Context first

Daniel Heesch, Robby Tan and Maria Petrou

Imperial College London, SW7 2AZ, UK

Abstract. We propose a probabilistic model that captures contextual information in the form of typical spatial relationships between regions of an image. We represent a region's local context as a combination of the identity of neighbouring regions as well as the geometry of the neighbourhood. We subsequently cluster all the neighbourhood configurations with the same label at the focal region to obtain, for each label, a set of configuration *prototypes*. We propose an iterative procedure based on belief propagation to infer the labels of regions of a new image given only the observed spatial relationships between the regions and the hitherto learnt prototypes. We validate our approach on a dataset of hand segmented and labelled images of buildings. Performance compares favourably with that of a boosted, non-contextual classifier.

1 Introduction

Object recognition in general scenes remains a formidable task for artificial systems. Part of the difficulty stems from the way humans categorise things: it is first and foremost shared functional or causal characteristics that define most classes of interest, not similarity in appearance [19]. The problem is compounded by the observation that even very similar objects may look very different under different viewing angles and partial occlusion. Amongst the most successful approaches is that of modelling objects in terms of large sets of discriminant keypoints in conjunction with representations that are invariant with respect to several types of transformation (e.g. [14], [15], [16], [22], [8], [25] and [20]). These appearance-based models have in common that the number of classes to be distinguished is very small and that images contain objects only of one class. It is not clear how to scale to the several thousands of categories humans discriminate without effort.

Promising alternatives are hierarchical models in which features are allowed to be shared between different classes thus reducing the computational burden over flat models, e.g. [9]. Another route is to employ information about other objects of the same scene or information about the type of scene. Such contextual information may reduce the space of plausible object hypotheses and suggest a smaller set of dedicated non-contextual classifiers. It is known that the gist of the scene or the relationships between objects can be captured by the low-frequency content of an image [18]. It has also been shown in neuro-physiological studies that low-frequency information is processed relatively early during visual recognition [2]. Combining these two observations suggests that context may play a pivotal role as an early facilitator during visual recognition.

While other authors have explored the use of scene information for object recognition [23] or region information for scene classification [4], [5], [10], our work investigates the question how and to what extent local geometric and topological relationships between objects can be exploited for object classification. Our approach is motivated by the discovery in [1] of cortical 'context networks that have been implicated in the storage of typical configurations of objects. Spatial relationships can arguably be extracted more easily than specific details of individual regions. Their extraction is also virtually insensitive to photometric variation. It may therefore not be surprising that these spatial networks also exhibit early activation during visual recognition tasks [1].

In contrast with other contextual models, we do not believe that contextual information should only be used to resolve tension between conflicting non-contextual evidence, e.g. [12] and more recently [21] in the context of probabilistic relaxation. Rather, we believe that context on its own can get us a long way and indeed may be the crucial ingredient to make object recognition scalable.

This paper makes three contributions: (i) we propose a fuzzy representation of the local neighbourhood of a region and a method to obtain *typical* neighbourhood configurations or *prototypes* (ii) we propose a way to use these prototypes in the formulation of a random field over image regions; (iii) we provide an optimisation technique based on belief propagation to relax the random field.

The paper is structured as follows. Section 2 describes related work. In Section 3 we formulate the graphical model, formalise the set of spatial relationships used and describe how representative configurations, or prototypes, are obtained from a training set of annotated images. Section 4 explains how inference is performed. Section 5 describes our experiments. Section 6 ends the paper with a discussion.

2 Related work

Several contextual models have been formulated that are concerned with dependencies between objects (as opposed to hierarchical dependencies). Amongst the probabilistic models, Markov random fields are the most popular, e.g. [17], [6], [13], [11], [20]. The authors in [11] and [20] define a conditional random field over individual pixels. In [20], contextual information is incorporated by using the joint boosting algorithm [24] for learning potential functions. Neither work explicitly considers spatial relationships, although [11] includes the absolute position of a site in the potential function.

