Experimental Evaluation of a Generative Probabilistic Image Retrieval Model on 'Easy' Data [*]

Thijs Westerveld
CWI
PO Box 94079, 1090 GB Amsterdam
The Netherlands
thijs@cwi.nl

Arjen P. de Vries
CWI
PO Box 94079, 1090 GB Amsterdam
The Netherlands
arjen@acm.org

ABSTRACT

We present evaluation results of a generative probabilistic image retrieval model using 'easy data'. Previous research into our model's retrieval effectiveness has used the test collection developed at TREC's Video Track, but as discussed in detail in [17], its search task has been too difficult to measure actual performance of the retrieval model. The 'easy data' experiments presented here evaluate our model under varying model parameters on the *Corel set*. The Corel data set is relatively easy because images are nicely grouped into coherent themes, the within theme similarity is high and the across theme similarity relatively low. These properties make Corel a nice vehicle for testing, presenting or selling new content based retrieval techniques and models. In contrast to the TREC data, the Corel collection gives statistically significant differences between varying experimental conditions, so we get more insight in the model's behaviour. We then discuss at length the limitations of the results obtained using this data set, comparing the experiments performed here to those on the TREC data.

1. INTRODUCTION

This paper presents another experimental evaluation of the generative probabilistic image retrieval model presented in [19], developed for (among others) our participation in TREC video track. Unfortunately, recent video tracks at TREC [14, 13] show merely that content based retrieval from generic collections still is too hard a task, for which metric-based evaluation on topical relevance is perhaps not the most useful evaluation methodology. Westerveld and De Vries have therefore argued in [17] that a deep analysis of results is a useful alternative for evaluation, while still using the TREC data. This paper pursues a supplementary study, by looking into the same model's effectiveness on a different collection, the *Corel set*.

The Corel set is a collection of stock photographs, which is divided into subsets of images each relating to a specific theme (e.g. *tigers*, *sunsets*, or *English pub signs*). This image collection is used to evaluate retrieval results or to illustrate the effectiveness of a given retrieval method in a large number of publications in the field of content based image retrieval (e.g., [3, 7, 4, 1, 2, 16, 8]). A recent study by Müller et al. [9] showed however that evaluations using Corel are highly sensitive to the subsets and evaluation measures used. In addition, as an image retrieval test collection, Corel data can be qualified as 'easy' because of the clear distinctions between themes and the high similarity within a theme. Therefore, good results on Corel do not guarantee good results in a more realistic setting. Yet, we were not certain whether this could explain the differences in conclusions on the feasibility of image retrieval from our negative experiences at TREC relative to the positive results presented in the 'Corel papers' cited above.

The goal of this paper is therefore twofold. First, to show that the Corel data set is indeed a relatively easy data set that can give misleadingly good results (Section 4). Second, to show that, because of its relative easiness, the Corel data is useful for testing different system settings to increase our understanding of the model's behaviour. In contrast to experiments with the TREC collection, Corel data gives significant differences for varying parameters (Section 5).

We start describing the model (Section 2) and experimental setup (Section 3). Section 6 finishes with some disclaimers on using Corel data, and compares results obtained with Corel and TREC respectively.

2. GENERATIVE DOCUMENT MODELS

Our retrieval model (see also [19]) is based on generative document models similar to the language models for information retrieval [11, 5]. Each document in our collection (i.e. each image) is modelled as a probabilistic process that generates visual samples (feature vectors representing small blocks of pixels). Thus, for each model ω_i we have a probability density function $P(x|\omega_i)$ defining the likelihood of samples x. These densities are assumed to be mixtures of Gaussian distributions (see Section 2.1).

Given a query (i.e. an example image), documents in the collection are ranked based on the ability of the corresponding models to explain the set of query samples. We smooth using background probabilities, calculated by marginalising over all documents in the collection $P(x) = \sum_i P(x|\omega_i)$.

[*] A colour version of this paper is available from `http://www.cwi.nl/~thijs/pub/SIGIR2003MMIR.html`

Thus, the retrieval status value (RSV) of a document model ω_i is computed as its probability of generating the query $\mathbf{X}$ consisting of N samples ($\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$):

$$\mathrm{RSV}(\omega_i) = \sum_{j=1}^{N} \log\left[\kappa P(x_j|\omega_i) + (1-\kappa)P(x_j)\right], \qquad (1)$$

where κ is a mixing parameter, which can be estimated on a training collection. Section 3 discusses how we convert query images to samples. Here we first describe the generative models.

2.1 Gaussian Mixture Models

We model each image in our collection as a random process that generates image samples (i.e. feature vectors describing pixel blocks). This is assumed to be a mixture of multivariate Gaussian processes, where the number of Gaussian components N_C is fixed for all images in the collection[15, 19]:

$$P(x_j|\omega_i) = \sum_{c=1}^{N_C} P(C_{i,c}) \frac{1}{\sqrt{(2\pi)^n |\Sigma_{i,c}|}} e^{-\frac{1}{2}\|x_j-\mu_{i,c}\|_{\Sigma_{i,c}}},$$

where

$$\|x_j - \mu_{i,c}\|_{\Sigma_{i,c}} = (x_j - \mu_{i,c})^T \Sigma_{i,c}^{-1} (x_j - \mu_{i,c}). \qquad (2)$$

Here $C_{i,c}$ is component c of class model ω_i and $(x_j - \mu)^T$ is the matrix transpose of $(x_j - \mu)$. The samples x_j are n-dimensional feature vectors describing an 8x8 pixel block (details in Section 3). A model is completely specified by the parameters of its components ($\mu_{i,c}$, $\Sigma_{i,c}$ and $P(C_{i,c})$) These parameters are estimated using standard EM.

3. EXPERIMENTAL SETUP

The experiments reported in this paper are carried out using a subset of the Corel data. One problem with evaluation using Corel is that the data is sold commercially on separate thematic CDs, and a single 'Corel set' does not exist. We have access to over 600 classes or themes. To improve the comparability across different publications, the experiments in this paper have used the intersection of these 600 with the classes used by Duygulu et al. [4] and Jeon et al. [7]. The resulting 39 classes are listed in table 1.

