Unsupervised Learning of a Finite Mixture Model Based on the Dirichlet Distribution and Its Application

Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt

***Abstract*—This paper presents an unsupervised algorithm for learning a finite mixture model from multivariate data. This mixture model is based on the Dirichlet distribution, which offers high flexibility for modeling data. The proposed approach for estimating the parameters of a Dirichlet mixture is based on the maximum likelihood (ML) and Fisher scoring methods. Experimental results are presented for the following applications: estimation of artificial histograms, summarization of image databases for efficient retrieval, and human skin color modeling and its application to skin detection in multimedia databases.**

***Index Terms*—Dirichlet distribution, Fisher's scoring method, image summarizing, maximum likelihood, mixture modeling, natural gradient, Riemannian space.**

I. Introduction

SCIENTIFIC pursuits and human activity in general generate data. These data may be incomplete, redundant, or erroneous. Probabilistic methods are particularly useful in understanding the patterns present in such data. One such method is the Bayesian approach, which can be roughly described as estimating the uncertainty of a model [1]. In fact, by the Bayesian approach, we can estimate the uncertainty of a model's fit and the uncertainty of the estimated parameters themselves. The Bayesian approach can be employed with finite mixture models, which constitute a powerful probabilistic modeling tool for univariate and multivariate data [2]. Finite mixture models have been used extensively to model a wide variety of important practical situations where data can be viewed as arising from several populations mixed in varying proportions. Nowadays, this kind of statistical model is used in a variety of domains. In computer vision applications, for example, we can use mixture models to organize image collections as well as for color image segmentation, restoration and texture processing, and content-based image retrieval [3]. The problem of estimating the parameters of the components of a mixture has been the subject of diverse studies [4]. The isotropic nature of Gaussian functions, along with their capability for representing the distribution compactly by a mean vector and covariance matrix, have made Gaussian mixture (GM) decomposition a popular technique [5]. The Gaussian mixture is not the best choice in all applications, however, and it will fail to discover *true* structure where the partitions are clearly non-Gaussian [6]. In fact, due to its definition, the Gaussian cannot approximate asymmetric distributions well [7]. Moreover, this distribution is defined on $\mathbb{R}$ and, thus, does not have a compact support, which is why parameters estimated by the moment method are not accurate in many cases. Indeed, having a compact support is an interesting property for a given density because of the nature of data in general. Generally, we estimate data which are compactly supported, such as data originating from videos, images, or text.

In this paper, we will show that the Dirichlet distribution can be a very good choice for modeling data. The Dirichlet distribution is the multivariate generalization of the Beta distribution, which offers considerable flexibility and ease of use. In contrast with other distributions, the Dirichlet distribution permits multiple symmetric and asymmetric modes [8]. In fact, the Dirichlet distribution may be skewed to the right, skewed to the left or symmetric (see Fig. 1). However, it has a negative covariance stucture [9].

For all these reasons, we are interested in the Dirichlet distribution. In contrast to the vast amount of theoretical work that exists on the Dirichlet distribution, however, very little work has been done on its practical applications, such as parameter estimation. The majority of the studies either consider a single distribution [10], [11] or are restricted to the two-parameter Beta distribution [12], [13]. This neglect may be due to the fact that this distribution is unfamiliar to many scientists. In this paper, we will propose an algorithm to estimate the parameters of a Dirichlet mixture. This mixture decomposition algorithm incorporates a penalty term in the objective function to find the number of components required to model the data. Our method is tested for different applications, such as artificial histogram estimation, summarization of image databases for efficient retrieval, and human skin detection in multimedia databases.

The paper is organized as follows. Section II describes the Dirichlet mixture in detail. In Section III, we propose a method for estimating the parameters of a Dirichlet mixture. In Section IV, we present a way of initializing the parameters and give the complete estimation algorithm. Section V is devoted to experimental results. We end the paper with some concluding remarks.

