Logistic Regression Models for a Fast CBIR Method Based on Feature Selection

R. Ksantini^1**, D. Ziou**^{1}**, B. Colin**^{2}**, and F. Dubeau**^{2}
University of Sherbrooke
(1) Computer Science Department
{riadh.ksantini, djemel.ziou}@usherbrooke.ca
(2) Mathematic Department
{bernard.colin, francois.dubeau}@usherbrooke.ca

Abstract

Distance measures like the Euclidean distance have been the most widely used to measure similarities between feature vectors in the content-based image retrieval (CBIR) systems. However, in these similarity measures no assumption is made about the probability distributions and the local relevances of the feature vectors. Therefore, irrelevant features might hurt retrieval performance. Probabilistic approaches have proven to be an effective solution to this CBIR problem. In this paper, we use a Bayesian logistic regression model, in order to compute the weights of a pseudo-metric to improve its discriminatory capacity and then to increase image retrieval accuracy. The pseudo-metric weights were adjusted by the classical logistic regression model in [Ksantini *et al.*, 2006]. The Bayesian logistic regression model was shown to be a significantly better tool than the classical logistic regression one to improve the retrieval performance. The retrieval method is fast and is based on feature selection. Experimental results are reported on the Zubud and WANG color image databases proposed by [Deselaers *et al.*, 2004].

1 Introduction

The rapid expansion of the Internet and the wide use of digital data in many real world applications in the field of medecine, security, communications, commerce and academia, increased the need for both efficient image database creation and retrieval procedures. For this reason, content-based image retrieval (CBIR) approach was proposed. In this approach, each image from the database is associated with a feature vector capturing certain visual features of the image such as color, texture and shape. Then, a similarity measure is used to compare these feature vectors and to find similarities between images with the assumption that images that are close to each other in the feature space are also visually similar. Distance measures like the Euclidean distance have been the most widely used for feature vector comparison in the CBIR systems. However, these similarity measures are only based on the distances between feature vectors in the feature space. Therefore, because of the lack of information about the relative relevances of the featurebase feature vectors and because of the noise in these vectors, distance measures can fail and irrelevant features might hurt retrieval performance. Probabilistic approaches are a promising solution to this CBIR problem, that when compared to the standard CBIR methods based on the distance measures, can lead to a significant gain in retrieval accuracy. In fact, these approaches are capable of generating probabilistic similarity measures and highly customized metrics for computing image similarity based on the consideration and distinction of the relative feature vector relevances. As to previous works based on these probabilistic approaches, [Peng *et al.*, 2004] used a binary classification to classify the database color image feature vectors as relevant or irrelevant, [Caenen and Pauwels, 2002] used the classical quadratic logistic regression model, in order to classify database image feature vectors as relevant or irrelevant, [Aksoy *et al.*, 2000] used weighted L_1 and L_2 distances, in order to measure the similarity degree between two images and [Aksoy and Haralick, 2001] measure the similarity degree between a query image and a database image using a likelihood ratio derived from a Bayesian classifier.

In this paper, we investigate the effectiveness of a Bayesian logistic regression model based on a variational method, in order to adjust the weights of a pseudo-metric used in [Ksantini *et al.*, 2006], and then to improve its discriminatory capacity and to increase image retrieval accuracy. This pseudo-metric makes use of the compressed and quantized versions of the Daubechies-8 wavelet decomposed feature vectors, and its weights were adjusted by the classical logistic regression. We will show that thanks to the variational method, the used Bayesian logistic regression model is a significantly better tool than the classical logistic regression model to compute the pseudo-metric weights and to improve the querying results. The retrieval method is fast, efficient and based on feature selection. The evaluation of the retrieval method using both models, separately, is performed using precision and scope curves as defined in [Kherfi and Ziou, 2006].

In the next section, we briefly define the pseudo-metric. In section 3, we briefly describe the pseudo-metric weight computation using the classical logistic regression model, while showing the limitations of this latter and that the Bayesian logistic regression model is more appropriate for the pseudo-metric weight computation. Then, we detail the Bayesian logistic regression model. Moreover, we will describe the data training performed for both models. The feature selection based image retrieval method and the feature vectors used to represent the database images are presented in section 4. Finally, in section 5, we will perform some experiments to validate the Bayesian logistic regression model and we will use the precision and scope, in order to show the advantage of the Bayesian logistic regression model over the classical logistic regression one, in terms of querying results.