To test the model from Section 2, we estimate generative models for each document in our collection using EM on the image samples. This is visualised in Figure 1. First, an image is converted to the YCbCr colour space, and we cut each channel into blocks of 8 by 8 pixels and compute the discrete cosine transform (DCT) for each block. Then, we take a fixed number of the most important DCT-coefficients from the Y-channel and a fixed number of the most important coefficients from the Cb and Cr channels.[1] These feature vectors are then fed to the EM algorithm to estimate a generative mixture model with 8 mixture components. After convergence, we describe the position in the image plane of each mixture component as a 2D Gaussian with mean and covariance computed from the positions of the pixel blocks associated with the component.

A query image is transformed into a set of samples using the same procedure without the EM step. Each query feature vector consists of a number of DCT coefficients and a position in the image plane. Documents can now be ranked by calculating the RSV conforming Equation 1.

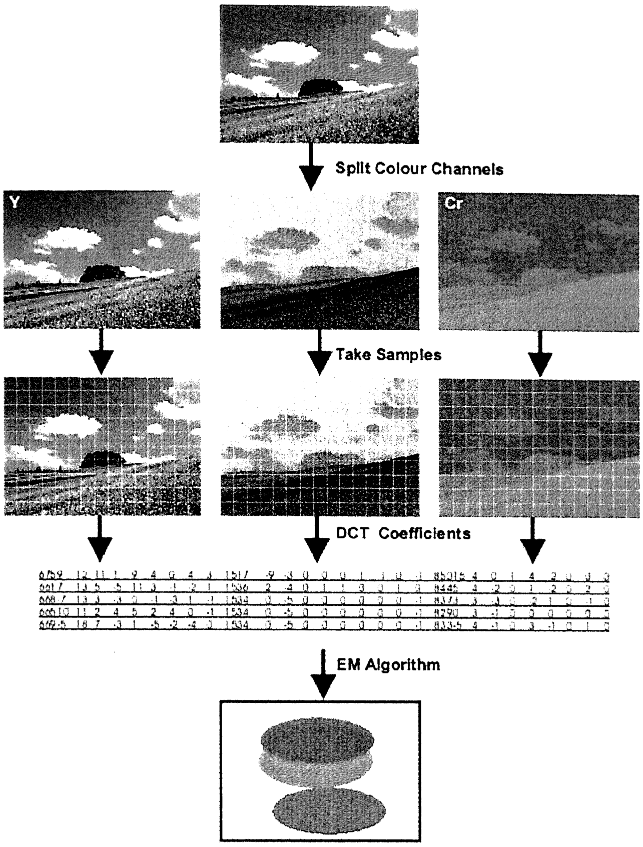


Figure 1: Building a Gaussian Mixture Model from an Image.

4. TESTING THE MODEL ON COREL

To evaluate the retrieval effectiveness of the model, we assume that a document is relevant to a query if and only if document and query are from the same class. We run each document in our collection as a query, rank the full document set (i.e., we do not take a top K) and compute average precision values for each query. We then compute mean average precision (MAP) scores per image class. Since there is some variety in the specificity of the classes, some classes might be harder than others (something as specific as *English pub signs* might be more easy than a generic class like *Israel*). Table 1 shows the MAP scores for the individual classes (sorted from high to low). Indeed, we see a fair amount of variation (.05 to .36). Figure 2 shows an example of a query with the top 5 documents from one of the classes with the highest scores: Arabian Horses.

[1] Usually, we take 10 coefficients from the Y channel and only 1 from Cb and Cr. These numbers are varied in Section 5.

Table 1: MAP per class.

Class	MAP
English Pub Signs	.36
English Country Gardens	.33
Arabian Horses	.31
Dawn & Dusk	.21
Tropical Plants	.19
Land of the Pyramids	.19
Canadian Rockies	.18
Lost Tribes	.17
Elephants	.17
Tigers	.16
Tropical Sea Life	.16
Exotic Tropical Flowers	.16
Lions	.15
Indigenous People	.15
Nesting Birds	.13
Images of Thailand	.13
Greek Isles	.10
Cowboys	.10
Mayan and Aztec Ruins	.09
Wildlife of Antarctica	.09
Israel	.09
Beaches	.09
Holland	.08
Hong Kong	.08
Sweden	.07
Ireland	.07
Wildlife of the Galapagos	.07
Hawaii	.07
Rural France	.07
Zimbabwe	.07
Images of Death Valley	.07
Nepal	.07
Foxes & Coyotes	.06
North American Deer	.06
California Coasts	.06
North American Wildlife	.06
Peru	.05
Alaskan Wildlife	.05
Namibia	.05
mean	.12

4.1 Class Confusion

In addition to computing the mean average precision per class, we can also look at confusion between classes by recomputing the scores with different sets of relevant documents. For example, to see how often we find lions when searching for tigers, we can simply rank the collection using a tiger query and then evaluate the results under the assumption that only images from the lions class are relevant. We computed these confusion MAP scores for all pairs of classes in the collection. The results are visualised in Figure 4, showing the log MAP scores for all pairs; darker squares indicate higher scores and the scores are printed inside the squares. Along the y-axis, the query class is plotted, the x-axes shows the class that was assumed relevant to the query. Thus looking at *row X* we can learn what we find if we search for X, and *column X* shows what would be a good query to retrieve X. The MAP scores are averaged over all documents in a given query class.

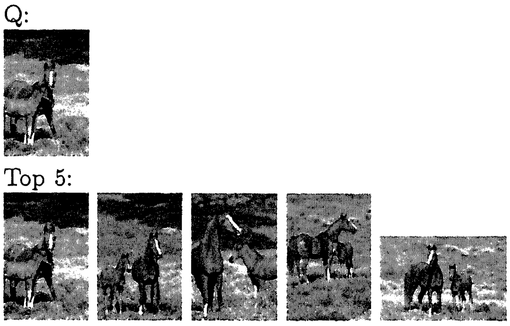


Figure 2: Example Query with top 5 documents.

