

CRF Based Region Classification Using Spatial Prototypes

Mohammad Jahangiri
Imperial College London
m.jahangiri@imperial.ac.uk

Daniel Heesch
Empora Research
daniel.heesch@empora.com

Maria Petrou
Imperial College London
maria.petrou@imperial.ac.uk

Abstract

This paper proposes a conditional random field (CRF) model that encodes and exploits the spatial context of a region for region labelling. Potential functions for a region depend on a combination of the labels of neighbouring regions as well as their relative location, and a set of typical neighbourhood configurations or prototypes. These are obtained by clustering neighbourhood configurations obtained from a set of annotated images. Inference is achieved by minimising the cost function defined over the CRF model using standard Markov Chain Monte Carlo (MCMC) technique. We validate our approach on a dataset of hand segmented and labelled images of buildings and show that the model outperforms similar such models that utilise either only contextual information or only non-contextual measures.

1. Introduction

The problems of object classification and recognition have long been at the core of computer vision and pattern recognition research. Much research has concentrated on the exclusive use of relatively simple non-contextual features for classification, with remarkable success. Examples include the work of Viola and Jones [22] for haar-wavelet based face recognition, and that of Dalal and Triggs [4] on human detection using gradient histograms. Much of the works that fall within the tradition of the visual word paradigm [17] can also be classed as non-contextual (e.g. [3], [21]). In an attempt to achieve greater robustness against occlusion and geometric transformations, images are represented as a large number of local and largely view-invariant features. The latter methods excel in particular when the task is the recognition and retrieval of specific object *instances* rather than the identification of the instance *class*. Contextual information in the computer vision systems was adopted by some early works in vision [5][1][18]. Over the last few years, the role of contextual information in particular for situations where there exist structural (spatial) constraints and sparse co-occurrence relationships between

object classes have been further investigated. Contextual information can be particularly useful when individual classes exhibit large appearance variability.

We propose a probabilistic model which makes use of spatial context for labelling regions in an image and apply it to highly structured scenes of buildings. Our guiding hypothesis is that parts of buildings relate to each other in very similar ways irrespective of the period, the architecture, or the country. We try to capture the spatial constraints by clustering neighbourhood configurations obtained from a set of hand-segmented images. These are subsequently used to evaluate the potential functions of a conditional random field (CRF) defined over the regions of an image. In section 2 we present related work that utilise context for classification. Section 3 describes our method for labelling regions through the integration within a CRF model of contextual and non-contextual (unary) information. In section 4 we present our experiments and we conclude in section 5.

2. Related Work

Context comes in different forms. At the pixel level it may take the form of a continuity constraint on the label map, such that neighbouring pixels are more likely to carry the same label. At higher levels, and more in the spirit of this work, context may encapsulate hierarchical relationships between scene type and objects and the spatial and co-occurrence relationships between objects. Motivated by the work of [15] on the gist of scenes, [19] and [20] proposed models that allow the scene type to inform subsequent object detection by suggesting the kinds of objects to expect as well as their scales and locations. Rabinovich et al. [16] suggest to model the co-occurrence relationships between labels as the pairwise potentials in a conditional random field with no consideration, however, of the spatial relationships. Galleguillos et al. [6] extended the work of [16] by parameterising the co-occurrence matrices over different spatial relationships.

While [23] implicitly models spatial relationships by allowing features outside the object to influence object detection, the work in [13] and [2] focuses on an explicit modelling of spatial context between labels using Markov random field

models. In [11], the proposed model aims at capturing the 3D layout of a street scene using scale and location constraints, and uses this 3D information for more accurate object detection. More recently, [7] suggested a two stage multi-class classification process in which the non-contextual classification achieved in the first stage is subsequently refined through the use of prior distributions over the relative distances between different labels. Li et al. [14] consider the task of understanding an image at scene and object level. Co-occurrence of objects and scene types are formalised in a hierarchical generative model which specifies a joint distribution over scene classes, objects, regions, image patches and annotation tags. Learning requires images in which at least a few object regions have been labelled with their corresponding tags. Spatial relationships between objects are not considered. Heitz et al. [10] distinguish between materials and objects. Object detections (using a sliding window approach) are linked to region type detections (found through unsupervised clustering) by a set of relationships, e.g. "detection i is about 100 pixels away from region j ".