Manuscript received October 1, 2002; revised January 2, 2004. This work was supported by Bell Canada's Bell University Laboratories R&D program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christine Guillemot.

N. Bouguila and D. Ziou are with the Département d'Informatique, Université de Sherbrooke, Sherbrooke, QC J1K 2R1 Canada (e-mail: nizar.bouguila@usherbrooke.ca; djemel.ziou@usherbrooke.ca).

J. Vaillancourt is with the Université du Québec en Outaouais, Hull, QC J8X 3X7 Canada (e-mail: jean.vaillancourt@uqo.ca).

Digital Object Identifier 10.1109/TIP.2004.834664

1057-7149/04$20.00 © 2004 IEEE

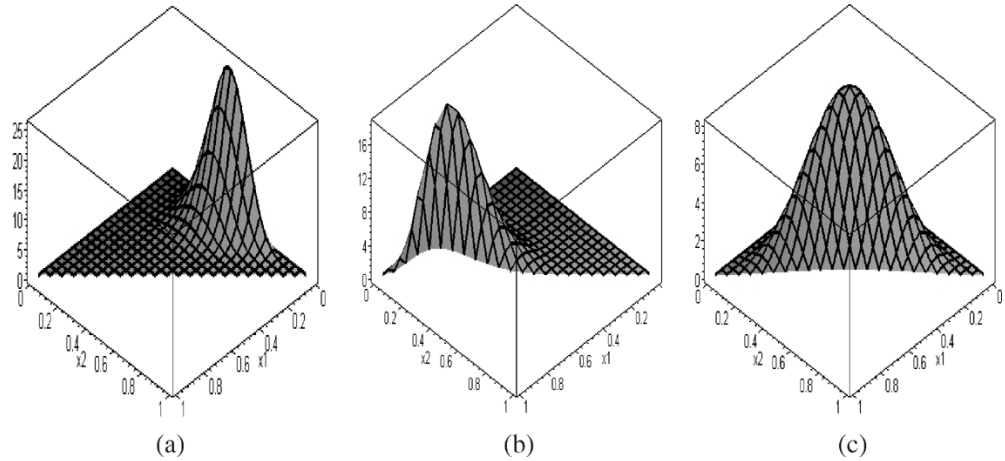


Fig. 1. Dirichlet distribution for different parameters. (a) $\alpha_1 = 8.5, \alpha_2 = 7.5, \alpha_3 = 1.5$. (b) $\alpha_1 = 10.5, \alpha_2 = 3.5, \alpha_3 = 3.5$. (c) $\alpha_1 = 3.5, \alpha_2 = 3.5, \alpha_3 = 3.5$.

II. Dirichlet Mixture

If the random vector $\vec{X} = (X_1, \ldots, X_{\dim})$ follows a Dirichlet distribution [14], [15] the joint density function is given by

$$p(X_1, \ldots, X_{\dim}) = \frac{\Gamma(|\vec{\alpha}|)}{\prod_{i=1}^{\dim+1} \Gamma(\alpha_i)} \prod_{i=1}^{\dim+1} X_i^{\alpha_i - 1} \quad (1)$$

where

$$\sum_{i=1}^{\dim} X_i < 1 \quad (2)$$

$$|\vec{X}| = \sum_{i=1}^{\dim} X_i, \quad 0 < X_i < 1 \quad \forall i = 1 \ldots \dim \quad (3)$$

$$X_{\dim+1} = 1 - |\vec{X}| \quad (4)$$

$$|\vec{\alpha}| = \sum_{i=1}^{\dim+1} \alpha_i, \quad \alpha_i > 0 \quad \forall i = 1 \ldots \dim + 1. \quad (5)$$

