A statistical translation model for contextual object recognition

Anonymous
Anonymout Department
Anonymous University
Anonymous Place

Abstract

We approach the problem of object recognition as the process of attaching meaningful labels to specific regions of an image, and propose a model that learns spatial relationships. Given a set of images and their labels (e.g. word captions), the objective is to segment the image, in either a crude or intelligent fashion, then to find the proper associations between words and regions. Previous methods are limited by the scope of the representation. In particular, they fail to exploit spatial context in the image and words. We develop a more expressive model that takes this into account. Using the analogy of building a lexicon using an aligned text, we formulate a spatially consistent probabilistic mapping between continuous image feature vectors and the supplied word tokens. To find the best hypothesis, we hill-climb the model log-posterior using an approximate EM algorithm. Spatial context introduces cycles to our probabilistic graphical model, so we use loopy belief propagation to compute the expectation of the complete log-posterior. Experimental results on a diverse array of images show that learning context considerably improves the accuracy of object recognition. Moreover, the results suggest that similar, or even better, performance can be obtained using crude image segmentations.

1. Introduction

Annotated image sets are not in short supply. Examples of labeled image data sets include the standard Corel data set, news photograph services, art galleries and, more recently, web services specializing in the provision or cataloging of vast stock photo databases (e.g. www.corbis.com, www.gettyimages.com, www.cinemage.com). Autonomous robots that acquire photographs, sounds and text as they explore an environment are another avenue for the collection of annotated images [19, 11]. Examples from different media sources are shown in figure 1.

A considerable body of research in computer vision is devoted to the exploitation of annotated image sets specifically for the purpose of image classification and clustering

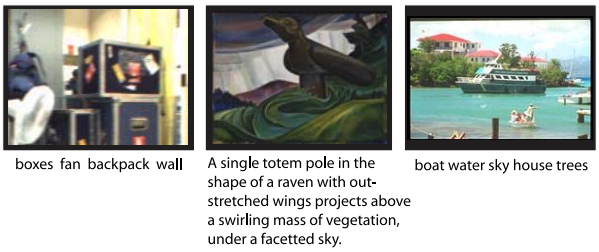

Figure 1: *Examples of text and image database entries: an annotated image taken by a mobile robot exploring an anonymous lab, a painting from the Vancouver Art Gallery, and an example from the Coral image database.*

[21, 15, 22, 3, 4]. The most direct and obvious application of image classification is image retrieval; once a classifier has learned to pair images of a certain description (e.g. color, texture) with the conjoined text, the classifier could function as a search engine. Moreover, with a representation of a joint distribution over words and images, the search engine could accept mixed media in a query to improve the the precision of retrieval [4, 8, 6].

The work of [10] takes image classification one step further. In particular, it builds a model for finding associations between words and individual objects within an image. That is, the image region labeling problem approaches that of object recognition. This approach allows us to pursue generic object recognition in a principled fashion using Bayesian theory. For a discussion of the variety of perspectives on object recognition, we refer the reader to [12].

By making the leap from image classification to object recognition, however, we do so at a cost — we are no longer blessed with the exact associations between image regions and words. In other words, the learning problem is now unsupervised. For a single training image and a particular word token, we have to learn both the probability of generating that word given an object description and the correct association to one of the regions within the image.

Fortunately, there is a simple and compelling parallel between our object recognition formulation and the statistical machine translation problem of building a lexicon from an

aligned bi text, for which there is a large body of research [5, 17, 2, 18]. Given a text in two languages (such as the Canadian parliamentary Mansard bi text which published in English and French) and assuming we know the correspondences at roughly a sentence level, the objective is to learn a dictionary of precise translations from English to French, which we call a lexicon. Similarly, in object recognition we build a lexicon of translation probabilities from image patches to words (from now on we use the word "patch" to refer to a contiguous region in an image).

We make two major contributions in this paper.

Our first contribution is to remedy an important omission from the object recognition models mentioned up to this point – namely, context. Humans make use of a considerable degree of multi-sensory low- and high-level information to make judgments about a scene. One of the important cues is contextual information [16, 1].