In [12], the authors construct a CRF with unary components from a random forest classifier and the interaction term reflecting the co-occurrence statistics of different labels. The CRF is defined on a set of regions with region neighbourhood determined by spatial adjacency. Whilst regions are determined automatically through a clever, entropy-based selection method, the contextual information does not include spatial relationships. In [9] and [8] relative location as well as labels of neighbourhood regions were used for labelling building scenes. These models use a Markov Random Field (MRF) for labelling different parts of a scene without incorporating unary features.

In this paper, we propose a conditional random field in order to incorporate unary features and contextual information in a single framework. In the proposed model, the interaction term of the CRF model is estimated using a set of typical neighbourhood configurations which are from a set of training data. A neighbourhood configuration represents a combination of the identity (labels) and the relative location of the neighbouring regions.

3. Methodology

We start by providing the basic notation used in the paper. Each segmented region, indexed by i , is associated with a random variable X_i which takes its value from a discrete set of class labels $\mathcal{L} = \{l_1, \dots, l_M\}$. The set of N segmented regions in an image I corresponds to a random field $\mathbf{X} = \{X_1, \dots, X_N\}$. Any possible assignment of labels to the random variables will be called a *labelling* (denoted by \mathbf{x}) which takes values from the set $L = \mathcal{L}^N$. The purpose of the proposed probabilistic framework is to estimate a labelling which maximises the following conditional proba-

bility:

$$\mathbf{x}^* = \arg \max_{\mathbf{x}} P(\mathbf{x}|\mathcal{R}, \mathbf{y}), \quad (1)$$

where \mathcal{R} is a set of typical neighbourhood configurations and comprises the labels and spatial relationships of regions that are within some radius r of a region, $\mathbf{y} = \{y_1, y_2, \dots, y_N\}$ is a set of description vectors y_i which are learnt from a set of training data and incorporate regions' size, shape, location, colour and texture features. Each region i is associated with a unary vector y_i . The probability of $P(\mathbf{X}=\mathbf{x})$ will be referred to as $P(\mathbf{x})$. By making the assumption that the set of unary features \mathbf{y} and set of prototypes \mathcal{R} are independent, the conditional probability in (1) can be written as:

$$P(\mathbf{x}|\mathcal{R}, \mathbf{y}) \propto P(\mathbf{y}|\mathbf{x}) P(\mathbf{x}|\mathcal{R}). \quad (2)$$

In section 3.1 we present our method for learning the set of prototypes and utilising them to estimate the joint probability distribution of $P(\mathbf{x}|\mathcal{R})$. In section 3.2 we describe our method for estimating $P(\mathbf{y}|\mathbf{x})$ and finally in section 3.3 we present the inference method for finding the labelling that maximises the joint distribution defined by equation (1).

3.1. Prototype discovery and estimating $P(\mathbf{x}|\mathcal{R})$

We assume that the set of random variables \mathbf{X} constitute an MRF. Thus,

$$P(x_i|x_{S-i}) = P(x_i|x_{\mathcal{N}_i}), \quad (3)$$

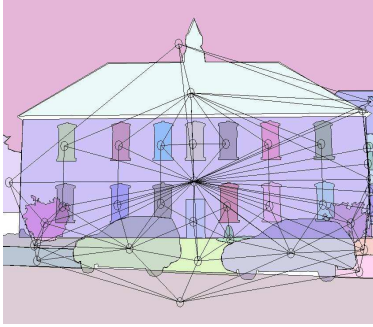
where S is the set of indices defined over segmented regions and $x_{\mathcal{N}_i}$ denotes the regions which are within the neighbourhood of the focal region i . In this paper, we define the Markov model over the random variables so that a region i not only depends on the labels of its neighbourhood regions, but also on the relative spatial relationships with its neighbourhood regions. This differs from other MRF models used in computer vision, such as the Potts model which only considers the labels of the neighbourhood regions or pixels. For incorporating relative spatial relationships in the MRF model, we use the method proposed in [9] and [8]:

$$P(x_i = l | \mathcal{N}_i, \mathcal{R}) = \frac{1}{Z} \exp(-\varphi(\mathcal{N}_i, \mathcal{R}_l)), \quad (4)$$

where \mathcal{N}_i is the *neighbourhood configuration* or simply *configuration* of region i , which comprises the labels and spatial relationships of regions that are within some radius r of region i and Z is a normalising constant. Function $\varphi(\mathcal{N}_i, \mathcal{R}_l)$ is the minimum distance between the neighbourhood configuration \mathcal{N}_i and a set of prototypes \mathcal{R} that have l at their focal regions. Prototypes \mathcal{R} are a subset of configurations which are learnt from a set of manually segmented and labelled regions. In figure 1 a hand segmented image of a building scene and its corresponding graphical model which we employ in the proposed probabilistic model are shown.