The diagonal of the Figure is darker than the rest, indicating that, on average, queries are better at retrieving images from their own class than images from a different class. Some interesting confusions are the following: When querying for *beaches* we also find *Greek Islands*; a query for *Tropical Plants* returns also *Tropical Sea life*, and searching for *Indigenous People* we find *Lost Tribes*. Moreover, we see some lighter and darker columns showing that some classes get retrieved hardly ever when using examples from outside that class (*Wildlife of Antarctica*, *Dawn & Dusk*) and others are returned more often for any query (*Indigenous People*, *Lost Tribes*). Also noticeable is the fact that country gardens and tropical plants get mixed up sometimes and that these two classes are retrieved relatively often when Arabian horses are used are used as query examples. The latter probably because of the similarity in background; all have green, grassy backgrounds.

This background matching is exactly why we often can find the correct documents – essentially on sheer luck. In the Corel set, photographs from one class are often taken in one or a few locations and therefore have highly similar backgrounds. In more heterogeneous image collections retrieving images from the same "class"[2] is not as easy. When experimenting with the Corel data set, one might conclude it is easy to retrieve horses, tropical fish or English pub signs, while in fact one has learnt to retrieve respectively, grass with yellow flowers, dark sea, and clear blue sky. Figures 3 and 5 illustrate this effect using subsets of the horses example from Figure 2 as a query. Clearly, we identify the green background rather than the horses.

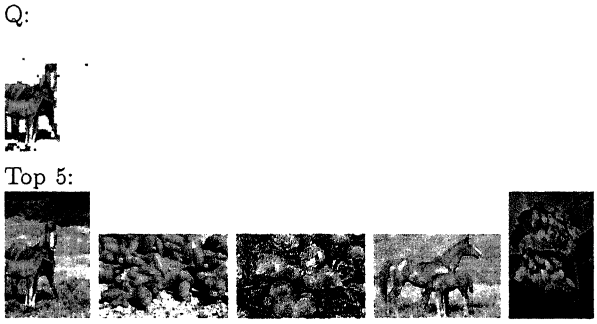


Figure 3: Horses query with top 5 results.