2 The pseudo-metric

Given a query feature vector Q and a featurebase of $|DB|$ feature vectors T_k ($k = 1, ..., |DB|$) having 2^J components each, our aim is to retrieve in the featurebase the most similar feature vectors to Q. To achieve this, Q and the $|DB|$ feature vectors are Daubechies-8 wavelets decomposed, compressed to m coefficients each and quantized. Then, to measure the similarity degree between Q and a target feature vector T_k of the featurebase, we use the one-dimensional version of the pseudo-metric used in [Ksantini *et al.*, 2006] and given by the following expression

$$\| Q, T_k \| = \tilde{w}_0 |\tilde{Q}[0] - \tilde{T}_k[0]| - \sum_{i:\tilde{Q}^c_q[i] \neq 0} w_{bin(i)} (\tilde{Q}^c_q[i] = \tilde{T}^c_{kq}[i]), \tag{1}$$

where

$$\left(\tilde{Q}^c_q[i] = \tilde{T}^c_{kq}[i]\right) = \begin{cases} 1 & \text{if } \tilde{Q}^c_q[i] = \tilde{T}^c_{kq}[i] \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

$\tilde{Q}[0]$ and $\tilde{T}_k[0]$ are the scaling factors of Q and T_k, $\tilde{Q}^c_q[i]$ and $\tilde{T}^c_{kq}[i]$ represent the i-th coefficients of their Daubechies-8 wavelets decomposed, compressed to m coefficients and quantized versions, $\tilde{w}_0$ and the $w_{bin(i)}$'s are the weights to compute, and the bucketing function $bin()$ groups these latters according to the J resolution levels, such as

$$bin(i) = \lfloor log_2(i) \rfloor \qquad \text{with} \qquad i = 1, ..., 2^J - 1. \tag{3}$$

3 The weight computation

In order to improve the discriminatory power of the pseudo-metric, we compute its weights $\tilde{w}_0$ and $\{w_k\}_{k=0}^{J-1}$ using a classical logistic regression model and a Bayesian logistic regression model, separately. We define two classes, the relevance class denoted by Ω_0 and the irrelevance class denoted by Ω_1, in order to classify the feature vector pairs as similar or dissimilar. The basic principle of using the Bayesian logistic regression model and the classical logistic regression one is to allow a good linear separation between Ω_0 and Ω_1, and then to compute the weights which represent the local relevances of the pseudo-metric components.

3.1 The classical logistic regression model

In this model, each feature vector pair is represented by an explanatory vector and a binary target variable. Specifically, for the i-th feature vector pair, we associate an explanatory vector $X_i = (\tilde{X}_{0,i}, X_{0,i}, ..., X_{J-1,i}, 1) \in \mathbb{R}^J \times \{1\}$ and a binary target S_i which is either 0 or 1, depending on whether or not the two feature vectors are intended to be similar. $\tilde{X}_{0,i}$ is the absolute value of the difference between the scaling factors of the Daubechies-8 wavelets decomposed, compressed and quantized versions of the two feature vectors and $\{X_{k,i}\}_{k=0}^{J-1}$ are the numbers of mismatches between the J resolution level coefficients of these latter. We suppose that we have n_0 pairs of similar feature vectors and n_1 pairs of dissimilar ones. Thus, the class Ω_0 contains n_0 explanatory vectors and their associated binary target variables $\{X^r_i, S^r_i = 0\}_{i=1}^{n_0}$ to represent the pairs of the similar feature vectors, and the class Ω_1 contains n_1 explanatory vectors and their associated binary target variables $\{X^{ir}_j, S^{ir}_j = 1\}_{j=1}^{n_1}$ to represent the pairs of the dissimilar feature vectors. The pseudo-metric weights $\tilde{w}_0$ and $\{w_k\}_{k=0}^{J-1}$ and an intercept v are chosen to optimize the following conditional log-likelihood.

