A Statistical Model for General Contextual Object Recognition

Peter Carbonetto[1], Nando de Freitas[1], and Kobus Barnard[2]

[1] Dept. of Computer Science, University of British Columbia
Vancouver, Canada
{pcarbo, nando}@cs.ubc.ca
[2] Dept. of Computer Science, University of Arizona
Tucson, Arizona
kobus@cs.arizona.edu

Abstract. We consider object recognition as the process of attaching meaningful labels to specific regions of an image, and propose a model that learns spatial relationships between objects. Given a set of images and their associated text (e.g. keywords, captions, descriptions), the objective is to segment an image, in either a crude or sophisticated fashion, then to find the proper associations between words and regions. Previous models are limited by the scope of the representation. In particular, they fail to exploit spatial context in the images and words. We develop a more expressive model that takes this into account. We formulate a spatially consistent probabilistic mapping between continuous image feature vectors and the supplied word tokens. By learning both word-to-region associations and object relations, the proposed model augments scene segmentations due to smoothing implicit in spatial consistency. Context introduces cycles to the undirected graph, so we cannot rely on a straightforward implementation of the EM algorithm for estimating the model parameters and densities of the unknown alignment variables. Instead, we develop an approximate EM algorithm that uses loopy belief propagation in the inference step and iterative scaling on the pseudo-likelihood approximation in the parameter update step. The experiments indicate that our approximate inference and learning algorithm converges to good local solutions. Experiments on a diverse array of images show that spatial context considerably improves the accuracy of object recognition. Most significantly, spatial context combined with a nonlinear discrete object representation allows our models to cope well with over-segmented scenes.

1 Introduction

The computer vision community has invested a great deal of effort toward the problem of recognising objects, especially in recent years. However, less attention has been paid to formulating an understanding of general object recognition; that is, properly isolating and identifying classes of objects (e.g. ceiling, polar bear) in an agent's environment. We say an object is *recognised* when it is labeled with a concept in an appropriate and consistent fashion. This allows us to propose a practical answer to the question of what is an object: an object is a semantic concept (in our case, a noun) in an image caption. Pursuing general object recognition may appear to be premature, given that good unconstrained object representations remain elusive. However, we maintain that a principled exploration using simple, learned representations can offer insight for further direction. Our approach permits examination of the relations between high-level computer vision and language understanding.

Ideally, we would train our system on images where the objects have been properly segmented and accurately labeled. However, the collection of supervised data by manually labeling semantically-contiguous regions of images is both time-consuming and problematic. We require captions at an image level, not at an image region level, and as a result we have large quantities of data at our disposal (e.g. thousands of Corel images with keywords, museum images with meta data, news photos with captions, and Internet photo stock agencies). Previous work shows that it is reasonable to use such loosely labeled data for problems in vision and image retrieval[1, 4, 13, 11, 2, 7]. We stress that throughout this paper we use annotations solely for testing — training data includes *only* the text associated with entire images. We do so at a cost since we are no longer blessed with the exact associations between objects and semantic concepts. In order to learn a model that *annotates*, *labels* or *classifies* objects in a scene, training implicates finding the *associations*, *alignments* or *correspondence* between objects and concepts in the data. As a result, the learning problem is unsupervised (or semi-supervised). We adapt the work in another unsupervised problem — learning a lexicon from an aligned bitext in statistical machine translation [9] — to general object recognition, as first proposed in [13].

The data consists of images paired with associated text. Each image consists of a set of blobs that identify the objects in the scene. A *blob* is a set of features that describes an object. Note that this does not imply that the scene is necessarily segmented, and one could easily implement scale-invariant descriptors to represent object classes, as in [14, 12]. Abstractly, a caption consists of a bag of semantic concepts that describes the objects contained in the image scene. For the time being, we restrict the set of concepts to English nouns (e.g. "bear", "floor"). See Fig. 1 for some examples of images paired with their captions.

We make three major contributions in this paper.

Our first contribution is to address a limitation of existing approaches for translating image regions to words: the assumption that blobs are statistically independent, usually made to simplify computation. Our model relaxes this assumption and allows for interactions between blobs through a Markov random

Fig. 1. Examples of images paired with their captions. A crucial point is that the model has to learn which word belongs with which part of the image.

field (MRF). That is, the probability of an image blob being aligned to a particular word depends on the word assignments of its neighbouring blobs. Due to the Markov assumption, we still retain some structure. One could further introduce relations at different scales using a hierarchical representation, as in [15].

Dependence between neighbouring objects introduces spatial context to the classification. Spatial context increases expressiveness; two words may be indistinguishable using low-level features such as colour (e.g. "sky" and "water") but neighbouring objects may help resolve the classification (e.g. "airplane"). Context also alleviates some of the problems caused by a poor blob clustering. For example, birds tend to be segmented into many parts, which inevitably get placed in separate bins due to their different colours. The contextual model can learn the co-occurrence of these blobs and increase the probability of classifying them as "bird" when they appear next to each other in a scene. Experiments in Sect. 4 confirm our intuition, that spatial context combined with a basic nonlinear decision boundary produces relatively accurate object annotations.

