Published in IET Image Processing
Received on 6th December 2009
Revised on 29th March 2010
doi: 10.1049/iet-ipr.2009.0402

ISSN 1751-9659

Relevance feedback approach for image retrieval combining support vector machines and adapted Gaussian mixture models

A. Marakakis[1] *G. Siolas*[1] *N. Galatsanos*[2]
A. Likas[3] *A. Stafylopatis*[1]

[1]*School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece*
[2]*Department of Electrical and Computer Engineering, University of Patras, 26500 Patras, Greece*
[3]*Department of Computer Science, University of Ioannina, 45110 Ioannina, Greece*
E-mail: amara@central.ntua.gr

Abstract: A new relevance feedback (RF) approach for content-based image retrieval (CBIR) is presented, which uses Gaussian mixture (GM) models as image representations. The GM of each image is obtained as an adaptation of a universal GM which models the probability distribution of the features of the image database. In each RF round, the positive and negative examples provided by the user until the current round are used to train a support vector machine (SVM) to distinguish between the relevant and irrelevant images according to the preferences of the user. In order to quantify the similarity between two images represented as GMs, Kullback–Leibler (KL) approximations are employed, the computation of which can be further accelerated taking advantage from the fact that the GMs of the images are all refined from a common model. An appropriate kernel function, based on this distance between GMs, is used to make possible the incorporation of GMs in the SVM framework. Finally, comparative numerical experiments that demonstrate the merits of the proposed RF methodology and the advantages of using GMs for image modelling are provided.

1 Introduction

In content-based image retrieval (CBIR), an image description based solely on low-level visual features (representing colour, texture, shape information, etc.) is usually adopted (e.g. [1–4]). This image description is subsequently used to compare the images of an image database to one or more query images submitted by a user as representative examples of his/her preferences and to rank the database images according to their similarity with the query. Then, the top images in the ranking are displayed to the user as the retrieval results. For the aforementioned comparison between the database images and the query provided by the user, a distance measure based on the specific image description used is needed. The final target of this task is to retrieve images relevant to the user query. To this end, and in order to improve the retrieval results, a lot of effort has been devoted in developing features and strategies appropriate to sufficiently describe the image content (e.g. [5–10]). Nevertheless, there is an intrinsic difficulty for low-level image features to capture the human perception of image similarity. Usually, it is the semantic content of an image that the user is interested in, and this semantic content cannot be described adequately using only low-level image features. This well-studied problem is widely known as the semantic gap.

Relevance feedback (RF) is an interactive supervised learning technique that has been proposed to bridge the semantic gap between the low-level image features used and the semantic content of the images and, thus, to improve the retrieval results (e.g. [11–15]). In particular, RF attempts to insert the subjective human perception of image similarity into a CBIR system. In order for this to be accomplished, the user is required to assess, in each RF round, the retrieved images as relevant or irrelevant to the initial query and to submit his/her assessment as a feedback to the system. Then, the system takes into account this feedback and updates in an appropriate way the image ranking criterion.

Several RF approaches have been proposed that can be classified into two main categories. The first category concerns methods that are based on some learning model. Among these methods are those that give the most promising results in the field of RF for CBIR (e.g. [13, 16–18]). It must be mentioned here that among all the learning models used for this task, the most popular one is the support vector machine (SVM). In what concerns SVMs, several SVM variations have been proposed and a number of different SVM kernels have been adopted, from typical ones to more sophisticated (e.g. [13, 16, 17, 19–21]). In the context of RF for CBIR, the learning models are trained in each RF round to discriminate between the positive and negative feedback examples provided by the user until the

current RF round. Thus, in each RF round, a new decision boundary between the classes of relevant and irrelevant images is formed by the exploitation of the new feedback. Then, the distance of each database image from the new decision boundary is used as an updated ranking criterion.

The second category of RF methods includes those that attempt to determine an appropriate representation for the user query in the image feature space. This can be interpreted as modelling the statistical distribution of feedback examples in the feature space. These methods can be further divided into two subcategories.

The first subcategory includes methods that make the assumption that the feedback examples form one cluster in the feature space, thus representing the query using the corresponding centroid and a covariance-like matrix. Then, the distance between the query and the database images is, usually, expressed in the form of a Mahalanobis distance. The cornerstone of such methods is MindReader [22]. Other methods that work under this assumption are presented in [11, 12, 14]. Unfortunately, in most cases, the single cluster assumption is very restrictive even for the set of positive examples. Moreover, under this assumption the negative feedback examples cannot be taken into account in an easy way, because they naturally spread over different semantic categories and, thus, it cannot be claimed that they form one cluster.

The second subcategory includes methods which assume that the feedback examples (either positive or negative) form more than one cluster (e.g. [15, 23–26]). In order to describe in a probabilistic manner a multi-cluster data set, a multi-modal distribution model is needed that can be obtained in two different ways. Either, when the images are represented as vectors in a multidimensional space, it can be used to approximate the distribution of the examples provided by the user (e.g. [23, 24]) or, it can be used to describe the distribution of the locally extracted feature vectors of each image (e.g. [15, 25, 26]). In the first case, a number of difficulties arise. In particular, in each RF round, a new distribution model must be computed based on both the new and the old examples. Furthermore, usually, the user participates in few RF rounds and does not provide the system with a sufficient number of examples in each round. Thus, in most cases, it is difficult to robustly estimate the distribution of, generally, very-high-dimensional representations of the examples in the feature space. In the second case, when one distribution model is used for each image based on locally extracted feature vectors, the situation is better, because a larger number of features will be available for the distribution estimation. However, when feature extraction techniques based on keypoint detection are employed to extract features from the image (which have been shown to be the most promising ones), the extracted features are rather few and of high dimensionality. Thus, the same question of robust distribution estimation arises. In this work, a special technique is used to alleviate this problem. Moreover, an efficiently computed distance measure between these distribution models is needed. In what concerns the multi-modal distribution models that can be used for the previously described task, the most popular and promising choice concerns Gaussian mixture (GM) models.

GMs are a well-established methodology to model probability density functions (pdfs), which is proven to have a lot of merits, such as adaptability to the data, modelling flexibility and robustness. Owing to these advantages, GM models have been employed for a wide range of applications (e.g. [15, 20, 21, 23, 27–32]). In what concerns the problem of CBIR, GM models have already been considered with promising results (e.g. [15, 23, 27, 30, 33]). The critical issue to be addressed when using GM models in RF framework for CBIR is what distance measure between GMs will be used. It must be one that separates well the different models and, in addition, can be computed efficiently. The standard distance measure between pdfs is the Kullback–Leibler (KL) divergence. Unfortunately, this measure cannot be computed in closed form for GM models. On the contrary, in order to approximate KL for GMs, random sampling Monte-Carlo methods have been proposed. However, these methods are extremely time consuming, making the use of KL divergence impractical for RF in CBIR, where the main issue is the real-time interaction between the system and the user. In order to overcome this difficulty, some alternatives have been proposed (e.g. [15, 33]). Furthermore, in [34, 35], two very similar KL approximations have been introduced, which can be computed in closed form for GMs. This work will depend mainly on these distance measures to obtain a fast estimation of the distance between GMs. Another issue regarding GM models is whether there are sufficient data for robust estimation of the model parameters. The standard algorithm used for estimation of the parameters of a GM is the expectation-maximisation (EM) algorithm, which estimates the aforementioned parameters in a maximum likelihood (ML) manner. Although, when a few samples of the underlying distribution are available, a more robust estimation, based on a maximum a posteriori (MAP) principle, can be computed using a variation of the standard EM algorithm, known as MAP-EM (e.g. [20, 21, 31, 32, 36]). In this work, the advantages of using this technique for the estimation of model parameters of images described by a relatively small number of locally extracted feature vectors will be demonstrated.