(a)



(b)

Figure 1. a) Original Image. b) Hand segmented regions and their corresponding graphical model (representing the MRF) which we employ in the proposed probabilistic framework.

3.1.1 Neighbourhood Configuration:

To determine the neighbourhood configuration of a region, we follow the same process as the one described in [9]. A pair of regions are considered neighbours if they are within a distance r from each other. After identifying a region's neighbourhood, spatial relationships between the focal region and its neighbouring regions are estimated. We define five different relations between region pairs. These are: relative vertical orientation, relative horizontal orientation, containment relation and the ratio of their widths and heights. The procedure for estimating these relationships is described in the following.

Vertical and Horizontal Relationships: Let p_{c_i} and p_{n_i} be points from a pair of regions, with subscript c indicating the points which belong to the central (focal) region and subscript n is indicating the points which belong to the neighbouring regions. First, the angle ϕ_i between vector $p_{n_i} - p_{c_i}$ and the reference direction (horizontal axis) is measured. The degree of aboveness (or belowness) of p_n with respect to p_c is then computed as:

$$f_{v_i}(p_{n_i}, p_{c_i}) = \sin \phi_i, \quad (5)$$

where f_{v_i} represents the vertical relationship of a point pair and reaches its maximum when p_n is exactly above point

p_c . For the horizontal relationship, we define:

$$f_{h_i}(p_{n_i}, p_{c_i}) = \cos \phi_i. \quad (6)$$

To represent the vertical and horizontal relations between two regions, the averages over point-wise membership values: $f_{vr} = \frac{1}{K} \sum_{i=1}^K f_{v_i}$ and $f_{hr} = \frac{1}{K} \sum_{i=1}^K f_{h_i}$ are computed. To be computationally efficient we consider randomly chosen subset of pairs.

Containment Relationships: To express whether region A includes region B , the following function is used:

$$f_{cr}(A, B) = \begin{cases} -1 & \text{if region } A \text{ contains region } B \\ +1 & \text{if region } B \text{ contains region } A \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Function f_{cr} is set to zero when the relationships of the two regions is neither "contains" nor "contained in".

Width and Height Relationships: The width ratio is the ratio between the width of region A and that of region B . Similarly, the height ratio is the ratio between the height of region A and that of region B . The width ratio is formulated as:

$$f_{wr}(A, B) = \begin{cases} 1 - w_A/w_B & \text{if } w_B/w_A \geq 1 \\ w_B/w_A - 1 & \text{otherwise.} \end{cases} \quad (8)$$

where w represents the width of a region's bounding box, and similarly for the height ratio. The formulation is intended to make the values of the ratios fall in the range $[-1, 1]$.

A *neighbourhood configuration*, \mathcal{N}_i , consists of the label of the focal region, the labels of the neighbours and the spatial relationships between the focal region and its neighbouring regions, i.e. $f_{vr}, f_{hr}, f_{cr}, f_{wr}, f_{hr}$. In our implementation, the neighbourhood configurations are encoded by $6 \times F$ matrices with F being the number of regions within a neighbourhood. Each column of this matrix is associated with one of the neighbouring regions and it encodes the region's label (1 component) and its spatial relationships (5 components) with respect to the focal region.

3.1.2 Prototype Discovery

The aim of prototype discovery is to identify for each label, a small set of typical neighbourhood configurations, or *prototypes*. A prototype, \mathcal{R}_l , is a neighbourhood configuration with label l at its focal region. In order to identify the prototypes, the set of configurations, which have been extracted from the *manually* annotated images, are first partitioned according to the label of their focal regions and subsequently each partition is clustered using the *k-medoids* algorithm. This clustering algorithm is based on the pair-wise

distances between the configurations' respective matrix representations. The distance between any two configurations is estimated using the pseudo-metric function described in [9].

After applying the *k-medoids* algorithm to a set of neighbourhood configurations, the centroids are considered as the prototypes. One example of an estimated prototype is shown in figure 2. For a training set in which a region may be assigned to M possible classes, a set \mathcal{R} of size $M \times k$ prototypes is estimated. In this set k prototypes are associated with each of the classes.