Second, we propose an approximate algorithm for estimating the parameters when the model is not completely observed and the partition function is intractable. Like previous work on detection of man-made structures using MRFs [16, 17], we use pseudo-likelihood for parameter estimation, although we go further and consider the unsupervised setting in which we learn both the potentials and the labels. As with most algorithms based on loopy belief propagation, our algorithm has no theoretical guarantees of convergence, but empirical trials show reasonably stable convergence to local solutions.

Third, we discuss how the contextual model offers purchase on the image segmentation problem. Segmentation algorithms commonly over-segment because low-level features are insufficient for forming accurate boundaries between objects. The object recognition data has semantic information in the form of captions, so it is reasonable to expect that additional high-level information could improve segmentations. Barnard *et al.* [3] show that translation models can suggest appropriate blob merges based on word predictions. For instance, high-level groupings can link the black and white halves of a penguin. Spatial consistency learned with semantic information smooths labellings, and therefore our proposed contextual model learns to cope with over-segmented images. In fact, with this model, a plausible strategy is to start with image grid patches and let segmentations emerge as part of the labeling process (see Fig. 6).

2 Specification of Contextual Model

First, we introduce some notation. The observed variables are the words $w_{n1}, \ldots, w_{nL_n}$ and the blobs $b_{n1}, \ldots, b_{nM_n}$ paired with each image (or document) n. M_n is the number of blobs or regions in image n, and L_n is the size of the image caption. For each blob b_{nu} in image n, we need to align it to a word from the attached caption. The unknown association is represented by the variable a_{nu}, such that $a_{nu} = i$ if and only if blob b_{nu} corresponds to word w_{ni}. The sets of words, blobs and alignments for all documents are denoted by w, b and a, respectively. Each w_{ni} represents a separate concept or object from the set $\{1, \ldots, W\}$, where W is the total number of word tokens.

Results in [11] suggest that representation using a mixture of Gaussians facilitates the data association task and improves object recognition performance. However, we retain the blob discretisation proposed by [13] because it scales better to large data sets and we will find model computation easier to manage. We use K-means to assign each blob b_{nu} in the feature space $\mathbb{R}^F$ to one of the B clusters. F is the number of features and B is the number of blob tokens.

The translation lexicon is a $B \times W$ table with entries $t(b^\star|w^\star)$, where $w^\star$ denotes a particular word token and $b^\star$ denotes a particular blob token. We define ψ to be a $W \times W$ table of potentials describing the "next to" relation between blob annotations. We define spatial context to be symmetric, so $\psi(w^\star, w^\diamond) = \psi(w^\diamond, w^\star)$. The set of model parameters is $\theta \triangleq \{t, \psi\}$. The set of cliques in document n is denoted by $\mathcal{C}_n$. The complete likelihood over all the documents is

$$p(b, a \mid w, \theta) = \prod_{n=1}^{N} \frac{1}{Z_n(\theta)} \prod_{u=1}^{M_n} \Phi(b_{nu}, a_{nu}) \prod_{(u,v) \in C_n} \Psi(a_{nu}, a_{nv}) \qquad (1)$$

where we define the translation and spatial context clique potentials to be

$$\Phi(b_{nu}, a_{nu}) = \prod_{i=1}^{L_n} t(b_{nu}, w_{ni})^{\delta(a_{nu}=i)}$$

$$\Psi(a_{nu}, a_{nv}) = \prod_{i=1}^{L_n} \prod_{j=1}^{L_n} \psi(w_{ni}, w_{nj})^{\delta(a_{nu}=i) \times \delta(a_{nv}=j)} \quad .$$

$Z_n(\theta)$ is the partition function for the disjoint graph of document n. δ is the indicator function such that $\delta(a_{nu} = i)$ is 1 if and only if $a_{nu} = i$, and 0 otherwise. An example representation for a single document is shown in Fig. 2.

3 Model Computation

Spatial context improves expressiveness, but this comes at an elevated computational cost due to cycles introduced in the undirected graph. We use a variation of Expectation Maximisation (EM) for computing an approximate maximum likelihood estimate. In the E Step, we use loopy belief propagation

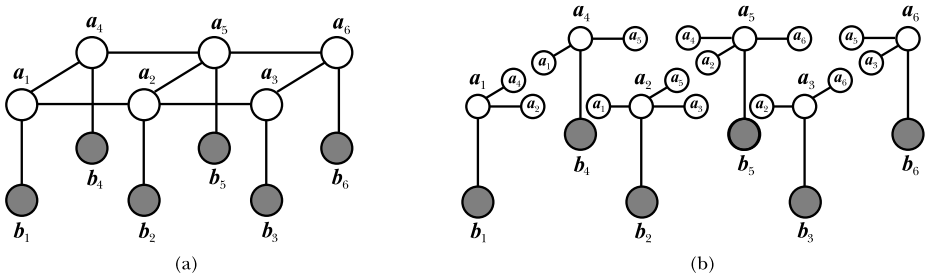

Fig. 2. *(a)* A sample Markov random field with 6 blob sites. We have omitted the n subscript. The Φ potentials are defined on the vertical lines, and Ψ on the horizontal lines. *(b)* The corresponding pseudo-likelihood approximation.

