Why can't José read? - The problem of learning semantic associations in a robot environment

Peter Carbonetto, Nando De Freitas

Why can't José read?
The problem of learning semantic associations in a robot environment

Peter Carbonetto
Department of Computer Science
University of British Columbia
pcarbo@cs.ubc.ca

Nando de Freitas
Department of Computer Science
University of British Columbia
nando@cs.ubc.ca

Abstract

We study the problem of learning to recognise objects in the context of autonomous agents. We cast object recognition as the process of attaching meaningful concepts to specific regions of an image. In other words, given a set of images and their captions, the goal is to segment the image, in either an intelligent or naive fashion, then to find the proper mapping between words and regions. In this paper, we demonstrate that a model that learns spatial relationships between individual words not only provides accurate annotations, but also allows one to perform recognition that respects the real-time constraints of an autonomous, mobile robot.

1 Introduction

In writing this paper we hope to promote a discussion on the design of an autonomous agent that learns semantic associations in its environment or, more precisely, that learns to associate regions of images with discrete concepts. When an image region is labeled with a concept in an appropriate and consistent fashion, we say that the object has been *recognised* (Duygulu et al., 2002). We use our laboratory robot, José (Elinas et al., 2002), as a prototype, but the ideas presented here extend to a wide variety of settings and agents.

Before we proceed, we must elucidate on the requirements for achieving semantic learning in an autonomous agent context.

Primarily, we need a model that learns associations between objects given a set of images paired with user input. Formally, the task is to find a function that separates the space of image patch descriptions into n_w semantic concepts, where n_w is the total number of concepts in the training set (from now on we use the word "patch" to refer to a contiguous region in an image). These supplied concepts could be in the form of text captions, speech, or anything else that might convey semantic information. For the time being, we restrict the set of concepts to English nouns (e.g. "face", "toothbrush", "floor"). See Figure 1 for examples of images paired with captions composed of nouns. Despite this restriction, we still leave ourselves open to a great deal of ambiguity and uncertainty, in part because objects can be described at several different levels of specificity, and at the same level using different words (e.g. is it "sea", "ocean", "wave" or "water"?). Ideally, one would like to impose a hierarchy of lexical concepts, as in WordNet (Fellbaum, 1998). We have yet to explore WordNet for our proposed framework,

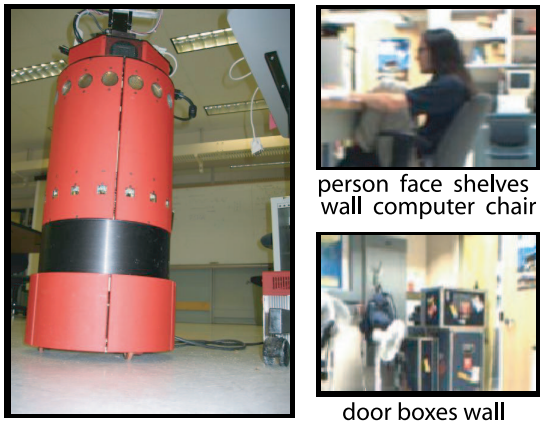

Figure 1: *The image on the left is José (Elinas et al., 2002), the mobile robot we used to collect the image data. The images on the right are examples the robot has captured while roaming in the lab, along with labels used for training. We depict image region annotations in later figures, but we emphasize that* ***the robot receives only the labels as input for training****. That is, the robot does not know what words correspond to the image regions.*

though it has been used successfully for image clustering (Barnard et al., 2001; Barnard et al., 2002).

Image regions, or patches, are described by a set of low-level features such as average and standard deviation of colour, average oriented Gabor filter responses to represent texture, and position in space. The set of patch descriptions forms an n_f-dimensional space of real numbers, where n_f is the number of features. Even complex low-level features are far from adequate for the task of classifying patches as objects — at some point we need to move to representations that include high-level information. In this paper we take a small step in that direction since our model learns spatial relations between concepts.

Given the uncertainty regarding descriptions of objects and their corresponding concepts, we further require that the model be probabilistic. In this paper we use Bayesian techniques to construct our object recognition model.

Implicitly, we need a thorough method for decomposing an image into conceptually contiguous regions. This is not only non-trivial, but also impossible without considering semantic associations. This motivates the segmentation of images and learning associations between patches and words as tightly coupled processes.