This work is based on the idea of combining the classifier-based RF methods with those based on probabilistic models to describe the image features. In this way, it will be able to exploit and amplify the merits of both the approaches. In particular, GMs are used as image representations and SVMs are employed for the task of RF. In order for this combination to be accomplished, an appropriate SVM kernel function is required to quantify the similarity between GMs. The kernel function used in this work is based on an efficiently computable distance measure between GMs, which is an approximation of the KL divergence [35]. To extract image features the scale invariant feature transform (SIFT) approach [7] is used, which is considered as a state-of-the-art feature extraction technique [37], and, in particular, the colour-SIFT variation [8] that exploits colour information. Furthermore, to cope with the problem of the relatively small number of keypoints (and thus feature vectors per image) detected by the SIFT technique, the MAP-EM algorithm [32] is employed to robustly estimate the parameters of the GM model of each image. This algorithm exploits a universal GM trained on information extracted from the whole image database for estimating a prior that is imposed on the parameters of each image GM model. This training strategy of the image GMs allows for an even faster computation of the KL approximations used, with minor losses regarding CBIR system performance.

The rest of this paper is organised as follows. In Section 2, GMs are described in the context of image modelling for

doi: 10.1049/iet-ipr.2009.0402

CBIR and details about the algorithms used to estimate the model parameters are provided. Furthermore, the issue of defining an efficiently computable distance measure between GMs is discussed further. In Section 3, the SVM methodology for binary classification is described and put in the framework of CBIR with RF. Moreover, an appropriate kernel function quantifying the similarity between GMs is presented. In Section 4, details about the implementation of the methods used in the experiments are provided. In Section 5, details are given about the simulations used to assess the validity of the proposed method, and the experimental results are presented and discussed. Finally, in Section 6, conclusions and directions for future research are provided.

2 GM models

As it has already been mentioned, GM models are very popular and promising pdf models and are characterised by many advantages. The pdf corresponding to a GM defined in a d-dimensional space is given by the following formulas

$$p(\boldsymbol{x}|\Theta) = \sum_{j=1}^{K} \pi_j \phi(\boldsymbol{x}|\theta_j) \tag{1}$$

$$\Theta = (\pi_1, \theta_1, \ldots, \pi_K, \theta_K) \tag{2}$$

$$\theta_j = (\mu_j, \boldsymbol{\Sigma}_j) \tag{3}$$

$$\phi(\boldsymbol{x}|\theta_j) = N(\boldsymbol{x}|\theta_j) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_j|}} \mathrm{e}^{-1/2(\boldsymbol{x}-\mu_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}-\mu_j)} \tag{4}$$

where $\boldsymbol{x} \in R^d$ is a vector in the previously mentioned space, K the number of Gaussian components in the model, $0 \leq \pi_j \leq 1$ the mixing probabilities with $\sum_{j=1}^{K} \pi_j$, and $\phi(\boldsymbol{x}|\theta_j)$ a Gaussian pdf with mean μ_j and covariance $\boldsymbol{\Sigma}_j$.

In this work, GMs will be used to model the distribution of feature vectors extracted from the images of an image database. In this framework, each image is described as a bag of feature vectors that are computed locally. Then, an iterative algorithm, for example the EM or some variation, is employed to estimate the parameters of a GM model that will represent the distribution of the features of the image in the feature space.

2.1 Parameter estimation of GM models

The standard procedure to estimate the parameters of a GM based on this set of feature vectors is the EM algorithm [38]. This algorithm is defined as an iterative process in which two steps, step E (expectation) and step M (maximisation), are repeatedly executed until convergence. The algorithm aims at estimating the model parameters that maximise the log-likelihood function. The iterative EM procedure monotonically converges to a local maximum of the usually-very-complex log-likelihood function, which depends on the initial estimates of the model parameters.

When the number of feature vectors available for training is relatively small compared to their dimension and to the number of components in the mixture, the robust estimation of the model parameters using the EM algorithm has been shown to be very problematic. Thus, in [32], and based on the analysis given in [36], a variation of the EM algorithm, called MAP-EM, is defined. Assume that there is somehow available a universal GM model on the feature space, representing the general distribution of the feature vectors in all database images. This universal model can be used as a prior distribution $p(\Theta)$ on the model parameters Θ of each image GM. Then, instead of maximising the log-likelihood, the MAP approach suggests the maximisation of the MAP likelihood function defined as

$$\begin{aligned} L_{\mathrm{MAP}}(\Theta) &= \log \prod_{i=1}^{N} p(\Theta) p(\boldsymbol{x}_i|\Theta) \\ &= N \log p(\Theta) + \sum_{i=1}^{N} p(\boldsymbol{x}_i|\Theta) \end{aligned} \tag{5}$$

For further information about this method one can refer to [36].

The resulting scheme of model parameter adaptation proposed in [32] is very similar to the standard EM algorithm. In particular, the E step is exactly the same for the two algorithms. However, in the new M step, the new estimates of the mixture parameters are computed by

$$\pi_j = \frac{a_j f_j + (1 - a_j)\pi_j^{pr}}{\sum_{k=1}^{K} [a_k f_k + (1 - a_k)\pi_k^{pr}]} \tag{6}$$

$$\mu_j = a_j \boldsymbol{c_j} + (1 - a_j)\mu_j^{pr} \tag{7}$$

$$\boldsymbol{\Sigma}_j = a_j \boldsymbol{R_j} + (1 - a_j)[\boldsymbol{\Sigma}_j^{pr} + \mu_j^{pr}(\mu_j^{pr})^{\mathrm{T}}] - \mu_j \mu_j^{\mathrm{T}} \tag{8}$$

In the above formulas, f_j, $\boldsymbol{c_j}$ and $\boldsymbol{R_j}$ are the same as those used in the standard EM algorithm

$$f_j = \frac{\sum_{i=1}^{N} \gamma_{ij}}{N}, \quad \boldsymbol{c_j} = \frac{\sum_{i=1}^{N} \gamma_{ij} \boldsymbol{x}_i}{\sum_{i=1}^{N} \gamma_{ij}}, \quad \boldsymbol{R_j} = \frac{\sum_{i=1}^{N} \gamma_{ij} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}}{\sum_{i=1}^{N} \gamma_{ij}} \tag{9}$$

where $\boldsymbol{x}_i$, $i = 1, \ldots, N$ are the feature vectors extracted from an image I. Furthermore, π_j^{pr}, μ_j^{pr} and $\boldsymbol{\Sigma}_j^{pr}$ correspond to the mixing weights, the means and the covariance matrices, respectively, of the components of the universal GM model, from which the parameters of the new GM model are adapted. From (6)–(8), it can easily be seen that the new estimations for the parameters of the model that is being adapted are not dependent solely on the new evidence derived by the training feature vectors $\boldsymbol{x}_i$, as this is expressed by f_j, c_j and R_j, but they are computed as a combination between the new evidence and the prior knowledge derived from the universal model. The relative weight with which the new evidence from the training feature vectors is taken into account for the determination of the new parameter estimates for the mixture component j is denoted by a_j. This coefficient is defined by

$$a_j = \frac{\sum_{i=1}^{N} \gamma_{ij}}{\sum_{i=1}^{N} \gamma_{ij} + \tau} \tag{10}$$

with $\sum_i \gamma_{ij}$ to be a measure about how many feature vectors are assigned to the component j of the mixture and τ to be a parameter controlling the weight with which the universal model parameters are taken into account in the determination of the new estimates. Obviously, the more the feature vectors correspond to component j (the larger the value of $\sum_i \gamma_{ij}$), the more the influence of the new evidence in the estimation of the new parameters for this component.

