Bayesian Models for Massive Multimedia Databases: a New Frontier *

Nando de Freitas and Eric Brochu
University of British Columbia
Vancouver, BC, Canada
{nando , ebrochu}@cs.ubc.ca

Kobus Barnard, Pinar Duygulu and David Forsyth
University of California Berkeley
Berkeley, California, USA
{kobus , duygulu, daf}@cs.berkeley.edu

February 18, 2003

Abstract

Modelling the increasing number of digital databases (the web, photo-libraries, music collections, news archives, medical databases) is one of the greatest challenges of statisticians in the new century. Despite the large amounts of data, the models are so large that they motivate the use of Bayesian models. In particular, the Bayesian perspective allows us to perform automatic regularisation to obtain sparse and coherent models. It also enables us to encode a priori knowledge, such as word, music and image preferences. The learned models can be used for browsing digital databases, information retrieval with image, music and/or text queries, image annotation (adding words to an image), text illustration (adding images to a text), and object recognition.

1 Introduction

A new frontier for Bayesian statistics has arisen with the expansion of data – in the form of images, video, text, sounds and other media – in digital databases and on the world-wide-web. A naive understanding of this frontier could lead us to think that the massive amounts of data warrant the sole use of frequentist techniques. However, the models used to describe these databases are also massive. We are in the realm of

*Presented at the 7th Valencia international meeting on Bayesian statistics, 2002.

models with thousands or millions of parameters. This motivates the use of Bayesian and information theoretic regularisation tools.

In addition, the type of problems and data – e.g. words – lend themselves to the incorporation of a priori knowledge through appropriate prior distributions. For example we can assign an advantageous prior probability to (application specific) special words. Natural language processing can also provide useful priors. When the models generate document items from a hierarchical set of nodes with more general "concepts" being emitted from higher levels, we can use information from WordNet, see Miller et al (1998) to encourage models to emit words at a level corresponding to their level of semantic abstraction.

In this frontier, there are many new, varied and exciting applications. These include the design of search engines for information retrieval with images, music and text; constructing browsing tools for digital databases; and combining different sources of information in useful multimedia applications, such as automatic annotation of text with images and automatic illustration of images with words. The latter application can be extended to labelling image regions with words. This is an important new direction in object recognition.

We have identified four areas of research within this frontier: designing more expressive and parsimonious models, defining utilities and performance measures, constructing algorithms that perform well in high-dimensions, and exploring new applications. In this paper, we provide a glimpse of our research in these areas. In particular, in the context of applying latent variable models to large digital databases containing documents with text, music and images.

2 Models

Let $\{\underline{x}_1, \ldots, \underline{x}_{n_x}\}$ denote a collection of n_x documents in a heterogeneous database. Each document $\underline{x}_i$ is assumed to have n_a different attributes, $\underline{x}_i = (\underline{x}_{i,1}, \ldots, \underline{x}_{i,a}, \ldots, \underline{x}_{i,n_a})$. Attributes may be categorical or continuous. Examples of categorical attributes include text, music notation symbols, and standard document meta-data. Features derived from images and other multimedia signals are typically continuous-valued attributes.

At present most of our work is based on multi-modal finite mixture models. The simplest model being the following:

$$\underline{x}_i|\varphi \sim \sum_{c=1}^{n_c} \lambda_c \prod_{a=1}^{n_a} p(\underline{x}_{i,a}|\theta_{a,c}),$$

where $\varphi = (\lambda, \theta)$ encompasses all the model parameters, λ denotes the mixing weights, θ denotes the parameters of the mixture component densities, and n_c denotes the number of components. We introduce the latent allocation variables $z_i \in \{1, \ldots, n_c\}$ to indicate that a particular document $\underline{x}_i$ belongs to a specific group c. That is, $p(z_i = c) = \lambda_c$ independently for $c = 1, \ldots, n_c$.