Prototype R
Focal Region Window

Label	Door	Facade	Window	Others
Vertical	-0.9	0.9	-0.6	0.3
Horizontal	0.1	0.3	-0.8	0.9
Containment	0	1	0	1
Width Ratio	0.9	0.2	1	0.9
Height Ratio	0.8	0.1	1	0.2

Figure 2. Example of an estimated prototype.

3.1.3 Estimating $P(\mathbf{x}|\mathcal{R})$

For estimating the joint probability of a labelling \mathbf{x} , $P(\mathbf{x}|\mathcal{R})$, we first determine a coding (colouring) of the set of regions. We use the *greedy colouring* strategy proposed in [8]. Because of the assumption of Markovianity (equation (3)), the likelihood over vertices with the same colour reduces to the product of the respective conditional probabilities. For estimating $P(\mathbf{x}|\mathcal{R})$ we use the pseudo-likelihood function introduced in [8]:

$$P(\mathbf{x}|\mathcal{R}) \approx \frac{1}{N} \sum_j |\mathcal{C}_j| \left[\prod_{i \in \mathcal{C}_j} P(x_i | \mathcal{N}_i, \mathcal{R}) \right]^{\frac{1}{|\mathcal{C}_j|}}, \quad (9)$$

where \mathcal{C}_j is the set of regions which are in the same colouring.

3.2. Estimating $P(\mathbf{y}|\mathbf{x})$

In order to estimate $P(\mathbf{y}|\mathbf{x})$ we know that vector descriptors y_i are conditionally independent from each other, given x_i . Therefore, $P(\mathbf{y}|\mathbf{x})$ can be written as:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^N P(y_i | x_i). \quad (10)$$

Furthermore, from Bayes theorem we know that $P(y_i | x_i) \propto P(x_i | y_i)$. To estimate $P(x_i | y_i)$ we use

a one-vs-all Adaboost classifier. We first computed a vector of appearance features using low level descriptors. For each segmented region we computed a vector of 63 dimensions, y_i , incorporating region size, location, shape and texture. These features are similar to a subset of those proposed in [7] and consist of the mean, standard deviation, skewness and kurtosis statistics over each segmented region of:

- RGB colour components-12
- HSV colour components-12
- Texture features computed from 8 filter responses. The filters were Gabor filters with 8 different orientations and a fixed scale. These were applied to the average of the RGB colour bands. -32.

In addition, we compute the normalised size (the area of the segmented region divided by the area of the image), convexity, eccentricity, and the ratio of the width and height of the enclosing bounding box, for describing the size and shape of each region. Also, we append the location of each region to the vector descriptor by including x and y offsets of the centre of each region from the image centre. We normalise these offsets to have values between -1 and 1 .

We used hand segmented images to train a set of one-vs-all Adaboost classifiers, A_l , for each class label l . For training an Adaboost classifier of label l , we consider all of the regions in our training set with label l as positive examples ($+1$) and the remaining regions as negative examples (-1). For estimating $P(x_i | y_i)$, after computing the vector descriptor y_i of a region, we use the trained Adaboost classifiers and estimate $P(x_i | y_i)$ using:

$$P(x_i = l_k | y_i) = \frac{\exp(A_{l_k}(y_i))}{\sum_{r=1}^M \exp(A_{l_r}(y_i))}. \quad (11)$$

3.3. Inference

Given a new set of segmented regions, the aim is to assign a label to each of the regions so that it maximises the joint probability distribution described by equation (2), or minimises the cost function which is defined by $-\log(P(\mathbf{x}|\mathbf{y}, \mathcal{R}))$. Having replaced (9) and (10) into the right hand side of equation (2) and taking the $-\log$ of the two sides, the cost function of a labelling \mathbf{x} can be written as:

$$E(\mathbf{x}) = \sum_{i=1}^N -\log(P(x_i | y_i)) - \log\left(\frac{1}{N} \sum_j |\mathcal{C}_j| \left[\prod_{i \in \mathcal{C}_j} P(x_i | \mathcal{N}_i, \mathcal{R}) \right]^{\frac{1}{|\mathcal{C}_j|}}\right), \quad (12)$$

and for inference, we are looking for a labelling \mathbf{x}^* which minimises $E(\mathbf{x})$, i.e.