On the contrary, if few feature vectors are assigned to component j, more emphasis is given to the universal model.

Thus, in order to train a GM model for each image in the database using the previously described MAP-EM algorithm, at first we need a universal GM model modelling the general distribution of feature vectors. Such a model can be easily acquired by using a sufficiently large sample of feature vectors from the image database and employing the standard EM algorithm to estimate the GM parameters. After the estimation of the parameters of the universal model, these can be used along with the feature vectors extracted by each image for the adaptation process described above, which will finally result in one adapted GM per image, with the same number of components as the universal one.

2.2 Distance measures between GM models

Given that each image is represented by one GM model, in order to describe the similarity between images, a distance measure between GMs must be defined.

The KL divergence [23] is the most commonly used distance measure between pdfs. In particular, for two pdfs, $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$, the KL divergence is defined by

$$\mathrm{KL}(p_1\|p_2) = \int p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \tag{11}$$

When $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ are Gaussian pdfs, it is known [35] that the KL divergence between them is given by

$$\mathrm{KL}(p_1\|p_2) = \frac{1}{2}\left[\mathrm{trace}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - \log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} - d\right] + \frac{1}{2}(\mu_1 - \mu_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\mu_1 - \mu_2) \tag{12}$$

where μ_1, $\boldsymbol{\Sigma}_1$ (μ_2, $\boldsymbol{\Sigma}_2$) are the mean and covariance matrix of p_1 (p_2) and d is the feature space dimension. Thus, a computation of KL in closed form and, thus, efficiently is possible when the pdfs under comparison are Gaussians. Unfortunately, this is not the case when GMs are considered. In particular, the KL divergence cannot be computed in closed form for GMs, because the integral in (11) cannot be computed analytically. Furthermore, the random sampling Monte-Carlo methods, which have been proposed to estimate this integral, are extremely time consuming and, thus, cannot be employed in the context of CBIR with RF.

In order to overcome these computational problems, some closed-form KL approximations have been proposed. Among these approximations, there are two popular distance measures that are very similar to each other. The first of them was introduced in [34] and is defined implicitly via a similarity measure called asymptotic likelihood approximation (ALA). More specifically, for two GMs, $p_1(\boldsymbol{x}) = \sum_{i=1}^{K_1} \pi_{1i}\phi(\boldsymbol{x}|\theta_{1i})$, $\theta_{1i} = (\mu_{1i}, \Sigma_{1i})$ and $p_2(\boldsymbol{x}) = \sum_{j=1}^{K_2} \pi_{2j}\phi(\boldsymbol{x}|\theta_{2j})$, $\theta_{2j} = (\mu_{2j}, \Sigma_{2j})$, the ALA measure is defined as

$$\mathrm{ALA}(p_1\|p_2) = \sum_{i=1}^{K_1} \pi_{1i}\left\{\log \pi_{2\beta_{\mathrm{mhl}}(i)} + \left[\log \phi(\mu_{1i}|\theta_{2\beta_{\mathrm{mhl}}(i)}) - \frac{1}{2}\mathrm{trace}(\boldsymbol{\Sigma}_{2\beta_{\mathrm{mhl}}(i)}^{-1}\boldsymbol{\Sigma}_{1i})\right]\right\} \tag{13}$$

where

$$\beta_{\mathrm{mhl}}(i) = \arg\min_{j=1,\ldots,K_2}[(\mu_{1i} - \mu_{2j})^{\mathrm{T}}\boldsymbol{\Sigma}_{2j}^{-1}(\mu_{1i} - \mu_{2j}) - \log \pi_{2j}] \tag{14}$$

is a correspondence function between the components of the two mixtures based on the Mahalanobis distance. It is proven that under certain conditions, the KL divergence between p_1 and p_2 can be computed using ALA as follows

$$\mathrm{KL}(p_1\|p_2) = \mathrm{ALA}(p_1\|p_1) - \mathrm{ALA}(p_1\|p_2) \tag{15}$$

Nevertheless, the conditions are too restrictive and they do not hold in the GM case. However, we can use the above equation as an approximation to the KL divergence. Moreover, taking into account that for $\mathrm{ALA}(p_1\|p_1)$ the correspondence function is $\beta_{\mathrm{mhl}}(i) = i$, the following approximation of the KL divergence can be obtained

$$\mathrm{KL}_{\mathrm{apprx}}(p_1\|p_2) = \sum_{i=1}^{K_1} \pi_{1i}\left[\mathrm{KL}(\phi(\boldsymbol{x}|\theta_{1i})\|\phi(\boldsymbol{x}|\theta_{2\beta(i)})) + \log\frac{\pi_{1i}}{\pi_{2\beta(i)}}\right] \tag{16}$$

with

$$\beta(i) = \beta_{\mathrm{mhl}}(i), \quad \forall i = 1, \ldots, K_1 \tag{17}$$

The second KL approximation has been proposed in [35]. It is defined in exactly the same way as the previous distance measure; thus, for its computation, (16) is used again. The only difference between this measure and the ALA-based KL approximation is that a different correspondence function is adopted. In particular, in [35] the following correspondence function is used

$$\beta_{\mathrm{gkl}}(i) = \arg\min_{j=1,\ldots,K_2}[\mathrm{KL}(\phi(\boldsymbol{x}|\theta_{1i})\|\phi(\boldsymbol{x}|\theta_{2j})) - \log \pi_{2j}] \tag{18}$$

and then, for the function β of (16), it holds

$$\beta(i) = \beta_{\mathrm{gkl}}(i), \quad \forall i = 1, \ldots, K_1 \tag{19}$$

Thus, the correspondence function of [35] is based on the KL divergence between the Gaussian components of the mixtures, instead of the Mahalanobis distance.

Both of these two KL approximations described above are based on the determination of a correspondence between the components of the two mixtures. Thus, for each component of p_1, the nearest component of p_2 is determined using (14) or (18), respectively. Then, both measures compute the final distance value by combining linearly some measures based on the KL divergence between the nearest components. It should be reminded that the KL divergence between Gaussians can be computed in closed form (12).