For clarity, we are restricting the presentation to a simple multi-modal mixture model, but it should be emphasized that we are currently using other models. In particular, we tend to use hierarchical mixtures as these provide a natural structure for database browsing (Barnard *et al* (2001)). We have also investigated the use of the aspect models of Hofmann (1999) and the very promising Dirichlet aspect models of Blei *et al* (2001). In other applications, such as statistical machine translation and object recognition, we have adopted data association models (see Duygulu *et al* (2002)).

We consider three types of mixture components for text, music and images (for simplicity, we drop the sub-index a).

Text: we assume that the data is available as a co-occurrence table of word counts $\underline{\mathrm{x}}$, where $x_{i,w}$ denotes the number of times word w appears in document $\underline{\mathrm{x}}_i$. For computational simplicity, we adopt a simple multinomial (naive Bayes) model:

$$p(\underline{\mathrm{x}}_i|\theta_c) = \prod_{w=1}^{n_w} \delta_{w,c}^{x_{i,w}},$$

where n_w, in this case, denotes the number of words in the vocabulary (dictionary), and $\delta_{w,c}$ denotes the probability of each word w in cluster c, with $\sum_w \delta_{w,c} = 1$. For notational simplicity, we have ignored the normalisation factor of the multinomial density (we assume that the document length is class-independent).

One can extend this text model by incorporating links into the table or word counts, see Cohn *et al* (2001). This information is of great relevance in the design of search engines. In Cohn *et al* (2001), the text and links are weighted by the heuristic constants. In the Bayesian framework, this weighting is performed by the prior and can be automatically computed using the data.

Music: musical scores, available in GUIDO format (see Hoos *et al* (2001)), are modelled with first order Markov chains with n_s states:

$$p(\underline{\mathrm{x}}_i|\theta_c) = \prod_{k=1}^{n_s} \nu_{k,c}^{\mathbb{1}_k(x_{i0})} \prod_{k=1}^{n_s} \prod_{l=1}^{n_s} \gamma_{k,l,c}^{x_{i,k,l}},$$

where $\underline{\mathrm{x}}_i$ denotes the transition matrix of the i-th score, $x_{i,k,l}$ denotes the transition probability from state k to state l, $x_{i,0}$ denotes the initial state, and $\mathbb{1}_k(x_{i0})$ is the set indicator function.

Images: images are segmented into homogeneous regions using standard segmentation algorithms. We treat the image features (colour, texture, etc.) of each segment as samples from multivariate Gaussian distributions (Barnard *et al* (2001)). That is, the a-th segment of image $\mathbf{x_i}$ is Gaussian distributed with density

$$p(\underline{\mathrm{x}}_{i,a}|\theta_{a,c}) = |2\pi\Sigma_{a,c}|^{-1/2} \exp\left(-\frac{1}{2}\left(\underline{\mathrm{x}}_{i,a} - \mu_{a,c}\right)' \Sigma_{a,c}^{-1}\left(\underline{\mathrm{x}}_{i,a} - \mu_{a,c}\right)\right),$$

Images with n_a segments are represented with a product of n_a Gaussians. That is, the image segments are assumed to be independent given the cluster variable.

2.1 Priors

Prior distributions are required for several reasons. First, they serve to reduce the ill-conditioning problem inherent to the maximum likelihood framework (the likelihood is unbounded). Second, one can use the prior to specify domain-specific knowledge (some rules derived from an expert) or subjective preferences (favouring simpler models). Third, maximum likelihood estimators often result in multiple mixture components with the same parameters and whose weights λ_c add up to the correct one. Bayesian estimation avoids this problem by specifying priors that favour sparse models.

We assign appropriate Dirichlet and normal-inverse Wishart priors, see for example Bernardo *et al* (1994). The hyper-parameters are computed via maximum likelihood type II (see Good (1983), Gelman *et al* (1995), and Narayanan (1991)). However, in some cases we fix the hyper-parameters using domain knowledge. For example, we can assign an advantageous prior probability to (application specific) special words. These priors allow us to, for example, bias search engines. Natural language processing can also provide useful priors. In this work, since the model generates document items from a hierarchical set of nodes with more general "concepts" being emitted from higher levels, we use information from WordNet to encourage the model to emit words at a level corresponding to their level of semantic abstraction.