$$L(\tilde{w}_0, w_0, ..., w_{J-1}, v) = \sum_{i=1}^{n_0} log(p^r_i) + \sum_{j=1}^{n_1} log(p^{ir}_j), \tag{4}$$

where p^r_i and p^{ir}_j are the relevance and irrelevance probabilities, respectively, and given by

$$\begin{aligned} p^r_i &= F(-\tilde{w}_0 \tilde{X}^r_{0,i} - \sum_{k=0}^{J-1} w_k X^r_{k,i} - v), \\ p^{ir}_j &= F(\tilde{w}_0 \tilde{X}^{ir}_{0,j} + \sum_{k=0}^{J-1} w_k X^{ir}_{k,j} + v), \end{aligned}$$

where $F(x) = \frac{e^x}{1+e^x}$ is the logistic function. For this reason, standard optimization algorithms such as Fisher scoring and gradient ascent algorithms [Clogg *et al.*, 1991], can be invoked. However, in several cases, especially because of the exponential in the likelihood function or because of the existence of many zero explanatory vectors, the maximum likelihood can fail and estimates of the parameters of interest (weights and intercept) may not be optimal or may not exist or may be on the boundary of the parameter space. Also, as there is complete or quasicomplete separation between Ω_0 and Ω_1, the function L is made arbitrarily large and standard optimization algorithms diverge [Krishnapuram *et al.*, 2005]. Moreover, as Ω_0 and Ω_1 are large and high-dimensional, these standard optimization algorithms have high computational complexity and take long time to converge. The first two problems can be solved by smoothing the parameter of interest estimates, assuming a certain prior distribution for the parameters, thereby reducing the parameter space, and the third problem can be solved by using variational transformations which simplify the computation of the parameter of interest estimates [Jaakkola and Jordan, 2000]. This motivates the adoption of a Bayesian logistic regression model based on variational methods.

3.2 The Bayesian logistic regression model

In the Bayesian logistic regression framework, there are three main components which are a chosen prior distribution over the parameters of interest, the likelihood function and the posterior distribution. These three components are formally combined by Bayes' rule. The posterior distribution contains all the available knowledge about the parameters of interest in the model. Among many priors having different distributional forms, gaussian prior has the advantage of having low computational intensity and of smoothing the parameter estimates toward a fixed mean and away from unreasonable extremes. However, when the likelihood function is not conjugate of the gaussian prior, the posterior distribution has no tractable form and its mean computation involves high-dimensional integration which has high computational cost. According to [Jaakkola and Jordan, 2000], it's possible to use accurate variational transformations in order to approximate the likelihood function with a simpler tractable exponential form. In this case, thanks to the conjugacy, with a gaussian prior distribution over the parameters of interest combined with the likelihood approximation, we obtain a closed gaussian form approximation to the posterior distribution. However, as the number of observations is large, the number of variational parameters updated to optimize the posterior distribution approximation is also large, thereby the computational cost is high. In the Bayesian logistic regression model that we propose, we use variational transformations and the Jensen's inequality in order to approximate the likelihood function with tractable exponential form. The explanatory vectors are not observed but instead are distributed according to two specific distributions. The posterior distribution is also approximated with a gaussian which depends only on two variational parameters. The computation of the posterior distribution approximation mean is fast and has low computational complexity. In this model, we denote the random vectors whose realizations represent the explanatory vectors $\{X_i^r\}_{i=1}^{n_0}$ of the relevance class Ω_0 and the explanatory vectors $\{X_j^{ir}\}_{j=1}^{n_1}$ of the irrelevance class Ω_1, by $\underline{\text{X}}_0 = (\underline{\tilde{\text{X}}}_{0,0}, \underline{\text{X}}_{0,0}, ..., \underline{\text{X}}_{J-1,0}, 1)$ and $\underline{\text{X}}_1 = (\underline{\tilde{\text{X}}}_{0,1}, \underline{\text{X}}_{0,1}, ..., \underline{\text{X}}_{J-1,1}, 1)$, respectively. We suppose that $\underline{\text{X}}_0 \sim q_0(\underline{\text{X}}_0)$ and $\underline{\text{X}}_1 \sim q_1(\underline{\text{X}}_1)$, where q_0 and q_1 are two chosen distributions. For $\underline{\text{X}}_0$ we associate a binary random variable $\underline{\text{S}}_0$ whose realizations are the target variables $\{S_i^r = 0\}_{i=1}^{n_0}$, and for $\underline{\text{X}}_1$ we associate a binary random variable $\underline{\text{S}}_1$ whose realizations are the target variables $\{S_j^{ir} = 1\}_{j=1}^{n_1}$. We set $\underline{\text{S}}_0$ equal to 0 for similarity and we set $\underline{\text{S}}_1$ equal to 1 for dissimilarity. Parameters of interest (weights and intercept) are considered as random variables and are denoted by the random vector $\underline{\text{W}} = (\underline{\tilde{w}}_0, \underline{w}_0, ..., \underline{w}_{J-1}, \underline{v})$. We assume that $\underline{\text{W}} \sim \pi(\underline{\text{W}})$, where π is a gaussian prior with prior mean μ and covariance matrix Σ. Using Bayes' rule, the posterior distribution over $\underline{\text{W}}$ is given by