[2] Here class means similar semantics; a generic image collection does not have classes like Corel.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
California Coasts 1	.06	.03	.02	.03	.02	.02	.02	.02	.02	.03	.03	.02	.03	.02	.02	.04	.02	.02	.02	.02	.03	.03	.02	.02	.02	.02	.02	.02	.03	.02	.02	.04	.03	.02	.03	.03	.02	.03	.02
Images of Death Valley 2	.02	.07	.02	.02	.02	.02	.03	.02	.02	.03	.02	.02	.02	.02	.02	.03	.02	.02	.02	.02	.03	.02	.02	.02	.02	.02	.02	.02	.03	.02	.02	.02	.02	.02	.04	.03	.02	.02	.03
Holland 3	.03	.02	.08	.04	.03	.02	.02	.02	.02	.03	.02	.02	.03	.02	.02	.03	.02	.02	.03	.02	.03	.03	.02	.03	.03	.03	.02	.02	.02	.03	.03	.03	.03	.03	.03	.04	.02	.03	.02
Ireland	.03	.03	.04	.07	.03	.02	.03	.02	.02	.02	.03	.02	.02	.02	.02	.03	.02	.02	.02	.02	.03	.02	.02	.03	.02	.03	.02	.03	.02	.03	.02	.03	.03	.02	.03	.04	.02	.03	.02
Peru 5	.03	.03	.04	.03	.05	.03	.03	.04	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.04	.03	.03	.03
Tropical Plants 6	.03	.02	.04	.03	.04	.19	.02	.04	[illegible]	.04	.04	.07	.05	.04	.03	.02	.03	.07	.04	.09	.02	.03	.03	.04	.04	.03	.07	.05	.03	.03	.06	.04	.02	.03	.03	.03	.05	.03	.03
Land of the Pyramids 7	.02	.04	.02	.03	.02	.02	.19	.02	.02	.03	.03	.02	.02	.02	.03	.02	.04	.02	.02	.02	.03	.03	.05	.03	.03	.03	.02	.03	.03	.03	.02	.03	.03	.03	.02	.02	.02	.02	.03
Rural France 8	.03	.03	.03	.03	.03	.02	.02	.07	.03	.02	.03	.03	.02	.03	.03	.03	.03	.03	.03	.03	.03	.02	.02	.03	.02	.02	.02	.03	.03	.02	.02	.02	.03	.02	.03	.03	.03	.03	.03
English Country Gardens 9	.04	.03	.04	.04	.04	.11	.03	.05	.33	.04	.04	.08	.03	.05	.03	.02	.04	.07	.03	.08	.03	.04	.04	.04	.03	.04	.04	.07	.03	.02	.04	.05	.02	.03	.03	.04	.08	.04	.04
Images of Thailand 10	.04	.05	.03	.03	.04	.03	.03	.03	.03	.13	.03	.02	.04	.03	.02	.04	.02	.03	.03	.02	.04	.04	.03	.03	.03	.03	.04	.03	.04	.04	.03	.03	.03	.08	.08	.02	.03	.03	.04
Alaskan Wildlife 11	.03	.03	.02	.03	.03	.02	.02	.03	.02	.02	.05	.03	.03	.03	.03	.03	.03	.03	.03	.02	.03	.02	.02	.03	.03	.02	.03	.03	.03	.02	.02	.03	.03	.02	.03	.03	.02	.03	.03
Nesting Birds 12	.03	.03	.03	.03	.03	.05	.02	.03	.05	.02	.03	.13	.03	.03	.03	.02	.03	.04	.03	.05	.03	.02	.03	.03	.03	.03	.03	.03	.02	.02	.03	.03	.02	.02	.02	.03	.04	.03	.03
Wildlife of the Galapagos 13	.03	.02	.02	.02	.02	.02	.02	.02	.02	.02	.03	.02	.07	.02	.02	.04	.02	.02	.03	.02	.02	.02	.02	.02	.02	.02	.03	.02	.02	.02	.02	.03	.03	.02	.03	.03	.02	.03	.02
North American Deer 14	.03	.03	.03	.03	.03	.03	.03	.03	.03	.02	.03	.03	.03	.08	.04	.03	.03	.03	.04	.03	.03	.02	.03	.02	.03	.02	.03	.03	.03	.02	.02	.03	.03	.02	.02	.03	.03	.03	.03
Lions 15	.02	.03	.02	.02	.03	.02	.04	.03	.02	.02	.03	.03	.03	.04	.15	.02	.05	.03	.04	.03	.03	.02	.03	.03	.03	.03	.03	.03	.02	.03	.03	.02	.04	.02	.02	.03	.03	.03	.04
Wildlife of Antarctica 16	.03	.02	.02	.02	.02	.02	.02	.02	.01	.02	.02	.02	.03	.02	.02	.09	.02	.02	.03	.01	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.03	.02	.04	.02	.02	.03	.02
Elephants 17	.03	.04	.03	.03	.03	.02	.05	.04	.03	.02	.04	.03	.03	.05	.05	.03	.17	.03	.04	.03	.03	.03	.04	.03	.03	.03	.03	.03	.03	.02	.02	.03	.04	.02	.03	.03	.04	.04	.05
Tigers 18	.04	.03	.04	.04	.04	.05	.03	.03	.05	.03	.04	.04	.04	.04	.05	.03	.04	.16	.05	.08	.04	.04	.04	.04	.04	.05	.04	.04	.03	.03	.04	.04	.03	.03	.02	.04	.04	.05	.04
Foxes & Coyotes 19	.03	.03	.03	.03	.03	.02	.02	.03	.02	.02	.04	.03	.04	.03	.03	.04	.03	.03	.06	.02	.03	.02	.02	.02	.03	.02	.03	.02	.03	.02	.02	.02	.03	.02	.03	.03	.02	.04	.03
Arabian Horses 20	.02	.02	.02	.02	.02	.05	.02	.03	.04	.02	.02	.04	.03	.03	.03	.02	.03	.05	.03	.31	.02	.02	.02	.02	.02	.02	.03	.06	.02	.02	.03	.02	.02	.02	.02	.02	.03	.02	.02
Greek Isles 21	.04	.05	.03	.04	.04	.02	.03	.03	.03	.04	.04	.03	.03	.03	.03	.04	.03	.03	.03	.02	.10	.05	.05	.04	.03	.04	.03	.03	.06	.03	.03	.04	.04	.04	.04	.04	.03	.04	.03
Hong Kong 22	.03	.03	.04	.03	.03	.03	.03	.03	.03	.05	.03	.03	.03	.03	.03	.03	.03	.03	.03	.02	.04	.08	.04	.04	.04	.04	.03	.03	.03	.04	.04	.03	.03	.05	.03	.03	.03	.03	.03
Israel 23	.03	.04	.03	.03	.03	.02	.08	.03	.03	.03	.03	.03	.02	.03	.04	.02	.04	.03	.02	.02	.04	.04	.09	.04	.04	.04	.03	.04	.03	.04	.03	.03	.03	.03	.02	.03	.03	.03	.04
Nepal 24	.03	.03	.04	.04	.03	.03	.04	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.02	.03	.03	.04	.06	.04	.04	.03	.03	.03	.03	.03	.04	.03	.03	.03	.04	.03	.03	.03
Indigenous People 25	.03	.04	.05	.04	.05	.05	.05	.04	.04	.05	.05	.05	.04	.05	.06	.04	.05	.05	.04	.04	.04	.05	.05	.05	.15	.05	.05	.05	.04	.05	.07	.04	.05	.06	.03	.03	.08	.05	.05
Sweden 26	.03	.03	.04	.04	.03	.03	.04	.03	.03	.04	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.04	.04	.05	.04	.03	.07	.03	.04	.03	.04	.03	.04	.03	.04	.03	.03	.03	.03	.03
Tropical Sea Life 27	.03	.03	.03	.03	.04	.07	.03	.04	.04	.04	.04	.04	.05	.04	.04	.03	.03	.04	.04	.05	.03	.03	.03	.03	.04	.03	.16	.04	.03	.04	.06	.03	.03	.03	.04	.02	.04	.03	.03
Cowboys 28	.03	.03	.03	.04	.03	.04	.04	.04	.04	.03	.03	.04	.03	.03.	.03	.03	.04	.04	.03	.05	.03	.03	.04	.04	.03	.04	.04	.10	.03	.04	.04	.04	.03	.03	.03	.03	.03	.03	.04
Beaches 29	.03	.03	.02	.02	.03	.02	.02	.03	.02	.03	.03	.02	.03	.02	.02	.03	.02	.02	.03	.02	.03	.03	.02	.02	.02	.02	.02	.02	.09	.02	.02	.03	.03	.02	.03	.03	.02	.02	.02
Dawn & Dusk 30	.02	.02	.02	.02	.02	.02	.03	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.01	.02	.02	.02	.02	.03	.02	.02	.02	.02	.02	.02	.21	.03	.03	.04	.02	.02	.01	.02	.02	.02
Exotic Tropical Flowers 31	.02	.02	.03	.02	.02	.04	.02	.02	.02	.03	.02	.03	.03	.02	.03	.02	.02	.02	.02	.03	.02	.03	.03	.02	.03	.02	.04	.03	.02	.03	.16	.03	.02	.03	.02	.02	.03	.02	.02
Hawaii 32	.03	.02	.02	.03	.02	.02	.02	.02	.02	.02	.02	.02	.03	.02	.02	.02	.02	.02	.02	.02	.02	.03	.02	.03	.02	.03	.02	.02	.03	.03	.02	.07	.03	.02	.02	.02	.02	.02	.02
Namibia 33	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.03	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.03	.02	.02	.05	.02	.02	.02	.02	.02	.02
English Pub Signs 34	.04	.04	.04	.03	.04	.04	.04	.04	.04	.07	.04	.03	.04	.03	.02	.05	.03	.03	.03	.03	.05	.05	.03	.04	.05	.04	.05	.03	.05	.05	.05	.04	.03	.36	.06	.03	.04	.03	.04
Canadian Rockies 35	.03	.04	.02	.03	.02	.02	.02	.02	.02	.04	.03	.02	.03	.02	.01	.06	.02	.02	.03	.01	.03	.03	.02	.02	.02	.02	.02	.02	.03	.02	.02	.02	.02	.03	.18	.02	.02	.03	.02
Mayan and Aztec Ruins 36	.03	.03	.03	.04	.03	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.03	.02	.02	.02	.02	.03	.02	.02	.03	.02	.03	.02	.02	.03	.02	.02	.02	.03	.02	.03	.09	.02	.03	.02
Lost Tribes 37	.04	.04	.05	.04	.05	.08	.04	.06	.08	.05	.05	.06	.04	.06	.06	.03	.05	.06	.04	.06	.04	.05	.04	.05	.09	.04	.05	.05	.04	.04	.05	.04	.04	.05	.03	.04	.17	.05	.06
North American Wildlife 38	.03	.02	.03	.03	.03	.02	.02	.02	.02	.02	.03	.02	.03	.03	.03	.03	.03	.03	.03	.02	.03	.02	.03	.03	.03	.03	.02	.02	.02	.02	.02	.02	.03	.02	.03	.03	.02	.06	.02
Zimbabwe 39	.03	.04	.03	.03	.03	.03	.04	.04	.03	.03	.03	.03	.03	.04	.04	.02	.04	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.03	.02	.03	.03	.03	.07