In [6], each image is assumed to be associated with a bag of words and the precise term-region associations have to be learnt from training data. The Markov random field is specified through single and pair-wise clique potential functions which are learnt on the assumption that they are symmetric. 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 [17], a graphical model is defined over image regions by specifying the clique functions for all types of single and pair-wise cliques. The potential functions are weighted sums of basis functions with the parameters being set manually. Our work has the same objectives as those of [6] and [17]. What sets it apart is that we allow neighbouring regions to influence each other differently depending on their relative spatial positions and topological relationships. This added complexity is best handled by specifying the field in terms of local conditional probability distributions which are obtained empirically from a training set.

3 Spatial Context Model

Let $S = \{1, \ldots, N\}$ index a set of regions in an image. We assume that each region is associated with a random variable x_i which takes its value from a discrete set of class labels. The neighbourhood configuration of the ith region, $\mathcal{N}_i$, comprises the labels and spatial relationships of regions that are within some radius r of the focal region. We define the probability with which label l is assigned to region i as

$$P(x_i = l|\mathcal{N}_i) \equiv \frac{1}{Z} \exp(-\psi(\mathcal{N}_i, R_l)), \tag{1}$$

where Z is a normalising constant, and $\psi(\mathcal{N}_i, R_l)$ is a function that measures the distance between neighbourhood configuration $\mathcal{N}_i$ and the set of prototypes with label l at their focal region. The next section will describe the set of relations we use. The subsequent section explains how ψ is defined and how prototypes R_l are obtained.

3.1 Pairwise Relations

Spatial relationships are often modelled such that only one relationship holds between any two objects, e.g. [21] and [5] in the context of scene understanding. The representational convenience of crisp relations comes at the cost of increased sensitivity to errors with respect to the spatial localisation and geometry of the input data. We believe that much can be gained by modelling relationships as fuzzy concepts [3]. A fuzzy relation holds to a variable *degree* determined by a membership function associated with that relation. We here consider five relations between region pairs. These are their relative vertical orientation, their relative horizontal orientation, their containment relation, and the ratio of their widths and heights, respectively. These are defined as follows.

Vertical and Horizontal Relationships. Let p_{cnt_i} and p_{n_i} be points from each of the two regions. We measure the angle ϕ_i between vector $(p_{n_i} - p_{cnt_i})$ and the unit vector $(-1, 1)^T$. The degree of aboveness (or belowness) of p_n with respect to p_{cnt} is then computed as

$$f_{v_i}(p_{n_i}, p_{cnt_i}) = \sin \phi_i \tag{2}$$

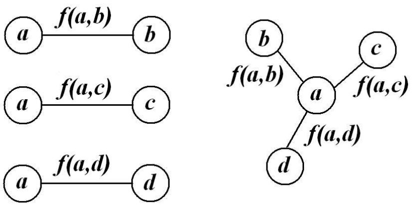

Fig. 1. The pictorial description of pairwise connections and a configuration, where a, b, c, d may correspond to wall, sky, roof and door, respectively. $f(a, b)$ in the diagram is a vector that consist of all components of the relationships between regions a and b.

where f_{v_i} represents the vertical relationship of a point pair. Similarly, we represent the horizontal relationship between point pairs as

$$f_{h_i}(p_{n_i}, p_{cnt_i}) = \cos\phi_i. \tag{3}$$

To represent the vertical and horizontal relationship between two regions, we compute the average over point-wise membership values: $f_v = \frac{1}{N}\sum_i^N f_{v_i}$ and $f_h = \frac{1}{N}\sum_i^N f_{h_i}$. To be computationally efficient, we generate (p_{n_i}, p_{cnt_i}) randomly within the respective regions.