$$P(\underline{\text{W}}|\underline{\text{S}}_0 = 0, \underline{\text{S}}_1 = 1) =$$

$$\frac{\left[\sum_{x_0 \in \Omega_0, x_1 \in \Omega_1} \prod_{i=0}^{1} P(\underline{\text{S}}_i = i|\underline{\text{X}}_i = x_i, \underline{\text{W}}) q_i(\underline{\text{X}}_i = x_i)\right] \pi(\underline{\text{W}})}{P(\underline{\text{S}}_0 = 0, \underline{\text{S}}_1 = 1)},$$

where $P(\underline{\text{S}}_i = i|\underline{\text{X}}_i = x_i, \underline{\text{W}}) = F((2i-1)\underline{\text{W}}^t x_i)$ for each $i \in \{0, 1\}$. Using a variational approximation [Jaakkola and Jordan, 2000] and the Jensen's inequality, the posterior distribution is approximated as follows

$$P(\underline{\text{W}}|\underline{\text{S}}_0 = 0, \underline{\text{S}}_1 = 1)$$

$$\geq \frac{\underline{\text{P}}\big(\underline{\text{W}}|\underline{\text{S}}_0 = 0, \underline{\text{S}}_1 = 1, \{\epsilon_i\}_{i=0}^{1}, \{q_i\}_{i=0}^{1}\big)\pi(\underline{\text{W}})}{P(\underline{\text{S}}_0 = 0, \underline{\text{S}}_1 = 1)},$$

$$\propto \underline{\text{P}}\big(\underline{\text{W}}|\underline{\text{S}}_0 = 0, \underline{\text{S}}_1 = 1, \{\epsilon_i\}_{i=0}^{1}, \{q_i\}_{i=0}^{1}\big)\pi(\underline{\text{W}})$$

where

$$\underline{\text{P}}\big(\underline{\text{W}}|\underline{\text{S}}_0 = 0, \underline{\text{S}}_1 = 1, \{\epsilon_i\}_{i=0}^{1}, \{q_i\}_{i=0}^{1}\big) =$$

$$\left[\prod_{i=0}^{1} F(\epsilon_i)\right] e^{\left[\sum_{i=0}^{1}\left[\frac{E_{q_i}[H_i] - \epsilon_i}{2}\right] - \sum_{i=0}^{1}\left[\varphi(\epsilon_i)\left(E_{q_i}[H_i^2] - \epsilon_i^2\right)\right]\right]},$$

where E_{q_0} and E_{q_1} are the expectations with respect to the distributions q_0 and q_1, respectively, $\varphi(\epsilon_i) = \frac{tanh(\frac{\epsilon_i}{2})}{4\epsilon_i}$ and $\{\epsilon_i\}_{i=0}^{1}$ are the variational parameters. Therefore, the approximation of the posterior distribution is considered as an adjustable lower bound and as a proper Gaussian distribution with a posterior mean μ_{post} and covariance matrix Σ_{post} which are estimated by the following Bayesian update equations

$$(\Sigma_{post})^{-1} = (\Sigma)^{-1} + 2\sum_{i=0}^{1}\left[\varphi(\epsilon_i) E_{q_i}[x_i(x_i)^t]\right], \quad (5)$$

$$\mu_{post} = \Sigma_{post}\left[(\Sigma)^{-1}\mu + \sum_{i=0}^{1}\left[(i - \frac{1}{2}) E_{q_i}[x_i]\right]\right]. \quad (6)$$