The subject of segmentation brings up another important consideration. A good segmentation algorithm such as Normalized Cuts (Shi and Malik, 1997) can take on the order of a minute to complete. For many real-time applications this is an unaffordable expense. It is important to abide by real-time constraints in the case of a mobile robot, since it has to simultaneously recognise and negotiate obstacles while navigating in its environment. Our experiments suggest that the costly step of a decoupled segmentation can be avoided without imposing a penalty to object recognition performance.

Autonomous semantic learning must be considered a supervised process or, as we will see later on, a partially-supervised process since the associations are made from the perspective of humans. This motivates a second requirement: a system for the collection of data, ideally in an on-line fashion. As mentioned above, user input could come in the form of text or speech. However, the collection of data for supervised classification is problematic and time-consuming for the user overseeing the autonomous agent, since the user is required to tediously feed the agent with self-annotated regions of images. If we relax our requirement on training data acquisition by requesting captions at an image level, not at a patch level, the acquisition of labeled data is suddenly much less challenging. Throughout this paper, we use manual annotations purely for testing only — we emphasize that the training data includes *only* the labels paired with images.

We are no longer exploring object recognition as a strict classification problem, and we do so at a cost since we are no longer blessed with the exact associations between image regions and nouns. As a result, the learning problem is now unsupervised. For a single training image and a particular word token, we must now learn both the probability of generating that word given an object description and the correct association to one of the regions with the image. Fortunately, there is a straightforward parallel between our object recognition formulation and the statistical machine translation problem of building a lexicon from an aligned bitext (Brown et al., 1993; Al-Onaizan et al., 1999). Throughout this paper, we reason about object recognition with this analogy in mind (Duygulu et al., 2002).

What other requirements should we consider? Since our discussion involves autonomous agents, we should pursue a dynamic data acquisition model. We can consider the problem of learning an object recognition model as an on-line conversation between the robot and the user, and it follows the robot should be able to participate. If the agent ventures into "unexplored territory", we would like it to make unprompted requests for more assistance. One could use active learning to implement a scheme for requesting user input based on what information would be most valuable to classification. This has yet to be explored for object recognition, but it has been applied to the related domain of image retrieval (Tong and Chang, 2001). Additionally, the learning process could be coupled with reinforcement — in other words, the robot could offer hypotheses for visual input and await feedback from user.

In the next section, we outline our proposed contextual translation model. In Section 3, we weigh the merits of several different error measures for the purposes of evaluation. The experimental results on the robot data are given in Section 4. We leave discussion of results and future work to the final section of this paper.

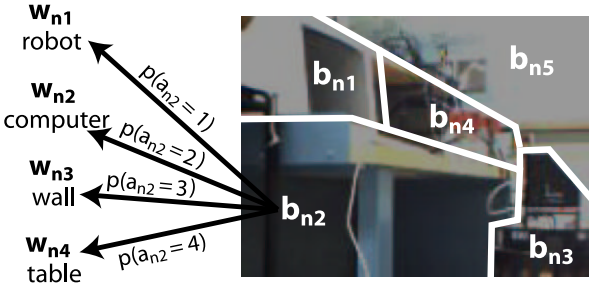

Figure 2: *The alignment variables represent the correspondences between label words and image patches. In this example, the correct association is* $a_{n2} = 4$.

2 A contextual translation model for object recognition

In this paper, we cast object recognition as a machine translation problem, as originally proposed in (Duygulu

et al., 2002). Essentially, we translate patches (regions of an image) into words. The model acts as a *lexicon*, a dictionary that predicts one representation (words) given another representation (patches). First we introduce some notation, and then we build a story for our proposed probabilistic translation model.

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 $\{w_{n1}, ..., w_{nL_n}\}$. M_n is the number of patches in image n and L_n is the number of words in the image's caption. 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 variable a_{nj}, such that $a^i_{nj} = 1$ if b_{nj} translates to w_{ni}; otherwise, $a^i_{nj} = 0$. Therefore, $p(a^i_{nj}) \triangleq p(a_{nj} = i)$ is the probability that patch b_{nj} is aligned with word w_{ni} in document n. See Figure 2 for an illustration. n_w is the total number of word tokens in the training set.

We construct a joint probability over the translation parameters and latent alignment variables in such a way that maximizing the joint results in what we believe should be the best object recognition model (keeping in mind the limitations placed by our set of features!). Without loss of generality, the joint probability is

$$p(\mathbf{b}, \mathbf{a}|\mathbf{w}) = \prod_{n=1}^{N} \prod_{j=1}^{M_n} p(a_{nj}|a_{n,1:j-1}, b_{n,1:j-1}, \mathbf{w_n}, \theta) \times p(b_{nj}|a_{n,1:j}, b_{n,1:j-1}, \mathbf{w_n}, \theta) \quad (1)$$

where $\mathbf{w_n}$ denotes the set of words in the nth caption, $a_{n,1:j-1}$ is the set of latent alignments 1 to $j-1$ in image n, $b_{n,1:j-1}$ is the set of patches 1 to $j-1$, and θ is the set of model parameters.