Containment Relationships To measure whether region r_n includes region r_{cnt}, we are guided by the following decision rule:

$$f_{ct}(r_n, r_c) = \begin{cases} -1 & \text{if } (r_n) \text{ contains } (r_{cnt}) \\ +1 & \text{if } (r_n) \text{ is contained in } (r_{cnt}) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Width and Height Relationships We define these as the ratios between the widths and heights, respectively, of region r_n and those of region r_{cnt}. In our particular application domain, these relationships are useful to distinguish between, for example, roofs and chimneys, which are indistinguishable under all other relations. Formally, the width ratio is

$$f_{wr}(r_n, r_c) = \begin{cases} 1 - w_{cnt}/w_n & \text{if } w_n/w_{cnt} \geq 1 \\ w_n/w_{cnt} - 1 & \text{otherwise} \end{cases} \tag{5}$$

where w represents the width of a region with respect to its principal axis. The height ratio is defined analogously.

The spatial relationship between two regions can then be modeled as a vector with values in $[-1, 1]$ each component of which is the membership value for the corresponding relation.

3.2 Configurations and Prototypes

A neighbourhood configuration consists of the labels of the neighbours and their spatial relationships with respect to the focal region. Formally, it is an ordered

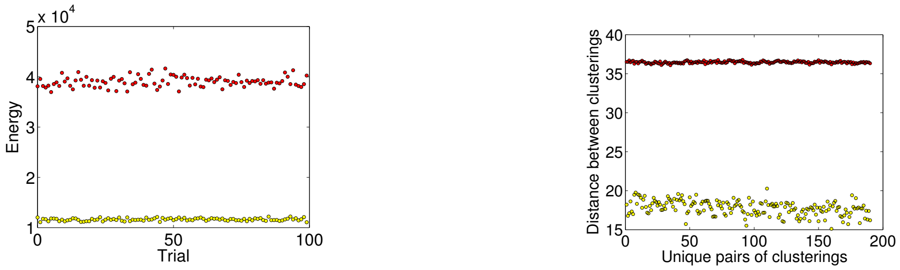

Fig. 2. Left: Value of the energy function at initialisation (top) and after convergence (bottom). Right: distances between clusterings for random clusters (top) and those obtained from the real data (bottom).

set of relationship vectors with each vector being associated with a particular label. See Figure 1 for an illustration.

Prototypes The purpose of the next step is to identify for each region label a small set of typical neighbourhood configurations, or prototypes. This is accomplished by clustering all those configurations that have the same label at the focal region. Clustering is based on the pair-wise distances between the configurations' respective matrix representations as described below.

Let P and Q denote the relation matrices of two configurations A and B of size M and N, respectively. Let the labels of the regions be represented by vectors p and q, respectively. For each region of configuration A, we determine its distance from all those regions of configuration B that bear the same label. This distance is computed by applying the l_1 metric to the respective row vectors of P and Q. We consider the closest region as the best match to the region of configuration A, add the distance to our overall cost and exclude the matching region from all subsequent comparisons. If configuration B does not have any region of that label, a fixed cost is applied to penalise label discrepancies. This is repeated for all other configurations of region A. The overall cost reflects both differences in the labels as well as differences in the geometry and topology of regions carrying the same label.

We employ the k-medoid algorithm to cluster configurations. Like k-means, the algorithm is guaranteed to converge because the sum of the distances between all points and their respective cluster centroid cannot increase and is also bounded from below. However, like any gradient descent algorithm, the final solution depends on the initialisation and is thus not guaranteed to be the global optimum. To assess the stability of the solution, we run the algorithm several times and compare the energy before and after convergence. As Figure 2 indicates, the final energy remains within narrow bounds and suggests that the final solutions come close to the global optimum.