A problem related to the above measures regards the fact that determination of the correspondence between components has quadratic complexity with the number of components in the mixtures. This can burden the computation, particularly when mixtures with many components are used. Nevertheless, in our case this problem can be alleviated [21], since all GMs are adapted from the same universal model via the MAP-EM algorithm.

doi: 10.1049/iet-ipr.2009.0402

In this case, first, all GMs have the same number of components and, second, at most of the times either the correspondence function provides the result $\beta(i) = i$ or, even if $\beta(i) \neq i$, the assumption that $\beta(i) = i$ does not result in a distance value much different from the correct one. Thus, the final formula for the approximate but efficient computation of the KL between two GMs adapted from a universal model is

$$\mathrm{KL}_{\mathrm{apprx}}(p_1 \| p_2) = \sum_{i=1}^{K} \pi_{1i} \left[\mathrm{KL}(\phi(\boldsymbol{x}|\theta_{1i}) \| \phi(\boldsymbol{x}|\theta_{2i})) + \log \frac{\pi_{1i}}{\pi_{2i}} \right] \tag{20}$$

It is easy to observe that, in this new approximation, the previous two KL approximations are unified, since their only difference lies in the correspondence function they use.

In this work, this new approximate KL distance measure (20) will be used in the context of RF for CBIR, since it is characterised by linear complexity on the number of mixture components and, thus, it is significantly faster than the initial KL approximation (16) which is of quadratic complexity.

3 Support vector machines

SVM [38] is a popular and successful learning model which in our case will be employed for binary classification. Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ be a training set corresponding to a binary classification problem, where $\boldsymbol{x}_i$ are the patterns and $y_i \in \{-1, +1\}$ are the corresponding labels. Given this training set, an SVM classifier can be constructed to discriminate between the two categories. The main advantage of the SVM methodology is the fact that the classification problem is solved in a high-dimensional kernel space (through an appropriate mapping of the original input vectors), where the problem becomes separable. Thus, a linear decision boundary can be computed in the kernel space, although the image of this boundary can be highly non-linear in the initial pattern space.

In our experiments, we will be using one of the most popular non-linear kernel functions for SVMs, Gaussian radial basis function (RBF), which is defined by

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) \tag{21}$$

After training the classifier by solving a quadratic optimisation problem (see [38] for more details), the decision function for a new pattern $\boldsymbol{x}$ is computed as

$$y(\boldsymbol{x}) = \sum_{i=1}^{N} a_i y_i k(\boldsymbol{x}, \boldsymbol{x}_i) + b \tag{22}$$

where $\boldsymbol{b}$ is a bias parameter, the value of which can be easily computed after the determination of the optimisation coefficients in the training phase.

In what concerns the classification of a new pattern x, it holds that the value $|y(\boldsymbol{x})|$ is proportional to the distance of the input pattern $\boldsymbol{x}$ from the decision boundary. Thus, the value $y(\boldsymbol{x})$ in (22) can be regarded as a measure of confidence about the class of $\boldsymbol{x}$, with large positive values (small negative values) strongly indicating that $\boldsymbol{x}$ belongs to the class denoted by '+1' ('−1'). On the contrary, values of $y(\boldsymbol{x})$ around zero provide little information about the class of x.

3.1 Using SVMs for RF in CBIR

In the framework of CBIR with RF, in each round of RF, a classification problem as the one described above needs to be solved. In particular, in each round of RF, we have a set of images that correspond to the feedback examples provided by the user until now. Each of these images is labelled by −1 or +1 in case the user considers it as irrelevant or relevant to the initial query, respectively. The initial query is considered to be one of the relevant images and is labelled by +1.

Using the training set of feedback examples, an SVM classifier can be trained to distinguish between the classes of relevant and irrelevant images. Each image in the database will be presented to the trained classifier, and the value of the decision function (22) will be used as the ranking criterion. The higher the value of the decision function for an image, the more relevant this image is considered by the system. In this context, in each RF round a new decision boundary is learned using the additional information derived from the new feedback examples provided by the user and, thus, the ranking criterion and the retrieval results are updated.

3.2 Combining GMs with SVMs

Assume now that each image is modelled using a GM. In the context of CBIR using RF, the question that immediately arises is how to employ an SVM classifier in each round of RF based on this image representation. In this context, both the training patterns and the new patterns presented for classification will be GM models.

Taking into account both the SVM optimisation problem and the form of decision function (22) used for category estimation after training, it becomes apparent that nowhere the patterns or their images in the kernel space are used explicitly. Only the values of the kernel function are needed. Thus, if an appropriate kernel function between GMs could be defined, GM representations for the patterns could be used in the SVM framework in exactly the same way as vectors.

The kernel function is defined as the inner product of the patterns in the kernel space, namely it is a similarity measure. Taking inspiration from the Gaussian RBF (21), a kernel function of similar form can be defined for GMs. In particular, if with $d(p_1, p_2)$ a symmetric distance measure between the GMs p_1 and p_2 is denoted, then a kernel function of the form $\exp(-\gamma d(p_1, p_2))$ can be used for GMs.

The KL approximations presented in Section 2.2 are not symmetric. However, this is not a real problem, because a symmetric version can be defined in the form

$$\mathrm{SKL}_{\mathrm{apprx}}(p_1, p_2) = \frac{1}{2}\mathrm{KL}_{\mathrm{apprx}}(p_1 \| p_2) + \frac{1}{2}\mathrm{KL}_{\mathrm{apprx}}(p_2 \| p_1) \tag{23}$$

where with $\mathrm{KL}_{\mathrm{apprx}}(p_1 \| p_2)$ either one of the two KL approximations proposed in [34, 35] or the KL approximation of (20) is denoted. Based on the above considerations, the function

$$k(p_1, p_2) = \exp(-\gamma \mathrm{SKL}_{\mathrm{apprx}}(p_1, p_2)) \tag{24}$$

can be used as a kernel function expressing the similarity between GMs.

4 Method implementation

Summarising the proposed method, SIFT-based image representations are estimated using GMs adapted from a universal model via the MAP-EM algorithm. Moreover, the KL approximations discussed in Section 2.2 are employed for comparison between the images and, particularly, the one that takes advantage from the fact that all the image GMs are adapted from a common universal model, to achieve linear complexity over the number of mixture components. This distance measure is used to define an appropriate SVM kernel, which makes possible the incorporation of GM representations in the SVM framework. Then, SVM classifiers trained using the feedback examples are employed, in each RF round, to update the image ranking criterion and, thus, to improve the retrieval results.

In order to test the validity of the proposed RF methodology, a number of experiments were conducted. In these experiments, a subset of Corel is used as image database, which is a common choice for the evaluation of CBIR systems. In particular, although the images of the Corel database are professionally annotated and categorised, many images containing the same semantic content are distributed across different Corel categories. Hence, a new semantic categorisation was defined by merging some Corel categories to form new more distinctive ones. Finally, an image database with 4500 images partitioned in 25 semantic categories was formed. Each of these images has a resolution of 384×256 pixels and its categorisation is considered as the ground truth in our experiments. We will refer to this database as DB.