Generally speaking, alignments between words and patches depend on all the other alignments in the image, simply because objects are not independent of each other. These dependencies are represented explicitly in equation 1. However, one usually assumes $p(a_{nj}|a_{n,1:j-1}, b_{n,1:j-1}, \mathbf{w_n}, \theta) = p(a_{nj} = i|\mathbf{w_n}, \theta)$ to guarantee tractability. In this paper, we relax the independence assumption in order to exploit spatial context in images and words. We allow for interactions between neighbouring image annotations through a pairwise Markov random field (MRF). That is, the probability of a patch being aligned to a particular word depends on the word assignments of adjacent patches in the image. It is reasonable to make the assumption that given the alignment for a particular patch, translation probability is independent from the other patch-word alignments. A simplified version of the graphical model for illustrative purposes is shown in Figure 3.

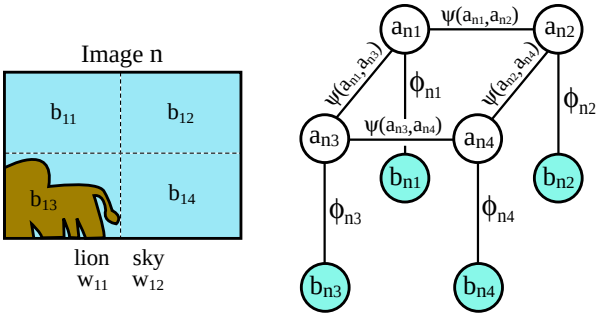

Figure 3: *The graphical model for a simple set with one document. The shaded circles are the observed nodes (i.e. the data). The white circles are unobserved variables of the model parameters. Lines represent the undirected dependencies between variables. The potential ψ controls the consistency between annotations, while the potentials ϕ_{nj} represent the patch-to-word translation probabilities.*

In Figure 3, the potentials $\phi_{nj} \triangleq p(b_{nj}|w^\star)$ are the patch-to-word translation probabilities, where $w^\star$ denotes a particular word token. We assign a Gaussian distribution to each word token, so $p(b_{nj}|w^\star) = \mathcal{N}(b_{nj}; \mu_{w^\star}, \Sigma_{w^\star})$. The potential $\psi(a_{nj}, a_{nk})$ encodes the compatibility of the two alignments, a_{nj} and a_{nk}. The potentials are the same for each image. That is, we use a single $W \times W$ matrix ψ, where W is the number of word tokens. The final joint probability is a product of the translation potentials and the inter-alignment potentials:

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

where $\delta_{w^\star}(w_{ni}) = 1$ if the ith word in the nth caption is the word $w^\star$; otherwise, it is 0.

To clarify the unsupervised model described up to this point, it helps to think in terms of counting word-to-patch alignments for updating the model parameters. Loosely speaking, we update the translation parameters $\mu_{w^\star}$ and $\Sigma_{w^\star}$ by counting the number of times particular patches are aligned with word $w^\star$. Similarly, we update $\psi(w^\star, w^\diamond)$ by counting the number of times the word tokens $w^\star$ and $w^\diamond$ are found in adjacent patch alignments. We normalize the latter count by the overall alignment frequency to prevent counting alignment frequencies twice.

In addition, we use a hierarchical Bayesian scheme to provide regularised solutions and to carry out automatic feature weighting or selection (Carbonetto et al., 2003).

In summary, our learning objective is to find good values for the unknown model parameters $\theta \triangleq \{\mu, \Sigma, \psi, \tau\}$,

where μ and Σ are the means and covariances of the Gaussians for each word, ψ is the set of alignment potentials and τ is the set of shrinkage hyper-parameters for feature weighting. For further details on how to compute the model parameters using approximate EM and loopy belief propagation, we refer the reader to (Carbonetto et al., 2003; Carbonetto and de Freitas, 2003)

3 Evaluation metric considerations

Before we discuss what makes a good evaluation metric, it will help if we answer this question: "what makes a good image annotation?" As we will see, there is no straightforward answer.