3 Computation

Our current computational methods of choice are the EM algorithm and its stochastic counterparts. As we mentioned, the hyper-parameters are computed by maximum likelihood type II (empirical Bayes). The detailed equations are available in a technical report (de Freitas *et al* (2001)).

We favour point estimates because empirical distribution estimates are currently beyond the available technology. Each point estimate requires megabytes of memory. A long Markov chain would be almost impossible to store. In addition, there are many other problems with running MCMC for finite mixture models that we would have to overcome (Celeux *et al* (2000)).

4 Applications

4.1 Browsing Multimedia Databases

Mixture models are very suitable for clustering items and studying coherence within groups. They allow us to visualise the data and identify hidden patterns. This feature is of great benefit when browsing digital databases.

We have performed several experiments on the Corel image database. This database contains approximately 40,000 images annotated with approximately 3 to 5 keywords each. The images were manually grouped into different themes of 100 images each by Corel. This available categorisation makes the database very suitable for testing our models and algorithms. For example, when applying the various algorithms to

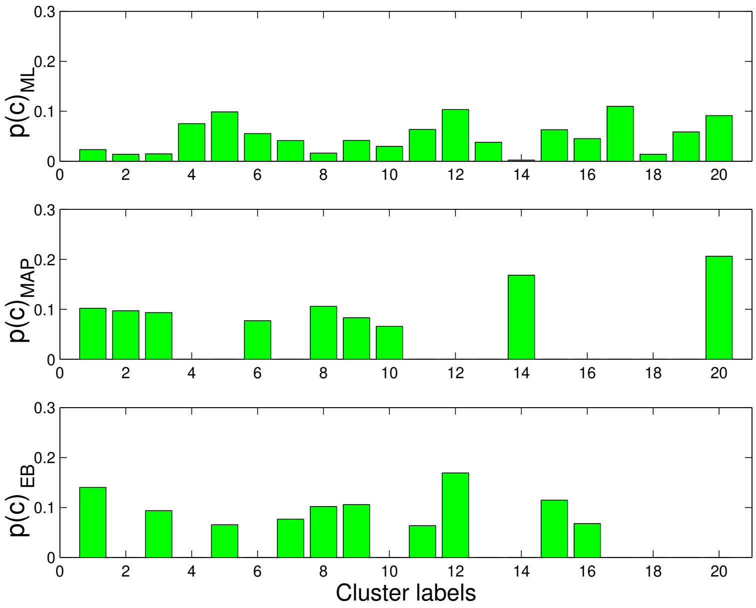

Figure 1: *Cluster probabilities in the Corel example. Maximum likelihood (top) maximum a posteriori with fixed hyper-parameters (middle) and maximum a posteriori with empirical Bayes (bottom). The lack of a bar indicates that the respective cluster has been automatically pruned by the regulariser.*

cluster a dataset consisting of 10 themes (1000 documents), and assuming 20 clusters initially, we obtained the cluster probabilities shown in Figure 1. Clearly, the Bayesian strategies allow us to obtain a number of clusters that is closer to the previous manual human sorting. Moreover, the clusters are more coherent and sparse, as shown in Figure 2. Note that we do not expect exactly 10 clusters as there is some overlap among the different themes.

One of the advantages of combining image and text attributes in the model is that people relate to images using both semantic and visual content. For example, if we want pictures of tigers on light green grasslands, we might do a search with the word “tiger” and a light grassland picture. This will hopefully not return images of tigers in dark places. Figure 3 shows an example where clustering images from themes containing tigers, using text and image features, results in the two separate, coherent clusters.

To test the Bayes model with text and music, we clustered on a database of musical scores with associated text documents. This will be used as the basis of information retrieval experiments which will allow users to enter a series of words and notes and return matches in the data base to sequences to songs that contain the sequence of notes and the lyrics entered by the user.

The database is composed of various types of musical scores – jazz, classical, television theme songs, and contemporary pop music – as well as associated text files.