In order to extract appropriate features for the images in the aforementioned database, the SIFT [7] framework was used. In this framework, image gradients in multiple image scales are approximated by differences between Gaussian convolutions and, for each image, a number of keypoints are detected by determining the local maxima and minima of the gradients in scale space. For each of these keypoints, an invariant descriptor is computed as the local histogram of gradient directions around the keypoint. As histogram quantisation 8 bins for direction and 16 bins for location are used, thus resulting in a 128-dimensional feature vector for each keypoint. These features have been proven to be highly distinctive and invariant to image scale, rotation and illumination changes, robust to noise addition and affine distortions of the images, etc. Nevertheless, they are extracted using grey images. Some variations have been proposed, in order to make possible the incorporation of colour information in the descriptors (e.g. [8, 9]). In [8], a number of colour characteristics invariant to varying lighting conditions have been introduced and studied. The extension of the SIFT methodology to coloured images is straightforwardly addressed by using, instead of grey gradients, colour gradients based on the aforementioned colour invariants. An open code implementation of these colour SIFT features, which constitute an adaptation of the standard SIFT algorithm to include colour image descriptions, is available in [39]. This is the method adopted in this work for feature extraction from the images. As in [31], prior to GM training, a reduction of the feature vector dimensionality to 50 features, using principal component analysis (PCA), was performed. This has significant advantages such as the decorrelation of features, noise removal, and efficient and robust estimation of GM parameters.

In the following experiments, in order to train a GM in an ML manner (as is the case of the universal model) a variation of EM, called greedy-EM [40], has been adopted. This method avoids the problems related to the strong dependence of the solution on parameter initialisation, by incrementally adding components to the mixture until the desirable number of components has been reached. Furthermore, in the experiments presented below, unless otherwise stated, for the universal GM training, about 65 000 feature vectors from the DB are images are randomly selected and used as input to the greedy-EM algorithm.

For the adaptation of each image GM, the MAP-EM iterative algorithm presented in Section 2.1 is used. In this case, there is no problem regarding the initialisation. The parameters of the universal GM model are always used as initial values for each image GM parameters. In this way, the resulting GM model for each database image has the same number of components with the universal model from which it has been adapted, and each component j of the image model has been adapted from the corresponding component j of the universal model. In what concerns the parameter τ used in MAP-EM algorithm, in our experiments it was set to a constant value $\tau = 15$. This value was determined empirically to be a reasonable choice resulting in good performance. Moreover, as stated in [32], the performance is rather insensitive for values between 8 and 20.

All GMs used in the experimental section, either the universal or the image-based ones, assume a diagonal covariance for each of their components. This choice has found to be the best, because it reduces dramatically the complexity of training and comparison of the models with no noticeable performance deterioration.

It is also important to note that, from an efficiency point of view, the GM parameters of all models (universal and image) are computed once off-line and serve as alternative representations to the histogram vectors. Consequently, the complexity of all further computations, namely SVM training and classification, will depend solely and linearly to the number of kernels used.

In order to provide comparative results for the proposed method, we implemented another RF methodology where each image is represented by a histogram vector. To derive such a representation, a universal vocabulary [20] of visual words is built, based on a corresponding universal GM trained as mentioned before. Each component of the universal GM model corresponds to one visual word and the posterior probabilities $p(j|\boldsymbol{x}_i)$ of the components j of the universal GM for the feature vectors $\boldsymbol{x}_i$ of an image are used to form a histogram for this image having the same number of bins as the number of the universal GM components. The frequency corresponding to bin j is given by

$$b_j = \frac{\sum_{i=1}^{N} p(j|\boldsymbol{x}_i)}{N} \tag{25}$$

assuming that N is the number of feature vectors of the image. After the histogram formation for each image, we use the vector representation to train in each RF round (i) a RBF Neural Network (RBF NN) and (ii) an SVM with the standard Gaussian RBF kernel (21).

For the RBF NN, the only parameter that needs to be set is the bias of the radial basis transfer function radbas of each first layer neuron

$$\text{radbas}(n) = \mathrm{e}^{-n^2}, \quad n = \|\boldsymbol{b} - \boldsymbol{w}\| \cdot \text{bias} \tag{26}$$

doi: 10.1049/iet-ipr.2009.0402

where $\boldsymbol{b}$ is the input vector and w is the radial basis centre point. For each one of our experiments, we determined the optimal bias value by cross validation of the results. In general, the performance of the RBF NN is highly sensitive to bias; nevertheless, we found that for our tests the optimal values ranged from 2.4 to 3. Furthermore, we used the standard Matlab Neural Network Toolbox implementation which combines linearly the first-layer RBF neurons to the output.

Two SVM parameters need to be determined when SVMs are used with both histogram and GM image representation. These parameters are the learning parameter C and the kernel function parameter γ. The performance of the methods is proven to be rather stable with respect to the first of these parameters, with the precondition that a value considerably larger than 1 has been selected. In all the experiments presented below, the choice $C = 100$ is made. The second parameter is related with the range of the distance measure used between patterns (either SKL_{apprx}, for SVMs handling patterns represented by GMs, or the squared Euclidean distance, for patterns represented by vectors). In [41], a value for this parameter equal to the inverse of the mean (computed for a sufficiently large set of patterns) of the distance used is adopted. On the experiments conducted in this work, the best choice for the value of this parameter seems not to diverge much from this rule. Moreover, the SVM implementation available in [42] is used for all of our experiments.

Furthermore, apart from comparing the proposed method with the one that uses the image representation based on visual vocabularies, a number of other comparisons have also been performed. In particular, the use of the approximate distance measure of (20) is compared with the use of one of the initial KL approximations, and particularly the one defined by (16) and (18). Moreover, the results of a comparison between the performances obtained for different choices for the number of components in the mixtures are presented. Finally, a study on the performance achieved for different choices regarding the source of the feature vectors used to train the universal GM is included.

5 Experiments

In this section, the results of the experiments conducted in order to quantify the performance of the proposed RF method are presented. For these experiments, an RF simulation scheme was designed.

In particular, a query set is formed by randomly selecting Q images from DB. Each image in this set is presented once as initial query to the system. Then, the system ranks the database images according to the distance measure used (squared Euclidean distance for vector representations, KL approximation for GMs). In this initial stage, there are no feedback examples and, thus, neither RBF NNs nor SVMs can be employed yet. From the M top-ranked images, a number P of relevant and a number N of irrelevant images are randomly selected (in case there are more than P or N such images, respectively) and are used as feedback examples (and RBF centre points), along with the initial query, to train an RBF NN and an SVM classifier to distinguish between relevant and irrelevant images. After training the NN and the classifier, the database images are re-ranked using the output of the NN or the decision function values of the new SVM model. Again, P relevant and N irrelevant images are selected from the M top images in the new ranking list and are used, along with the previous examples, to re-train the NN (with additional RBF centre points) and the classifier, thus updating the NN output and the SVM ranking function. This process is repeated, until a number R of RF rounds have been completed. As a measure of performance, the precision in top T images in the ranking list is used. This measure is, simply, the ratio of relevant images in the first T images of the ranking. Images that belong to the same database category with the initial query are considered relevant, and those that belong to a different category are considered irrelevant. For the experiments presented in this section, the following choices have been made for the values of the simulation parameters: $Q = 1000$, $M = 50$, $P = 5$, $N = 5$ and $R = 6$. For the parameter of scope, T, results for $T = 20$ and $T = 30$ are presented.