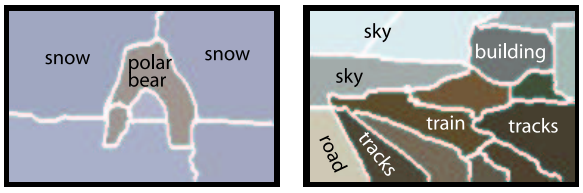

Figure 4: *Examples of images which demonstrate that the importance of concepts has little or no relation to the area these concepts occupy. On the left, "polar bear" is at least pertinent as "snow" even though it takes up less area in the image. In the photograph on the right, "train" is most likely the focus of attention.*

It is fair to say that certain concepts in an image are more prominent than others. One might take the approach that objects that consume the most space in an image are the most important, and this is roughly the evaluation criterion used in previous papers (Carbonetto et al., 2003; Carbonetto and de Freitas, 2003). Consider the image on the left in Figure 4. We claim that "polar bear" is at least as important as snow. There is an easy way to test this assertion – pretend the image is annotated either entirely as "snow" or entirely as "polar bear". In our experience, people find the latter annotation as appealing, if not more, than the former. Therefore, one would conclude that it is better to weight all concepts equally, regardless of size, which brings us to the image on the right. If we treat all words equally, having many words in a single label obfuscates the goal of getting the most important concept,"train", correct.

Ideally, when collecting user-annotated images for the purpose of 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 an burdensome task.

For lack of a good metric, we evaluate the proposed translation models using two error measures. *Error measure 1* reports an error of 1 if the model annotation with the highest probability results in an incorrect patch annotation. The error is averaged over the number of patches in each image, and then again over the number of images in the data set. *Error measure 2* is similar, only we average the error over the patches corresponding to word (according to the manual annotations). The equations are given by

$$E.m.\ 1 \triangleq \frac{1}{N}\sum_{n=1}^{N}\frac{1}{M_n}\sum_{j=1}^{M_n}\left(1-\delta_{\tilde{a}_{nj}}(\widehat{a}_{nj})\right) \quad (2)$$

$$E.m.\ 2 \triangleq \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L_n}\sum_{i=1}^{L_n}\frac{1}{|\mathcal{P}_{ni}|}\sum_{j}^{\mathcal{P}_{ni}}\left(1-\delta_{\tilde{a}_{nj}}(\widehat{a}_{nj})\right) \quad (3)$$

where $\mathcal{P}_{ni}$ is the set of patches in image n that are manually-annotated using word i, $\widehat{a}_{nj}$ is the model alignment with the highest probability, $\tilde{a}_{nj}$ is the provided "true" annotation, and $\delta_{\tilde{a}_{nj}}(\widehat{a}_{nj})$ is 1 if $\tilde{a}_{nj} = \widehat{a}_{nj}$.

Our intuition is that the metric where we weight all concepts equally, regardless of size, is better overall. As we will see in the next section, our translation models do not perform as well under this error measure. This is due to the fact that the joint probability shown in equation 1 maximises the first error metric, not the second. Since the agent cannot know the true annotations beforehand, it is difficult to construct a model that maximises the second error measure, but we are currently pursuing approximations to this metric.

4 Experiments

We built a data set by having José the robot roam around the lab taking pictures, and then having laboratory members create captions for the data using a consistent set of words. For evaluation purposes, we manually annotated the images. The *robomedia* data set is composed of 107 training images and 43 test images [1]. The training and test sets contain a combined total of 21 word tokens. The word frequencies in the labels and manual annotations are shown in figure 5.

In our experiments, we consider two scenarios. In the first, we use Normalized Cuts (Shi and Malik, 1997) to segment the images into distinct patches. 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 different segmentations are shown along with the anecdotal results in Figure 8. For the crude segmentation, we used patches of height and width approximately $1/6$th the size of the image. We found that smaller patches introduced too much noise to the features and resulted in poor test performance, and larger patches contained too many objects at once. In future work, we

[1] Experiment data and Matlab code are available at `http://www.cs.ubc.ca/~pcarbo`.