This distribution is the multivariate extension of the two-parameter Beta distribution. The mean and the variance of the Dirichlet distribution are given by

$$E(X_i) = \frac{\alpha_i}{|\vec{\alpha}|} \quad (6)$$

$$\mathrm{Var}(X_i) = \frac{\alpha_i (|\vec{\alpha}| - \alpha_i)}{|\vec{\alpha}|^2 (|\vec{\alpha}| + 1)} \quad (7)$$

and the variance between X_i and X_j is

$$\mathrm{Cov}(X_i, X_j) = \frac{-\alpha_i \alpha_j}{|\vec{\alpha}|^2 (|\vec{\alpha}| + 1)}. \quad (8)$$

Note that all the pairwise correlations are negative. The Dirichlet distribution with parameter vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{\dim+1})$ can be represented either as a distribution on the hyperplane $B_{\dim+1} = \{(X_1, \ldots, X_{\dim+1}), \sum_{i=1}^{\dim+1} X_i = 1\}$ in $\mathbb{R}_+^{\dim+1}$, or as a distribution inside the simplex $A_{\dim} = \{(X_1, \ldots, X_{\dim}), \sum_{i=1}^{\dim} X_i < 1\}$ in $\mathbb{R}_+^{\dim}$.

A Dirichlet mixture with M components is defined as

$$p\left(\vec{X}|\Theta\right) = \sum_{j=1}^{M} p\left(\vec{X}|j, \Theta_j\right) P(j) \quad (9)$$

where $P(j)$ ($0 < P(j) < 1$ and $\sum_{j=1}^{M} P(j) = 1$) are the mixing proportions and $p(\vec{X}|j, \Theta_j)$ is the Dirichlet distribution. The symbol Θ refers to the entire set of parameters to be estimated

$$\Theta = (\vec{\alpha_1}, \ldots, \vec{\alpha_M}, P(1), \ldots, P(M))$$

where $\vec{\alpha_j}$ is the parameter vector for the j^{th} population. In the following developments, we use the notation $\Theta_j = (\vec{\alpha_j})$ for $\mathrm{j} = 1 \ldots \mathrm{M}$.

III. Maximum Likelihood Estimation

The problem of estimating the parameters which determine a mixture has been the subject of diverse studies [16]. During the last two decades, the method of maximum likelihood (ML) [17]–[19] has become the most common approach to this problem. Of the variety of iterative methods which have been suggested as alternatives to optimize the parameters of a mixture, the one most widely used is expectation maximization (EM). EM was originally proposed by Dempster *et al.* [20] for estimating the maximum likelihood estimator (MLE) of stochastic models. This algorithm gives us an iterative procedure and the practical form is usually simple. The EM algorithm can be viewed as an approximation of the Fisher scoring method [21]. This algorithm suffers from the following drawback: the need to specify the number of components each time. In order to overcome this problem, criterion functions have been proposed, such as the Akaike information criterion (AIC) [22], minimum description length (MDL) [23], and Schwartz's Bayesian inference criterion (BIC) [24]. A maximum likelihood estimate associated with a sample of observations is a choice of parameters which maximizes the probability density function of the sample. Thus, with ML estimation, the problem of determining Θ becomes

$$\max_{\Theta} p\left(\frac{\vec{X}}{\Theta}\right) \quad (10)$$

with the constraints $\sum_{j=1}^{M} P(j) = 1$ and $P(j) > 0 \; \forall j \in [1, M]$. These constraints permit us to take into consideration *a*

priori probabilities $P(j)$. Using Lagrange multipliers, we maximize the following function:

$$\Phi(\vec{X},\Theta,\Lambda) = ln\left(p\left(\vec{X}|\Theta\right)\right) + \Lambda\left(1-\sum_{i=1}^{M}P(i)\right)+\mu\sum_{j=1}^{M}P(j)ln\left(P(j)\right) \quad (11)$$

where Λ is the Lagrange multiplier. For convenience, we have replaced the function $p(\vec{X}/\Theta)$ in (10) by the function $ln(p(\vec{X}/\Theta))$. If we assume that we have N random vectors $\vec{X}_i$ which are independent, we can write

$$p\left(\vec{X}|\Theta\right) = \prod_{i=1}^{N} p\left(\vec{X}_i|\Theta\right) \quad (12)$$

$$p\left(\vec{X}_i|\Theta\right) = \sum_{j=1}^{M} p\left(\vec{X}_i|j,\Theta_j\right)P(j). \quad (13)$$