[19] on the complete likelihood (1) to compute the marginals $\widetilde{p}(a_{nu} = i)$ and $\widetilde{p}(a_{nu} = i, a_{nv} = j)$. Since the partition function is intractable and the potentials over the cliques are not complete, parameter estimation in the M Step is difficult. Iterative scaling (IS) works on arbitrary exponential models, but it is not a saving grace because convergence is exponentially-slow. An alternative to the maximum likelihood estimator is the pseudo-likelihood [6], which maximises local neighbourhood conditional probabilities at sites in the MRF, independent of other sites. The conditionals over the neighbourhoods of the vertices allow the partition function to decouple and render parameter estimation tractable. The pseudo-likelihood neglects long-range interactions, but empirical trials show reasonable and consistent results [20].

Essentially, the pseudo-likelihood is a product of undirected models, where each undirected model is a single latent variable a_{nu} and its observed partner b_{nu} conditioned on the variables in its Markov blanket. See Fig. 2 for an example. The pseudo-likelihood approximation of (1) is

$$p\ell(b, a \,|\, w, \theta) = \prod_{n=1}^{N} \prod_{u=1}^{M_n} \frac{1}{Z_{nu}(\theta)} \Phi(b_{nu}, a_{nu}) \prod_{v \,\in\, \mathcal{N}_{nu}} \Psi(a_{nu}, a_{nv}) \tag{2}$$

where $\mathcal{N}_{nu}$ is the set of blobs adjacent to node u and $Z_{nu}(\theta)$ is the partition function for the neighbourhood at site u in document n.

Iterative scaling allows for a tractable update step by bounding the log pseudo-likelihood. As in [5], we take the partial derivative of a tractable lower bound, $\Lambda(\theta)$, with respect to the model parameters, resulting in the equations

$$\begin{aligned} \frac{\partial \Lambda}{\partial t(b^\star, w^\star)} &= \sum_{n=1}^{N} \sum_{u=1}^{M_n} \sum_{i=1}^{L_n} \delta(b_{nj} = b^\star)\, \delta(w_{ni} = w^\star)\, \widetilde{p}(a_{nu} = i) \\ &\quad + \sum_{n=1}^{N} \sum_{u=1}^{M_n} \triangle t(b^\star, w^\star)^{|\mathcal{N}_{nu}|+1} \sum_{i=1}^{L_n} \delta(w_{ni} = w^\star) p(b_{nu} = b^\star, a_{nu} = i | \theta) \end{aligned} \tag{3}$$

$$\frac{\partial \Lambda}{\partial \psi(w^\star,w^\diamond)} = \sum_{n=1}^{N}\sum_{u=1}^{M_n}\sum_{v\,\in\,\mathcal{N}_{nu}}\sum_{i=1}^{L_n}\sum_{j=1}^{L_n}\delta(w_{ni}=w^\star)\,\delta(w_{nj}=w^\diamond)\,\widetilde{p}(a_{nu}=i,a_{nv}=j)$$
$$+\sum_{n=1}^{N}\sum_{u=1}^{M_n}\sum_{v\in\mathcal{N}_{nu}}\triangle\psi(w^\star,w^\diamond)^{|\mathcal{N}_{nu}|+1}\sum_{i=1}^{L_n}\sum_{j=1}^{L_n}\delta(w_{ni}=w^\star)\delta(w_{nj}=w^\diamond)p(a_{nu}=i|\widetilde{a}_{nv}=j,\theta) \quad (4)$$

where we take $p(a_{nu} = i \,|\, \widetilde{a}_{nv} = j, \theta)$ to be the estimate of alignment $a_{nu} = i$ conditional on the empirical distribution $\widetilde{p}(a_{nv} = j)$ and the current parameters. To find the conditionals for (4), we run universal propagation and scaling (UPS) [23] at each pseudo-likelihood site nu with the neighbours $v \in \mathcal{N}_{nu}$ clamped to the current marginals $\widetilde{p}(a_{nv})$. UPS is exact because the undirected graph at each neighbourhood is a tree. Also note that (3) requires estimation of the blob densities in addition to the alignment marginals.

The partial derivatives do not decouple because we cannot expect the feature counts (i.e. the number of cliques) to be the same for every site neighbourhood. Observing that (3) and (4) are polynomial expressions where each term has degree $|\mathcal{N}_{nu}| + 1$, we can find new parameter estimates by plugging the solution for (3) or (4) into the IS update $\theta_i^{(new)} = \theta_i \times \triangle\theta_i$. Cadez and Smyth [10] prove that the gradient of the pseudo-likelihood with respect to a global parameter is indeed well-conditioned since it has a unique positive root.