The weight and intercept computation algorithm is in two phases. The first phase is the initialization of q_0, q_1 and the gaussian prior $\pi(\underline{\text{W}})$, and the second phase is iterative and allows the computation of Σ_{post} and μ_{post} through the Bayesian update equations (5) and (6), respectively, while using an EM type algorithm [Jaakkola and Jordan, 2000], in order to find the variational parameters $\{\epsilon_i\}_{i=0}^{1}$ at each iteration to have an optimal approximation to the posterior distribution. In the initialization phase, q_0 and q_1 are chosen to model Ω_0 and Ω_1, respectively, and because of the absence of prior knowledge about the weights and the intercept, $\pi(\underline{\text{W}})$ is chosen univariate with zero mean and large variances [Congdon, 2001]. The values of μ_{post} components are the desired estimates of the pseudo-metric weights $\tilde{w}_0$ and $\{w_k\}_{k=0}^{J-1}$ and the intercept v. Once the parameters of the posterior distribution approximation are computed, its magnitude is given by the term $\prod_{i=0}^{1} F(\epsilon_i)$. This latter becomes very close to 1 as Ω_0 and Ω_1 are linearly separated or quasi separated and tends towards 0 as Ω_0 and Ω_1 become more and more overlapped. Analogically, in the classical logistic regression model, the term e^{2L} has almost the same characteristics as $\prod_{i=0}^{1} F(\epsilon_i)$ [Caenen and Pauwels, 2002]. These two terms will be used to perform feature selection in the retrieval method.

3.3 Training

Let us consider a color image database which consists of several color image sets such that each set contains color images which are perceptually close to each other in terms of object shapes and colors. In order to compute the pseudo-metric weights and the intercept by the classical logistic regression model, we have to create the relevance class Ω_0 and the irrelevance class Ω_1. To create Ω_0, we draw all possible pairs of feature vectors representing color images belonging to the same database color image sets, and for each pair we compute an explanatory vector and we associate to this latter a binary target variable equal to 0. Similarly, to create Ω_1, we draw all possible pairs of feature vectors representing color images belonging to different database color image sets, and for each pair we compute an explanatory vector and we associate to this latter a binary target variable equal to 1. For the Bayesian logistic regression model, we create the Ω_0 and Ω_1 with the same way, but instead of associating a binary target variable value to each explanatory vector of Ω_0 and Ω_1, we associate a binary target variable $\underline{\mathrm{S}}_0$ equal to 0 to all Ω_0 explanatory vectors and we associate a binary target variable $\underline{\mathrm{S}}_1$ equal to 1 to all Ω_1 explanatory vectors.

4 Color image retrieval method

The querying method is in two phases. The first phase is a preprocessing phase done once for the entire database containing $|DB|$ color images. The second phase is the querying phase.

4.1 Color image database preprocessing

We detail the preprocessing phase done once for all the database color images before the querying in a general case by the following steps.

1. Choose N feature vectors for comparison.
2. Compute the N feature vectors T_{li} $(l \in \{1, ..., N\})$ for each i-th color image of the database, where $i \in \{1, ..., |DB|\}$.
3. The feature vectors representing the database color images are Daubechies-8 wavelets decomposed, compressed to m coefficients each and quantized.
4. Organize the decomposed, compressed and quantized feature vectors into search arrays Θ_+^l and Θ_-^l $\big(l = 1, ..., N\big)$ which are used to optimize the pseud-metric computation process [Ksantini *et al.*, 2006].
5. Adjustment of the metric weights $\tilde{w}_0^l$ and $\{w_k^l\}_{k=0}^{J-1}$ for each featurebase T_{li} $(i = 1, ..., |DB|)$ representing the database color images, where $l \in \{1, ..., N\}$.

4.2 The querying algorithm

We detail the querying algorithm in a general case by the following steps.

1. Given a query color image, we denote the feature vectors representing the query image by Q_l $\big(l = 1, ..., N\big)$.
2. The feature vectors representing the query image are Daubechies-8 wavelets decomposed, compressed to m coefficients each and quantized.
3. The similarity degrees between Q_l $\big(l = 1, ..., N\big)$ and the database color image feature vectors T_{li} $\big(l = 1, ..., N\big)$ $(i = 1, ..., |DB|)$ are represented by the arrays $Score_l$ $\big(l = 1, ..., N\big)$ such that $Score_l[i] = \parallel Q_l, T_{li} \parallel$ for each $i \in \{1, ..., |DB|\}$. These arrays are returned by the procedure Retrieval(Q_l, m, Θ_+^l, Θ_-^l) $\big(l = 1, ..., N\big)$, respectively. The procedure Retrieval is used to optimize the querying process [Ksantini *et al.*, 2006].
4. The similarity degrees between the query color image and the database color images are represented by a resulted array $TotalScore$, such as, $TotalScore[i] = \sum_{l=1}^{N} \gamma_l Score_l[i]$ for each $i \in \{1, ..., |DB|\}$, where $\{\gamma_l\}_{l=1}^{N}$ are weightfactors used to down-weight the feature which has low discriminatory power. $\gamma_l = e^{2L_l}$ when the weights are computed by the classical logistic regression model, and $\gamma_l = \prod_{i=0}^{1} F(\epsilon_i^l)$ when the weights are computed by the Bayesian logistic regression model.
5. Organize the database color images in order of increasing resulted similarity degrees of the array $TotalScore$. The most negative resulted similarity degrees correspond to the closest target images to the query image. Finally, return to the user the closest target color images to the query color image and whose number is denoted by RI and chosen by the user.

