JMLR: Workshop and Conference Proceedings 17 (2011) 41–47 2nd Workshop on Applications of Pattern Analysis

Bayesian Probabilistic Models for Image Retrieval

Vassilios Stathopoulos VASSILIOS@STATS.UCL.AC.UK
Department of Statistical Science
University College London.

Joemon M. Jose JOEMON.JOSE@GLASGOW.AC.UK
School of Computing Science
University of Glasgow.

Editor: Tom Diethe, José L. Balcázar, John Shawe-Taylor, and Cristina Tîrnăucă

Abstract

In this paper we present new probabilistic ranking functions for content based image retrieval. Our methodology generalises previous approaches and is based on the predictive densities of generative probabilistic models modelling the density of image features. We evaluate the proposed methodology and compare it against two state of the art image retrieval systems using a well known image collection.

Keywords: Information Retrieval, Generative Models, Image Retrieval, Mixture Models, Variational Inference.

1. Introduction

Probabilistic models for information retrieval are based on decision and probability theory and thus they provide guidance on how to optimally rank documents with respect to user queries (Robertson and Zaragoza, 2009). Despite their successful applications on web and text document retrieval, their application on retrieving multimedia documents such as images and videos with no associated text meta-data has not been widely explored until recently. Early content based image retrieval systems where based on similarity and distance functions designed specifically for the underlying image representation and feature extraction method (Smeulders et al., 2000).

Recently, Chum et al. (2008) proposed a methodology to represent images as unordered sets of discrete descriptive salient features which are analogous to terms for text documents and thus indexing and ranking models for information retrieval can be directly applied. Vasconcelos (2001); Westerveld et al. (2003) have generalised the methodology of Chum et al. (2008) and instead of creating a representation similar to text documents they employ generative probabilistic models to directly model the density of continuous features. This methodology has been shown to be very general and has also been applied for audio retrieval (Turnbull et al., 2008).

In this paper we present a framework based on Bayesian inference for deriving probabilistic ranking functions for multimedia information retrieval. We employ this framework to derive two new ranking functions for the *bag of terms* representation of Chum et al. (2008) and the generative model representation of Westerveld et al. (2003) and Vasconcelos (2001)

and show how previous ranking functions can be seen as approximations to those presented here.

2. Probabilistic Image Retrieval

Similar to text information retrieval, probabilistic models for content based image retrieval are based on the probabilistic ranking principle, i.e an image I is ranked w.r.t. to a user query image Q using $p(Q|I)$. This probability however is not estimated directly and thus a parametric model $p(\boldsymbol{x}|\boldsymbol{\theta}_I)$ is employed to model the density of image features. The image specific parameters $\boldsymbol{\theta}_I$ are often estimated using a maximum likelihood, $\hat{\boldsymbol{\theta}}_I = \arg\max_{\boldsymbol{\theta}_I} p(I|\boldsymbol{\theta}_I)$, or a maximum a posteriori procedure $\hat{\boldsymbol{\theta}}_I = \arg\max_{\boldsymbol{\theta}_I} p(I|\boldsymbol{\theta}_I)p(\boldsymbol{\theta}_I)$. Assuming the same parametric model for query images, ranking is then based on the query likelihood $p(Q|\hat{\boldsymbol{\theta}}_I)$. Probabilistic image retrieval systems differ on the type of features extracted from images, i.e. the image representation, and on the model assumptions defined by the parametric models $p(I|\boldsymbol{\theta}_I)$.

A popular methodology for image retrieval is to create a representation of images that is similar to that of text documents and then apply directly information retrieval ranking models. For example, Chum et al. (2008) extract local SIFT features (Lowe, 2004) from a collection of images and quantise them using K-means to form a visual vocabulary. SIFT features from an image are mapped to their closest visual term from the vocabulary and an image is then represented as an unordered set of *visual terms*. The distribution of terms in an image under this representation is modelled as a multinomial distribution $\mathcal{M}(x|\boldsymbol{\theta}_I)$ and the ML estimates of the parameters is $\hat{\boldsymbol{\theta}}_I = n_{t,I}/\sum_{t'} n_{t',I}$ where $n_{t,I}$ denotes the frequency of term t in image I. A MAP estimate can be obtained by assuming a Dirichlet prior distribution over the parameters and results in $\hat{\boldsymbol{\theta}}_I = (n_{t,I} + \alpha_t - 1)/\sum_{t'}(n_{t',I} + \alpha_{t'} - 1)$. The prior hyper-parameters α_t are commonly set to the frequency of terms in the collection (Zhai and Lafferty, 2001).