On large data sets, the IS updates are slow. Optionally, one can boost the M Step with an additional iterative proportional fitting (IPF) step, which converges faster than IS because it doesn't have to bound the gradient of the log likelihood. We are permitted to perform an IPF update on t because it is associated with only one clique in each neighbourhood. The IPF update for t is

$$t^{(new)}(b^\star,w^\star) = t(b^\star,w^\star) \times \frac{\sum_{n=1}^{N}\sum_{u=1}^{M_n}\sum_{i=1}^{L_n}\delta(w_{ni}=w^\star)\delta(b_{nu}=b^\star)\widetilde{p}(a_{nu}=i)}{\sum_{n=1}^{N}\sum_{u=1}^{M_n}\sum_{i=1}^{L_n}\delta(w_{ni}=w^\star)p(b_{nj}=b^\star,a_{nu}=i|\theta)} \quad . \quad (5)$$

To stabilise the parameter updates, we place weak priors on t and ψ of 10^{-5} and 10^{-4}, respectively. We find a near-uninformative prior for ψ works well, although we caution that prior selection in MRFs is notoriously difficult [6].

4 Experiments

The experiments compare two models. *dInd* is the discrete translation model proposed in [13] and assumes object annotations are independent. *dMRF* is the contextual model developed in this paper. We evaluate object recognition results on data sets composed of a variety of images, and examine the effect of two different segmentations on the performance of our models.

We composed two sets, denoted by *CorelB* and *CorelC*[3]. The first data set, *CorelB*, has 199 training images and 100 test images, with 38 words in the training set. The *CorelB* data set contains a total of 504 annotated images,

[3] The experiment data is available at `http://www.cs.ubc.ca/~pcarbo`.

divided into training and test sets numbering 336 and 168 in size. The training set has a total of 55 distinct concepts. The frequencies of words in the *CorelB* labels and manual annotations are shown in Fig. 4.

We consider two scenarios. In the first, we use Normalized Cuts [22] to segment the images. In the second scenario, we take on the object recognition task without the aid of a sophisticated segmentation algorithm, and instead construct a uniform grid of patches over the image. Examples of the segmentations are shown in Fig. 6. The choice of grid size is important since the features are not scale invariant. We use patches approximately 1/6th the size of the image; smaller patches introduce too much noise to the features, and larger patches contain too many objects. The two scenarios are denoted by *NCuts* and *Grid*.

The blobs are described using simple colour features. Vertical position was found to be a simple and useful feature [11], but it does not work well with the discrete models because the K-means clustering tends to be poor. The number of blob clusters, B, is a significant factor; too small, and the classification is non-separable; too large, and finding the correct associations is near impossible. As a rule of thumb, we found $B = 5W$ to work well.

The relative importance of objects in a scene is task-dependent. Ideally, when collecting user-annotated images for evaluation, we should tag each word with a weight to specify its prominence in the scene. In practice, this is problematic because different users focus their attention on different concepts, not to mention the fact that it is a burdensome task. Rewarding prediction accuracy over blobs — not objects — is a reasonable performance metric as it matches the objective functions of the translation models. We have yet to compare our models using the evaluation procedures proposed in [8, 2]. The prediction error is given by

$$\frac{1}{N}\sum_{n=1}^{N}\frac{1}{M_n}\sum_{u=1}^{M_n}\left(1-\delta\left(\widehat{a}_{nu}=a_{nu}^{(max)}\right)\right) \tag{6}$$

where $a_{nu}^{(max)}$ is the model alignment with the highest probability and $\widehat{a}_{nu}$ is the ground-truth annotation.

One caveat regarding our evaluation procedure: the segmentation scenarios are not directly comparable because the manual annotation data is slightly different for *NCuts* and *Grid*. For testing purposes, we annotated the image segments in the two scenarios by hand. Since we cannot expect the segmentation methods to perfectly delineate concepts in a scene, a single region may contain several subjects, all deemed to be correct. We found that the Normalized Cuts segments frequently encompassed several objects, whereas the uniform grid segments, by virtue of being smaller, more regularly contained a single object. As a result, our evaluation measure can report the same error for the two scenarios, when in actual fact the uniform grid produces more precise object recognition. To address the role of segmentation in object recognition more faithfully, we are in the process of building data sets with ground-truth annotations and segmentations.

Figure 3 compares model annotations over 12 trials with different initialisations. Model *dInd* took on the order of a few minutes to converge to a local minimum of the log-likelihood, whereas model *dMRF* generally took several