4.3 Used feature vectors

In order to describe the luminance, colors and the edges of a color image, we use luminance histogram and weighted histograms. The image texture description is performed by kurtosis and skewness histograms. Given an $M \times N$ pixel LAB color image, its luminance histogram h_L contains the number of pixels of the luminance L, and can be written as follows

$$h_L(c) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta(I_L(i,j) - c), \tag{7}$$

for each $c \in \{0, ..., 255\}$, where I_L is the luminance image and δ is the Kronecker symbol at 0. The weighted histograms are the color histogram constructed after edge region elimination and the multispectral gradient module mean histogram. The former is given by

$$h_k^h(c) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta(I_k(i,j) - c)\chi_{[0,\eta]}\Big(\lambda_{max}(i,j)\Big), \tag{8}$$

and the latter is given by

$$\bar{h}_k^e(c) = \frac{h_k^e(c)}{N_{p,k}(c)}, \tag{9}$$

where $N_{p,k}(c)$ is the number of the edge region pixels and is defined as

$$N_{p,k}(c) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta(I_k(i,j) - c)\chi_{]\eta,+\infty[}\Big(\lambda_{max}(i,j)\Big), \tag{10}$$

and

$$h_k^e(c) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta(I_k(i,j) - c)\lambda_{max}(i,j)\ \chi_{]\eta,+\infty[}\Big(\lambda_{max}(i,j)\Big), \tag{11}$$

for each $c \in \{0, ..., 255\}$ and $k = a, b$, where λ_{max} represents the multispectral gradient module [Ksantini *et al.*, 2006], η is a threshold defined by the mean of the multispectral gradient modules computed over all image pixels, I_a and I_b are the images of the chrominances a red/green and b yellow/blue, respectively, and χ is the characteristic function. The multispectral gradient module mean histogram provides information about the overall contrast in the chrominance and the edge region elimination allows the avoidance of overlappings or noises between the color histogram populations caused by the edge pixels. The LAB color image kurtosis and skewness histograms are given by

$$h_k^{\kappa}(c) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta(I_k^{\kappa}(i,j) - c), \tag{12}$$

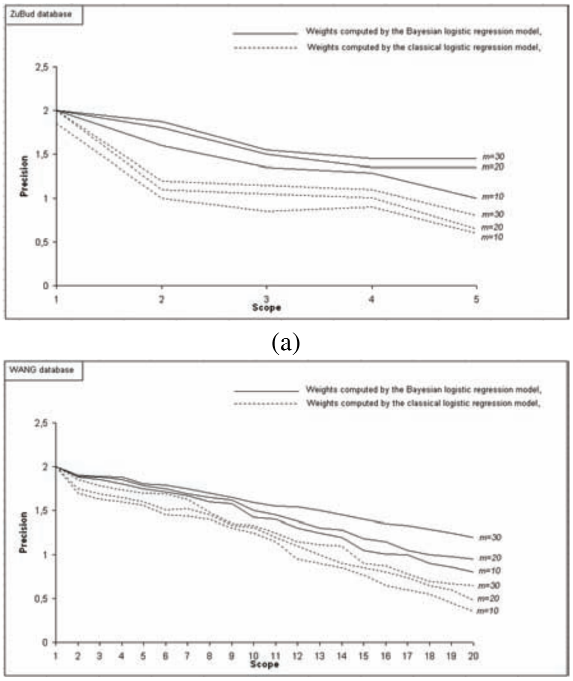
and

$$h_k^{s}(c) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta(I_k^{s}(i,j) - c), \tag{13}$$

respectively, for each $c \in \{0, ..., 255\}$ and $k = L, a, b$, where I_L^{κ}, I_a^{κ} and I_b^{κ} are the kurtosis images of the luminance L and the chrominances a and b, respectively, and I_L^s, I_a^s and I_b^s are the skewness images of these latter. They are obtained by local computations of the kurtosis and skewness values at the luminance and chrominance image pixels. Then, a linear interpolation is used to represent the kurtosis and skewness values between 0 and 255. Since each used feature vector is a histogram having 256 components, we set J equal to 8 in the following section.

