows as these two classes exhibit very similar spatial relationships with other building parts whilst having markedly different priors.

	wi	ch	ro	do	wa	do	sk	ot
window	2848	50	5	81	0	0	25	131
chimney	20	151	50	5	0	5	10	15
roof	25	20	101	0	30	10	25	76
door	348	5	0	20	5	0	0	96
wall	40	0	25	5	292	10	10	91
dormer	30	15	20	5	5	15	5	0
sky	15	10	10	0	5	5	192	30
other	217	15	15	40	30	5	25	343

Table 2: Confusion matrix for NG-MRF labelling. The top row entries are indexed by the first two letters of the respective label. The matrix element a_{ij} gives the number of regions of the ith class that have been classified as belonging to the jth class.

5.2 Robustness to initialisation

We investigate two different initialisation schemes to assess the robustness of the contextual inference to initial conditions. The first scheme assigns each region the most frequently occuring label (in this case 'window'), the second draws labels randomly from the prior distribution, i.e. it will result in a similar initial distribution of classes within the image but with random assignment of classes to regions. The results are shown in Table 3. While we notice a performance degradation compared to non-contextual initialisation, the contextual model continues to improve over the new baselines of 0.52 and 0.32, respectively.

Initialisation scheme	Initial	NG-MRF
Non-contextual	0.690	0.729 (0.124)
Max Prior	0.521	0.654 (0.127)
Random	0.315	0.621 (0.135)

Table 3: Dependence of contextual classification on initial conditions. The second column shows the accuracy after initialisation with the three different schemes discussed in the text. The initial accuracy of the random assignment is $1-\sum_c p_c(1-p_c)$ where p_c is the prior of the cth class.

6 Conclusions

We presented a Markov random field model for contextual labelling of objects in structured scenes. In our model the context of a region consists not only of the identity of neighbouring regions but also, crucially, on their relative spatial and topological relationships. By incorporating what are typically asymmetric relationships, the Markov random field is capable of modelling the non-isotropic nature of typical scenes. The asymmetry makes the field non-Gibbsian as it no longer admits to a factorisation into cliques, so that the model is formulated in terms of conditional distributions that are learned from training data.

Given a new scene, the Markov random field is relaxed by iteratively sampling from conditional probability distributions. We proposed an objective function to help us identify good labelling solutions. The objective function is based on the vertex colouring of the region neighbourhood graph and is not the global cost function usually associated with Gibbsian MRFs. A comparison with a non-contextual and an alternative contextual classifier suggests the validity of the approach.

There are several ways how to take the work further. For this study we hand-segmented and hand-labelled several hundred images. To demonstrate the robustness of the technique, a next step is to learn configurations from automatically segmented, but possibly hand-labelled training exemplars. Also, we currently make no attempt to generalise from observed configurations to new ones. As some configurations are supersets of smaller configurations, or are otherwise similar to each other, endowing the configuration space with some distance metric would allow more accurate label distributions to be inferred for previously unseen configurations.

References

[1] M Bar and E Aminoff. Cortical analysis of visual context. *Neuron*, 38:347–358, 2003.

[2] M Bar, K Kassam, A Ghuman, J Boshyan, A Schmidt, A Dale, M Hämäläinen, K Marinkovic, D Schacter, B Rosen, and E Halgren. Top-down facilitation of visual recognition. *Proceedings National Academy of Sciences*, 103(2):449–454, 2006.

[3] J Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal Royal Statistical Society, B*, 36:192–236, 1974.

[4] P Carbonetto, N de Freitas, and K Barnard. A statistical model for general contextual object recognition. In *Proc European Conf Computer Vision*, pages 350–362, 2004.

[5] G Csurka, C Bray, C Dance, and L Fan. Visual categorization with bags of keypoints. In *Proc European Conf Computer Vision*, 2004.

[6] X He, R Zemel, and D Ray. Learning and incorporating top-down cues in image segmentation. In *Proc European Conf Computer Vision*, 2006.

[7] S Kumar and H Hebert. Discriminative random fields: a discriminative framework for contextual interaction in classification. In *Proc Int'l Conf Computer Vision*, 2003.

[8] S Li. *Markov Random Field Modeling in Computer Vision*. Springer, New York, 1995.

[9] Z Li and P Dayan. Computational differences between asymmetrical and symmetrical networks. *Network: Computation in Neural Systems*, 10(1):59–77, 1999.

[10] J Modestino and J Zhang. A Markov Random Field model-based approach to image interpretation. *IEEE Trans Pattern Analysis and Machine Intelligence*, 14(6):606–615, 1992.

[11] A Oliva and A Torralba. Modelling the shape of the scene: a holistic representation of the spatial envelope. *Int'l Journal Computer Vision*, 42(3):145–175, 2001.

[12] A Rosenfeld, A Hummel, and S Zucker. Scene labeling by relaxation operations. *IEEE Trans Systems, Man and Cybernetics*, 6(6):420–433, 1976.

[13] T Serre, A Oliva, and T Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings National Academy of Science*, 104(15):6424–6429, 2007.

[14] J Shotton, J Winn, C Rother, and A Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc European Conf Computer Vision*, 2006.

[15] J Sivic and A Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proc Int'l Conf Computer Vision*, pages 1–8, 2003.

[16] A Torralba, K Murphy, and W Freeman. Sharing fatures: efficient boosting procedures for multiclass object detection. In *Proc Int'l Conf Computer Vision and Pattern Recognition*, pages 762–769, 2004.

[17] P Viola and M Jones. Rapid object detection using a boosted cascade of simple features. In *Proc Int'l Conf Computer Vision and Pattern Recognition*, 2001.

[18] J Winn, A Criminisi, and T Minka. Object categorization by learned universal visual dictionary. In *Proc Int'l Conf Computer Vision*, pages 1800–1807, 2005.