A COMPARISON OF CONTINUOUS VS. DISCRETE IMAGE MODELS FOR PROBABILISTIC IMAGE AND VIDEO RETRIEVAL

Arjen P. de Vries and Thijs Westerveld

Centrum voor Wiskunde en Informatica (CWI)
Amsterdam
The Netherlands

ABSTRACT

The language modeling approach to retrieval is based on the philosophy that the language in a relevant document follows the same distribution as that in the query. This same philosophy can also be applied to content-based image and video retrieval, where the only difference lies in the definition of 'language'. Previous results on the TRECVID 2003 corpus have demonstrated that the visual content can be captured successfully by a continuous Gaussian Mixture Model. This paper investigates whether modeling the visual content by a discrete multinomial model (as used in full-text retrieval) is also viable. We compare the retrieval effectiveness obtained on the TRECVID 2003 corpus when using continuous vs. discrete keyframe models.

1. INTRODUCTION

Content-based image retrieval systems are usually based on a vector-space model [1]. The collection images are represented as vectors in a high-dimensional feature space, and similarity between images is estimated by a distance metric defined on this space. A drawback of this model is that it is far from obvious how to combine similarity in one representation (e.g., color histograms) with that of another one (e.g., texture); especially when we consider the combination of different modalities, such as video shots represented by their visual, audio, and speech content.

This paper follows an alternative approach based on a probabilistic model of retrieval. The idea is to guess the relevance of a collection image by computing its 'probability of relevance' to the user's request. Under the assumption that the relevance of one image is independent of the relevance of all other images, the Probabilistic Ranking Principle [2] states optimality of ranking the images in (reverse) order of their probability of relevance to the user.

The research presented here compares two different approaches to estimating this probability, based on the representations of query image and collection image in feature space. We use the TRECVID test collection, treating the video search task as an image retrieval problem by modeling each of the 32,318 shots with its representative keyframe. Refer to [3] for a version that incorporates the spatio-temporal aspect in the shot representation.

The paper is organised as follows. Section 2 explains the assumptions underlying the probabilistic approach to information retrieval. Section 3 introduces a discrete and a continuous approach to estimating the probability of relevance given a query and a collection image. Section 4 discusses experimental results on the TRECVID 2003 test collection.

2. PROBABILISTIC RETRIEVAL MODELS

A probabilistic image retrieval system attempts to answer the following 'basic question' (cf. [4, 5]): *What is the probability that* this *image is relevant to* this *query example?* Now, let random variable Q represent an example image (the query), random variable I represent an image from a collection of images $\mathcal{I}$, and event r denote 'relevance'. Answering the basic question is then equivalent to estimating the *probability of relevance* $\mathrm{P}(r|I,Q)$.

This probability can be estimated indirectly using Bayes' rule: $\mathrm{P}(r|I,Q) = \mathrm{P}(I,Q|r)\mathrm{P}(r)/\mathrm{P}(I,Q)$. For ranking images, estimation of $\mathrm{P}(I,Q)$ is avoided by using the *odds of relevance* instead of the likelihood. The odds are defined as $\mathrm{P}(r|I,Q)/\mathrm{P}(\bar{r}|I,Q)$ (where $\bar{r}$ denotes irrelevance). Assuming Q and I independent in the irrelevant case gives $\mathrm{P}(Q,I|\bar{r}) = \mathrm{P}(Q|\bar{r})\mathrm{P}(I|\bar{r})$.

Following the *query generation* approach [5], we factor $\mathrm{P}(I,Q|r)$ as $\mathrm{P}(Q|I,r)\mathrm{P}(I|r)$ to obtain the following equation for the odds of relevance:

$$\frac{\mathrm{P}(r|I,Q)}{\mathrm{P}(\bar{r}|I,Q)} = \frac{\mathrm{P}(I,Q|r)\mathrm{P}(r)}{\mathrm{P}(I,Q|\bar{r})\mathrm{P}(\bar{r})} \tag{1}$$

$$= \mathrm{P}(Q|I,r) \cdot \underbrace{\frac{\mathrm{P}(I|r)}{\mathrm{P}(I|\bar{r})}}_{\text{prior odds}} \cdot \underbrace{\frac{\mathrm{P}(r)}{\mathrm{P}(Q|\bar{r})\mathrm{P}(\bar{r})}}_{\text{independent of } I} \tag{2}$$

Since the goal is to rank the collection images, we can safely ignore the terms that are independent from a specific image. We further assume that, a priori, all images are equally

likely. This results in the following retrieval status value (RSV) for image I:

$$\mathrm{RSV}(I) = \mathrm{P}(Q|I,r) \tag{3}$$

2.1. A 'Language' Modeling Approach

So far, nothing has been said on the pragmatics of *how* this probability $\mathrm{P}(Q|I,r)$ should be determined from two given images Q and I.