	LABEL%		ANNOTATION%‡		PRECISION	
WORD	**TRAIN**	**TEST†**	**TRAIN**	**TEST**	**TRAIN**	**TEST**
backpack	0.019	0.011	0.008	0.002	0.158	0.115
boxes	0.022	0.011	0.038	0.028	0.218	0.081
cabinets	0.080	0.066	0.118	0.081	0.703	0.792
ceiling	0.069	0.066	0.061	0.063	0.321	0.347
chair	0.131	0.148	0.112	0.101	0.294	0.271
computer	0.067	0.071	0.052	0.065	0.149	0.144
cooler	0.004	n/a	0.002	n/a	0.250	n/a
door	0.084	0.055	0.067	0.042	0.291	0.368
face	0.011	0.022	0.001	0.002	0.067	0.042
fan	0.022	0.011	0.012	0.005	0.114	0.133
filers	0.030	0.033	0.028	0.019	0.064	0.077
floor	0.004	n/a	0.004	n/a	0.407	n/a
person	0.022	0.049	0.018	0.040	0.254	0.340
poster	0.037	0.033	0.026	0.021	0.471	0.368
robot	0.011	0.016	0.008	0.011	0.030	0.014
screen	0.041	0.049	0.042	0.051	0.289	0.263
shelves	0.082	0.071	0.115	0.120	0.276	0.281
table	0.032	0.038	0.027	0.049	0.160	0.121
tv	0.026	0.027	0.007	0.007	0.168	0.106
wall	0.103	0.104	0.109	0.122	0.216	0.216
whiteboard	0.105	0.120	0.146	0.171	0.274	0.278
Totals	**1.000**	**1.000**	**1.000**	**1.000**	**0.319**	**0.290**

Figure 5: *The first four columns list the probability of finding a particular word in a label and a manually annotated patch, in the* robomedia *training and test sets. The final two columns show the precision of the translation model* tMRF *using the grid segmentation for each token, averaged over the 12 trials. Precision is defined as the probability the model's prediction is correct for a particular word and patch. Since precision is 1 minus the error of equation 3, the total precision on both the training and test sets matches the average performance of* tMRF-patch *on Error measure 2, as shown in in Figure 7. While not presented in the table, the precision on individual words varies significantly from one one trial to the next. Note that some words do not appear in both the training and test sets, hence the* n/a.
†*The model predicts words* ***without*** *access to the test image labels. We provide this information for completeness.*
‡*We can use the manual annotations for evaluation purposes, but we underline the fact that an agent would not have access to the information presented in the "*Annotation %*" column.*

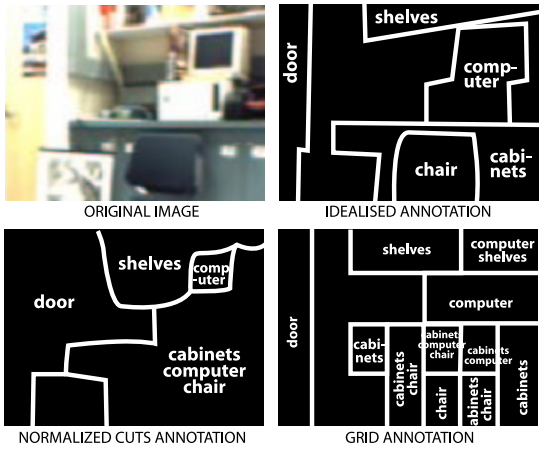

Figure 6: *Correct annotations for Normalized Cuts, grid and manual segmentations. When there are multiple annotations in a single patch, any one of them is correct. Even when both are correct, the grid segmentation is usually more precise and, as a result, more closely approximates generic object recognition.*

will investigate a hierarchical patch representation to take into account both short and long range patch interactions, as in (Freeman and Pasztor, 1999).

We compare two models. The first is the translation model where dependencies between alignments are removed for the sake of tractability, called *tInd*. The second is the translation model in which we assume dependences between adjacent alignments in the image. This model is denoted by *tMRF*. We represent the sophisticated and crude segmentation scenarios by *-seg* and *-patch*, respectively.

One admonition regarding the evaluation procedure: a translation is deemed correct if at least one of the patches corresponds to the model's prediction. In a manner of speaking, when a segment encompasses several concepts, we are giving the model the benefit of the doubt. For example, according to our evaluation the annotations for both the grid and Normalized Cuts segmentations shown in Figure 6 correct. However, from observation the grid segmentation provides a more precise object recognition. As a result, evaluation can be unreliable when Normalized Cuts offers poor segmentations. It is also important to remember that the *true result* images shown in the second column of Figure 8 are idealisations.