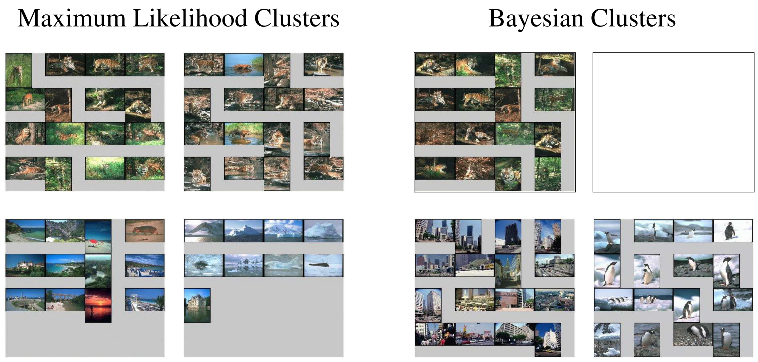

Figure 2: *Some of the clusters in the Corel example obtained by maximum likelihood and maximum a posteriori estimation. The Bayesian model gives us more coherent and sparse clusters (note that some of the clusters are empty).*

Figure 3: Result of clustering documents using both images and keywords. The two example clusters have obvious text and image semantics. That is, at the text level both groups relate to the word "tiger" and at the image level there are tigers on light green grasslands and tigers in dark places.

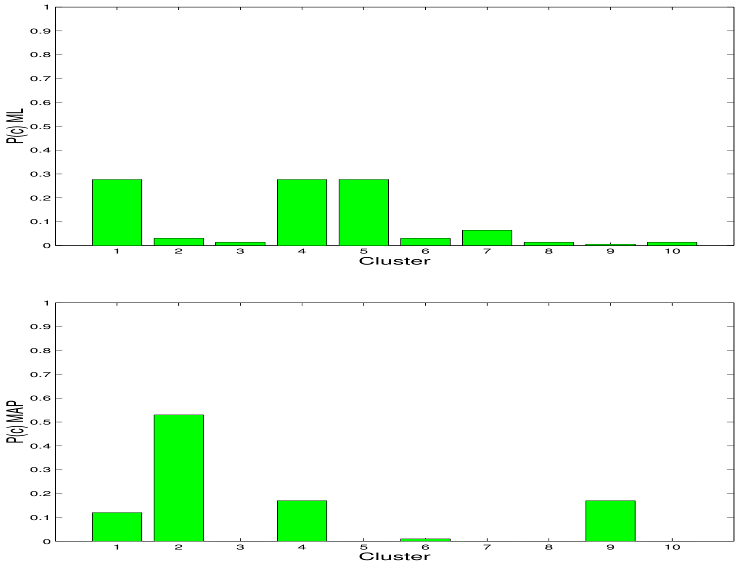

Figure 4: *Cluster probabilities of combined score-text database using EM with ML and MAP.*

The scores are represented in GUIDO notation, a powerful language for representing musical scores in an HTML-like notation. The associated text files are a song's lyrics, where applicable, or commentary on the score for instrumental pieces. The experimental database contains 100 scores, each with a single associated text document.

Clustering was again done via EM, using maximum likelihood (ML) and maximum a posteriori (MAP) estimation. The resulting probability mass distribution over the clusters is shown in Figure 4. Representative datum cluster probability assignments are shown in Table 1.

By and large, the clusters generated using MAP are intuitive – all of the 15 pieces by J. S. Bach are assigned to the same cluster, for instance – though a few curious anomalies exist. The Beatles' song "The Yellow Submarine" is included in the same cluster as the Bach pieces, though all the other Beatles songs are assigned to other clusters.

4.2 Information Retrieval

An important facility for image, music and text databases, such as the World-Wide Web, is retrieval based on user queries. We wish to support queries based on text, music, image features, categorical variables or combinations of these. We also would like the queries to be soft in the sense that the combinations of items is taken into consideration, but documents which do not have a given item should still be considered. Our probabilistic models enable us to overcome these problems.