The language modeling approach to information retrieval (IR) takes the view that the 'language' in a *relevant* (text) document follows the same distribution as that in the query. A statistical 'language model' is created for each document to represent its content; usually a simple unigram model capturing the frequency distribution of words occurring in the document. Several approaches using this paradigm for text retrieval have been collected in [6].

Interpreting 'language' in a more generic sense, the same IR paradigm can be applied to image retrieval. Comparing the image feature distribution between the query image and the collection image provides then the basis to estimate $\mathrm{P}(Q|I,r)$. So, applying the language modeling approach to image retrieval consists of the same three steps: (1) build a statistical model $\mathcal{M}_I$ for each image in the collection, (2) compute the likelihood of observing the query image given each and any of these models, and (3) present to the user the images with highest likelihood.

3. IMAGE REPRESENTATION

This philosophy considers an image to be the outcome of a random process generating n-dimensional feature vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$: the observed *samples*. In our current retrieval system, each sample corresponds to a small patch of the image represented in YCbCr color space. These patches are non-overlapping square 'pixel blocks' of 8×8 pixels, each represented by a feature vector consisting of the first 10 DCT coefficients of the Y-channel, the first DCT coefficient of both the Cb and the Cr channels, as well as its x and y position. Together, these fourteen dimensions describe colour, texture and position of the pixel block.

Summarising, we consider an image I as a bag of samples $\mathcal{X}_I = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_{N_S}}\}$; the observations obtained from a statistical source $\mathcal{M}_I$. We will compute the probability of relevance $\mathrm{P}(Q|I,r)$ by determining the likelihood of observing the query samples, $\mathrm{P}(\mathcal{X}_Q|\mathcal{M}_I)$. The probability of drawing a bag of samples is simply the joint probability of drawing the individual samples:

$$\mathrm{P}(\mathcal{X}_Q|\mathcal{M}_I) = \prod_{\boldsymbol{x}\in\mathcal{X}_Q} \mathrm{P}(\boldsymbol{x}|\mathcal{M}_I) \tag{4}$$

In practice, using $\mathrm{P}(\mathcal{X}_Q|\mathcal{M}_I)$ is prone to crediting the non-discriminative information in $\mathcal{X}_Q$ too much. In some choices for $\mathcal{M}_I$, the resulting estimate is also too sensitive to rare events, a problem known as the zero-frequency problem. A common solution to both problems is to *smooth* the probability distribution with a model based on a large, representative collection of images. We use the collection itself as reference collection for smoothing, using linear interpolation of the image model $\mathcal{M}_I$ with collection model $\mathcal{M}_\mathcal{M}$ (a technique known as Jelinek–Mercer smoothing). This results in the following estimate for the query likelihood:

$$\mathrm{P}(\mathcal{X}_Q|\mathcal{M}_I) = \prod_{\boldsymbol{x}\in\mathcal{X}_Q} \left(\lambda \cdot \mathrm{P}(\boldsymbol{x}|\mathcal{M}_I) + (1-\lambda) \cdot \mathrm{P}(\boldsymbol{x}|\mathcal{M}_\mathcal{M})\right) \tag{5}$$

The question that remains to be answered is the choice of the statistical model to describe the distribution of samples $\mathcal{X}_I$ in a collection image. The paper investigates two alternative statistical models for $\mathcal{M}_I$ and $\mathcal{M}_\mathcal{M}$. The first describes the bags of samples using a discrete model, assuming the samples can be generated from a multinomial distribution of grid cells defined by a regular partitioning of the feature space. The latter applies a continuous model, modeling the image feature density as a Gaussian Mixture Model.

3.1. Discrete Model

Past research has demonstrated that text retrieval models can be applied successfully as a basis for image retrieval. The Viper group were the first to recognise the benefits of applying the inverted file data structure to image retrieval, allowing a query-specific subspace in which to perform the ranking [7]. First, the feature space is discretised using a regular grid, and then the cosine metric is applied for ranking the images. The Mirror DBMS modeled a variety of multimedia retrieval problems in a uniform retrieval framework, proposing to cluster the various feature spaces with an unsupervised clustering algorithm, and then apply a text retrieval model based on the Inquery system [8]. The approach proved feasible for image as well as music retrieval.

Jin and Hauptmann presented in [9] the first language modeling approach to image retrieval, using a Multinomial Model (MNM) to capture the distribution of the image's pixels in Munsell color space. The skew in pixel distribution was handled by defining the grid cell boundaries from the data, assigning an equal number of pixels to each cell (instead of partitioning the color space uniformly).

