LEARNING MARKOVIAN DEPENDENCIES FROM ANNOTATED IMAGES

Daniel Heesch and Maria Petrou

Imperial College London
Communications and Signal Processing Group
Department of Electrical and Electronic Engineering
London SW7 2AZ, UK

ABSTRACT

In this paper we propose to model structural knowledge about scenes by a Markov random field whose conditional probabilities are learned from the spatial and topological relationships observed between regions in a set of training images. A locally consistent labelling of new scenes is achieved by relaxing the Markov random field directly, using conditional probabilities rather than a Gibbs formulation. We validate our approach on several hundreds of hand-segmented photographs of buildings.

1. INTRODUCTION

Recent years have seen a considerable improvement in the quality of object classification and recognition. To a large extent, these improvements are the result of modelling objects as loose sets of local and largely view-invariant features such as SIFT [9], maximally stable regions [10] and several others [11]. The resulting classifiers exhibit greater robustness against occlusion and allow fast recognition of objects in large image collections, e.g. [15], [5], [17], and [14].

However, a naïve application of non-contextual object detection models to a multi-class setting is not only inefficient (as such a system would scale linearly with the number of classes to be detected), but is also likely to suffer from low accuracy. 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 [13]. 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]. Much like the gist of a scene, the spatial relationships between objects can be determined without high frequency information. In fact, 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 over the set of regions that correspond to objects. From training data we learn the conditional probability distribution over labels for a region, given the objects in its neighbourhood. These probability distributions are used during an iterative relaxation scheme to find a probable realisation of the MRF 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 network of 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 [8]. The human brain somehow is then able to select from all possible interpretations of a scene 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 253 photographs of building scenes.

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

2. RELATED WORK

Several contextual models for object recognition have been formulated in recent years. We consider here only those that are concerned with modelling peer-to-peer dependencies.

A natural choice for probabilistic modelling of local dependencies are Markov random fields, defined either on a segmentation of the image as in [12, 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 [6] includes the absolute position of a site in the potential function.

In [4], it is assumed that there are no explicit associations between terms and image regions in the training data; rather each image is associated with a bag of words and the precise term-region associations have to be learned from training data. On the one hand, this makes the learning task more difficult. On the other hand, however, it gives access to a much larger volume of training data as publicly available collections of annotated photos now abound on the World Wide Web. 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 therefore does not capture asymmetries in the dependency relationship. The model also does not take into account spatial relationships and thus is indifferent to whether, for example, a blue patch is above (sky) or below (sea) another.

In [12], 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 a weighted sum of basis functions whose parameters are set manually.

Our work shares the same objectives with [4] and [12]. Unlike these two, however, we allow neighbouring blobs to influence each other differently depending on their relative spatial positions and topological relationships. 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. Learning dependencies of non-Gibbsian MRFs

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 and topological relationships between regions, and R_i only 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 and topological 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 illustrates these notions.

We refer to the model as a non-Gibbsian Markov random field (NG-MRF) as it cannot be expressed in terms of cliques.

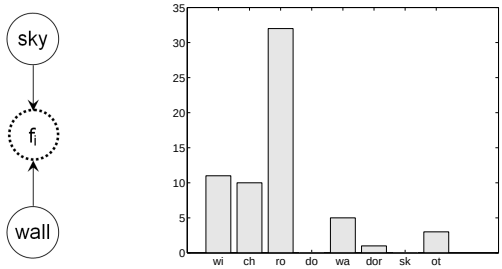

Fig. 1. Schematic representation of the configuration $(f_{\mathcal{N}_i}, R_i)$ (left); the conditional probability distribution over all labels for the focal region (dotted circle), $P(f_i|f_{\mathcal{N}_i}, R_i)$, as obtained from training images (right). The most frequent label associated with the focal region is 'roof', followed by 'window' and 'chimney'.

For this work, we manually segment all our images into regions using an extension of the LabelMe tool[1]. The long-term goal, of course, is to learn configurations and associated probabilities from automatically derived segmentations.

3.2. Learning optimal neighbourhood sizes

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 regions contained

[1] http://labelme.csail.mit.edu

in and containing region i (e.g. windows embedded in a facade). 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

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

where a_x represents the x coordinate of point a.

Other choices of a distance function are conceivable. This particular one has the effect that two regions need not be the same to score a zero distance (thus, it is not a metric). In particular, regions with zero distance can 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). The optimal distance cutoff, learned through cross-validation, turns out to be zero. For a zero distance cutoff, 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. Figure 2 (left) 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 (Equation 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.