The method presented by Westerveld et al. (2003) and Vasconcelos (2001) can be seen as a generalisation of the method of Chum et al. (2008) that avoids quantisation errors by using a semi-parametric model to model directly the density of continuous image features. For each image a finite multivariate Gaussian mixture model of the form $p(\boldsymbol{x}|\boldsymbol{\theta}_I) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is employed to model the density of Discrete Cosine Transform (DCT) coefficients extracted from a uniform grid over an image. In this setting an image I is represented by an unordered set of vectors in $\mathbb{R}^D$ where D is the number of DCT coefficients. ML estimates are obtained by using the EM algorithm for finite mixture models while MAP estimates can also be obtained by assuming conjugate priors. In Westerveld et al. (2003) smoothing with a background model is also discussed as an alternative in order to obtain a regularised estimate for the parameters.

The *bag of terms* approach of Chum et al. (2008) is a very efficient method for retrieval and can scale to large collections since an inverted index data structure can be used due to the sparse nature of the representation. However, retrieval performance is greatly affected by the quantisation errors induced by the K-means procedure. On the other hand, the approach of Vasconcelos (2001) and Westerveld et al. (2003) is also sensitive to the number of mixture components which in both studies is set empirically and to a constant value

across the collection. This can lead to severe over, or under, fitting for images with less, or more, complex densities.

3. Predictive Densities for Image Ranking

The ranking functions using the query likelihood based on ML or MAP point estimates are in fact approximations to a ranking function employing the *predictive densities* of image models. In particular we can write $p(\boldsymbol{x}|I)$ as

$$p(\boldsymbol{x}|I) = \int_{\boldsymbol{\theta}_I} p(\boldsymbol{x}|\boldsymbol{\theta}_I)p(\boldsymbol{\theta}_I|I)d\boldsymbol{\theta}_I \tag{1}$$

where $p(\boldsymbol{\theta}_I|I)$ is the posterior of the model parameters obtained by Bayes' theorem $p(\boldsymbol{\theta}_I|I) = p(I|\boldsymbol{\theta}_I)p(\boldsymbol{\theta}_I)/p(I)$. In cases where the posterior is sharply peaked around some value $\hat{\boldsymbol{\theta}}_I$ then $p(\boldsymbol{x}|I) \approx p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_I)$ and thus it is equivalent to the ML or MAP functions. However, when data is scarce the posterior is broad and the uncertainty is taken into account providing regularised estimates of relevance. Moreover, the ranking functions obtained by the predictive densities in Equation (1) are no longer sensitive to parameter estimates as they are not dependent on $\boldsymbol{\theta}_I$. However, they rely on the ability to accurately estimate the integral in Equation (1) and the posteriors $p(\boldsymbol{\theta}_I|I)$ for all images in the collection.

3.1. The Multinomial Dirichelt model

For the *bag of terms* model discussed in the previous section, the posterior and predictive densities can be easily calculated in closed form provided that a Dirichlet prior is specified. In particular, the posterior is also a Dirichlet of the form $\mathcal{D}(\boldsymbol{\theta}_I|\boldsymbol{n}_{\cdot,I} + \boldsymbol{\alpha} - 1)$, where $\boldsymbol{n}_{\cdot,I}$ is the vector of term frequencies in image I, and the predictive density for a query image Q is

$$p(Q|I) = \frac{(\sum_t n_{t,Q})!}{\prod_t n_{t,Q}!} \frac{\Gamma\left(\sum_t n_{t,I} + \alpha_t\right)}{\Gamma\left(\sum_t n_{t,Q} + n_{t,I} + \alpha_t\right)} \prod_t \frac{\Gamma\left(n_{t,Q} + n_{t,I} + \alpha_t\right)}{\Gamma\left(n_{t,I} + \alpha_t\right)} \tag{2}$$