Experimental results on 12 trials are shown in Figure 7, and selected annotations predicted by the *tMRF* model on the test set are shown in Figure 8. The most significant result is that the contextual translation model performs the best overall, and performs equally well when supplied with either Normalized Cuts or a naive segmentations. We stress that even though the models trained using both the grid and Normalized Cuts segmentations are displayed on the same plots, in Figure 6 we indicate that *object recognition using the grid segmentation is generally more precise, given the same evaluation result in Figure 7.* Learning contextual dependencies between alignment appears to improve performance, despite the large amount of noise and the increase in the number of model parameters that have to be learned. The contex-

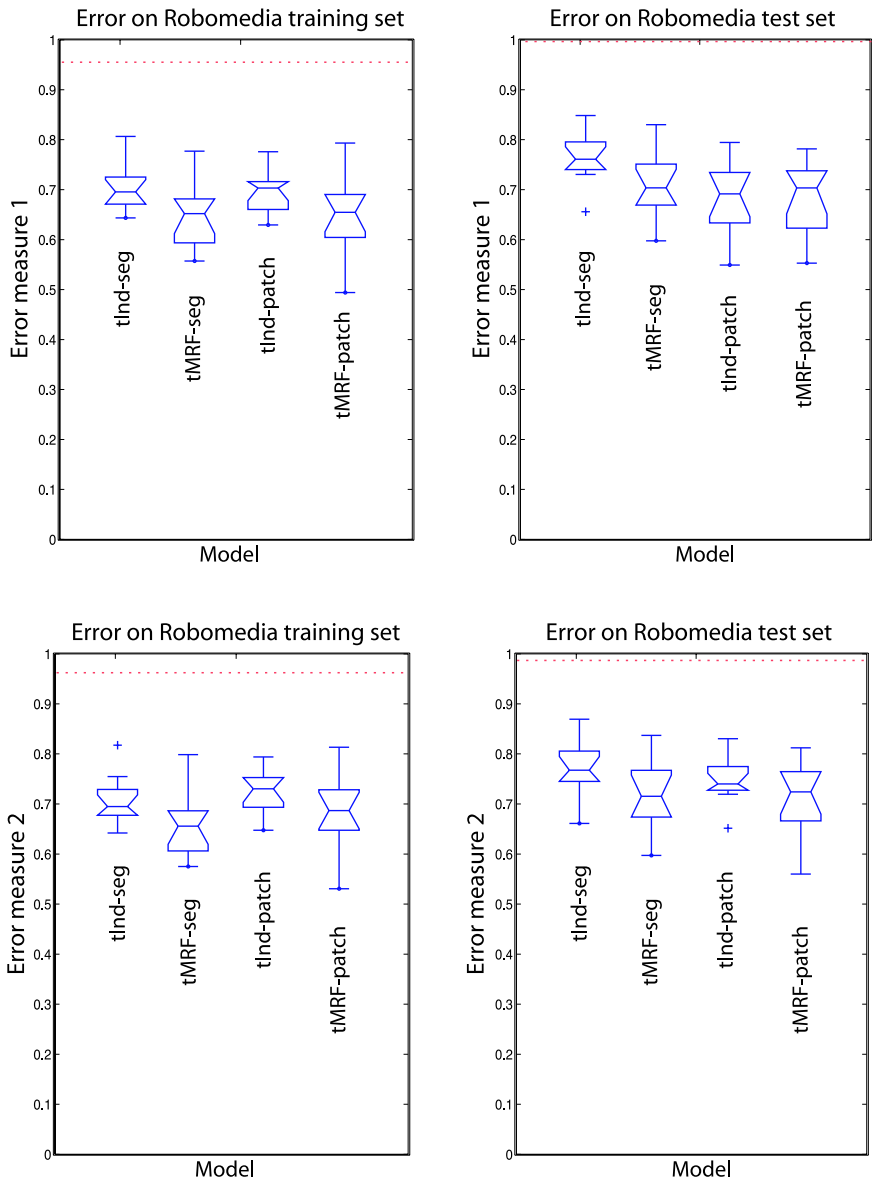

Figure 7: *Results using Error measures 1 and 2 on the* robomedia *training and test sets, displayed using a Box-and-Whisker plot. The middle line of a box represents the median. The central box represents the values from the 25 to 75 percentile, using the upper and lower statistical medians. The horizontal 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 prediction upper bound. Overall, the contextual model* tMRF *is an improvement over the independent model,* tInd. *On average,* tMRF *tends to perform equally well using the sophisticated or naive patch segmentations.*

tual model also tends to produce more visually appealing annotations since they the translations smoothed over neighbourhoods of patches.

The performance of the contextual translation model on individual words on the training and test sets is shown in Figure 5, averaged over the trials. Since our approximate EM training a local maximum point estimate for the joint posterior and the initial model parameters are set to random values, we obtain a great deal of variance from one trial to the next, as observed in the Box-and-Whisker plots in Figure 7. While not shown in Figure 5, we have noticed considerable variation in what words are predicted with high precision. For example, the word "ceiling" is predicted with an average success rate of 0.347, although the precision on individual trials ranges from 0 to 0.842.