CLUSTER	**SONG**	$p(c\|k)$
2	*Moby – Porcelain*	1
2	*Nine Inch Nails – Terrible Lie*	1
2	*other – 'Addams Family' theme*	1
⋮	⋮	⋮
4	*J. S. Bach – Invention #1*	1
4	*J. S. Bach – Invention #8*	1
4	*J. S. Bach – Invention #15*	1
4	*The Beatles – Yellow Submarine*	0.9975
⋮	⋮	⋮
6	*other – 'Wheel of Fortune' theme*	1
⋮	⋮	⋮
7	*The Beatles – Taxman*	1
7	*The Beatles – Got to Get You Into My Life*	0.7247
7	*The Cure – Saturday Night*	1
⋮	⋮	⋮
9	*R.E.M – Man on the Moon*	1
9	*Soft Cell – Tainted Love*	1
9	*The Beatles – Got to Get You Into My Life*	0.2753

Figure 5: *Representative probabilistic cluster allocations using MAP estimation.*

Figure 5 shows that for a query consisting of a conjunction of words that does not appear in the Corel database, we are able to retrieve images where it is clear that the models are generalising reasonably. Figures 6 shows a few retrieved images for a query consisting of a single image. By adding words to the query we can bias the result to retrieve images that look like the image query but contain the elements specified by the words (Figure 7). The same results can be obtained in Music databases. This enables users to query for songs using words that they might remember and hummed tunes.

4.3 Auto-Illustration and Auto-Annotation

Within our framework, one can build an application that takes text selected from a document, and suggests images to go with the text (Barnard *et al* (2001)). This "auto-illustrate" application is essentially a process of linking pictures to words. However, it should be clear by symmetry that we can just as easily go the other way, and link words to pictures. This "auto-annotate" process is very interesting for a number of reasons. First, given an image, if we can produce reasonable words for an image feature, then we can use existing text search infrastructure to broaden searches beyond the confines of our system. For example, consider image search by user sketch. If the sketch contains an orange ball in the upper right corner, the annotation model might return the words "sun" and "sunset". These words can, in turn, be used to search for images that match the sketch using a text based search engine.

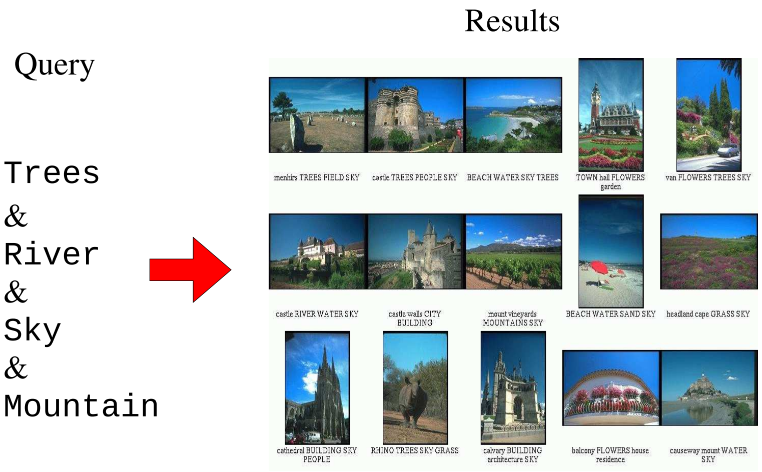

Figure 6: *Results using text query with conjunction of words that does not appear in the database. The model generalises to a reasonable extent.*