Replacing (12) and (13), we obtain

$$\Phi(\vec{X},\Theta,\Lambda) = \sum_{i=1}^{N} ln\left(\sum_{j=1}^{M} p\left(\vec{X}_i|j,\Theta_j\right)P(j)\right) + \Lambda\left(1-\sum_{j=1}^{M}P(j)\right)+\mu\sum_{j=1}^{M}P(j)ln\left(P(j)\right). \quad (14)$$

In order to automatically find the number of components needed to model the mixture, we use an entropy-based criterion previously used in the case of Gaussian mixtures [25], [26]. Thus, the first term in (14) is the log-likelihood function, and it assumes its global maximum value when each component represents only one of the feature vectors. The last term (entropy) reaches its maximum when all of the feature vectors are modeled by a single component, i.e., when $P(j1) = 1$ for some $j1$ and $P(j) = 0$, $\forall j, j \neq j1$. The algorithm starts with an over-specified number of components in the mixture, and as it proceeds, components compete to model the data. The choice of μ is critical to the effective performance of the algorithm, since it specifies the tradeoff between the required likelihood of the data and the number of components to be found. We choose μ to be the ratio of the first term to the last term in (14) of each iteration t, i.e.

$$\mu(t) = \frac{\sum_{i=1}^{N} ln\left(\sum_{j=1}^{M} p^{(t-1)}\left(\vec{X}_i|j,\Theta_j\right)P^{(t-1)}(j)\right)}{\sum_{j=1}^{M} P^{(t-1)}(j)ln\left(P^{(t-1)}(j)\right)}. \quad (15)$$

We will now try to resolve this optimization problem. To do so, we must determine the solution to the following equations:

$$\frac{\partial}{\partial\Theta}\Phi(\vec{X},\Theta,\Lambda) = 0 \quad (16)$$

$$\frac{\partial}{\partial\Lambda}\Phi(\vec{X},\Theta,\Lambda) = 0. \quad (17)$$

Calculating the derivative with respect to Θ_j, we obtain [27]

$$\frac{\partial}{\partial\Theta_j}\Phi(\vec{X},\Theta,\Lambda) = \sum_{i=1}^{N} p\left(j|\vec{X}_i,\Theta_j\right)\frac{\partial}{\partial\Theta_j}ln\left(p\left(\vec{X}_i|j,\Theta_j\right)\right) \quad (18)$$

where $p(j|\vec{X}_i,\Theta_j)$ is the *a posteriori* probability. In what follows, we will estimate the parameters.

A. Estimation of the a priori Probability

Since $p(\vec{X}_i|j,\vec{\alpha_j})$ is independent of $P(j)$, straightforward manipulations yield

$$\frac{\partial}{\partial P(j)}\Phi(\vec{X},\Theta,\Lambda) = \frac{1}{P(j)}\sum_{i=1}^{N} p\left(j|\vec{X}_i,\vec{\alpha_j}\right) - \Lambda + \mu\left(1+ln\left(P(j)\right)\right). \quad (19)$$

Now, we take the derivative of (14) with respect to Λ. We find

$$\frac{\partial}{\partial\Lambda}\Phi(\vec{X},\Theta,\Lambda) = 1-\sum_{j=1}^{M}P(j) = 0. \quad (20)$$

Thus, the *a priori* probability can be shown to be [27]

$$P(j)^{\text{new}} = \frac{\sum_{i=1}^{N} p^{\text{old}}\left(j|\vec{X}_i,\vec{\alpha_j}\right)+\mu\left(p(j)^{\text{old}}\left(1+ln\left(p(j)^{\text{old}}\right)\right)\right)}{N+\mu\sum_{j=1}^{M}\left(p(j)^{\text{old}}\left(1+ln\left(p(j)^{\text{old}}\right)\right)\right)}. \quad (21)$$