In Figs. 1 and 2 the proposed method is compared with the RBF NN and SVM-based methods that use as image representations the histograms obtained from visual vocabularies. In Fig. 1, precision in scope $T = 20$, averaged on all the initial queries, is depicted and Fig. 2 depicts the same measure in scope $T = 30$. For each image GM, eight Gaussian components are used. The proposed method is denoted as GMM8_MAP_FKLA. The label VW256

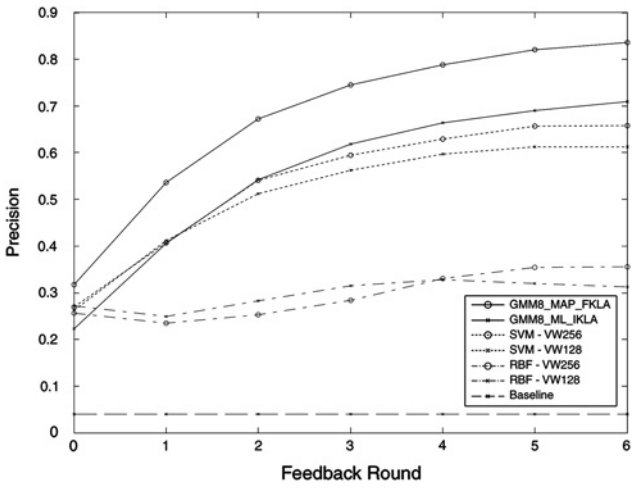

Fig. 1 *Comparison between GMs and visual vocabularies (VW) with SVM and RBF: average precision in scope $T = 20$*

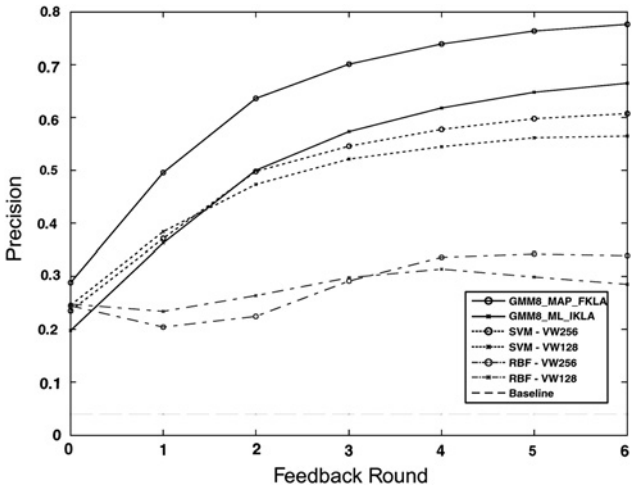

Fig. 2 *Comparison between GMs and visual vocabularies (VW) with SVM and RBF: average precision in scope $T = 30$*

(VW128) denotes that visual vocabularies of 256 (128) visual words are used. On the contrary, GMM8_ML_IKLA refers to a variation of the proposed method, in which the GM models of the images have been trained based on the standard ML manner. In this case, the distance measure of (20) cannot be used, because the components of the image models have been trained independently. Thus, only one of the initial KL approximations proposed in [34, 35] can be used. In particular, for the experiments presented in this section, distance measure [35] is adopted, because, after some preliminary experiments, it seems to lead to slightly better performance. The baseline shows how a random ranking system would have performed, since the average precision would have been 4% (as there are 25 one-against-all classes). All methods clearly perform better than the baseline. As it is apparent from the figures, the proposed method based on adapted GMs constantly outperforms all other methods. Furthermore, doubling the size of the visual vocabulary used, from 128 visual words to 256 visual words, leads to a relatively small improvement in performance. The most interesting point is that even use of GMs, trained in an ML manner as image representations, results in better performance compared to that of the visual vocabulary representation. As demonstrated in the figures, the GMM8_ML_IKLA method starts from a lower precision level compared to that of VW256 and VW128, but after a few RF rounds results in a precision level higher than that of both VW256 and VW128. These results are a strong indication of the strength of GMs as image models. Furthermore, it would be interesting to test whether improvement or deterioration in terms of precision will result from using standard SIFT instead of colour SIFT, since, in the case of colour SIFT, the same number of features is used to describe the properties of all three RGB channels. In contrast, in plain SIFT, all features are used for encoding one channel. We also note that, even though the RBF NNs improve precision by about 30% over the baseline, they are clearly outperformed by the state-of-the-art SVM approach. Thus, we conduct the rest of our comparisons only between GMs and SVMs.

Fig. 3 demonstrates the average precision in scope $T = 20$ per database category, after the sixth RF round. The performance of the proposed method is compared to that of VW256, category by category. It can be seen that there is no one category in the database in which the visual words representation results in superior performance compared to that obtained using adapted GMs. On the contrary, there are many database categories in which the adapted GMs give significantly better performance.

Fig. 4 concerns the influence of the number of GM components used for the image models. Again, average precision in 20 first images of the ranking is demonstrated. The proposed method, GMM8_MAP_FKLA, is compared with a number of variations using image models with different number components. In particular, GMM4_MAP_FKLA, GMM16_MAP_FKLA and GMM32_MAP_FKLA use adapted GMs with 4, 16 and 32 components, respectively. As it can be seen, the best performance is obtained when GMs of eight components are used. GMs with four components result in slightly worse performance. On the other hand, increasing the number of components does not seem to improve performance. On the contrary, it leads to degradation, particularly in the first retrieval rounds, when few feedback examples have been collected.

Furthermore, in Fig. 5, the degradation in performance that results from using the fast form of the KL approximation (20) is

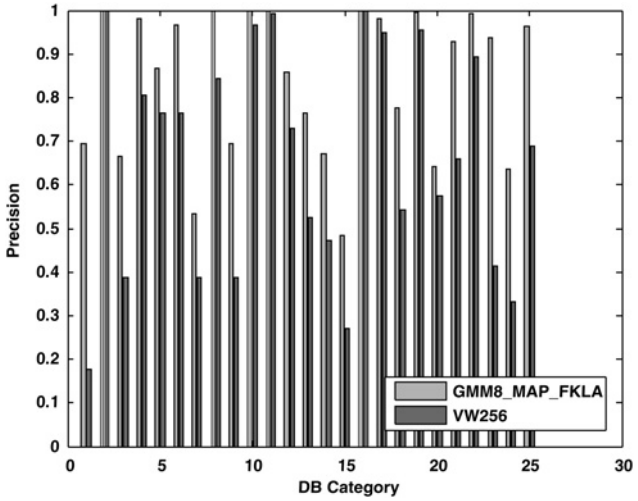

Fig. 3 *Comparison between GMs and visual vocabularies: average precision per DB category after sixth RF round, in scope $T = 20$*