5 Experimental results

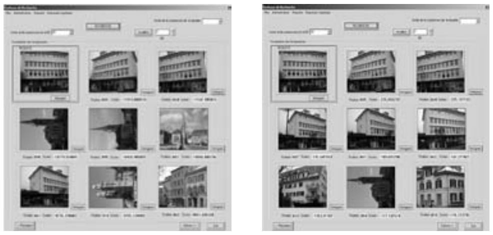
In this section, we will discuss the choices of the distributions q_0 and q_1, in order to validate the Bayesian logistic regression model in the image retrieval context. Finally, we will use the precision and scope as defined in [Kherfi and Ziou, 2006], to evaluate the querying method using both models separately. The choices of the distributions q_0 and q_1 and the querying evaluation will be conducted on the WANG and Zubud color image databases proposed by [Deselaers *et al.*, 2004]. The WANG database contains $|DB| = 1000$ color images which were selected manually to form 10 sets (e.g. Africa, beach, ruins, food) of 100 images each. The Zurich Building Image Database (ZuBuD) contains a training part of $|DB| = 1005$ color images and query part of 115 color images. The training part consists of 201 building image sets, where each set contains 5 color images of the same building taken from different positions. Before the feature vector extractions, we represent the WANG and Zubud database color images in the perceptually uniform LAB color space. Since from each color image of the Zubud and WANG databases we extract $N = 11$ histograms which are given by (7), (8), (9), (12) and (13) respectively, each database is represented by eleven featurebases. The choices of q_0 and q_1 will be separately performed for each featurebase. For each featurebase, we assume that $\tilde{\underline{X}}_{0,0}$ and $(\underline{X}_{0,0}, ..., \underline{X}_{J-1,0})$ are independent. We make the same assumption for $\tilde{\underline{X}}_{0,1}$ and $(\underline{X}_{0,1}, ..., \underline{X}_{J-1,1})$. Moreover, we suppose that the random vector $(\underline{X}_{0,0}, ..., \underline{X}_{J-1,0})$ random variables whose realizations are positive integers, are independent and each one of them follows a truncated poisson distribution at its greatest realization, to have a best fit. Analogically, we make the same choice for $(\underline{X}_{0,1}, ..., \underline{X}_{J-1,1})$. Also, we assume that the random variable $\tilde{\underline{X}}_{0,0}$ whose realizations are positive reals, follows a gaussian mixture distribution, which is the same choice for $\tilde{\underline{X}}_{0,1}$. Generally, to carry out an evaluation in the image retrieval field, two principal issues are required: the acquisition of ground truth and the definition of performance criteria. For ground truth, we use human observations. In fact, three external persons participate in the below evaluation. Concerning performance criteria, we represent the evaluation results by the precision-scope curve $Pr = f(RI)$, where the scope RI is the number of images returned to the user. In each querying performed in the evaluation experiment, each human subject is asked to give a goodness score to each retrieved image. The goodness score is 2 if the retrieved image is almost similar to the query, 1 if the retrieved image is fairly similar to the query and 0 if there is no similarity between the retrieved image and the query. The precision is computed as follows: $Pr =$ the sum of goodness scores for retrieved images$/RI$. Therefore, the curve $Pr = f(RI)$ gives the precision for different values of RI which lie between 1 and 20 when we perform the querying evaluation on the WANG database, and lie between 1 and 5 when we perform the querying evaluation on the ZuBuD database. When the human subjects perform different queryings in the evaluation experiment, we compute an average precision for each value of RI, and then we construct the precision-scope curve. In our evaluation experiment, each color image of the WANG and Zubud databases is represented by $N = 11$ histograms which are h_L, h_a^h, h_b^h, $\bar{h}_a^e$, $\bar{h}_b^e$, h_L^{κ}, h_a^{κ}, h_b^{κ}, h_L^s, h_a^s and h_b^s. In order to evaluate the querying in the WANG database, each human subject is asked to formulate a query from the database and to execute a querying, using weights computed by the classical logistic regression model, and to give a goodness score to each retrieved image, then to reformulate a query from the database and to execute the querying, using weights computed by the Bayesian logistic regression model, and to give a goodness score to each retrieved image. Each human subject performs the querying fifty times by choosing a new query from the database each time. We repeat this experience for different orders of compression $m \in \{30, 20, 10\}$. To evaluate the querying in the ZuBuD database, each human subject is asked to follow the preceding steps, while formulating the queries from the database query part. For the WANG and Zubud databases, the resulted precision-scope curves are given in Figure 1 for compression orders $m \in \{30, 20, 10\}$. The Figure 2 illustrates two retrieval examples in the Zubud database comparing the performances of the regression models for $m = 30$. In each example the query is located at the top-left of the dialog box.