Equation (2) can be simplified by calculating its log, as it is a convex function and thus it does not affect ranking; removing terms which depend only on $n_{t,Q}$, as they are constant for all images in the collection; and finally using the fact that $\Gamma(n) = (n-1)!$ for all positive integers n to give the following ranking function

$$\log p(Q|I) \propto \sum_{t:n_{t,Q},n_{t,I}>0} \sum_{g}^{n_{t,Q}} \log\left(1 + \frac{n_{t,I}}{a_t + g - 1}\right) - \sum_{j=1}^{\sum_{t'} n_{t',Q}} \log\left(\sum_{t'} n_{t',I} + a_{t'} + j - 1\right) \tag{3}$$

3.2. Variational inference for finite Gaussian mixture models

Unfortunately the posterior and the predictive density do not have a closed form for mixture models and thus the above methodology cannot be applied directly to the methods of Westerveld et al. (2003). We therefore resort to the framework of variational inference (Attias, 2000) in order to obtain analytical approximations.

We start by imposing a conjugate prior over the model parameters, i.e. Dirichlet for the mixing coefficients $\boldsymbol{\pi}$, Gaussian for the means $\boldsymbol{\mu}$ and inverse Wishart for the covariance matrices $\boldsymbol{\Sigma}$, of the form $p(\boldsymbol{\theta}_I) = p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k)p(\boldsymbol{\Sigma}_k)$ where

$$p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{a}_0), \quad p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\boldsymbol{m}_0, \beta^{-1}\boldsymbol{\Sigma}_k), \quad p(\boldsymbol{\Sigma}_k) = \mathcal{IW}(\boldsymbol{\Sigma}_k|\boldsymbol{W}_0, v_0)$$

Furthermore, we introduce the latent variables $\boldsymbol{Z}_I$ where $z_{i,k} = 1$ iff the i^{th} vector of an image is allocated to the k^{th} mixture component otherwise $z_{i,k} = 0$ and re-write the likelihood as $p(I|\boldsymbol{\theta}_I) = \prod_i \prod_{k=1}^{K} [\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{i,k}}$ where i indexes vectors in image I. We augment model parameters and latent variables in an extended space $\boldsymbol{\Theta}_I = \{\boldsymbol{\theta}_I, \boldsymbol{Z}_I\}$ and by assuming an approximate posterior q for the augmented set of parameters $\boldsymbol{\Theta}_I$ which factorizes as $q(\boldsymbol{\Theta}_I) = q(\boldsymbol{\theta}_I)q(\boldsymbol{Z}_I)$ the marginal likelihood can be written as

$$p(I) = \underbrace{\int_{\boldsymbol{\Theta}_I} q(\boldsymbol{\Theta}_I) \log \frac{p(I, \boldsymbol{\Theta}_I)}{q(\boldsymbol{\Theta}_I)} d\boldsymbol{\Theta}_I}_{\text{Lower Bound}} - \underbrace{\int_{\boldsymbol{\Theta}_I} q(\boldsymbol{\Theta}_I) \log \frac{p(\boldsymbol{\Theta}_I|I)}{q(\boldsymbol{\Theta}_I)} d\boldsymbol{\Theta}_I}_{\text{KL}} \tag{4}$$

From Equation (4) we can see that by maximising the *lower bound* the KL divergence between the true and the approximate posterior is minimised. By optimising the *lower bound* for each of the approximate densities separately while considering the other fixed we arrive at the following result for the approximate posteriors of mixture model parameters

$$\begin{aligned} q(\boldsymbol{z}_i) &= \mathcal{M}(\boldsymbol{z}_i|1, \rho_{i,1}, \ldots, \rho_{i,K}), & q(\boldsymbol{\mu}_k) &= \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (\beta + n_k)^{-1}\boldsymbol{\Sigma}_k\right) \\ q(\boldsymbol{\Sigma}_k) &= \mathcal{IW}\left(\boldsymbol{\Sigma}_k|\boldsymbol{W}_k, n_k + v_0\right), & q(\boldsymbol{\pi}) &= \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{a}_0 + n_k) \end{aligned}$$

The parameters $\boldsymbol{\rho}, \boldsymbol{m}, \boldsymbol{W}, n_k$ of the approximate posteriors in the above equations are found by the *variational* EM algorithm. See Bishop and Corduneanu (2001) for more details.