B. Estimation of the $\vec{\alpha}$ Parameters

In order to estimate the $\vec{\alpha}$ parameters, we will use the Fisher scoring method. This approach is a variant of the Newton–Raphson [28] method. In fact, (16) can be approximated by expanding it into a power series around a point Θ_{j_0}

$$\frac{\partial}{\partial\Theta_j}\Phi(\vec{X},\Theta,\Lambda) \simeq \frac{\partial}{\partial\Theta_j}\Phi(\vec{X},\Theta,\Lambda)\left(\Theta_{j_0}\right) + \left(\Theta_j-\Theta_{j_0}\right)\frac{\partial^2}{\partial\Theta_j^2}\Phi(\vec{X},\Theta,\Lambda)\left(\Theta_{j_0}\right). \quad (22)$$

Since $(\partial/\partial\Theta_j)\Phi(\vec{X},\Theta,\Lambda) = 0$, then

$$\Theta_j \simeq \Theta_{j_0} - \left(\frac{\partial^2}{\partial\Theta_j^2}\Phi(\vec{X},\Theta,\Lambda)\left(\Theta_{j_0}\right)\right)^{-1} \times\frac{\partial}{\partial\Theta_j}\Phi(\vec{X},\Theta,\Lambda)\left(\Theta_{j_0}\right). \quad (23)$$

Thus, an updated estimate $\hat{\Theta}_j^{(k+1)}$ of a current estimate $\hat{\Theta}_j^{(k)}$ is given by

$$\hat{\Theta}_j^{(k+1)} = \hat{\Theta}_j^{(k)} - \left(\frac{\partial^2}{\partial\Theta_j^2}\Phi(\vec{X},\Theta,\Lambda)(\Theta_j)\right)^{-1}_{\Theta_j=\hat{\Theta}_j^{(k)}} \times\left(\frac{\partial}{\partial\Theta_j}\Phi(\vec{X},\Theta,\Lambda)(\Theta_j)\right)_{\Theta_j=\hat{\Theta}_j^{(k)}} \quad (24)$$

which is equivalent to

$$\hat{\Theta}_j^{(k+1)} = \hat{\Theta}_j^{(k)} - H^{-1}\left(\hat{\Theta}_j^{(k)}\right) \times\left(\frac{\partial}{\partial\Theta_j}\Phi(\vec{X},\Theta,\Lambda)(\Theta_j)\right)_{\Theta_j=\hat{\Theta}_j^{(k)}} \quad (25)$$

where H is the Hessian matrix evaluated at the current estimate. One variant of this approach is Fisher's scoring method, where the Hessian matrix H is replaced by the negative of Fisher's information matrix. The scoring method is based on the first, second, and mixed derivatives of the log-likelihood function.

Thus, we will compute these derivatives. According to (18), we have

$$\frac{\partial}{\partial\alpha_{jl}}\Phi(\vec{X},\Theta,\Lambda)=\sum_{i=1}^{N}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)\partial|\partial\alpha_{jl}ln\left(p\left(\vec{X}_i|j,\vec{\alpha_j}\right)\right). \tag{26}$$

Calculating the derivative of $ln(p(\vec{X}_i/j,\vec{\alpha_j})$ with respect to Θ_j, we obtain [27]

$$\frac{\partial}{\partial\alpha_{jl}}ln\left(p\left(\vec{X}_i|j,\vec{\alpha_j}\right)\right)=\Psi\left(|\vec{\alpha_j}|\right)-\Psi(\alpha_{jl})+ln(X_{il}) \tag{27}$$

where $\Psi(.)$ is the digamma function. Thus

$$\frac{\partial}{\partial\alpha_{jl}}\Phi(\vec{X},\Theta,\Lambda)=\sum_{i=1}^{N}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)(ln(X_{il}))+\left(\Psi\left(|\vec{\alpha_j}|\right)-\Psi(\alpha_{jl})\right)\sum_{i=1}^{N}p\left(j|r\vec{X}_i,\vec{\alpha_j}\right). \tag{28}$$