4. LABELLING OF NEW SCENES

This section details how to obtain probable realisations of the MRF given a new scene in which we observe certain spatial and topological relationships (R in Equation 1). We make the assumption that scenes have been segmented into regions where each region corresponds to an object to be recognised. How these regions may be obtained automatically in the first place is a problem in its own right and outside the scope of this work. We obtain regions by manually segmenting images.

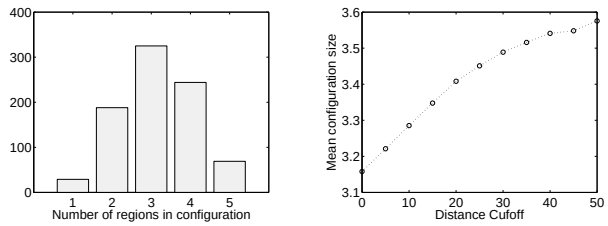

Fig. 2. Frequency distribution of different configuration sizes for $d = 0$ for which the majority involves three regions (left). As we increase the distance threshold, the configurations become larger (right).

4.1. Local 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, i.e.

$$f_i^{(n+1)} \sim P(f_i | f_{\mathcal{N}_i}^{(n)}, R_i).$$

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 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) and each vertex is assigned the first possible colour from a list of colours (see Figure 3 for an example).

4.2. Choosing a solution

Regions are updated within each coding by retrieving and sampling from the probability distribution corresponding to

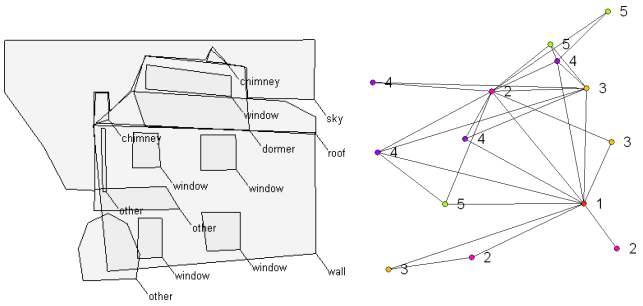

Fig. 3. Hand-segmented and hand-labelled training image (left). Vertex colouring of the neighbourhood graph (right), vertices with the same number have non-overlapping neighbourhoods.

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 settle on a particular solution (left graph of figure 4). Following each update, we compute for each coding $\mathcal{C}_j$

$$P(f_{\mathcal{C}_j}|R) = \prod_{i \in \mathcal{C}_j} P(f_i|f_{\mathcal{N}_i}, R). \tag{3}$$

Our estimate of the overall probability of the data is obtained by averaging over $P(f_{\mathcal{C}_j}|R)$. Because the codings are generally of different size, we cannot employ the arithmetic average that is suitable for inference on regular MRFs. Instead we estimate the joint probability as

$$P(f_1, \ldots, f_N) \approx \frac{1}{N} \sum_j |\mathcal{C}_i| \left[\prod_{i \in \mathcal{C}_j} P(f_i|f_{\mathcal{N}_i}, R) \right]^{\frac{1}{|\mathcal{C}_j|}}. \tag{4}$$

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 Equation 4 increases over successive iterations (one iteration involving the update of all labels).

5. EXPERIMENTS

We collected 253 images of buildings from the World Wide Web. Each image was manually segmented into regions that corresponded to parts of a building or parts of the environment such as sky or vegetation. Each region was labelled manually using an annotation tool similar to LabelMe. The complete dataset contained nearly 6,000 regions covering a dozen of classes.[2]

[2]http://www.commsp.ee.ic.ac.uk/∼dheesch/academic/ngmrf/data

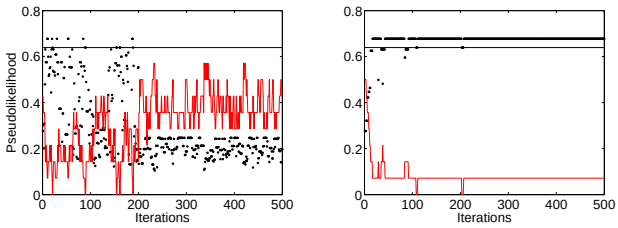

Fig. 4. Dynamics of stochastic updating process with and without maximisation of $P(f_i, \ldots, f_N)$ (Equation 4). The dotted line marks the value associated with the true labelling. The continuous line shows the proportion of misclassified regions. In both diagrams, regions are updated based on the conditional probabilities. For the left diagram, a new labelling is always accepted. For the right diagram, a labelling is accepted if it improves the current solution or if it is worse by no more than a value that decreases with time.