Figure 4: Confusion between classes. MAP scores for different classes of relevant images. Darker squares indicate higher scores, the scores are listed inside the squares.

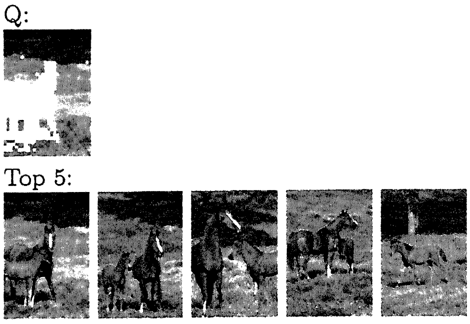


Figure 5: Grass query with top 5 results.

5. TUNING THE MODELS

Summarising the previous, results on the Corel data may cause overestimation of retrieval capabilities and one should be very careful with generalising conclusions drawn from experiments performed on Corel. Still, this data set can be useful for tuning a model (although Section 6 will give some disclaimers). This section uses the Corel data to investigate the performance effects of various model parameters. In order to be able to try a large number of different settings in a fair amount of time, we only use each tenth image from a theme, thus reducing the data set to contain only 10 images per theme. We then index this reduced collection for each setting, and use each of the 390 images as a query and calculate average precision scores for them. These scores are averaged over all queries and we report the MAP scores per setting.

We use the general procedure described in Section 4 to build models, but we vary the following parameters[3]:

NY: Number of DCT coefficients from Y channel (1, 3, 6, 10, 15 or 21).

NCbCr: Number of DCT coefficients from Cb and Cr channels (0, 1 or NY).

XYpos: Way of using sample position information of pixel blocks (0, do not use, 1, add to feature vector before training or 2, add to mixture components after training on DCT coefficients only)

c: Number of Gaussian mixture components (1, 2, 4, 8, 16 or 32).

Since we are mainly interested in the influence of different ways of using colour and position information, we fix the number of DCT coefficients from the Y channel (NY) at 10 while varying the other parameters. We leave the variation of the number of coefficients from the Y channel for future research.

Table 2 shows the results for different values of NCbCr, XYpos and c. Figure 6 shows visualisations of models estimated using different parameter settings.[4] The images show mean colour and texture representations of the components at the position in the image plane where the standard deviation from the mean position is below 2; prior probabilities of the components are not visualised. Figure 7 shows come images constructed by randomly sampling from the Gaussian mixture model (Equation 2). We then perform the inverse discrete cosine transform on the DCT part of the feature vectors to get pixel blocks and use the position information in the feature vector to place the pixel block in the image plane. For models without position information, we scatter the blocks randomly across the image plane.

5.1 Statistical significance

The scores in Table 2 do not differ a lot. One could conclude that as long as one uses a mixture (c>1) rather than a single Gaussian (c=1), it does not matter much which model one chooses. However, a small difference in average scores might still be significant: Run A might be consistently better than run B, but a few outliers or errors for which run B is better can cancel out this effect resulting in a similar MAP score for both runs. For this reason, researchers in Information Retrieval have recommended using a statistical test for significance instead of the aggregated run score. This Section highlights some of the significant differences between the parameter settings, but first we describe the statistical test we use to compare two runs.