During iterations, the α_{jl} can become negative. In order to overcome this problem, a suggestion was given by Ronning [11] for the case of one Dirichlet distribution. His suggestion was to set all $\alpha_{jl}=min\{X_{il}\}, i=1\ldots N$. These initial estimates only prevent the α_{jl} from becoming negative during the first few iterations, however. Besides, the method gives good results only in the case of one distribution, because of sensitivity to initialization in the case of a mixture (see next section). Here, we give a better solution for keeping the α_{jl} positive during all the iterations. Since we require that the α_{jl} be strictly positive, and we want the parameters upon which we will derive to be unconstrained, we reparametrize, setting $\alpha_{jl}=e^{\beta_{jl}}$, where β_{jl} is an unconstrained real number. Then, the partial derivative of Φ (14) with respect to β_{jl} is as follows [27]:

$$\frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda)=\alpha_{jl}\left[\sum_{i=1}^{N}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)(ln(X_{il}))+\left(\Psi\left(|\vec{\alpha_j}|\right)-\Psi(\alpha_{jl})\right)\sum_{i=1}^{N}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)\right]. \tag{29}$$

By computing the second and mixed derivatives of the log-likelihood function, we obtain [27]

$$\begin{aligned}\frac{\partial^2}{\partial^2\beta_{jl}}&\Phi(\vec{X},\Theta,\Lambda)\\&=\frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda)\\&\quad+\alpha_{jl}^2\left(\Psi'\left(|\vec{\alpha_j}|\right)-\Psi'(\alpha_{jl})\right)\sum_{i=1}^{N}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)\\&\quad+\alpha_{jl}\left[\left(\Psi\left(|\vec{\alpha_j}|\right)-\Psi(\alpha_{jl})\right)\sum_{i=1}^{N}\frac{\partial}{\partial\beta_{jl}}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)+\sum_{i=1}^{N}\frac{\partial}{\partial\beta_{jl}}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)(ln(X_{il}))\right]\end{aligned} \tag{30}$$

$$\begin{aligned}\frac{\partial^2}{\partial\beta_{jl_1}\partial\beta_{jl_2}}&\Phi(\vec{X},\Theta,\Lambda)\\&=\alpha_{jl_1}\alpha_{jl_2}\Psi'\left(|\vec{\alpha_j}|\right)\sum_{i=1}^{N}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)\\&\quad+\alpha_{jl_1}\left[\left(\Psi\left(|\vec{\alpha_j}|\right)-\Psi(\alpha_{jl_1})\right)\sum_{i=1}^{N}\frac{\partial}{\partial\beta_{jl_2}}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)+\sum_{i=1}^{N}\partial|\partial\beta_{jl_2}p\left(\frac{j}{\vec{X}_i},\vec{\alpha_j}\right)\left(ln\left(X_{il_1}\right)\right)\right]\end{aligned} \tag{31}$$

where $\Psi'(.)$ is the trigamma function. Note that we need to compute the derivative of the a posteriori probability $p(j|\vec{X}_i,\vec{\alpha_j})$ with respect to β_{jl} [27]

$$\frac{\partial}{\partial\beta_{jl}}p\left(j|\vec{X}_i,\vec{\alpha_j}\right)=\alpha_{jl}\times p\left(j|\vec{X}_i,\vec{\alpha_j}\right)\left(1-p\left(j|\vec{X}_i,\vec{\alpha_j}\right)\right)\times\left(\Psi\left(|\vec{\alpha_j}|\right)-\Psi(\alpha_{jl})+ln(X_{il})\right). \tag{32}$$