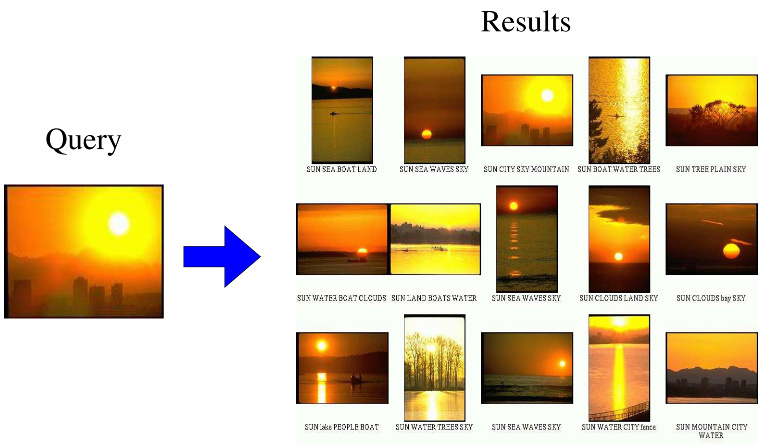

Figure 7: *Results using image query.*

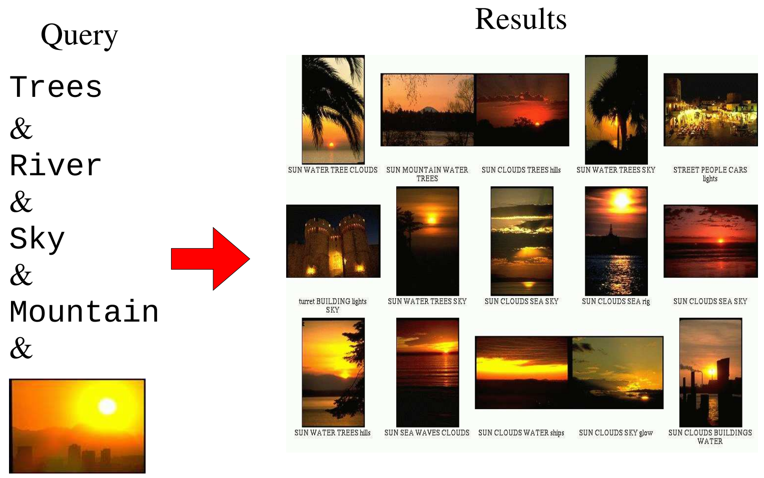

Figure 8: *Results using image query of Figure 5 with added text. The model still returns images of sunsets, but this time there are trees.*

4.4 Object Recognition

The association of text with images is even more interesting from a computer vision perspective because it is a form of minimally supervised learning of semantic labels for image features (Duygulu *et al* (2002)). Many databases contain images and text (typically annotations). We can exploit this data to build mixture models that translate image regions to words. The models are similar to the ones used in traditional statistical machine translation. As shown in Figure 8, we are able to label segments in arbitrary test-set images. This represents a significant advance in the field of computer vision.

5 Discussion and Further Work

This paper showed that there are sound reasons for adopting the Bayesian paradigm to model multimedia databases. We have only provided a brief overview of our work. There are many more interesting problems facing us. First, we need to develop variable selection strategies for clustering models that scale well. Second, we need to concentrate on scaling the software and algorithms. Third, we need to develop ways of processing the data on-line, as it can seldom be loaded into memory. Fourth, new utilities and loss functions are required. Much of our recent work is based on using word prediction on images (recognition) as a measure of performance. There is also room for improving our models and algorithms.