Secondly, we demonstrate empirically that we can supply equally good or improved annotations using the contextual model *without* segmenting the image. Instead, we compose a naive segmentation using a uniform grid (see figure 4 for an example). This eliminates a significant bottleneck in the object recognition process, and at no cost to precision. Disposing of the segmentation step is of great importance for the application of object recognition to mobile robots. After training the recognition models, we want the robots to navigate while automatically placing labels on the regions of the observed images. Standard segmentation algorithms in this test phase tend to be too slow. A crude grid-based segmentation enables us to move to real-time labeling.

In the next section we motivate the need for context information. Section 3 outlines the basic probabilistic model for "object recognition as machine translation". Section 4 specifies the contextual model. In section 5 we discuss how to resolve computation of the model. As it turns out, this is not a trivial matter because there are many parameters and cycles in the probabilistic graphical model. In the final two sections we present the results of our experiments and concluding remarks.

2. Motivation for adding context: an illustrative example

Before examining the empirical evidence in succeeding sections, we first pose the question: why do we think introducing spatial context to the model might benefit object recognition? In this section, we briefly reason about this question with an anecdotal example.

Say we are trying to learn, among other words, the correct patch associations for two animals: the impala, a large brown mammal that inhabits wooded savannas in eastern Africa, and the white-tailed deer, its habitat distributed throughout eastern North America. See figure 2 for some example images. Unless one has access to a very sophisticated (or very specific) object recognition algorithm, it would be difficult to classify these two animals solely based on low-level vision information. In fact, the small differences in colour between the impala and deer are unreliable indicators since the deer has a coat that varies from red-brown to gray-brown depending on the time of year. While snow is not always found in the habitat of the white-tailed deer, contextual information could still be a useful indicator since the impala is rarely found near snow.

Figure 2: *The two images on the left are instances of the impala in its native habitat. The image on the far right shows a white-tailed deer, again in its native surroundings.*

Context will, in many cases, give us another dimension to resolve recognition conflcts. Learning context in the model can, of course, deteriorate recognition results in some cases. However, we believe that we can expect an improvement most of the time. Our intuition is confirmed in the experiments.

3. Statistical models for object recognition as machine translation

We can cast object recognition as a machine translation problem, as was originally proposed in [10]. In particular, we translate image regions (patches) into words. The model acts as a *lexicon*, a dictionary that predicts one representation (words) given another representation (patches).

We consider a set of N images paired with their captions. Each training example n is composed of a set of patches $\{b_{n1}, ..., b_{nM_n}\}$ and a set of words $\{w_{n1}, ..., w_{nL_n}\}$, where M_n is the number of patches in image n and L_n is the corresponding number of words. Each $b_{nj} \epsilon \mathbb{R}^{n_f}$ is a vector containing a set of feature values representing colour, texture, position, *etc*, where n_f is the number of features. For each patch b_{nj}, our objective is to align it to a word from the attached caption. We represent this unknown association by a latent variable a_{nj}, such that $a^i_{nj} = 1$ if b_{nj} translates to w_{ni} and 0 otherwise. Thus, $p(a^i_{nj}) \triangleq p(a_{nj} = i)$ is the probability that patch b_{nj} is aligned with word w_{ni} in document n. W is the total number of word tokens in the training set.

In [7], the approach of [10] was extended to deal with continuous patch distributions. In particular, the following model was adopted:

$$p(\mathbf{b},\mathbf{a}|\mathbf{w}) = \prod_{n=1}^{N}\prod_{j=1}^{M_n}\prod_{i=1}^{L_n}[\, p(a_{nj}=i) \times \mathcal{N}(b_{nj};\mu_{w^\star},\Sigma_{w^\star})\,\delta_{w^\star}(w_{ni})\,]^{a^i_{nj}}$$

where $\delta_{w^\star}(w_{ni}) = 1$ if $w^\star$ appears in the n-th annotation, otherwise it is 0. $w^\star$ denotes the particular word token. Hence, the model assigns a Gaussian to each word token.

The ultimate objective is to generate words from patches. We accomplish word prediction through a a straightforward application of Bayes rule, $p(w|b) \propto p(b|w)p(w)$. We assume the prior probability $p(w)$ is uniform over all words.