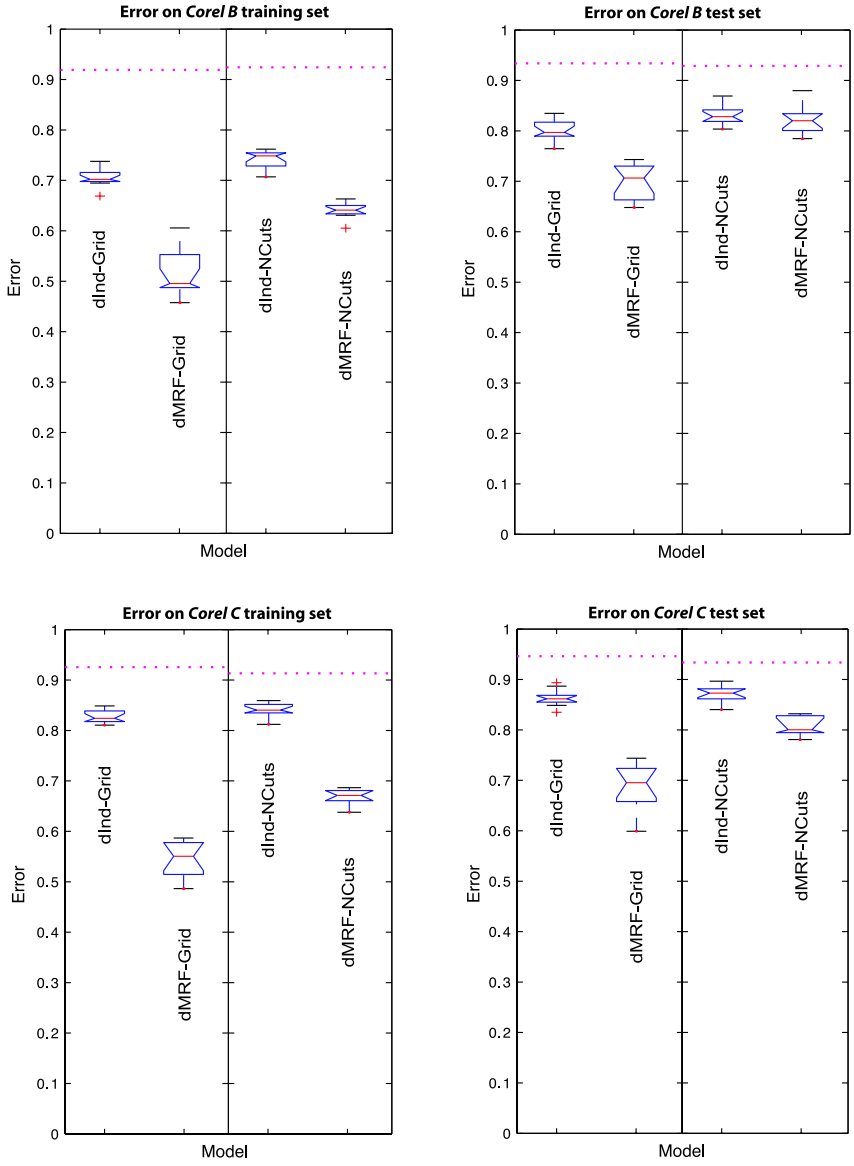

Fig. 3. Prediction error of the two models using the *Grid* and *NCuts* segmentations on the *CorelB* and *CorelC* data sets. The results are displayed using a Box-and-Whisker plot. The middle line of a box is the median. The central box represents the values from the 25 to 75 percentile, using the upper and lower statistical medians. The horizontal line extends from the minimum to the maximum value, excluding outside and far out values which are displayed as separate points. The dotted line at the top is the random upper bound. The contextual model introduced in this paper substantively reduces the error over *dInd* in the grid segmentation case.

Precision on *CorelB* data set using *Grid*

	LABEL %		ANNOTATION %†		*dInd* PR.		*dMRF* PR.	
WORD	TRAIN	TEST†	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
airplane	0.060	0.055	0.036	0.028	0.135	0.102	0.290	0.187
astronaut	0.003	0.003	0.001	0.002	0.794	0.087	0.000	0.135
atm	n/a	0.003	n/a	0.006	n/a	0.000	n/a	0.000
bear	0.031	0.017	0.021	0.013	0.192	0.092	0.452	0.272
beluga	n/a	0.003	n/a	0.005	n/a	0.000	n/a	0.000
bill	0.019	0.017	0.046	0.031	0.269	0.175	0.335	0.146
bird	0.017	0.010	0.009	0.004	0.148	0.111	0.556	0.458
building	0.014	0.007	0.006	0.002	0.368	0.013	0.408	0.137
cheetah	0.012	0.017	0.010	0.013	0.833	0.683	0.710	0.395
cloud	0.050	0.045	0.050	0.048	0.222	0.152	0.300	0.239
coin	0.005	0.007	0.008	0.008	0.611	0.213	0.767	0.017
coral	0.005	n/a	0.011	n/a	0.815	n/a	0.738	n/a
crab	0.002	0.003	0.001	0.002	0.802	0.663	1.000	0.833
dolphin	0.014	0.003	0.006	0.001	0.606	0.899	0.916	0.000
earth	0.003	0.007	0.004	0.002	0.543	0.000	0.732	0.142
fish	0.007	n/a	0.003	n/a	0.236	n/a	0.695	n/a
flag	0.005	0.007	0.004	0.008	0.617	0.831	0.888	0.890
flowers	n/a	0.003	n/a	0.003	n/a	0.000	n/a	0.000
fox	0.010	0.017	0.011	0.010	0.246	0.052	0.691	0.008
goat	0.003	n/a	0.001	n/a	0.704	n/a	0.994	n/a
grass	0.129	0.120	0.177	0.157	0.165	0.176	0.172	0.229
hand	n/a	0.003	n/a	0.002	n/a	0.000	n/a	0.000
map	n/a	0.003	n/a	0.003	n/a	0.000	n/a	0.000
mountain	0.012	0.003	0.007	0.000	0.204	0.060	0.671	0.057
person	0.003	0.014	0.001	0.004	0.170	0.037	1.000	0.000
polarbear	0.021	0.024	0.016	0.015	0.510	0.625	0.681	0.634
rabbit	0.002	n/a	0.001	n/a	0.489	n/a	1.000	n/a
road	0.026	0.024	0.016	0.008	0.190	0.062	0.526	0.213
rock	0.019	0.038	0.018	0.033	0.127	0.078	0.446	0.130
sand	0.034	0.024	0.023	0.024	0.246	0.150	0.330	0.185
shuttle	0.007	0.007	0.006	0.005	0.504	0.268	0.305	0.107
sky	0.156	0.137	0.172	0.173	0.190	0.138	0.190	0.208
snow	0.036	0.062	0.064	0.110	0.358	0.296	0.435	0.356
space	0.007	0.010	0.008	0.018	0.000	0.000	0.326	0.071
tiger	0.021	0.034	0.015	0.030	0.450	0.233	0.623	0.285
tracks	0.024	0.017	0.010	0.009	0.351	0.163	0.575	0.315
train	0.026	0.021	0.021	0.014	0.165	0.164	0.396	0.272
trees	0.095	0.076	0.075	0.069	0.169	0.094	0.227	0.134
trunk	0.003	0.010	0.001	0.006	0.553	0.023	0.910	0.000
water	0.091	0.089	0.120	0.095	0.214	0.133	0.212	0.137
whale	0.007	0.007	0.003	0.004	0.476	0.268	0.854	0.405
wolf	0.009	0.038	0.007	0.029	0.512	0.166	0.660	0.102
zebra	0.012	0.010	0.009	0.007	0.652	0.667	0.903	0.710
Totals	**1.000**	**1.000**	**1.000**	**1.000**	**0.294**	**0.199**	**0.486**	**0.301**