Given a set of initial estimates (see Section IV), Fisher's scoring method can now be used. The iterative scheme of the Fisher method is given by the following equation:

$$\begin{aligned}\begin{pmatrix}\hat{\beta}_{j1}\\\vdots\\\hat{\beta}_{j\dim+1}\end{pmatrix}^{\text{new}}&=\begin{pmatrix}\hat{\beta}_{j1}\\\vdots\\\hat{\beta}_{j\dim+1}\end{pmatrix}^{\text{old}}\\&+\begin{pmatrix}\text{Var}(\hat{\beta}_{j1})&\cdots&\text{Cov}(\hat{\beta}_{j1},\hat{\beta}_{j\dim+1})\\\vdots&\ddots&\vdots\\\text{Cov}(\hat{\beta}_{j\dim+1},\hat{\beta}_{j1})&\cdots&\text{Var}(\hat{\beta}_{j\dim+1})\end{pmatrix}^{\text{old}}\\&\times\begin{pmatrix}\frac{\partial}{\partial\hat{\beta}_{j1}}\Phi\\\vdots\\\frac{\partial}{\partial\hat{\beta}_{j\dim+1}}\Phi\end{pmatrix}^{\text{old}}\end{aligned} \tag{33}$$

where j is the class number: $1<j<M$.

The variance–covariance matrix is obtained as the inverse[1] of the Fisher information matrix $\mathbb{I}$. The information matrix $\mathbb{I}$ is

$$\mathbb{I}=I_{l_1l_2}=-E\left[\frac{\partial^2}{\partial\beta_{jl_1}\partial\beta_{jl_2}}\Phi(\vec{X},\Theta,\Lambda)\right]. \tag{34}$$

Comparing this iterative scheme based on Fisher's scoring method with a quasi-Newton method presented by the following [16], [29]:

$$\beta_{jl}^{\text{new}}:=\beta_{jl}^{\text{old}}-\eta\frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda) \tag{35}$$

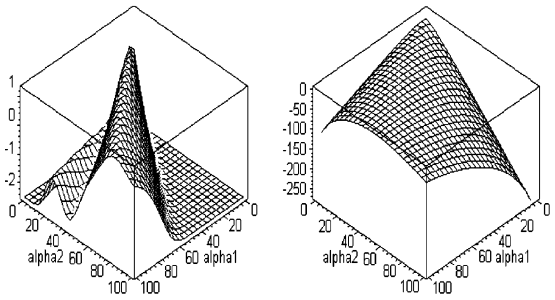
where $0<\eta\leq 1$; we note the presence of an extra term which is the inverse of the Fisher information matrix. Let us now focus on the geometrical interpretation of (33)

$$\beta_{jl}^{\text{new}}:=\beta_{jl}^{\text{old}}\underbrace{-\left(\text{Var}(\hat{\beta}_{jl})\ \ldots\ \text{Cov}(\hat{\beta}_{jl},\hat{\beta}_{j\dim+1})\right)^{\text{old}}\begin{pmatrix}\frac{\partial\Phi}{\partial\hat{\beta}_j}\\\vdots\\\frac{\partial\Phi}{\partial\hat{\beta}_{j\dim+1}}\end{pmatrix}^{\text{old}}}_{NG}. \tag{36}$$

Thus, the ordinary gradient $(\partial/\partial\beta_{jl})\Phi(\vec{X},\Theta,\Lambda)$ is replaced by the term NG, which is called the *natural gradient* or *contravariant gradient* by Amari [30]. Let $S=\{\vec{\beta}=(\beta_1,\ldots,\beta_{\dim+1})\in\mathbb{R}^{\dim+1}\}$ be a parameter space in which a given likelihood function (Φ) is defined. When S is a Euclidean space with an orthonormal coordinate system, the squared length of a small incremental vector $\partial\vec{\beta}=(\partial\beta_1,\ldots,\partial\beta_{\dim+1})$ connecting $\vec{\beta}$ and $\vec{\beta}+\partial\vec{\beta}$ is given by

$$|\partial\vec{\beta}|^2=\sum_{l_1=1}^{\dim+1}(\partial\beta_i)^2. \tag{37}$$

[1]We have made some approximations to avoid inverting the information matrix in each iteration. See [27] for more details.