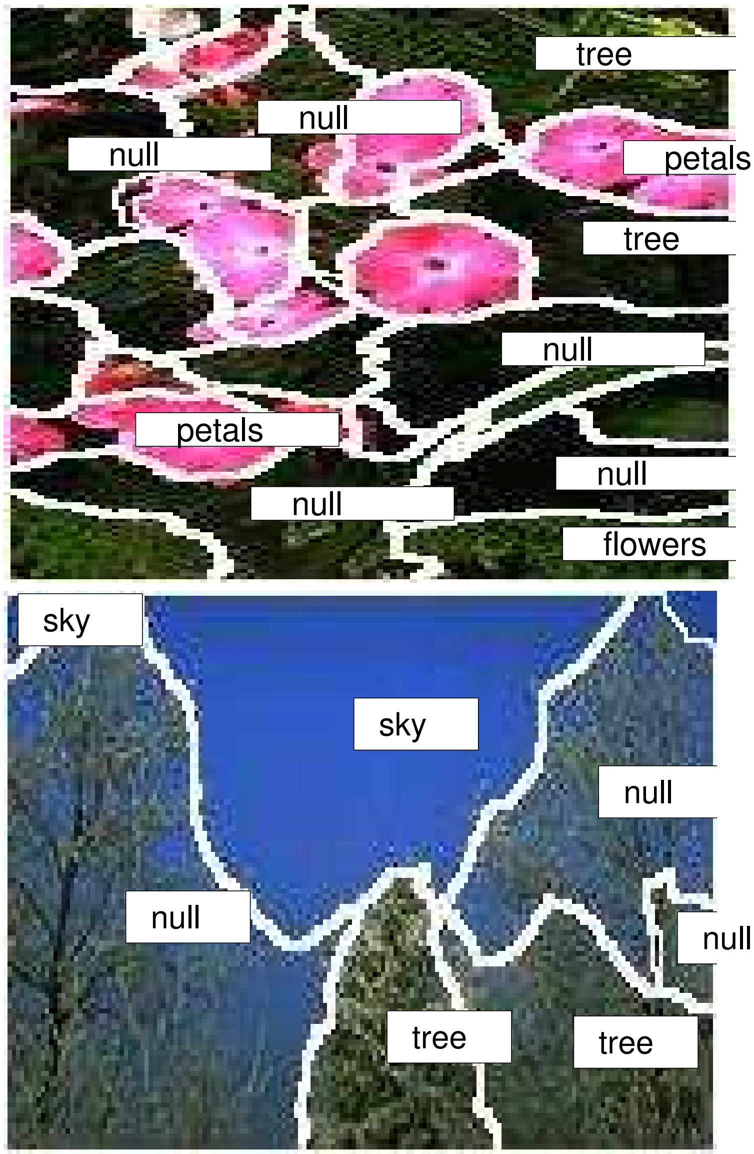

Figure 9: *Two examples of how our methods can be used to automatically recognise regions in arbitrary test set images. The null word indicates that the algorithm is not confident predicting a word.*

References

Barnard, K., Duygulu P., de Freitas, N., Forsyth, D., Blei, D. M. and Jordan, M. I. (2001). Matching Words and Pictures. Under Review.

Barnard, K. and Forsyth, D. (2001). Learning the Semantics of Words and Pictures, *International Conference on Computer Vision*, vol 2, pp. 408-415, 2001.

Barnard, K., Duygulu P., and Forsyth, D. (2001). Clustering Art, *Computer Vision and Pattern Recognition, 2001*, pp. II:434-439.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*, Wiley Series in Applied Probability and Statistics.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2001) Latent Dirichlet allocation. *Advances in Neural Information Processing Systems*.

Celeux, G., Hurn, M. and Robert, C. P. (2000) Computational and Inferential Difficulties with Mixture Posterior Distributions, *J. Amer. Statist. Assoc.*, vol 95: 957–970.

Cohn, D. and Hofmann, T. (2001). The missing link - a probabilistic model of document content and hypertext connectivity, *in* T. K. Leen, T. G. Dietterich and V. Tresp (eds), *Advances in Neural Information Processing Systems*, Vol. 10.

de Freitas, N. and Barnard, K. (2001). Bayesian Latent Semantic Analysis of Multimedia Databases. UBC TR 2001-15. Department of Computer Science, University of British Columbia.

Duygulu, P., Barnard, K., de Freitas, N. and Forsyth D. (2002). Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. *ECCV 2002*. Best Paper prize on Cognitive Computer Vision.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*, Chapman and Hall.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and its Applications*, Minnesota Press, Minneapolis.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis, *Uncertainty in Artificial Intelligence*.

Hoos, H. H., Renz, K. and Gorg, M. (2001). GUIDO/MIR - an Experimental Musical Information Retrieval System based on GUIDO Music Notation. *Proceedings of the 2nd International Symposium on Music Information Retrieval.*

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. (1998). Introduction to WordNet: An On-Line Lexical Database, *International Journal of Lexicography* **3**: 235–244.

Narayanan, A. (1991). Maximum Likelihood Estimation of the Parameters of the Dirichlet Distribution, *Applied Statistics* **40**(2): 365–374.