Fig. 4. The first four columns list the probability of finding a particular word in an image caption and manually-annotated image region in the *CorelB* data using the grid segmentation. The final four columns show the precision of models *dInd* and *dMRF* averaged over the 12 trials. Precision is defined as the probability the model's prediction is correct for a particular word and blob. Since precision is 1 minus the error of (6), the total precision on both the training and test sets matches the average performance shown in Fig. 3. The variance in the precision on individual words is not presented in this table. Note that some words do not appear in both the training and test sets, hence the "n/a".
†We underline the fact that an agent would not have access to the test image labels and information presented in the ANNOTATION % column.

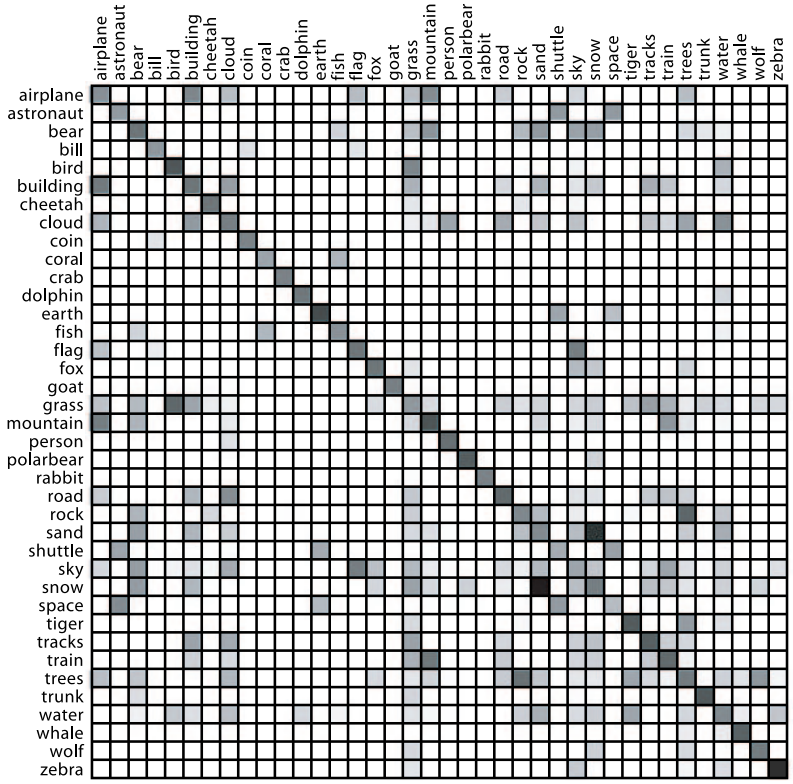

Fig. 5. Our algorithm learned the above contextual relations for the *CorelB* data using the grid segmentation (the matrix is averaged over all the learning trials). Darker squares indicate a strong neighbour relationship between concepts. White indicates that the words were never observed next to each other. For example, fish goes with coral. It is intriguing that planes go with buildings!

hours to learn the potentials. The first striking observation is that the contextual model shows consistently improved results over *dInd*. Additionally, the variance of *dMRF* is not high, despite an increase in the number of parameters and a lack of convergence guarantees.