Finally, substituting the above expressions into Equation (1) we can analytically evaluate the integral and obtain the predictive density which takes the form of a mixture of Student-t densities. The ranking function for a query image is then

$$p(Q|I) = \prod_{x \in Q} \frac{1}{\sum_{k=1}^{K} p_k} \sum_{k=1}^{K} p_k \mathrm{St}(\boldsymbol{x}|\boldsymbol{m}_k, \boldsymbol{\Lambda}_k, v_k + 1 - D) \tag{5}$$

where $p_k = a_0 + n_k$, $v_k = v_0 + n_k$ and $\boldsymbol{\Lambda}_k = \frac{(v_k+1-D)(\beta+n_k)}{1+\beta+n_k} \boldsymbol{W}_k^{-1}$ and $\mathrm{St}(\cdot)$ is the Student-t density.

To estimate the number of mixture components K, we follow a method similar to that of Bishop and Corduneanu (2001). The method is based on initially over-estimating the number of components and setting K to a large value. Selecting a Dirichlet prior for the mixing coefficients $\boldsymbol{\pi}$ such that sparse solutions are preferred, i.e. setting α_0 to a value close to zero, the *variational* EM algorithm converges to solutions where many of the components are identical to the prior distribution with $n_k = 0$ and thus they can be removed as they do not affect the result in Equation (5).

4. Experiments

In this section we present experimental results in order to validate the proposed methodology and compare it with previous approaches. We will be using the Corel 5K dataset which has been widely used in the literature to evaluate image retrieval and classification systems. The Corel 5K dataset consists of 5,000 images from 50 thematic categories such as images of tigers or images of cars. The collection is further divided into a training set of 4,500 images and a test set of 500 images where in the test set there are 10 images from each thematic category. We index only the 4,500 images of the training set and use the test set as user queries. Retrieval performance is evaluated using the thematic category information of each query. That is, for each query image we except the retrieval systems to rank higher the 90 images in the training set from the same category.

We follow the same methodology as in (Westerveld et al., 2003) and (Vasconcelos, 2001) in order to extract features from images. All images are rescaled to 192×128 pixels and a sliding window of 8×8 pixels with an overlap of 4 pixels is used to segment images into local regions. From each region we use the DCT coefficients after a transformation from the RGB colour space to the Luminance-Colour space. Exploiting the compression properties of the DCT coefficients we use only the first 3 DCT coefficients from the colour bands of the image and all DCT coefficients from the Luminance band resulting into 70 dimensional vectors.

For the *bag of terms* representation we apply K-means with the Euclidean distance to cluster the feature vectors from the 4,500 images in the training set into 2,000 clusters and map each vector to its closest centroid. Two ranking function were then obtained by using a MAP estimate of the Multinomial parameters and the predictive densities as discussed in the previous sections. For the rest of the section we will denote them by BOT-MAP and BOT-PD respectively. In both cases the prior hype-parameters α_t are set to the frequency of visual terms in the training set.

For the mixture model representation we reproduce the experiments in (Westerveld et al., 2003) and (Vasconcelos, 2001) and use the EM algorithm to obtain ML and MAP estimates. The number of mixture components was fixed to 8 and as reported in both studies results where not significantly affected by different settings while 8 produced the higher retrieval performance. For the *variational* EM algorithm we used 40 components as an initial estimate and after convergence removed all components with $n_k = 0$. For the prior hyper-parameters we followed (McLachlan and Peel, 2000, Chap. 4) and used the following settings, $\alpha_0 = 10E^{-4}$, $\beta = 1$, $v_0 = 5$ while $\boldsymbol{m}_0$ and $\boldsymbol{W}_0$ where set to the mean and covariance of the feature vectors in the training set. Both the EM and *variational* EM algorithms where initialised using a random assignment of the latent variables $\boldsymbol{Z}_I$. The three ranking functions obtained by the ML, MAP and the predictive densities, will be denoted as GMM-ML, GMM-MAP and GMM-PD respectively for the rest of this section.