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} E(\mathbf{x}). \quad (13)$$

	Fac	Chi	Doo	Dor	Roo	Sta	Sky	Veg	Gro	Balc	Win	Oth
Facade	244	0	3	0	6	0	0	7	6	4	22	0
Chimney	0	32	0	4	4	0	0	0	0	0	25	0
Door	3	0	53	0	0	0	0	1	1	0	65	1
Dormer	1	1	0	28	5	0	0	0	0	4	31	0
Roof	4	1	0	2	100	0	0	0	0	4	9	0
Stairs	2	0	3	0	0	12	0	2	0	2	15	0
Sky	0	2	0	1	1	0	72	0	0	0	1	0
Vegetation	11	2	1	0	5	0	1	103	3	1	10	1
Ground	6	0	2	0	0	0	0	7	77	0	5	4
Balcony	1	3	0	6	7	0	0	3	5	121	55	0
Window	2	8	25	3	3	0	0	14	1	18	1273	2
Others	1	0	0	0	0	0	0	1	0	3	4	59

Table 1. Confusion matrix obtained from the Adaboost classifier.

We optimise stochastically using the MCMC technique. We first assign random labels to each of the segmented regions by drawing labels from a uniform distribution. We then run the MCMC procedure to sample and minimise the cost function defined in equation 12.

4. Experiments

This section presents an experimental evaluation of the proposed labelling technique. The proposed approach is evaluated using manually segmented images. We start by describing the dataset used for training and testing. In section 4.2 we present our experimental results and compare the proposed labelling framework with a unary based classifier and the “context first” classifier of [9].

4.1. Dataset

We used 372 annotated images of the 12-class dataset of the eTRIMS repository¹ in our experimental evaluation. These images were collected from different European cities including London, Prague, Barcelona, Hamburg, Bonn and Basel and thus encompass a variety of different architectural styles. The object classes which we used in our experiments include, “window”, “door”, “facade”, “balcony”, “dormer”, “stairs”, “chimney”, “roof”, “ground”, “sky”, “vegetation” and class “others” which contains objects like “pedestrians” and “cars”. For training we employed 310 annotated images and for testing we used the remaining 62 images. The training data were used for learning both the neighbourhood prototypes and parameters of the Adaboost classifiers.

4.2. Evaluation

We compared the proposed method with both a multi-class Adaboost classifier, which was trained with the unary features discussed in section 3.2 and the “context first” model proposed in [9]. Tables 1,2 and 3 show the confusion matrices obtained from each of these classifiers. In all three matrices, objects “door” and “window” have been confused most frequently.

	Fac	Chi	Doo	Dor	Roo	Sta	Sky	Veg	Gro	Balc	Win	Oth
Facade	213	1	12	2	3	0	4	24	13	0	9	11
Chimney	2	42	0	1	3	0	6	3	0	0	8	0
Door	2	3	57	0	0	1	0	3	4	0	37	17
Dormer	0	9	0	44	2	0	6	0	0	0	9	0
Roof	1	3	0	5	68	0	30	1	3	1	7	1
Stairs	1	0	1	0	0	17	0	3	7	3	0	4
Sky	1	3	0	0	7	0	61	1	1	0	2	1
Vegetation	21	4	10	0	4	7	5	27	15	11	9	25
Ground	4	0	3	0	0	8	0	1	74	1	0	10
Balcony	2	2	2	5	12	0	0	18	4	128	17	11
Window	15	56	97	19	9	10	22	42	24	41	1003	11
Others	3	0	7	0	0	10	0	3	19	1	2	23

Table 2. Confusion matrix obtained from the “context first” algorithm proposed in [9].

	Fac	Chi	Doo	Dor	Roo	Sta	Sky	Veg	Gro	Balc	Win	Oth
Facade	246	0	2	0	5	0	0	11	6	4	17	1
Chimney	0	38	0	2	4	0	0	0	0	0	21	0
Door	1	0	58	0	0	0	0	1	0	0	63	1
Dormer	0	0	0	42	2	0	0	0	0	3	23	0
Roof	2	1	0	3	107	0	0	0	0	0	7	0
Stairs	0	0	4	0	0	17	0	4	1	2	8	0
Sky	0	2	0	1	0	0	72	0	0	0	2	0
Vegetation	10	3	3	0	7	0	1	96	4	2	10	2
Ground	4	0	3	0	0	0	0	7	79	0	5	3
Balcony	1	0	0	3	6	0	0	0	5	155	31	0
Window	2	3	24	1	3	0	0	8	0	10	1296	2
Others	0	0	0	0	0	0	0	1	1	2	4	60

Table 3. Confusion matrix obtained from the proposed CRF model.