Recognition using the grid segmentation tends to do better than the Normalized Cuts results, keeping in mind that we require the *Grid* annotations to be more precise, as discussed above. This suggests we can achieve comparable results without an expensive segmentation step. However, we are cautious not to make strong claims about the utility of sophisticated low-level segmentations in object recognition because we do not have a uniform evaluation framework nor have we examined segmentation methods in sufficient variety and detail.

Figure 4 shows the precision on individual words for the *CorelB Grid* experiment, averaged over the 12 trials. While not shown in the figures, we have noticed considerable variation among individual trials as to what words are pre-

dicted with high precision. For example, model *dMRF* with a grid segmentation predicts the word "train" with average success 0.396, although the precision on individual trials ranges from 0.102 to 0.862 in the training set. Significantly, the spatial context model tends to do better on words that cannot be described using simple colour features, such as "building", "bear" and "airplane".

Figure 5 depicts the potentials ψ for the *CorelB Grid* experiment. Note the table is symmetric. The diagonal is dark because words appear most often next to themselves. The strong diagonal acts as a smoothing agent on the segmented scenes. Most of the high affinities are logical (e.g. "airplane" and "cloud") but there are a few that defy common sense (e.g. "trees" and "trunk" have a weak affinity), likely due to incorrect associations between the blobs and words.

Selected annotations predicted by the *dMRF* model on the *CorelB* training and test sets are displayed in Fig. 6. In several instances, observers found *dInd* annotations more appealing than those of *dMRF*, even though precision using the latter model was higher. This is in part because *dMRF* tends to be more accurate for the background (e.g. sky), whereas observers prefer getting the principal subject (e.g. airplane) correct. This suggests that we should explore alternative evaluation measures based on decision theory and subjective prior knowledge.

5 Discussion and Conclusions

We showed that spatial context helps classify objects, especially when blob features are ineffective. Poorly-classified words may be easier to label when paired with easily-separable concepts. Spatial context purges insecure predictions, and thus acts as a smoothing agent for scene annotations. The pseudo-likelihood approximation allows for satisfactory results, but we cannot precisely gauge the extent to which it skews parameters to suboptimal values. Our intuition is that it gives undue preference to the diagonal elements of the spatial context potentials.

Normalized Cuts is widely considered to produce good segmentations of scenes, and surprisingly our experiments indicate that crude segmentations work equally well or better for object recognition. Upon further consideration, our results are indeed sensible. We are attempting to achieve an optimal balance between loss of information through compression and adeptness in data association through mutual information between blobs and labels. The Normalized Cuts settings we use tend to fuse blobs containing many objects, which introduces noise to the classification data. *dMRF* can cope with lower levels of compression, and hence it performs much better with smaller segments even if they ignore object boundaries. Since model *dMRF* fuses blobs with high affinities, we claim it is a small step towards a model that learns both scene segmentations and annotations concurrently. A couple considerable obstacles in the development of such a model are the design of efficient training algorithms and the creation of evaluation schemes that uniformly evaluate the quality of segmentations combined with annotations.