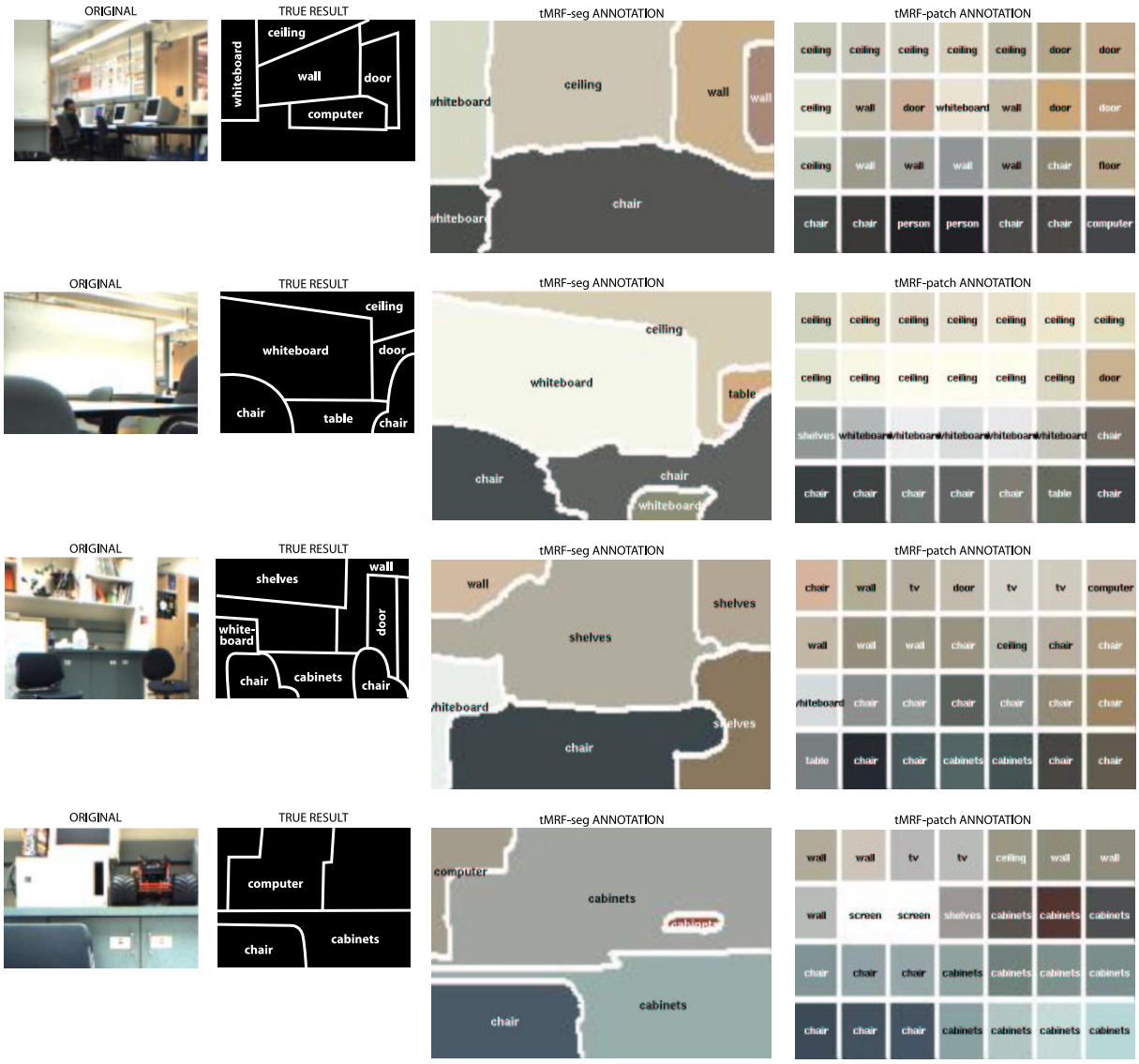

Figure 8: *Selected annotations on the* robomedia *test data predicted by the contextual (tMRF) translation model. We show our model's predictions using both sophisticated and crude segmentations. The "true" annotations are shown in the second column. Notice that the annotations using Normalized Cuts tend to be more visually appealing compared to the rectangular grid, but intuition is probably misleading: the error measures in Figure 7 demonstrate that both segmentations produce equally accurate results. It is also important to note that these annotations are probabilistic;* ***for clarity we only display results with the highest probability****.*

From the Bayesian feature weighting priors τ placed on the word cluster means, we can deduce the relative importance of our feature set. In our experiments, luminance and vertical position in the image are the two most important features.