The approach in [7] also introduces a hierarchical Bayesian scheme to provide regularised solutions and to carry out automatic feature weighting (or selection). In particular, the following model is adopted:

$$\begin{aligned}\mu_{w^\star} &\sim \mathcal{N}(\mu^*, T)\\ \Sigma_{w^\star} &\sim \mathcal{IW}(\alpha,\Sigma^*)\\ \tau_k^2 &\sim \mathcal{I}g(\nu,\varepsilon)\end{aligned}$$

where $T = diag(\tau_1^2,\ldots,\tau_k^2,\ldots,\tau_{n_f}^2)$, $\mathcal{IW}(\cdot)$ denotes the Inverse-Wishart distribution and $\mathcal{I}g(\nu,\varepsilon)$ denotes the Inverse-Gamma distribution. Ordinarily, μ^* and Σ^* are set to the mean and variance of the data.

The shrinkage parameters τ_k^2 are priors on the cluster means, and act to control the variance of the means (cluster inter-distance) and the variance of each cluster (cluster intra-distance). The regularisation prior α on the variances of the word clusters stabilises the algorithm by preventing ill-conditioning arising from an unbounded joint likelihood.

The shrinkage variances $\{\tau_1^2,...,\tau_{n_f}^2\}$ on the cluster means allows one to carry out feature selection within the translation procedure. This method places low importance on features that have little or no relevance to the concepts they describe, and thus prevents the model from over-fitting on irrelevant features [7].

In practice, we found that setting the prior on the cluster variances to the sample variance caused the Gaussians to spread out too much. We achieved superior results with $\Sigma^* = \overline{\Sigma}/W$, where W in our case is the number of clusters and $\overline{\Sigma}$ is the sample variance.

4. Specification of contextual model

One of the limitations of the existing approaches for translating image patches to words is the assumption that patches are statistically independent. This assumption is usually made to simplify the translation model to a simple mixture. Here we relax this assumption. We allow for interactions between neighbouring image patches through a pairwise Markov random field (MRF). That is, the probability of an image patch being aligned to a particular word depends on the word assignments of its neighbouring patches. This is illustrated In figure 3. Note that the potentials ϕ_{nj} represent the patch-to-word translation probabilities.

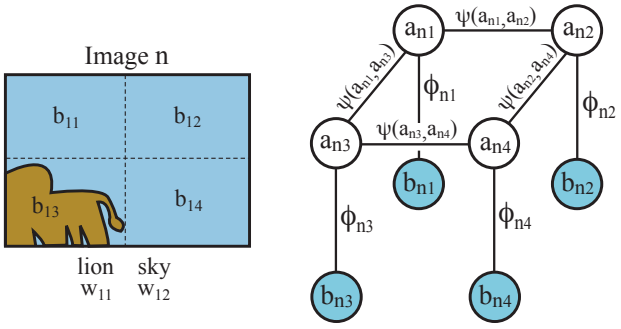

Figure 3: *The graphical model for a very simple example with one document. The shaded circles are the observed nodes (i.e. the data), and the white circles are unobserved variables of the model parameters. Lines represent the undirected dependencies between variables. The potentials ψ control the consistency between annotations, while the potentials ϕ_{nj} represent the independent translation probabilities.*

In the contextual model, the latent alignments of patches to words, $\{a_{nj}\}$, are no longer independent of one another. If any two patches b_{nj} and b_{nk} are adjacent to each other in the image, the potential $\psi(a_{nj},a_{nk})$ encodes the compatibility of the two associations. The potentials are assumed to be the same for each image. That is, we use a single $W \times W$ matrix $\{\psi(w^\star,w^\diamond)\}$, which ranges over the word tokens.

The entry $(w^\star,w^\diamond)$ represents the probability that two words will be next to each other spatially in the image. To ensure that the none of the entries for ψ collapse to 0, which occurs in situations where two words never appear in the same label, we place a Dirichlet prior over the alignment potentials. In mathematical terms, our contextual translation model is:

$$p(\mathbf{b},\mathbf{a}|\mathbf{w}) = \prod_{n=1}^{N}\frac{1}{Z_n}\left\{\prod_{j=1}^{M_n}\prod_{i=1}^{L_n}[\mathcal{N}(b_{nj};\mu_{w^\star},\Sigma_{w^\star})\delta_{w^\star}(w_{ni})]^{a^i_{nj}} \times \prod_{(r,s)\,\epsilon\,\mathcal{C}_n}\prod_{i=1}^{L_n}\prod_{j=1}^{L_n}[\psi(w^\star,w^\diamond)\delta_{w^\star}(w_{ni})\delta_{w^\diamond}(w_{nj})]^{a^i_{nr}\times a^j_{ns}}\right\}$$

where $\mathcal{C}$ denotes the clique of neighbour patches and Z_n is the standard partition function for image n.