A similar energy upon convergence does not imply, however, that the clusterings are the same or similar. Let $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$ and $\mathcal{B} = \{B_1, B_2, \ldots, B_n\}$ denote two clusterings. The members of each clustering are themselves sets of indices denoting a particular configuration. To assess the quality of the result of the k-medoid algorithm, we run it several times on the real data and compute a distance measure for all pairs of clusterings thus obtained. We then generate random clusterings consisting of the same number of clusters and the same cluster size distribution as the real clustering, and measure their pair-wise distances. We compute the distance between two clusterings $\mathcal{A}$ and $\mathcal{B}$ as

$$\sum_{i=1}^{m} d(A_i, \mathcal{B}) = \sum_{i=1}^{m} \left(|A_i| - \max_j |A_i \cap B_j| \right) . \tag{6}$$

The distances between the random clusterings and the true clusterings are plotted in Figure 2 on the right. The plot demonstrates that the set of clusterings obtained by running k-medoid repeatedly on the same distance matrix are more similar to each other than a set of random clusterings. We take this as circumstantial evidence that k-medoid does capture intrinsic structure in the space of configurations.

Cluster centroids are those configurations for which the sum of the distances to all other members of the respective cluster is minimal. Prototypes correspond to the cluster centroids and thus are themselves configurations.

4 Inference

Given a set of regions in an image, we intend to label them using the prototypes we have generated. To arrive at correct labellings, we define a cost function that is based on the distance between the observed configurations and their closest prototype. Formally, we define the potential function as

$$\psi(\mathcal{N}_i, R_l) = \min_{R \in R_l} d(\mathcal{N}_i, R) \tag{7}$$

where $d(N_i, R)$ is the distance between a configuration and a prototype R as defined in Section 3.2. We intend to obtain the closest distance of all configurations from the corresponding prototypes, that is we want to minimise

$$E(\mathbf{x}) \equiv E(x_1 = l_1, ..., x_N = l_N) = \sum_{i \in S} \psi(\mathcal{N}_i, R_l) \tag{8}$$

In order to apply the technique of belief propagation, we change the undirected configurations into directed ones and generate a factor graph. Figure 3 shows an example of the transformation. We may subsequently use the following equations to optimise the cost function.

$$b(x_i) = \prod_{c \in N_i} m_{c \to i}(x_i) \tag{9}$$

$$m_{c \to i}(x_i) = \sum_{\{x_c\} - x_i} \psi \prod_{i' \in N_{c-i}} n_{i' \to c}(x_{i'}) \tag{10}$$

$$n_{i \to c}(x_i) = \prod_{c' \in N_{i-c}} m_{c' \to i}(x_i) \tag{11}$$

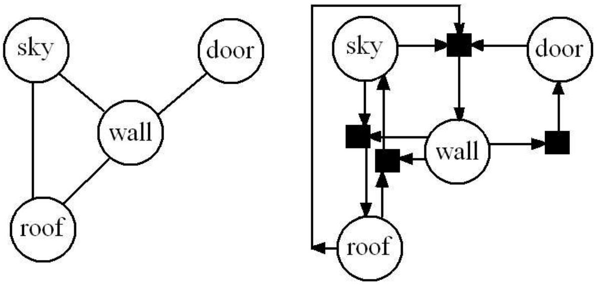

Fig. 3. The pictorial description of the transformation from an undirected graph into a directed graph.

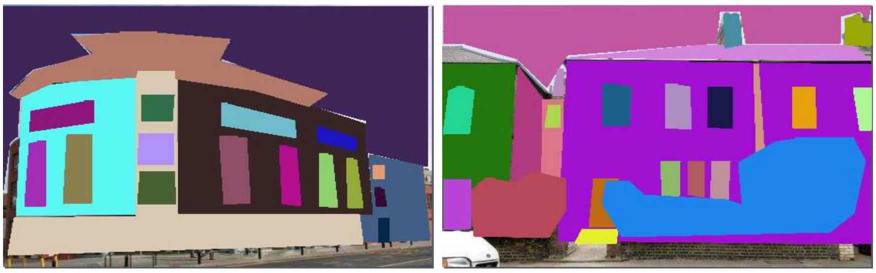

Fig. 4. Examples of manually segmented images

where $b(x_i)$ is the belief that x_i has a particular label. $m_{c \to i}(x_i)$ is the message from the neighbouring nodes of i with respect to the label of x_i. In Figure 3, $m_{c \to i}(x_i)$ is represented by the black boxes that connect the regions or nodes. Since the configurations in the factor graph are directed configurations, we may simplify the equations thus