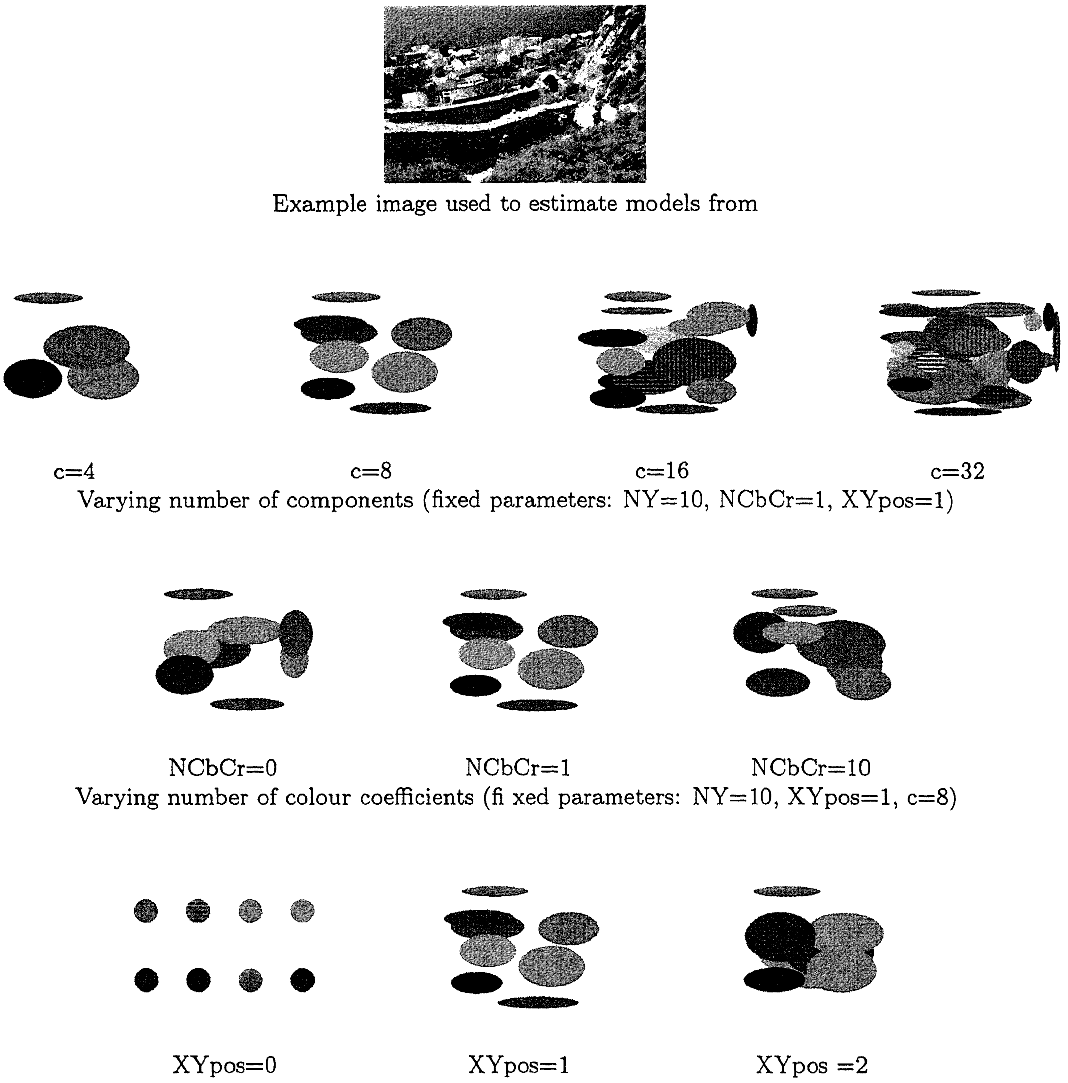
Different statistical tests for significance make different assumptions about the process in which the measurements were acquired, and the distribution of the values measured. We preferred to use non-parametric tests (see [10]), to stay away from too many assumptions of normality. We have chosen to use the non-parametric 'Wilcoxon matched-pairs signed-ranks test' to test for significance on the outcomes of our experiments, the decision mainly based on the discussions in [20] and [6]. Zobel found that the Wilcoxon test gave best reliability and greater power than its alternatives, and Hull also argues in favour of this test. The test analyses the differences between measurements per query, replacing the difference by the rank of its absolute value. These ranks are multiplied by the sign of the difference, and the sum of the ranks for each group is compared to its expected value under the assumption that the two groups are equal. The test assumes that the errors come from a continuous distribution symmetric about 0, and can only be used if the number of measurements is sufficiently large. According to [6] however, these constraints should not cause too many worries as long as one remains pragmatic about interpreting the results carefully.

So, we applied the paired Wilcoxon signed-rank test at a significance level of 5% to each pair of parameter settings. For each setting we then counted the number of other models that are significantly better, the number of significantly worse models and the number of other models that did not differ significantly. Table 3 shows the best models according to these counts (i.e., the models with the lowest number of significantly better models). While the top scoring models share a few interesting properties (like the number of CbCr coefficients used), it is more interesting to look at the influence of changes in the individual parameters. Consecutively, we discuss the influence of changing c, NCbCr and XYpos.

When varying the number of components (c) in the mixture model, we expect that a low number gives insufficient resolution to describe well all image samples, whereas a high number of components is bound to result in overfitting. Ta-

[3] In Section 4 we used the following setting: NY=10, NCbCr=1, XYpos=2, c=8.

[4] Figures 6 and 7 are best viewed in colour. For a colour version of this paper see `http://www.cwi.nl/~thijs/pub/SIGIR2003MMIR.html`



Example image used to estimate models from

c=4 c=8 c=16 c=32

Varying number of components (fixed parameters: NY=10, NCbCr=1, XYpos=1)

NCbCr=0 NCbCr=1 NCbCr=10

Varying number of colour coefficients (fi xed parameters: NY=10, XYpos=1, c=8)