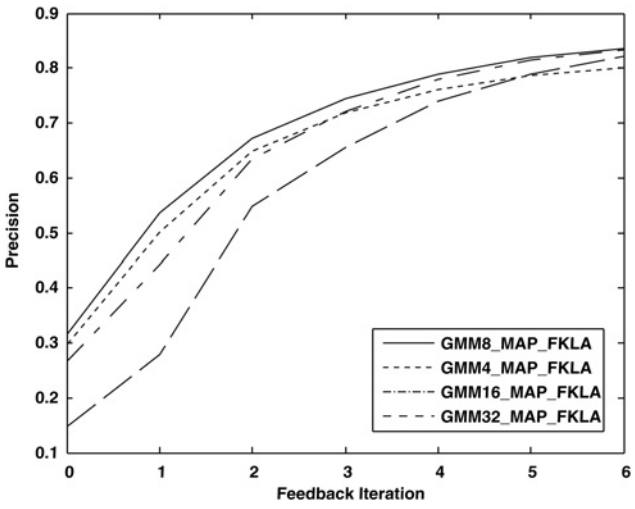

Fig. 4 *Study on the influence of the number of mixture components: average precision in scope $T = 20$*

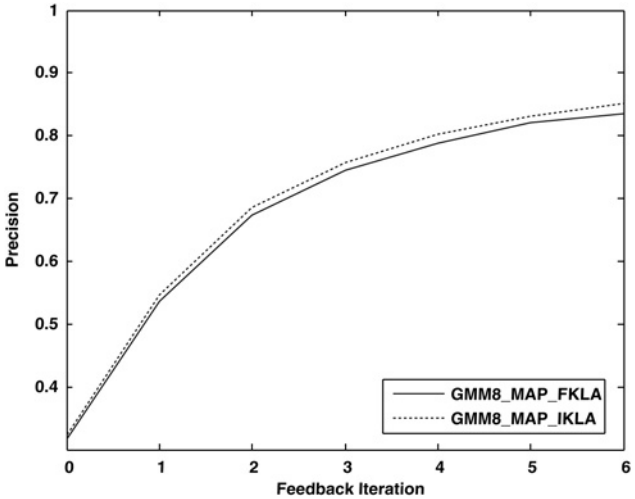

Fig. 5 *Comparison between distance measures for GMs: average precision in scope $T = 20$*

doi: 10.1049/iet-ipr.2009.0402

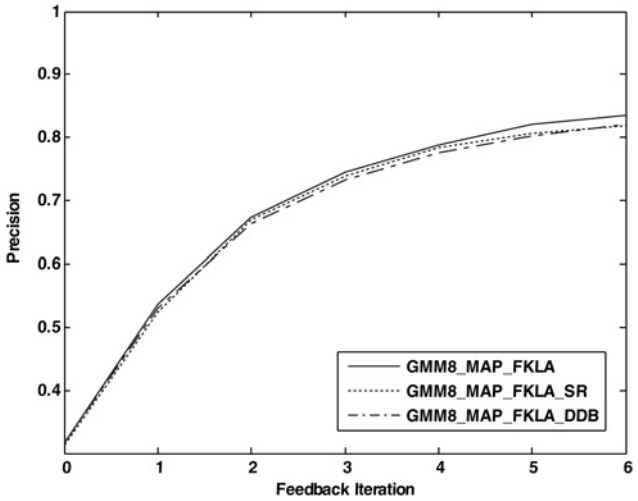

Fig. 6 *Study on the influence of training set used for the universal GMs: average precision in scope $T = 20$*

investigated. As discussed previously, with GMM8_MAP_FKLA the proposed method is denoted. On the other hand, GMM8_MAP_IKLA is a variation using the initial form of distance measure (equations (16) and (18)). As one could predict, when the initial form of the distance measure is used, better results are obtained. But, as it is apparent, the difference in performance is very small. Taking into account that the average time needed to compute the distance between two GMs is 0.004 s (on a 3 GHz PC) when the initial form of distance measure is used, and only 0.0008 s for the fast KL approximation, this minimal loss in performance is compensated.

Finally, in Fig. 6, the robustness of the proposed method is shown when reducing the number of images used to train to universal GM. For the proposed method GMM8_MAP_FKLA, the universal GM used for the adaptation of the image GMs was trained using about 65 000 feature vectors from the entire database DB. In this figure, GMM8_MAP_FKLA_SR denotes the proposed method, when a reduced set of only 25 randomly selected images from DB have been used for the training of universal GM. Furthermore, with GMM8_MAP_FKLA_ DDB, the training of the universal GM on feature vectors from a database different than DB is denoted. In particular, for the training of the universal GM, about 25 000 feature vectors from the images of database [43] were used. As it can be seen from the corresponding results, the performance of the method is rather robust with respect to variations in the feature set used for the training of the universal GM, since only small performance degradation can be observed in both experiments.

6 Conclusions – future work

An RF methodology based on adapted GM models and SVMs has been proposed. This methodology uses a variation of the standard EM algorithm to adapt, in a Bayesian manner, the GM models of the images based on a universal GM model trained using the standard EM algorithm and information extracted from the entire database. Then, a fast KL approximation is used as a distance measure between GMs, and an SVM classifier with an appropriate kernel function is employed in each RF round to perform the RF task.

As indicated by our experiments, the proposed method results in significantly higher performance compared to that obtained using visual vocabularies for image representation and the standard RBF NN and SVM methodology for RF. Furthermore, even if GMs trained in the standard ML manner are used as representations for the images, the resulting performance is superior compared to that of visual vocabularies. This is a proof of the efficiency and flexibility of GMs as image representations for RF. Moreover, the presented experimental results indicate that a small number of mixture components are required to achieve satisfactory performance. Additional experiments proved that the use of a fast KL approximation, enabled by the particular algorithm used to train the GMs (instead of that defined in [35]), results in no significant degradation in performance. Finally, by providing experimental results regarding the performance of the method when the universal GM was trained on a set of feature vectors not completely representative of the database used for retrieval, the noticeable robustness of the method was demonstrated.

In future, we intend to test our method using other distance measures between GMs. Furthermore, we aim to attempt to apply techniques for determining automatically the best number of mixture components. Moreover, we plan to modify our method to incorporate some feature selection techniques in each RF round, in order to improve the results of the corresponding classification problem. Additionally, it would be interesting to try to use, in the same context, other learning models appropriate for RF. Finally, it is in our plans to test the scalability of the proposed method using even larger image databases.

7 Acknowledgments

This work was partly supported by Public Funds under the PENED 2003 Project. The Project is co-funded by the European Social Fund (80%) and National Resources (20%) from the Hellenic Ministry of Development – General Secretariat for Research and Technology.