$$b(x_i) = m_{c \to i}(x_i) \tag{12}$$

$$m_{c \to i}(x_i) = \sum_{\{x_c\} - x_i} \psi \prod_{i' \in N_{c-i}} n_{i' \to c}(x_{i'}) \tag{13}$$

$$n_{i \to c}(x_i) = m_{c' \to i}(x_i) \tag{14}$$

5 Evaluation

The image collection used for training and testing consists of photographs depicting buildings from different cities of several countries, mostly taken from street-level. Each image was manually segmented and labeled with one of the following nine classes: ‘window’, ‘chimney’, ‘roof’, ‘door’, ‘wall’, ‘stairs’, ‘pipe’, ‘sky’, and ‘vegetation’. Figure 4 shows two examples. The training set contains 197 images with a total of 3,675 regions. The test set comprises 80 images with a total of 1,372 regions. Images were randomly assigned to one of the two sets.

We shall note that we do not address the challenge of segmenting regions automatically. Instead, we assume that for learning and testing regions have been manually segmented. The recent work of [5] indicates that one may profitably start with an automated segmentation when the regions to be separated are visually distinct (e.g. material types like water, sky, sand). In our case, however, a general-purpose segmentation routine is unlikely to achieve sufficient accuracy for our contextual model.

Table 1 shows the confusion matrix for the proposed method. We compare performance with a non-contextual AdaBoost classifier which is trained to find the optimal combination of unary attributes pertaining to the shape of regions including the ratio of principal axes, compactness, variances, and elliptical variances. The confusion matrix for the labelling based on the adaboost algorithm is shown in Table 2. The total error is 576 (42%) compared to 473 (34%) for the contextual labelling.

	Win	Chi	Roo	Doo	Wal	Dor	Sta	Pip	Sky	Veg
Window	487	11	9	77	2	11	1	1	23	17
Chimney	10	70	1	0	0	24	0	0	2	0
Roof	18	0	31	0	0	2	1	0	31	8
Door	27	3	0	84	0	9	1	0	1	10
Wall	30	8	2	5	45	3	1	5	10	17
Dormer	7	1	0	1	0	2	0	0	3	1
Stairs	0	0	0	0	0	0	19	0	0	1
Pipes	12	16	0	2	3	0	0	48	1	1
Sky	11	2	1	0	0	0	0	1	79	0
Vegetation	7	2	2	3	1	1	6	5	1	35

Table 1. Confusion matrix for contextual classification (rows: true labels; columns: hypothesised labels).

	Win	Chi	Roo	Doo	Wal	Dor	Sta	Pip	Sky	Veg
Window	435	56	9	49	4	78	4	0	1	4
Chimney	71	21	2	7	0	5	0	0	1	0
Roof	5	1	45	0	0	2	0	1	37	0
Door	71	3	0	53	3	5	0	0	0	0
Wall	2	9	3	1	83	10	1	2	14	1
Dormer	7	3	3	0	0	2	0	0	0	0
Stairs	15	1	2	0	0	2	0	0	0	0
Pipes	0	0	0	0	0	4	0	79	0	0
Sky	7	1	6	0	5	1	0	0	72	2
Vegetation	10	0	10	3	19	4	0	3	6	8

Table 2. Confusion matrix for boosting on shape descriptors (rows: true labels; columns: hypothesised labels).

6 Discussion

In most computer vision applications that take into account contextual information, regions are initially assigned labels on the basis of their unitary attributes, and the label assignment is subsequently refined through the use of contextual information (e.g. [7]). Inspired by neuro-physiological findings about human visual processing, we here advocate the view that the order in which information is processed ought to be reversed. When humans view a scene, they first view it as a whole before focussing on particular details that merit further interpretation. Anecdotal evidence supporting this idea is the preference people show for seats in trains that look forward over seats that look backward in relation to the travelling direction. When looking backwards, one sees first the detail and then the context of the object; when looking forward, the global picture is captured first.