In the following section, we derive an approximate EM algorithm for learning the model parameters. The algorithm

uses loopy belief propagation to deal with the intractability that arises in the E step due to cycles in the graph.

5. Model computation

The presence of loops in the graphical model destroys the tractability of the problem. We overcome this by adopting a loopy belief propagation algorithm [23] to infer the marginal posterior probabilities of the assignments. Conditioning on these marginals, we learn the model parameters $\theta \triangleq \{\mu, \Sigma, \psi, \tau\}$ by iteratively performing hill-climbing on the log-posterior. What we have, therefore, is something that resembles EM. However, one should be cautious. The E step is only approximate and hence we cannot assure convergence. In practice, however, we have noticed that the algorithm behaves well.

The E step updates the expected complete log-posterior given the estimate of the model parameters at the previous time step. We use the superscript $\theta^{(t)}$ to denote the parameter estimate at time t. To find the new alignment probabilities, we make use of Bayes' theorem:

$$\begin{aligned}\xi_{n,i,j} &\triangleq p(a_{nj} = i | w_{ni}, b_{nj}, \theta^{(t-1)}) \\ &= \frac{p(a_{nj} = i, b_{nj} | w_{ni}, \theta^{(t-1)})}{p(b_{nj} | w_{ni}, \theta^{(t-1)})}\end{aligned}$$

The normalising constant in this expression is prohibitively expensive to evaluate. Hence, we adopt loopy belief propagation to infer the marginal probabilities of the hidden variables $p(a_{nj} = i | w_{ni}, b_{nj}, \theta^{(t-1)})$.

Once we know the marginal probabilities of the hidden variables, we can compute the parameters by taking derivatives of the log-posterior in order to obtain a gradient ascent algorithm. The M step equations for μ, Σ and τ are the same as in [7].

To simplify the equations, we adopt the notation $f_{n,i,j,w^\star}(x) \triangleq \sum_{j=1}^{M_n} \sum_{i=1}^{L_n} [\, \xi_{n,i,j}\, x\, \delta_{w^\star}(w_{ni}) \,]$. μ_{wk} denotes the kth dimension of μ_w.

$$\widehat{\tau}_k^2 = \frac{\varepsilon}{\nu + \frac{1}{2}W + 1} + \frac{1}{2\nu + W + 2} \sum_{w=1}^{W} (\mu_{wk} - \mu_k^*)^2 \quad (1)$$

$$\begin{aligned}\widehat{\mu}_{w^\star} &= T \left(\Sigma_{w^\star} + T \sum_{n=1}^{N} f_{n,i,j,w^\star}(1) \right)^{-1} \sum_{n=1}^{N} f_{n,i,j,w^\star}(b_{nj}) \\ &\quad + \Sigma_{w\star} \left(\Sigma_{w^\star} + T \sum_{n=1}^{N} f_{n,i,j,w^\star}(1) \right)^{-1} \mu^* \quad (2)\end{aligned}$$

$$\widehat{\Sigma}_{w^\star} = \frac{\sum_{n=1}^{N} f_{n,i,j,w^\star}((b_{nj} - \mu_{w^\star})(b_{nj} - \mu_{w^\star})') + \alpha \Sigma^*}{\sum_{n=1}^{N} f_{n,i,j,w^\star}(1) + \alpha + F + 1} \quad (3)$$

The update equation for the potentials for the alignments is:

$$\begin{aligned}\widehat{\psi}(w^\star, w^\diamond) \propto \sum_{n=1}^{N} \sum_{(r,s)\,\epsilon\,\mathcal{C}_n} \sum_{u=1}^{L_n} \sum_{v=1}^{L_n} [\xi_{n,u,r}\, \xi_{n,v,s} \\ \delta_{(w^\star, w^\diamond)}(w_{nu}, w_{nv})] + \eta \quad (4)\end{aligned}$$

where η is the Dirichlet hyper-parameter. This is an intuitive update. We are simply counting co-occurrences and normalising so labeling is not biased toward more frequent words.