4.1. Results

Table 1 summarises the results for the 500 queries in the test set using the standard information retrieval evaluation measures. Average Precision (AP) is the average of precisions computed at the point of each relevant image in a ranking list. Mean Average Precision

(MAP) is the mean AP across all queries. R-Prec is the precision calculated at the position of the last relevant image in a ranking list and P@N is the precision calculated at the Nth position of the ranking list.

Table 1: Retrieval results for 500 query images in the test set. $*$ indicates statistical significance using a Wilcoxon rank-sum test with 1% significance level.

Method	**MAP**	**R-Prec.**	**P@5**	**P@10**	**P@20**
BOT-MAP	0.0333	0.0364	0.0441	0.0429	0.0383
BOT-PD	0.0341	0.0375	0.0477	0.0431	0.0387
GMM-ML	**0.0975***	**0.1280***	**0.3038***	**0.2599***	**0.2179***
GMM-MAP	0.0999	0.1308	0.3070	0.2645	0.2210
GMM-PD	**0.1165***	**0.1457***	**0.3315***	**0.2836***	**0.2370***

From Table 1 we can see that despite the efficiency and scalability of the *bag of terms* representation, quantisation errors can negatively impact retrieval performance. Directly modelling the density of continuous features in images using semi-parametric models significantly improves retrieval performance at the cost of the additional computations for calculating the query likelihood for all images in the collection. Using regularised estimates of the parameters of Gaussian mixture models also improves retrieval performance although results are not statistically significant. Finally, the superior performance of GMM-PD method can be attributed to the following two reasons. Firstly, in contrast to a MAP estimate which provides a regularised point estimate, the predictive densities provide a regularised estimate of relevance where the uncertainty associated with model parameters is marginalised. The two approaches will be equivalent if the posterior is sharply peaked around some values, but when the posterior is broad the predictive density averages all possible solutions weighted by their posterior probability. Secondly, the number of mixture components in GMM-PD is automatically determined by the output of the *variational* EM algorithm. In contrast, previous approaches (Westerveld et al., 2003; Vasconcelos, 2001) set the number of components empirically to a fixed value for all images in the collection which can result in images with more complex densities to be under-fitted while others with more simple densities to be over-fitted.

5. Conclusions and Future Work

We have presented a methodology for deriving retrieval functions for multimedia documents based on the predictive density of generative models. The method does not make particular assumptions about the representation of documents in the collection but requires the specification of a probabilistic generative model for the density of the documents' features. Despite the superior retrieval performance compared to previous approaches scalability to large scale collections remains an important issue since the predictive densities for all images in the collection have to be evaluated.

Designing efficient indexing data structures such as the inverted index for the Multinomial Dirichlet model is not trivial for models such as mixtures of Gaussians. We believe that a more general methodology such as Locality Sensitive Hashing applied to kernel functions (Kulis and Grauman, 2009) for generative models, such as Probability Product Kernels (Jebara et al., 2004), is an interesting future direction.

References

Hagai Attias. A variational bayesian framework for graphical models. In *In Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

C. M. Bishop and A. Corduneanu. Variational Bayesian model selection for mixture distributions. In *Artificial Intelligence and Statistics*, 2001.

O. Chum, J. Philbin, and A. Zisserman. Near duplicate image detection: min-hash and tf-idf weighting. In *British Machine Vision Conference*, 2008.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *J. Mach. Learn. Res.*, 5:819–844, December 2004. ISSN 1532-4435.

Brian Kulis and Kristen Grauman. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*, pages 2130–2137, 2009.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. ISSN 0920-5691.

Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, October 2000.

Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389, 2009.

Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):467 –476, 2008.

Nuno Vasconcelos. Image indexing with mixture hierarchies. In *CVPR '01*, pages 3–10, 2001.

Thijs Westerveld, Arjen P. de Vries, Alex van Ballegooij, Franciska de Jong, and Djoerd Hiemstra. A probabilistic multimedia retrieval model and its evaluation. *EURASIP: J. of Applied Signal Processing*, 2003(1):186–198, 2003.

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings ACM SIGIR*, SIGIR '01, pages 334–342. ACM, 2001.