XYpos=0 XYpos=1 XYpos =2

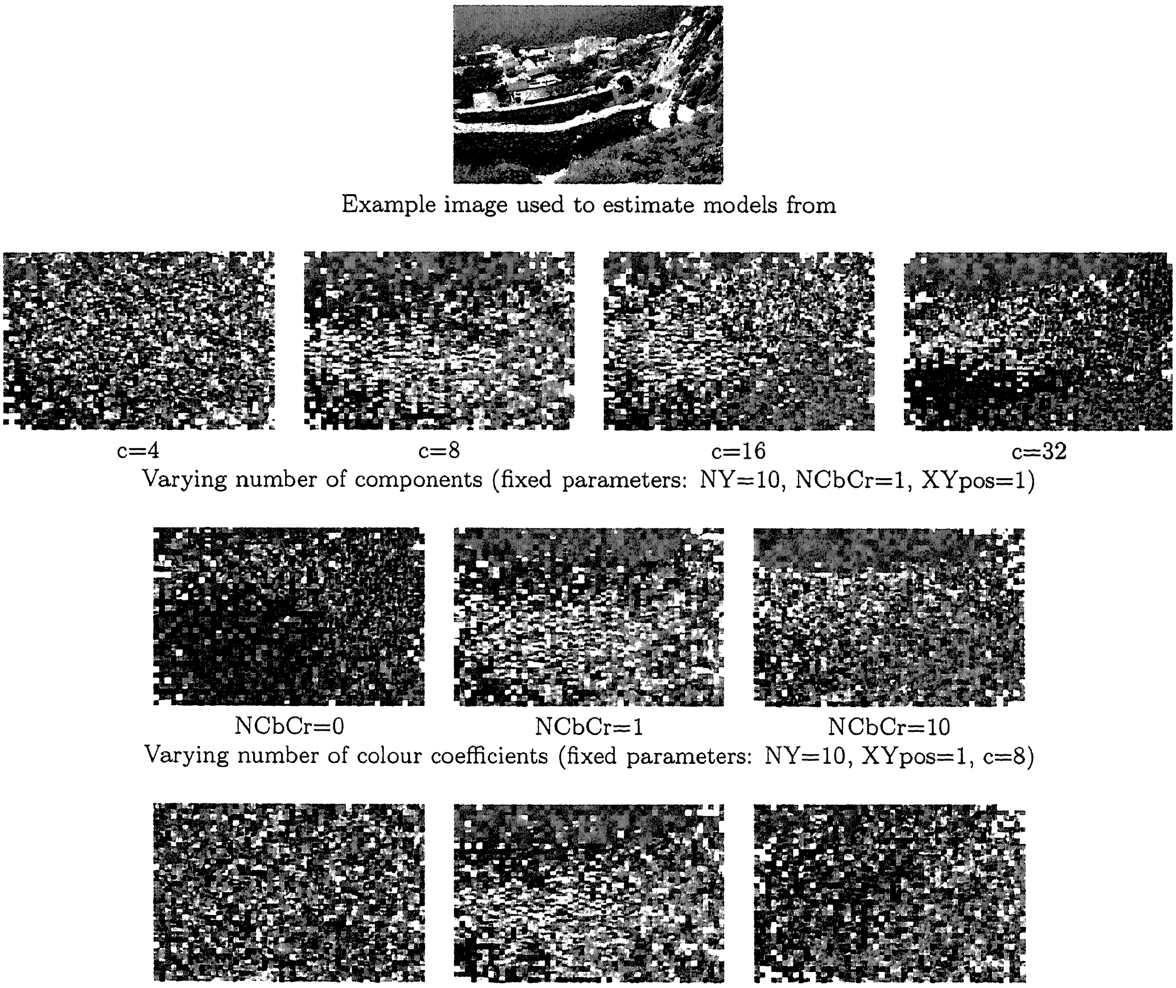
Varying position information, (fixed parameters: NY=10, NCbCr=1, c=8)

For visualisation purposes, the components in the XYpos=0 setting are distributed uniformly across the image plane, while in fact no position information is available in this setting.

Figure 6: Example image with different models estimated using various parameter settings. The images show mean colour and mean texture of the components where the standard deviation from the mean position in the image plane is below 2. Note that prior probabilities of the components and the variance in colour and texture are not represented in these visualisations.

Table 2: MAP scores for different parameter settings (fixed: NY=10).

NY	NCbCr	XYpos	c=1	c=2	c=4	c=8	c=16	c=32
10	0	0	0.08	0.18	0.20	0.21	0.21	0.21
10	0	1	0.09	0.19	0.21	0.21	0.21	0.20
10	0	2	0.09	0.19	0.21	0.21	0.22	0.21
10	1	0	0.13	0.22	0.23	0.23	0.23	0.23
10	1	1	0.13	0.22	0.23	0.23	0.23	0.22
10	1	2	0.13	0.22	0.23	0.24	0.23	0.23
10	10	0	0.12	0.22	0.23	0.24	0.24	0.23
10	10	1	0.13	0.21	0.24	0.24	0.24	0.23
10	10	2	0.13	0.22	0.23	0.24	0.24	0.23



Example image used to estimate models from

c=4 c=8 c=16 c=32

Varying number of components (fixed parameters: NY=10, NCbCr=1, XYpos=1)

NCbCr=0 NCbCr=1 NCbCr=10

Varying number of colour coefficients (fixed parameters: NY=10, XYpos=1, c=8)

XYpos=0 XYpos=1 XYpos =2

Varying position information, (fixed parameters: NY=10, NCbCr=1, c=8)

Figure 7: Random samples from the models visualised in Figure 6. The images are constructed by randomly sampling from the Gaussian mixture model (Equation 2) and then transforming the thus obtained feature vectors to pixel blocks.

Table 3: Top scoring parameter settings based on pair wise comparisons.

NY	NCbCr	XYpos	c	#better	#equal	#worse
10	10	1	8	0	8	45
10	10	2	8	0	9	44
10	10	0	8	0	11	42
10	10	0	16	0	12	41
10	10	1	16	0	12	41
10	10	2	32	0	20	33
10	10	1	4	0	23	30
10	10	0	32	1	19	33
10	1	2	8	1	19	33
10	10	1	32	3	18	32

Table 4: Comparing different models, varying the number of components c (with fixed NCbCr=10, XYpos=2). The numbers indicate if we see a significant difference when changing from row to column setting: 0 for no significant difference; 1 for significant improvement; -1 for significant deterioration.

c	1	2	4	8	16	32
1	0	1	1	1	1	1
2	-1	0	1	1	1	1
4	-1	-1	0	1	1	0
8	-1	-1	-1	0	0	0
16	-1	-1	-1	0	0	-1
32	-1	-1	0	0	1	0

Table 5: Comparing different models, varying the number of DCT coefficients from the colour channels NCbCr (with fixed XYpos=1, c=8). See Table 4 for explanation.

NCbCr	0	1	10
0	0	1	1
1	-1	0	1
10	-1	-1	0

ble 4 shows that this intuition is confirmed by the experiments. We initially find significant improvements when using more components. But we reach an optimum at c=8. After that no significant improvements are measured and sometimes using more than 8 components even harms results (because of over-fitting). Comparable results are found for settings of NCbCr and XYpos not shown in Table 4. For some settings we already reach an optimum at c=4.

We expect the colour information in the image representation to be a valuable source for the purpose of image matching. When we vary the number of coefficients used from the Cb and Cr channels (NCbCr), we see that colour information is important for each setting of XYpos and c. Both NCbCr=1 and NCbCr=10 yield significantly better MAP scores than NCbCr=0. Thus it is important to use at least 1 DCT coefficient from each colour channel and thus to encode colour information in the models. For some settings in which we use more components (c>=8), we see that using 10 coefficients from the colour channels is significantly better than using only 1 (see Table 5 for an example). So, it seems wise to use as much information as possible for describing the images, as long as the models can accommodate all this information (i.e., as long as we have enough components).