6. Experimental Results

We trained the models on annotated images from the Corel image database. We composed two separate sets, each which contain a wide variety of images, such as airplanes, trains, lions, birds, bears, currency and landscapes. The first data set, *Corel A*, has 199 training images and 100 test images. The *Corel B* data set contains 504 labeled images, divided into training and test sets of size 336 and 168. The results showed in succeeding experiments are for these two data sets.

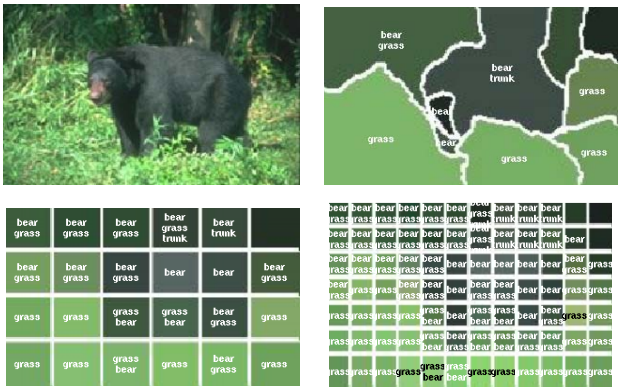

Figure 4: *Patches, constructed for the same image but using three different methods, are shown with the correct labelings: Normalized Cuts segmentation on the top-right corner, and rectangular grids using patches of size 32 and 16 on the bottom-left and bottom-right, respectively. The top-left image is the original.*

For each set, we consider two scenarios. In the first scenario, we use Normalized Cuts [20] to segment the images into distinct patches. In the second scenario, we take on the object recognition task without the aid of a segmentation algorithm, and instead construct a uniform grid of patches over the image. Figure 4 shows an example of how image patches might be constructed under the two scenarios. Experimentation over these different situations gives us an indication of the importance of a good segmentation for object recognition. We used patches of size 32 x 32. We found that

smaller patches introduced too much noise into the features and resulted in poor test performance, and larger patches contained too many objects at once. In future work, we will investigate a hierarchical patch representation to take into account both short and long range patch interactions, as in [13].

Each of the patches is represented by a real-valued feature vector representing colour and position information.

To provide a ground truth for the proper measurement of the success of our translation models, we tediously annotated the several hundred images by hand. In some cases, there might be more than one annotation deemed correct because the patch consists of several subjects. We uses a straightforward error metric, which reports an error of 1 if the word with the highest probability results in an incorrect patch annotation (in the graphs, we use $pr1$ to denote this metric). The error measure is averaged over the number of patches in each image, and then over the number of documents in the data set.

In our experiments we compare two models. The first is the original translation model with shrinkage [7], in which we assume alignments are independent of each other. We call this *tInd*. The second model, *tMRF*, is the contextual translation model with a weak prior $\eta = 0.01$.

Experimental results are shown in figures 7 and 8 for the *Corel A* and *Corel B* data sets. We achieve the best annotations when the contextual model is trained on the naively segmented images.

Removing the independence assumption between annotations is especially helpful when we make naive, uninformed segmentations — the extra contextual information leads to more robust results. It's interesting to observe that while the NCuts segmentations are well-intentioned, this expensive step in the translation process does not improve performance.

We provide some typical results in figures 5 and 6. Notice that the $tInd$ model finds the correct words for some of the patches, but also produces a great deal of convoluted and inconsistent annotations. On the other hand, $tMRF$ finds more satisfying labelings because it enforces a degree of consistency over regions of patches. The contextual model is not a panacea, but it does leads to more consumable annotations.

An attractive feature of our contextual model is that we recover the same results as the independent model by setting η, the Dirichlet prior on the alignment potentials, to a suitably large number.

7. Conclusions

In this paper, we introduced a contextual translation model for object recognition. We presented an approximate EM algorithm, with loopy belief propagation in the E step, to compute the model parameters. Our results are very encouraging for two reasons. First, they indicate that learning context in the model improves recognition performance. Second, they suggest that there is no need for a sophisticated initial segmentation process. This eliminates the main computational bottleneck when testing the model. Starting with a crude grid segmentation, our model allows us to place appropriate words on the image patches. Subsequently, neighbouring image patches with the same word can be merged. We are therefore solving the segmentation (using bottom-up and top-down information) and recognition problems simultaneously.