(a)

(b)

Figure 1: Evaluation ((a) ZuBud database and (b) WANG database): precision-scope curves for retrieval using weights computed by the classical logistic regression model and weights computed by the Bayesian logistic regression model.



(a) (b)

Figure 2: Comparison (ZuBud database): a) first 8 color images retrieved using weights computed by the classical logistic regression model, b) first 8 color images retrieved using weights computed by the Bayesian logistic regression model.

6 Conclusion

We presented a simple, fast and effective color image querying method based on feature selection. In order to measure the similarity degree between two color images both quickly and effectively, we used a weighted pseudo-metric which makes use of the one-dimensional Daubechies decomposition and compression of the extracted feature vectors. A Bayesian logistic regression model and a classical logistic regression one were used to improve the discriminatory capacity of the pseudo-metric and to allow feature selection. Evaluations of the querying method showed that the Bayesian logistic regression model is a better tool than the classical logistic regression one to compute the pseudo-metric weights and to improve the querying results.

References

[Aksoy and Haralick, 2001] S. Aksoy and R. M. Haralick. Feature Normalization and Likelihood-based Similarity Measures for Image Retrieval. *Pattern Recognition Letters*, 22(5):563-582, 2001.

[Aksoy *et al.*, 2000] S. Aksoy, R. M. Haralick, F. A. Cheikh, and M. Gabbouj. A Weighted Distance Approach to Relevance Feedback. In *15th International Conference on Pattern Recognition*, page 4812, Barcelona, Spain, 2000.

[Caenen and Pauwels, 2002] G. Caenen and E. J. Pauwels. Logistic Regression Models for Relevance Feedback in Content-Based Image Retrieval. In *Storage and Retrieval for Media Databases 2002, Proceedings of SPIE*, pages 49-58, San Jose, California, USA, 2002.

[Clogg *et al.*, 1991] C. C. Clogg, D. B. Rubin, N. Schenker, B. Schultz, and L. Widman. Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression. *Journal of the American Statistical Association*, 86:68-78, 1991.

[Congdon, 2001] P. Congdon. *Bayesian Statistical Modelling*. John Wiley, Chichester, UK, 2001.

[Deselaers *et al.*, 2004] T. Deselaers, D. Keysers, and H. Ney. Classification Error Rate for Quantitative Evaluation of Content-based Image Retrieval Systems. In *17th International Conference on Pattern Recognition*, pages 505-508, Cambridge, UK, 2004.

[Jaakkola and Jordan, 2000] T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25-37, 2000.

[Kherfi and Ziou, 2006] M. L. Kherfi and D. Ziou. Relevance feedback for CBIR: a new approach based on probabilistic feature weighting with positive and negative examples. *IEEE Transactions on Image Processing*, 15(4):1017-1030, 2006.

[Krishnapuram *et al.*, 2005] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957-968, 2005.

[Ksantini *et al.*, 2006] R. Ksantini, D. Ziou, and F. Dubeau. Image Retrieval Based on Region Separation and Multiresolution Analysis. *International Journal of Wavelets, Multiresolution and Information Processing*, 4(1):147-175, 2006.

[Peng *et al.*, 2004] J. Peng, B. Bhanu, and S. Qing. Learning Feature Relevance and Similarity Metrics in Image Databases. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 14-18, Santa Barbara, California, USA, 2004.