For our experiments, we use 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 other labels used during annotation but not used explicitly in the classification task. 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 cope with data not included in the training set, 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 three other classification models, a simple maximum prior classifer, a non-contextual Bayes classifier and a Markov random field that ignores spatial relationships.

5.1.1. Maximum prior

Our simplest benchmark is a maximum prior classifier that assigns to each region the most frequently occuring label (i.e. 'window').

5.1.2. 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, $\mathcal{C}_c$

$$p(x|c) \propto \sum_{x_i \in \mathcal{C}_c} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|x - x_i|}{2\sigma^2}\right), \qquad (5)$$

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.

5.1.3. Isotropic MRF (ISO-MRF)

To assess the added value one gains by considering explicitly the spatial and topological relationships between neighbouring regions, we implemented a simpler Markov random field in which the Markovian dependencies depend only on the labels of neighbouring regions but not on their spatial relationships. The isotropic MRF model thus assumes that a neighbour of a certain focal region has the same effect on the latter regardless of its relative position, i.e. in the notation of section 3.1,

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

5.1.4. Results

Table 1 shows the results for the Parzen classifier and the ISO-MRF and NG-MRF models. Note that we use the output of the Parzen classifier to initialise the labelling for the MRF models. 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 5,682 regions.

The best results are obtained by the non-Gibbsian MRF, followed closely by the non-contextual Parzen classifier. The performance of the isotropic MRF is notably worse even in comparison with the simple maximum prior method.

Model	Accuracy	Std Dev	Known Configurations
Max Prior	0.521	0.0006	—
Parzen	0.690	0.125	—
ISO-MRF	0.434	0.145	0.9769
NG-MRF	0.729	0.124	0.8949

Table 1. Comparison of different methods using 5,682 blobs for training. Prior: each region is given the same, most frequently occuring label; Parzen: non-contextual classification; ISO-MRF: isotropic Markov random field; NG-MRF: non-Gibbsian Markov random field. Performance is measured in terms of the fraction of regions classified correctly. The last column gives the fraction of configurations in the test data observed in the training set.

The confusion matrix (Table 2) reveals that the greatest accuracy is achieved for windows. That many other classes are misclassified as windows may be attributed to the strong 'window' prior that influences the result through the non-contextual Parzen initialisation. 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.

True label	Predicted label							
	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.

5.2. Robustness to initialisation

We investigate two different initialisation schemes in addition to the Parzen initialisation to assess the robustness of the contextual inference to initial conditions. The first scheme assigns to each region the most frequently occuring label (i.e. '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 with non-contextual initialisation, the contextual model continues to improve over the new baselines of 0.52 and 0.32, respectively.

Initialisation	Initial	NG-MRF (Mean and Std Dev)
Parzen	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 performance on initial conditions. The second column shows the accuracy after initialisation with the three different schemes described in the text. The initial accuracy of the random assignment can be shown to be $1 - \sum_c p_c(1 - p_c)$ where p_c is the prior of the cth class.

6. CONCLUSIONS

We presented a Markov random field model for contextual scene labelling. A region's context includes not only the

identity of neighbouring regions but also their relative spatial and topological relationships, thus rendering the model capable of capturing the non-isotropic nature of typical scenes. The asymmetry makes the field non-Gibbsian as it no longer admits to a factorisation into cliques. The model is therefore formulated directly in terms of conditional distributions that are learned from a training set of annotated and segmented images. Given a new scene, the Markov random field is relaxed by iteratively sampling from these conditional probability distributions. We proposed an objective function to help identify good labellings. 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 a contextual classifier demonstrates the validity of the approach and the importance of utilising relational information for scene labelling.

There are several ways to take this work further. First, learning is presently based on manually segmented and labelled images. A next step is to work on automatically segmented, but possibly hand-labelled images. Second, relations are modelled as crisp concepts which leads to quantisation error and sensitivity to small translations of individual regions. The idea of fuzzy relations might prove valuable here. Third, the proposed model assumes that all labels are equally likely for the focal region, should the associated configuration not have been seen before. We thus 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, we believe that by endowing the configuration space with some distance metric, more accurate label distributions could be inferred for previously unseen configurations.

Acknowledgments: This work was supported by the FP6 European project eTRIMS.

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

[9] D Lowe. Distinctive image features from scale-invaraitn keypoints. *Int'l Journal of Computer Vision*, 60:91–110, 2004.

[10] J Matas, O Chum, M Urban, and T Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *British Machine Vision Conference*, pages 384–393, 2002.

[11] K Mikolajczyk and C Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1615, 2005.

[12] 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.

[13] 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.

[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] 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.