The discrete model partitions the feature space of Section 3 using a regular grid, following the Viper approach. The grid cells are called 'clusters', and they are treated just like the word tokens in text retrieval, using the 'standard' unigram MNM (better known as the 'urn model'). The probability of observing a cluster c_i given an image model $\mathcal{M}_I$ is estimated by its normalised frequency of occurrence

in that image. Analogously, the background probability equals the normalised cluster occurrence frequency as observed in the complete collection. We discretised the dimensions corresponding to the first DCT coefficient of each YCbCr channel in ten, equisized partitions, and assigned five cells per dimension for the remaining DCT coefficients. The x and y dimensions are ignored to avoid an underpopulated discrete space.

3.2. Continuous Model

Vasconcelos has been the first to represent the images for retrieval by fitting a Gaussian Mixture Model (GMM) on image samples [10]. We follow this approach, modeling the images as mixtures of Gaussians with a fixed number of components ($C = 8$). The GMMs are trained using the standard EM algorithm, assuming diagonal covariance matrices. Details are provided in [11].

Vasconcelos has proposed to compare the statistical models of query and collection images using Kullback Leiber divergence, and developed an approximation that is cheaper to compute. Our research has shown however that the approximation assumptions were violated in the TRECVID data, and that ranking by the likelihood of the query image samples improved retrieval effectiveness [11]. Also, mean average precision (MAP) of search results improved when the image models were smoothed with a background model, emphasising the typicalities in the query samples. The background probabilities are computed by marginalisation over all collection image models, assuming uniform prior $\mathrm{P}(I)$:

$$\mathrm{p}(\boldsymbol{x}|\mathcal{M}_{\mathcal{M}}) = \sum_{I \in \mathcal{I}} \mathrm{p}(\boldsymbol{x}|\mathcal{M}_I)\mathrm{P}(I) \qquad (6)$$

4. EXPERIMENTS

A significant drawback of the GMM approach is the cost of both indexing time, when an iterative algorithm is needed to determine the model parameters, as well as retrieval time, because each and every image has to be accessed to determine the best results. Conversely, the MNM model restricts processing to those images that contain one or more of the query clusters. A disadvantage is however the partitioning of the feature space in discrete grid cells, fixing the number of grid cells per dimension at indexing time.

Figure 1 shows a collection image, and visualisations of its representations based on the MNM and GMM model. The 'ghost image' generated from the GMM shows how the baseball player is not well represented in the image model. Still, it results in reasonable search results – it captures well the baseball field itself. While the MNM model should be expected to generalise less well from the specific query image, it could perform better for highly focused image queries; the 'image spots' of [12].

Table 1. TRECVID 2003 Search Task Results (MAP)

	Full image		Selected regions	
	All	Designated	All	Designated
MNM	0.0044	0.0085	0.0066	0.0036
GMM	0.0281	0.0245	0.0264	0.0217

We evaluate the effectiveness of both models on the TRECVID 2003 search task. The query images are rescaled to at most 272x352 pixels and then JPEG compressed at a quality level of 20%, to match size and quality of the video collection. The mixing parameters have been set to $\lambda = 0.15$ (MNM) and $\lambda = 0.9$ (GMM) respectively. Table 1 summarises the results for using all example images per query or only a single 'designated' image, and using all samples in the query images or only the manually selected most important regions.

The best λ for the GMM was found using the TRECVID 2002 search task. The value for the discrete model has been determined a posteriori though. Somewhat surprisingly, the optimal value proved identical to the setting performing best in a variety of text retrieval experiments based on the same unigram model. Naturally, the influence of the background model is higher for the best MNM model, because of the zero-frequency problem. This is not an issue in the GMM, because Gaussians have infinite support; here, the role of smoothing is *only* to reduce the influence of non-discriminative query samples on the ranking.

Clearly, representing the samples from the keyframe images by a Gaussian Mixture Model of the image feature density directly in the 14-d feature space gives better retrieval effectiveness than the multinomial representation in the discretised feature space. Unfortunately, the experiments have not given sufficient evidence to conclude whether the discrete representation is indeed better at focusing on the specific information in an image (the 'image spots'). As argued in [13], a deeper analysis than just comparing the MAP scores.

The experimental results obtained with the multinomial representation are not yet optimal however, as only minimal efforts have been invested to choose the partitioning of the feature space. For instance, using a regular grid is not perfectly adequate for the skew present in the distribution of the higher DCT coefficients. Another experiment to be performed would use partially overlapping grid cells. Another improvement would be to apply Vector Quantisation in the feature space. But, each of these steps increases the cost of indexing and retrieval, and the potential advantages over the GMM approach are lost. Also note that Vasconcelos showed that the mixture modeling approach is equivalent to the VQ coding approach, when we assume separated Gaussians and replace image blocks by cluster means [14].