8 References

1 Rui, Y., Huang, T.S., Chang, S.F.: 'Image retrieval: current techniques, promising directions, and open issues', *J. Vis. Commun. Image Represent.*, 1999, **10**, pp. 39–62
2 Datta, R., Li, J., Wang, J.Z.: 'Content-based image retrieval: approaches and trends of the new age'. Proc. 7th ACM SIGMM Int. Workshop on Multimedia Information Retrieval (MIR'05), 2005, pp. 253–262
3 Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: 'Content-based image retrieval at the end of the early years', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000, **22**, (12), pp. 1349–1380
4 Vasconcelos, N.: 'Minimum probability of error image retrieval', *IEEE Trans. Signal Process.*, 2004, **52**, (8), pp. 2322–2336
5 Manjunath, B.S., Ohm, J.R., Vasudevan, V.V., Yamada, A.: 'Color and texture descriptors', *IEEE Trans. Circuits Syst. Video Technol.*, 2001, **11**, (6), pp. 703–715
6 Da Cunha, A.L., Jianping, Z., Do, M.N.: 'The nonsubsampled contourlet transform: theory, design, and applications', *IEEE Trans. Image Process.*, 2006, **15**, (10), pp. 3089–3101
7 Lowe, D.G.: 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vis.*, 2004, **60**, (2), pp. 91–110
8 Burghouts, G.J., Geusebroek, J.M.: 'Performance evaluation of local colour invariants', *Comput. Vis. Image Underst.*, 2009, **113**, pp. 48–62
9 Abdel-Hakim, A.E., Farag, A.A.: 'CSIFT: a SIFT descriptor with color invariant characteristics'. Proc. IEEE Computer Social Conf. on Computer Visual Pattern Recognition. 2006, vol. 2, pp. 1978–1983
10 Ke, Y., Sukthankar, R.: 'PCA-SIFT: a more distinctive representation for local image descriptors'. Proc. IEEE Computer Social Conf. on Computer Visual Pattern Recognition. 2004, vol. 2, pp. 506–513

11 Su, Z., Zhang, H., Li, S., Ma, S.: 'Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning', *IEEE Trans. Image Process.*, 2003, **12**, (8), pp. 924–937
12 Hsu, C.T., Li, C.Y.: 'Relevance feedback using generalized Bayesian framework with region-based optimization learning', *IEEE Trans. Image Process.*, 2005, **14**, (10), pp. 1617–1631
13 Tong, S., Chang, E.: 'Support vector machine active learning for image retrieval'. Proc. ACM Int. Conf. on Multimedia, 2001, pp. 107–118
14 Kherfi, M.L., Ziou, D., Bernardi, A.: 'Combining positive and negative examples in relevance feedback for content-based image retrieval', *J. Vis. Commun. Image Represent.*, 2003, **14**, pp. 428–457
15 Marakakis, A., Galatsanos, N., Likas, A., Stafylopatis, A.: 'Probabilistic relevance feedback approach for content-based image retrieval based on Gaussian mixture models', *IET Image Process.*, 2009, **3**, (1), pp. 10–25
16 Guo, G.D., Jain, A.K., Ma, W.Y., Zhang, H.J.: 'Learning similarity measure for natural image retrieval with relevance feedback', *IEEE Trans. Neural Netw.*, 2002, **13**, (4), pp. 811–820
17 Jing, F., Li, M., Zhang, H.J., Zhang, B.: 'An efficient and effective region-based image retrieval framework', *IEEE Trans. Image Process.*, 2004, **13**, (5), pp. 699–709
18 Jiang, W., Er, G., Dai, Q., Gu, J.: 'Similarity-based online feature selection in content-based image retrieval', *IEEE Trans. Image Process.*, 2006, **15**, (3), pp. 702–712
19 Vasconcelos, N., Ho, P., Moreno, P.: 'The Kullback–Leibler kernel as a framework for discriminant and localized representations for visual recognition'. European Conf. on Computer Vision, 2004, pp. 430–441
20 Farquhar, J., Szedmak, S., Meng, H., Shawe-Taylor, J.: 'Improving 'bag-of-keypoints' image categorisation'. Technical report, University of Southampton, 2005
21 Liu, Y., Perronnin, F.: 'A similarity measure between unordered vector sets with application to image categorization'. IEEE Conf. on Computer Vision Pattern Recognition, 2008, pp. 1–8
22 Ishikawa, Y., Subramanya, R., Faloutsos, C.: 'MindReader: querying databases through multiple examples'. Proc. Conf. on Very Large Data Bases, 1998
23 Qian, F., Li, M., Zhang, L., Zhang, H.J., Zhang, B.: 'Gaussian mixture model for relevance feedback in image retrieval'. Proc. IEEE ICME, August 2002
24 Kherfi, M.L., Ziou, D.: 'Relevance feedback for CBIR: a new approach based on probabilistic feature weighting with positive and negative examples', *IEEE Trans. Image Process.*, 2006, **15**, (4), pp. 1017–1030
25 Vasconcelos, N., Lippman, A.: 'Learning over multiple temporal scales in image databases'. European Conf. on Computer Vision, Dublin, Ireland, 2000
26 Vasconcelos, N., Lippman, A.: 'Learning from user feedback in image retrieval systems'. Proc. Advances in Neural Information Processing Systems (NIPS'99), 1999, pp. 977–986
27 Carson, C., Belongie, S., Greenspan, H., Malik, J.: 'Blobworld: Image segmentation using expectation-maximization and its application to image querying', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, **24**, (8), pp. 1026–1038
28 Bishop, C.M.: 'Neural networks for pattern recognition' (Oxford University Press Inc., New York, 1995)
29 Mclachlan, G.M., Peel, D.: 'Finite mixture models' (John Wiley & Sons Inc., New York, 2001)
30 Goldberger, J., Gordon, S., Greenspan, H.: 'Unsupervised image-set clustering using an information theoretic framework', *IEEE Trans. Image Process.*, 2006, **15**, (2), pp. 449–458
31 Perronnin, F., Dance, C., Csurka, G., Bressan, M.: 'Adapted vocabularies for generic visual categorization'. Proc. European Conf. on Computer Vision, Springer Verlag, 2006
32 Reynolds, D., Quatieri, T., Dunn, R.: 'Speaker verification using adapted Gaussian mixture models', *Digit. Signal Process.*, 2000, **10**, (1–3), pp. 19–41
33 Greenspan, H., Dvir, G., Rubner, Y.: 'Context-dependent segmentation and matching in image databases', *Comput. Vis. Image Underst.*, 2004, **93**, pp. 86–109
34 Vasconcelos, N.: 'On the efficient evaluation of probabilistic similarity functions for image retrieval', *IEEE Trans. Inf. Theory*, 2004, **50**, (7), pp. 1482–1496
35 Goldberger, J., Gordon, S., Greenspan, H.: 'An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures'. Int. Conf. Computer Vision, 2003, vol. 1, pp. 487–493
36 Gauvain, J.-L., Lee, C.-H.: 'Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains', *IEEE Trans. Speech Audio Process.*, 1994, **2**, (2), pp. 291–298
37 Mikolajczyk, K., Schmid, C.: 'A performance evaluation of local descriptors', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (10), pp. 1615–1630
38 Bishop, C.M.: 'Pattern recognition and machine learning' (Springer, 2006)
39 Homepage of Jan-Mark Geusebroek. http://staff.science.uva.nl/~mark/downloads.html
40 Vlassis, N., Likas, A.: 'A greedy EM algorithm for Gaussian mixture learning', *Neural Process. Lett.*, 2002, **15**, pp. 77–87
41 Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: 'Local features and kernels for classification of texture and object categories: an in-depth study'. Technical report RR-5753, INRIA, 2005
42 LIBSVM – A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/
43 Microsoft Research Cambridge Object Recognition Image Database, version 1.0. http://research.microsoft.com/research/downloads/Details/b94de342-60dc-45d0-830b-9f6eff91b301/Details.aspx

doi: 10.1049/iet-ipr.2009.0402