5 Discussion and conclusion

Our experiments suggest that we can eliminate the costly step of segmentation without incurring a penalty to the object recognition task. This realisation allows us to remove the main computational bottleneck and pursue real-time learning in a mobile robot setting. Moreover, by introducing spatial relationships into the model, we maintain a degree of consistency between individual patch annotations. We can consider this to be an early form of segmentation that takes advantage of both high-level and low-level information. Thus, we are solving both the segmentation and recognition problems simultaneously. However, we emphasize the need for further investigation to pin down the role of segmentation in the image translation process.

Our translation model is disposed to predicting certain words better than others. However, at this point we cannot make make any strong conclusions as to why certain words easy to classify (e.g. cabinets), while others are difficult (e.g. filers). From Figure 5, it appears to be the case that words that occur frequently and possess a consistent set of features tend to be more easily classified.

Initially, we were doubtful that spatial context in the model would improve results given that the robot roams in a fairly homogeneous environment. This contrasts with experiments on the *Corel* data sets (Carbonetto and de Freitas, 2003), whereby the photographs were captured from a wide variety of settings. However, the experiments on the *robomedia* data demonstrate that there is something to be gained by introducing inter-alignment dependencies in the model, even in environments with relatively noisy and unreliable data.

Generic object recognition in the context of robotics is a challenging task. Standard low-level features such as colour and texture are particularly ineffective in a laboratory environment. For example, chairs can come in a variety of shapes and colours, and "wall" refers to a vertical surface that has virtually no relation to colour, texture and position. Moreover, it is much more difficult to delineate specific concepts in a scene, even for humans — does a table include the legs, and where does one draw the line between shelves, drawers, cabinets and the objects contained in them? (This explains why many of the manually-annotated patches in Figures 6 and 8 are left empty.) Object recognition on the *Corel* data set is comparatively easy because the photos are captured to artificially delineate specific concepts. Colour and texture tend to be more informative in natural scenes.

In order to tackle concerns mentioned above, one approach would be to construct a more sophisticated representation of objects. A more realistic alternative would be to reinforce our representation with high-level features, including more complex spatial relations.

One important criterion we did not address explicitly is on-line learning. Presently, we train our models assuming that all the images are collected at one time. Research shows that porting batch learning to an on-line process using EM does not pose significant challenges (Smith and Makov, 1978; Sato and Ishii, 2000; Brochu et al., 2003). With the discussion presented in this paper in mind, real-time interactive learning of semantic associations in José's environment is very much within reach.

Acknowledgements

We would like to acknowledge the help of Eric Brochu in revising and proofreading this paper, Kobus Barnard and David Forsyth for enlightening discussions, and the José team, in particular Don Murray and Pantelis Elinas, for helping us collect invaluable data. Additionally, the workshop reviewer committee offered very insightful suggestions and criticisms, so we would like to thank them as well.

References

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. 1999. Statistical machine translation: final report. *Johns Hopkins University Workshop on Language Engineering.*

Kobus Barnard, Pinar Duygulu and David Forsyth. 2001. Clustering art. *Conference on Computer Vision and Pattern Recognition.*

Kobus Barnard, Pinar Duygulu and David Forsyth. 2002. Modelling the statistics of image features and associated text. *Document Recognition and Retrieval IX, Electronic Imaging.*

Eric Brochu, Nando de Freitas and Kejie Bao. 2003. The Sound of an album cover: probabilistic multimedia and IR. *Workshop on Artificial Intelligence and Statistics.*

P. Brown, S. A. Della Pietra, V.J. Della Pietra and R. L. Mercer. 1993. The Mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

Peter Carbonetto and Nando de Freitas. 2003. A statistical translation model for contextual object recognition. *Unpublished manuscript.*

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

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

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

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database.* MIT Press.

William T. Freeman and Egon C. Pasztor. 1999. Learning low-level vision. *International Conference on Computer Vision.*

Masa-aki Sato and Shin Ishii. 2000. On-line EM algorithm for the Normalized Gaussian Network. *Neural Computation*, 12(2):407-432.

Jianbo Shi and Jitendra Malik. 1997. Normalized cuts and image segmentation. *Conference on Computer Vision and Pattern Recognition.*

A. F. M. Smith and U. E. Makov. 1978. A Quasi-Bayes sequential procedure for mixtures. *Journal of the Royal Statistical Society*, Series B, 40(1):106-111.

Simon Tong and Edward Chang. 2001 Support vector machine active learning for image retrieval. *ACM Multimedia.*