We proposed a probabilistic model in which a region's local neighbourhood is represented in the form of fuzzy relationship matrices. Because of the inability to define cliques in directed graphical models, the joint label probability over all regions cannot be expressed as a Gibbs distribution. Instead, we defined local conditional probabilities of a region's label that depend on the neighbourhood through a potential function that takes into consideration differences in the geometry of and the labels present in the neighbourhood. Central to our approach is the idea of typical neighbourhood configurations or prototypes which are obtained from a training set through clustering. Every label is associated with a small number of prototypes and inference aims to find a labelling of all regions such that the observed configurations are close the labels' closest prototypes. Comparison with a non-contextual Adaboost classifier trained on a variety of shape features support the view that contextual information can provide powerful information for labelling structured scenes.

Acknowledgements: This work was supported by the EU project eTRIMS.

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. I Bloch. *Fuzzy Representations of Spatial Relations for Spatial Reasoning*. John Wiley & Sons, 2007.
4. M Boutell, C Brown, and J Luo. Learning spatial configuration models using modified Dirichlet priors. In *Workshop on Statistical Relational Learning*, 2004.
5. M Boutell, J Luo, and C Brown. Factor graphs for region-based whole-scene classification. In *Proc IEEE Conf Computer Vision and Pattern Recognition, Semantic Learning Workshop*, 2006.
6. 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.

7. W J Christmas, J Kittler, and M Petrou. Structural matching in computer vision using probabilistic relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):749–764, 1995.
8. G Csurka, C Bray, C Dance, and L Fan. Visual categorization with bags of keypoints. In *Proc European Conf Computer Vision*, 2004.
9. S Fidler, M Boben, and A Leonardis. Similarity-based cross-layered hierarchical representation for object categorization. In *Proc Int'l Conf Computer Vision and Pattern Recognition*, 2008.
10. D Gökalp and S Aksoy. Scene classification using bag-of-regions representations. In *Proc IEEE Conf Computer Vision and Pattern Recognition, Beyond Patches Workshop*, 2007.
11. X He, R Zemel, and D Ray. Learning and incorporating top-down cues in image segmentation. In *Proc European Conf Computer Vision*, 2006.
12. J Kittler and S Hancock. Combining evidence in probabilistic relaxation. *Journal of Pattern Recognition and Artificial Intelligence*, 3:29–51, 1989.
13. S Kumar and H Hebert. Discriminative random fields: a discriminative framework for contextual interaction in classification. In *Proc Int'l Conf Computer Vision*, 2003.
14. D Lowe. Object recognition from local scale-invariant features. In *Proc Int'l Conf Computer Vision*, pages 1150–1157, 1999.
15. D Lowe. Distinctive image features from scale-invariant keypoints. *Int'l Journal of Computer Vision*, 60:91–110, 2004.
16. 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.
17. 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.
18. 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.
19. M Petrou. Learning in Computer Vision: some thoughts. *Progress in Pattern Recognition, Image Analysis and Applications. The 12th Iberoamerican Congress on Pattern Recognition, CIARP 2007*, Vina del Mar-Valparaiso, November, L Rueda, D Mery and J Kittler (eds), LNCS 4756, Springer, pages 1–12, 2007.
20. 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.
21. A Singhal, J Luo, and W Zhu. Probabilistic spatial context models for scene content understanding. In *Proc IEEE Conf on Computer Vision and Pattern Recognition*, 2003.
22. 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.
23. A Torralba. Contextual priming for object detection. *Int'l Journal of Computer Vision*, 52(2):169–191, 2003.
24. A Torralba, K Murphy, and W Freeman. Sharing fatures: efficient boosting procedures for multiclass object detection. In *Proc IEEE Conf Computer Vision and Pattern Recognition*, pages 762–769, 2004.
25. 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.