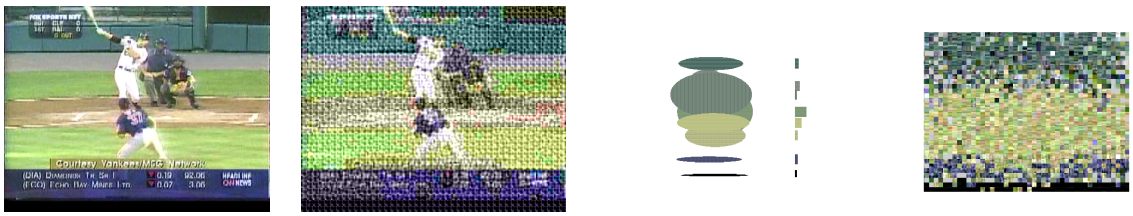

Fig. 1. A collection image, a visualisation of its discrete and continuous models, and a sample drawn from the GMM.

5. CONCLUSIONS

This paper applied the language modeling approach to information retrieval to the problem of image retrieval. Two different retrieval models were compared, a discrete model based on multinomials (the prevalent method in text retrieval) and a continuous model using Gaussian Mixture Models. The latter performed significantly better on the TRECVID 2003 test collection. For some queries however, the discrete model proved more effective.

We plan additional (though limited) research into improving the results obtained with the discrete model – especially regarding the definition of the grid cells. Improving the sampling process may be beneficial for both models. E.g., it might be worthwhile to investigate a different texture representation, such as Gabor filters. Another possible extension is to draw feature vectors from overlapping image patches of varying sizes, resulting in a multi-scale representation of the image. Finally, our TRECVID 2003 results on modeling the spatial-temporal information in the shots, instead of just a single keyframe image, have been very promising.

The most significant improvements for our retrieval system are however expected by applying advanced techniques to handle relevance feedback in an interactive setting. With a human in the loop, giving ample opportunity for relevance feedback, the balance between the costly computations used for the GMM approach vs. the relatively cheap processing based on the multinomial models may be reversed in favour of the discrete models, because these give more opportunity to adapt the search strategy on the fly to the user's decisions.

6. REFERENCES

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval: the end of the early years," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000, invited review.

[2] S.E. Robertson, "The Probability Ranking Principle in IR," *Journal of documentation*, vol. 33, no. 4, pp. 294–304, Dec. 1977.

[3] A.P. de Vries, T. Westerveld, and T. Ianeva, "Combining multiple representations on the TRECVID search task," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, Montreal, Quebec, Canada, May 2004, to appear.

[4] K. Sparck Jones, W. Walker, and S.E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments, part 1 & 2," *Information Processing & Management*, vol. 36, no. 6, pp. 779–840, 2000.

[5] J. Lafferty and Ch. Zhai, "Probabilistic IR models based on document and query generation," In Croft and Lafferty [6].

[6] W.B. Croft and J. Lafferty, Eds., *Language Modeling for Information Retrieval*, Kluwer Academic Publishers, Dordrecht, May 2003.

[7] D. McG. Squire, W. Müller, H. Müller, and T. Pun, "Content-based query of image databases: inspirations from text retrieval," *Pattern Recognition Letters*, vol. 21, no. 13-14, pp. 1193–1198, 2000, B.K. Ersboll, P. Johansen, Eds.

[8] A.P. de Vries, *Content and multimedia database management systems*, Ph.D. thesis, University of Twente, Enschede, The Netherlands, Dec. 1999.

[9] R. Jin and A.G. Hauptmann, "Using a probabilistic source model for comparing images," in *IEEE 2002 International Conference on Image Processing (ICIP'02)*, Rochester, NY, Sept. 2002.

[10] N. Vasconcelos, *Bayesian Models for Visual Information Retrieval*, Ph.D. thesis, Massachusetts Institut of Technology, 2000.

[11] T. Westerveld, A.P. de Vries, A. van Ballegooij, F.M.G. de Jong, and D. Hiemstra, "A probabilistic multimedia retrieval model and its evaluation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 2, pp. 186–198, 2003.

[12] H.G.P. Bosch, A. van Ballegooij, A.P. de Vries, and M.L. Kersten, "Exact matching in image databases," in *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo (ICME2001)*, Tokyo, Japan, August 22–25 2001, pp. 513–516.

[13] 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, pp. 135–142.

[14] N. Vasconcelos and A. Lippman, "Library-based coding: A representation for efficient video compression and retrieval," in *Proceedings of the 7th Data Compression Conference (DCC '97)*, J.A. Storer and M. Cohn, Eds., Snowbird, Utah, Mar. 1997, pp. 121–130, IEEE Computer Society Press.