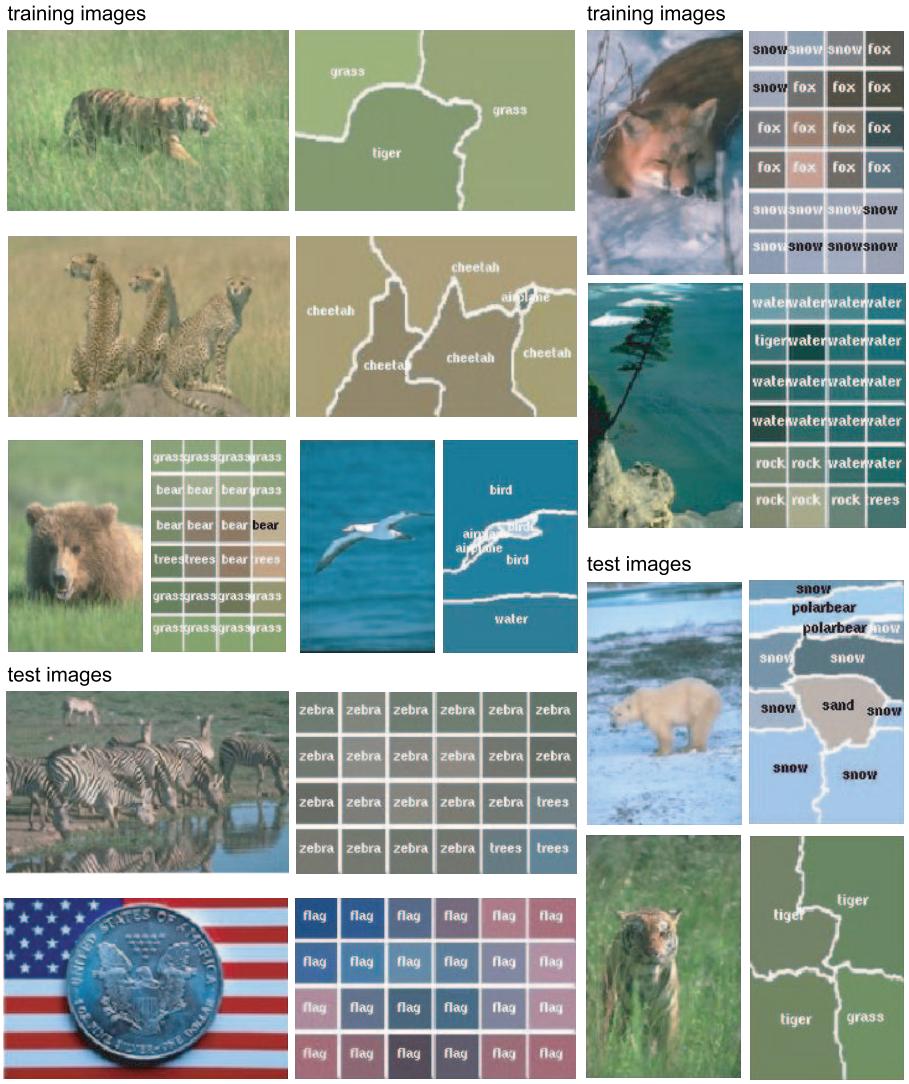

Fig. 6. Selected *dMRF* model annotations on the *CorelB* training (top) and test sets (bottom). It is important to emphasize that the model annotations are probabilistic, and for clarity we only display the classification with the highest probability. Also, the the predictions are made using only the image information.

Acknowledgements

We thank Yee Whye Teh for his discussions on parameter estimation in graphical models and Kevin Murphy for his advice on belief propagation. We would also like to acknowledge invaluable financial support from IRIS ROPAR and NSERC.

References

1. Barnard, K., Duygulu, P., Forsyth, D.A.: Clustering art. IEEE Conf. Comp. Vision and Pattern Recognition (2001)
2. Barnard, K., Duygulu, P., Forsyth, D.A., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. J. Machine Learning Res., Vol. 3 (2003) 1107-1135
3. Barnard, K., Duygulu, P., Guru, R., Gabbur, P., Forsyth, D.A.: The Effects of segmentation and feature choice in a translation model of object recognition. IEEE Conf. Comp. Vision and Pattern Recognition (2003)
4. Barnard, K., Forsyth, D.A.: Learning the semantics of words and pictures. Intl. Conf. Comp. Vision (2001)
5. Berger, A.: The Improved iterative scaling algorithm: a gentle introduction. Carnegie Mellon University (1997)
6. Besag, J.: On the Statistical analysis of dirty pictures. J. Royal Statistical Society, Series B, Vol. 48, No. 3 (1986) 259–302
7. Blei, D.M., Jordan, M.I.: Modeling annotated data. ACM SIGIR Conf. on Research and Development in Information Retrieval (2003)
8. Borra, S., Sarkar, S.: A Framework for performance characterization of intermediate-level grouping modules. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, No. 11 (1997) 1306–1312
9. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The Mathematics of statistical machine translation: parameter estimation. Computational Linguistics, Vol. 19, No. 2 (1993) 263–311
10. Cadez, I., Smyth, P.: Parameter estimation for inhomogeneous Markov random fields using PseudoLikelihood. University of California, Irvine (1998)
11. Carbonetto, P., de Freitas, N., Gustafson, P., Thompson, N.: Bayesian feature weighting for unsupervised learning, with application to object recognition. Workshop on Artificial Intelligence and Statistics (2003)
12. Dorkó, G., Schmid, C.: Selection of scale invariant neighborhoods for object class recognition. Intl. Conf. Comp. Vision (2003)
13. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.A.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. European Conf. Comp. Vision (2002)
14. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. IEEE Conf. Comp. Vision and Pattern Recognition (2003)
15. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. Intl. J. of Comp. Vision, Vol. 40, No. 1 (2000) 23–47
16. Kumar, S., Hebert, H.: Discriminative Random Fields: a discriminative framework for contextual interaction in classification. Intl. Conf. Comp. Vision (2003)
17. Kumar, S., Hebert, H.: Discrminative Fields for modeling spatial dependencies in natural images. Adv. in Neural Information Processing Systems, Vol. 16 (2003)
18. Lowe, D.G.: Object recognition from local scale-invariant features. Intl. Conf. Comp. Vision (1999)
19. Murphy, K., Weiss, Y., Jordan, M.I.: Loopy belief propagation for approximate inference: an empirical study. Conf. Uncertainty in Artificial Intelligence (1999)
20. Seymour, L.: Parameter estimation and model selection in image analysis using Gibbs-Markov random fields. PhD thesis, U. of North Carolina, Chapel Hill (1993)
21. Mikolajczk, K., Schmid, C.: A Performance evaluation of local descriptors. IEEE Conf. Comp. Vision and Pattern Recognition (2003)
22. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Conf. Comp. Vision and Pattern Recognition (1997)
23. Teh, Y.W., Welling, M.: The Unified propagation and scaling algorithm. Advances in Neural Information Processing Systems, Vol. 14 (2001)