In the future, we plan to carry out further testing on other image databases. We also need to address various problems, such as estimating the size of the image patches, adding trivial context nodes (e.g. a binary node stating that "sky" tends to be "above" the "sea") and adopting more sophisticated language models.

Acknowledgments

Anonymous for review purposes.

References

[1] Yael Adini, Dov Sagi and Misha Tsodyks. "Context-enabled learning in the human visual system." *Nature*, Vol. 415, 7 Feb 2002, pp. 790 - 793.

[2] Y. Al-Onaizan, J. Curin, Michael Jahr, K. Knight, J. Lafferty, I. D. Melamed, F.-J. Och, D. Purdy, N. A. Smith and D. Yarowsky. "Statistical machine translation: final report." *Johns Hopkins University Workshop on Language Engineering*, 1999.

[3] Kobus Barnard and David A. Forsyth. "Clustering art." *IEEE Conference on Computer Vision*, 2001.

[4] Kobus Barnard and David A. Forsyth. "Learning the semantics of words and pictures." *IEEE Conference on Computer Vision*, 2001.

[5] P. Brown, S. A. Della Pietra, V.J. Della Pietra and R. L. Mercer. "The Mathematics of Statistical Machine Translation." *Computational Linguistics*, Vol. 19, No. 2, 1993, pp. 263-311.

[6] Eric Brochu, Nando de Freitas and Kejie Bao. "The sound of an album cover: probabilistic multimedia and information retrieval." *Artificial Intelligence and Statistics*, 2003.

[7] P. Carbonetto, N. de Freitas, P. Gustafson and N. Thompson. "Bayesian feature weighting for unsupervised learning, with application to object recognition." *Artificial Intelligence and Statistics*, 2003.

[8] Chad Carson, Serge Belongie, Hayit Greenspan and Jitendra Malik. “Blobworld: color- and texture-based image segmentation using EM and its application to image querying and classification.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 8, August 2002, pp. 1026-1038.

[9] A. P. Dempster, N. M. Laird and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm.” *Journal of the Royal Statistical Society B*, Vol. 39, 1977, pp. 1-38.

[10] P. Duygulu, K. Barnard, N. de Freitas and D. A. Forsyth. “Object recognition as machine translation: learning a lexicon for a fixed image vocabulary.” *European Conference on Computer Vision*, 2002.

[11] P. Elinas, J. Hoey, D. Lahey, J. D. Montgomery, D. Murray, S. Se and J. J. Little. “Waiting with José, a vision-based mobile robot.” *International Conference on Robotics and Automation*, 2002.

[12] David A. Forsyth and Jean Ponce. *Computer vision: a moden approach.* Prentice Hall, 2002.

[13] William T. Freeman and Egon C. Pasztor. “Learning low-level vision.” *International Conference on Computer Vision*, 1999.

[14] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall, 1995.

[15] Jing Huang and Ramin Zabih. “Combining color and spatial information for content-based image retrieval.” *European Conference on Digital Libraries*, 1998.

[16] Laurent Itti, Christof Koch. “Computational modelling of visual attention.” *Nature Reviews Neuroscience*, Vol. 2, 2 March 2001, pp. 194 - 203.

[17] Kevin Knight. “A Statistical MT tutorial workbook.” *Johns Hopkins University Workshop on Language Engineering*, 1999.

[18] I. Dan Melamed. *Empirical Methods for Exploiting Parallel Texts.* MIT Press, 2001.

[19] Eric Paulos and John Canny. “Social tele-embodiment: understanding presence.” *Autonomous Robots 11*, Vol. 11, No. 1, July 2001, pp. 87-95.

[20] Jianbo Shi and Jitendra Malik. “Normalized cuts and image segmentation.” *Computer Vision and Pattern Recognition*, 1997.

[21] John R. Smith and Shih-Fu Chang. “VisualSEEk: a fully automated content-based image query system.” *ACM Multimedia*, 1996.

[22] Kinh Tieu and Paul Viola. “Boosting image retrieval.” *Computer Vision and Pattern Recognition*, 2000.

[23] Jonathan S. Yedidia, William T. Freeman and Yair Weiss. “Generalized Belief Propagation.” *Neural Information Processing Systems*, 2000.