Finally, it is unclear how varying the use of position information (XYpos) influences the MAP scores. For many settings of c and NCbCr, there is no significant difference between different values of XYpos. When there is, it is sometimes an improvement, sometimes a deterioration. Only when we use just a single component (c=1), we see a consistent significant improvement for models that do use position information. But when we use a single component to describe an image, all samples must be assigned to the same component and the position of this single component must be the centre of the image plane with a variance related to the size of the image. Still, the position information is different for portrait and landscape images since the position of the blocks in these will have different variance. Thus, adding position information in the single Gaussian case, acts as a portrait vs. landscape classifier, which, not surprisingly[5], improves retrieval results. One thing we *can* learn from analysing the results for different XYpos settings is that it never harms to use position information: in all cases using it either significantly improves results or it does not change results. Still, the experimental results do not clarify whether we should incorporate this information directly (XYpos=1) or after training the models (XYpos=2).

6. DISCUSSION: COREL VS TREC

As stated before, the Corel data set is a relatively easy dataset and we have to be careful to carry over results to other data sets. This section compares the results on the Corel data to the ones obtained on TREC data using the same model [18, 17].

First of all, the MAP score on Corel (.12, see Table 1) is much higher than the .03 obtained using TREC data[6], indicating that indeed Corel is a much easier collection (at least for the retrieval model under study). However, the Corel dataset is not a very realistic one; real-life unannotated multimedia collections are often not as well organised into disjoint and coherent themes and the data is often of much lower quality. Of course the classes in Corel are not strictly disjoint either: for example, the class *Thailand* contains images of beaches, elephants, and tropical flowers. In the experiments reported in this paper, we ignored these additional relevant images. This fact can be used to claim that the results are underestimated (we may have found more elephants then we think), but we think actual scores will be lower if full relevance judgements are available, since we probably will not retrieve many of the extra documents because of the lack of high similarity to the query image (the additional relevant images will have different backgrounds). Thus, we will miss more relevant images, and MAP will drop.

Section 5 showed that different parameter settings cause measurable (and significant) differences in retrieval scores. One has to take into account though that statistical signif-

[5] Images within a class tend to have the same orientation.

[6] For full example queries. With the TREC data we had the additional problem of combining multiple example images, but even using the best scoring single example for each query, MAP never exceeded .04 [18].

icance does not imply practical significance. If one parameter setting is only slightly (but significantly) better than another, but a user will not notice this in practise when looking at say the top 20 results, it is of not much use. On the other hand if we are able to find the right parameters this might be a small step in the right direction. Many of these small, but statistically significant, steps can in the end lead to an improvement that is noticeable in practise.

The fact that we did find measurable differences for different parameter settings, contrasts our findings with the TREC data [17], where we concluded that content based multimedia retrieval from generic archives is not mature enough for metric based evaluation. The question now is, how well the optimal settings found in Section 5 transfer to other collections. An argument in favour of this transferability is the fact that the Corel data set might be easy, but it does not have a clear bias toward one of the tested models. Furthermore, in experiments not reported in this paper, we saw that the Corel data suffers from the same colour prevalence we found on TREC data [17].

Still, results from one collection cannot directly be generalised for all collections. To show one technique is better than another one needs to test this using a variety of different collections, user types and relevance judgements [12]. Thus, on the one hand it is important to have a test collection which is representative of a certain realistic search task. This is what TREC tries to build. On the other hand we need to be able to measure (differences in) results, like we can when using Corel. It remains an open question still what kind of collections and retrieval scenarios are at once realistic and doable.

7. REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color- and texture-based image segmentation using EM and its application to content-based image retrieval. In *Proceedings of the Sixth International Conference on Computer Vision*, 1998.

[3] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03)*, 2003.

[4] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Seventh European Conference on Computer Vision*, pages 97–112, 2002.

[5] D. Hiemstra. *Using language models for information retrieval.* PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2001.

[6] D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the sixteenth annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'93)*, pages 329–338, 1993.

[7] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03)*, 2003.

[8] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10), 2003.

[9] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about Corel – evaluation in image retrieval. In *Proceedings of The Challenge of Image and Video Retrieval (CIVR2002)*, London, UK, July 2002.

[10] H. Neave and P. Worthington. *Distribution-Free Tests.* Unwyn Hyman Ltd., 1988.

[11] J. Ponte and W. Croft. A language modelling approach to information retrieval. In *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, 1998.

[12] G. Salton. The state of retrieval system evaluation. *Information Processing and Management*, 28(4):441–449, 1992.

[13] A. F. Smeaton and P. Over. The trec-2002 video track report. In E. M. Voorhees and D. K. Harman, editors, *The Eleventh Text REtrieval Conference (TREC-2002)*, 2002.

[14] A. F. Smeaton, P. Over, C. J. Costello, A. P. de Vries, D. Doermann, A. Hauptmann, M. E. Rorvig, J. R. Smith, and L. Wu. The trec2001 video track: Information retrieval on digital video information. In *Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002, Proceedings*, volume 2458 of *Lecture Notes in Computer Science*, pages 266–275, Rome, Italy, Sept. 2002. Springer.

[15] N. Vasconcelos. *Bayesian Models for Visual Information Retrieval.* PhD thesis, Massachusetts Institut of Technology, 2000.

[16] N. Vasconcelos and A. Lippman. A probabilistic architecture for content-based image retrieval. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2000)*, pages 216–221, 2000.

[17] T. Westerveld and A. P. de Vries. Experimental result analysis for a generative probabilistic image retrieval model. In *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03)*, 2003.

[18] T. Westerveld, A. P. de Vries, and A. van Ballegooij. CWI at the TREC-2002 Video Track. In E. M. Voorhees and D. K. Harman, editors, *The Eleventh Text REtrieval Conference (TREC-2002)*, 2002.

[19] T. Westerveld, A. P. de Vries, A. van Ballegooij, F. M. G. de Jong, and D. Hiemstra. A probabilistic multimedia retrieval modela and its evaluation. *EURASIP Journal on Applied Signal Processing, special issue on U nstructured Information Management from Multimedia Data Sources*, 2003.

[20] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 307–314, 1998.