We also report the per-class accuracy of each of the classifiers (labelling algorithms) in figure 3. From this figure it can be seen that the “context first” approach outperforms the other two approaches for labelling instances of “chimney”, “dormer” and “stairs”. These are objects which tend to have a unique neighbourhood configuration in a scene. For instance, a region which is at the top of a building within the roof and below the sky it is most probably a chimney. However, for labelling regions which associate with objects like “vegetation”, context cannot play a significant role, as these regions usually do not construct a discriminative neighbourhood configuration in an image. The average per-class and overall accuracy of each of the three labelling approaches are also listed in table 4. In general, we can see from the table that, at least for this application, the integration of unary features and contextual features in a single framework outperforms both the unary based classifier and the context-based classifier.

	Context First	Adaboost	CRF model
Overall Accuracy	66.5%	82.3%	85.8%
Average Per-Class Accuracy	57.8%	68.1%	74.5%

Table 4. Average per-class and overall accuracy of the three different approaches.

¹<https://www.ipb.uni-bonn.de/svn/etrim-img-dbs/>

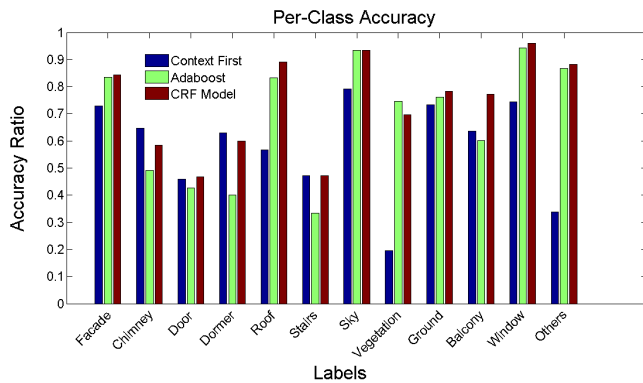


Figure 3. The per-class accuracy of different labelling approaches.

5. Conclusions

We proposed a conditional random field model for labelling parts of building scenes, like chimneys, windows, doors, etc. The interaction term of the CRF model was estimated using a set of typical neighbourhood configurations which were learnt from training data. We defined a cost function based on MAP estimation that combines unary measurements and contextual information and which was minimised effectively using MCMC. The proposed model combines unary measurements and contextual information in a unified framework and outperforms classifiers that used only contextual information or only unary measurements. This paper was more concerned with high level vision and we did not apply our method to automatically segmented images. To extend the current model and also for comparing our results with the other contextual models like [6] which spatial context has been used, in future we are aiming to apply our model to automatically segmented regions. As segmentation is actually a significant problem in its own right, ideally, segmentation and labelling should be integrated in a system with feedback loops.

References

- [1] E. R. A. Hanson. Visions: A computer vision system for interpreting scenes. pages 303–334, 1978.
- [2] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.
- [3] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorisation with bags of keypoints. In *ECCV*, 2001.
- [4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.
- [5] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973.
- [6] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorisation using co-occurrence, location and appearance. In *CVPR*, 2008.
- [7] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *IJCV*, 80(3):300–316, 2008.
- [8] D. Heesch and M. Petrou. Markov random fields with asymmetric interactions for modelling spatial context in structured scenes. *Journal of Signal Processing Systems*, 2009.
- [9] D. Heesch, R. Tan, and M. Petrou. Context first. In *Proc Int'l Workshop on Structural and Syntactic Pattern Recognition*, 2008.
- [10] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.
- [11] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.
- [12] S. Kluckner, T. Mauthner, P. Roth, and H. Bischof. Semantic image classification using consistent regions and individual context. In *BMVC*, 2009.
- [13] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005.
- [14] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009.
- [15] A. Oliva and A. Torralba. Modelling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.
- [16] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.
- [17] J. Sivic and A. Zisserman. Video data mining using configurations of viewpoint invariant regions. In *CVPR*, 2004.
- [18] T. Strat and M. Fischler. Context-based vision: recognising objects using information from both 2d and 3d imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1050–1065, 1991.
- [19] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, 2003.
- [20] A. Torralba, K. Murphy, W. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *ICCV*, 2003.
- [21] A. Torralba, K. P. Murphy, and T. Freeman. Sharing features: Efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.
- [22] P. Viola and M. Jones. Robust real-time face detection. In *ICCV*, 2001.
- [23] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, 69(2):251–261, 2006.