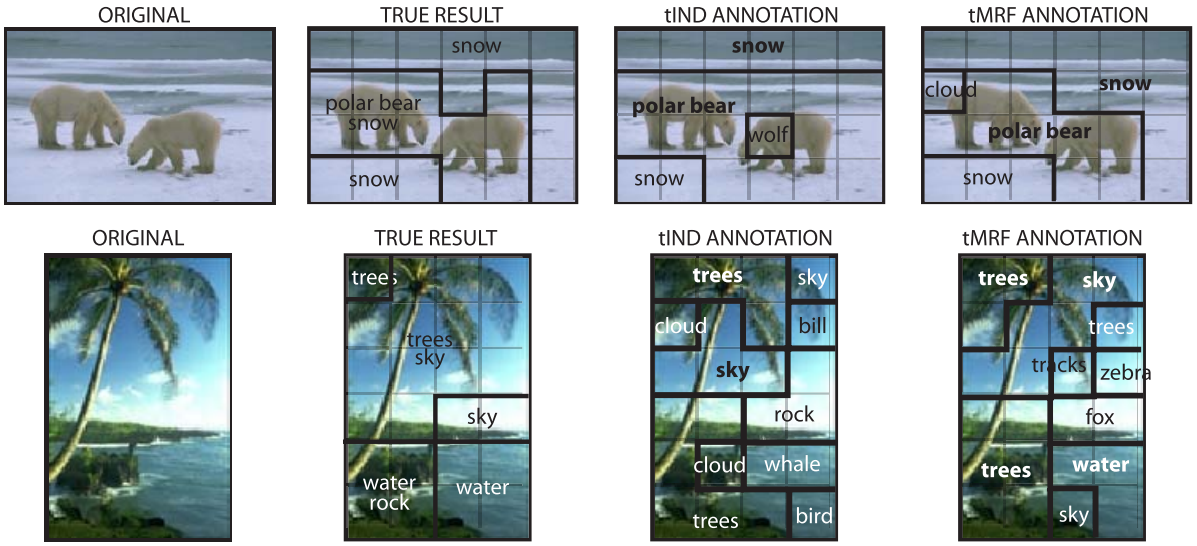

Figure 5: *Selected annotations on the* Corel A *training set for both the independent (tInd) and contextual (tMRF) translation models. For clarity, a single word is used to represent a uniformly-annotated region. The manual annotation is shown second to the left. The patch grid is depicted with faint lines.*

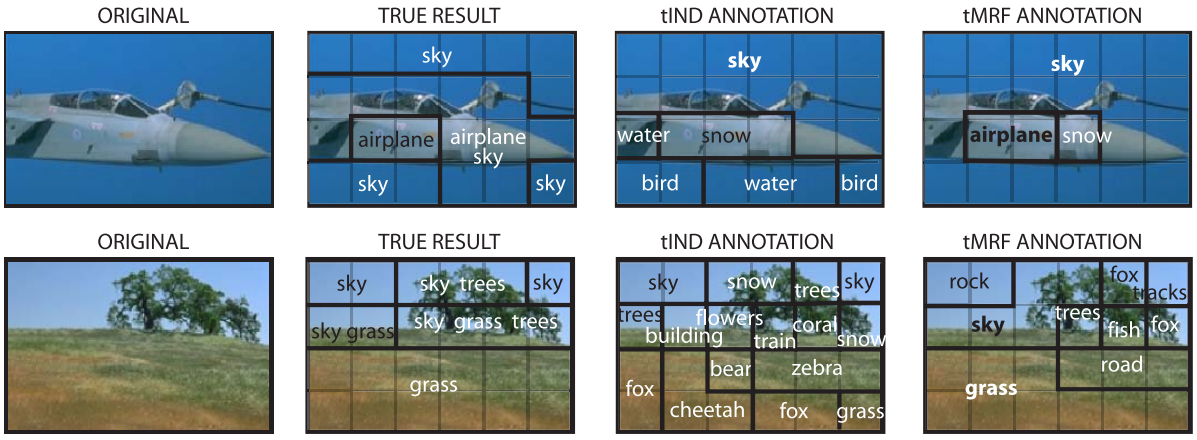

Figure 6: *Selected annotations on the* Corel B *test set for the the independent (tInd) and contextual (tMRF) translation models. For clarity, a single word is used to represent a uniformly-annotated region. The manual annotation is shown second to the left. The patch grid is depicted with faint lines.*

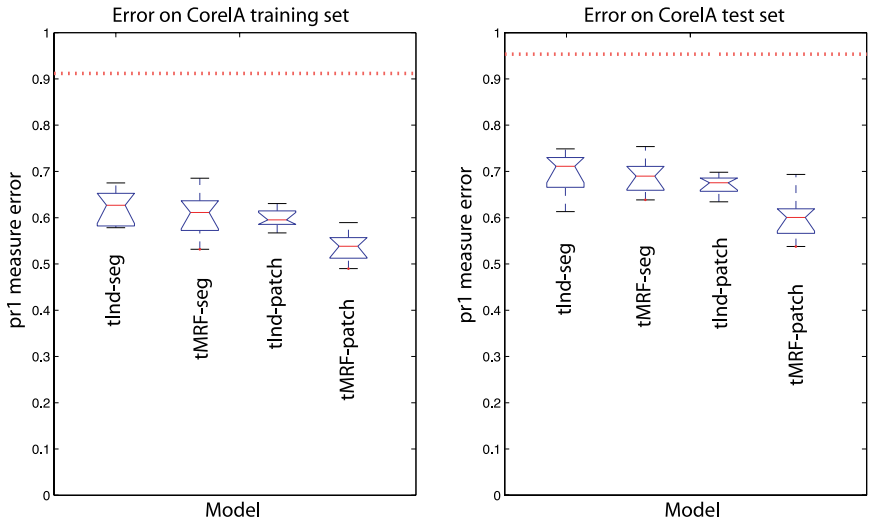

Figure 7: *Results using the* prn1 *measure for the* Corel A *training and test sets, displayed using a Box-and-Whisker plot. The middle line of a box represents the median. The central box represents the values from the 25 to 75 percentiles. The horizontal line extends from the minimum to the maximum value, excluding outside and far out values which are displayed as separate points. The first two models of each plot are results for the segmented images and the last two are results on the images with uniform patch grids. The dotted line at the top represents the random prediction upper bound. Overall, the contextual model $tMRF$ is an improvement over the independent model, $tInd$, and both perform better using a naive patch segmentation.*

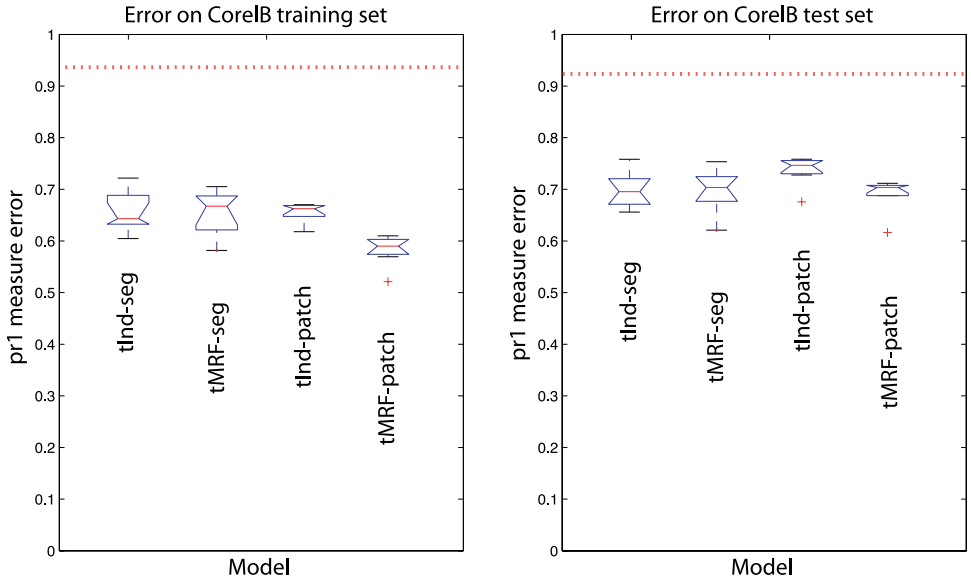

Figure 8: *Results using the* prn1 *measure for the Corel B training and test sets, displayed using a Box-and-Whisker plot. See figure 8 for an explanation of the graph. The contextual model $tMRF$ outperforms the independent model, $tInd$. We achieve the best annotations when $tMRF$ is trained on the crude segmentations.*