(a) (b)

Fig. 2. Representation of Φ as a function of two parameters α_1 and α_2. (a) $M > 1$. (b) $M = 1$.

However, when the coordinate system is nonorthonormal, the squared length is given by the following quadratic form [30]:

$$|\vec{\partial\beta}|^2 = \sum_{l_1=1}^{\dim+1} \sum_{l_2=1}^{\dim+1} g_{l_1 l_2} (\partial\beta_{l_1})(\partial\beta_{l_2}). \tag{38}$$

The $(\dim+1)\times(\dim+1)$ matrix $G = (g_{l_1 l_2})$ is called the Riemannian metric tensor. This matrix is reduced to the unit matrix $I_{\dim+1\times\dim+1}$ in the Euclidean orthonormal case. Knowing that the Dirichlet is an exponential density [14], we can affirm that the parameter space of our likelihood function Φ, described by (14), is a curved manifold. In fact, according to Amari, the exponential family of probabilities forms a manifold which is equipped with a Riemannian metric given by the Fisher information matrix [30]. Thus, we do not have an orthonormal linear coordinate system, and the length of $\vec{\partial\beta}$ is written as (38). Knowing that the steepest direction of a function Φ at $\vec{\beta}$ is defined by the vector $\vec{\partial\beta}$ that minimizes $\Phi(\beta + \vec{\partial\beta})$, where $|\vec{\partial\beta}|$ has a fixed length and is sufficiently small, we deduce that the gradient of Φ cannot be the same in a Euclidean space and a Riemannian one. The relation between the natural gradient $(\check{\partial}\phi/\partial\beta_l)$ and the ordinary gradient is given by the following equation [31]:

$$\frac{\check{\partial}\phi}{\partial\beta_l} = G^{-1}\frac{\partial\phi}{\partial\beta_l} \tag{39}$$

where G is the Fisher information matrix. This result has been confirmed by experiments. Indeed, we have implemented these two methods and observed that the method given by (35) does not give good results compared with the Fisher scoring method.

IV. Initialization and Convergence Test

The maximum likelihood function presented by (14) is globally concave [11] in the case of one distribution $(\mathrm{M} = 1)$. However, this particular advantage is not preserved when $M > 1$, as shown in Fig. 2. In order to make our algorithm less sensitive to local maxima, we have used some initialization schemes, including the fuzzy C means [32] and the method of moments (MM). In fact, the MM gives really good estimates because of the compact support of the Dirichlet distribution. From an examination of (6) and (7), we see that there are $\dim$ first-order moments and $\dim$ second-order moments, yielding a total of $C_{\dim+1}^{2(\dim+1)}$ possible combinations of equations to solve for the $\dim$ parameters. According to Fieltiz and Myers [12], a symmetrical way of proceeding would be to choose the first $\dim$ first-order equations and the first second-order equation. The reason for not choosing the $(\dim+1)$th first-order equation is that the $(\dim+1)$th equation is a linear combination of the others and together they do not form an independent set of equations. Thus, we have

$$\alpha_l = \frac{(x'_{11} - x'_{21})\,x'_{1l}}{x'_{21} - (x'_{11})^2} \quad l = 1, 2\ldots, \dim \tag{40}$$

and

$$\alpha_{\dim+1} = \frac{(x'_{11} - x'_{21})\left(1 - \sum_{l=1}^{\dim} x'_{1l}\right)}{x'_{21} - (x'_{11})^2} \tag{41}$$

where

$$x'_{1l} = \frac{1}{N}\sum_{i=1}^{N} x_{il} \quad l = 1, 2\ldots, \dim+1 \tag{42}$$

$$x'_{21} = \frac{1}{N}\sum_{i=1}^{N} x_{i1}^2. \tag{43}$$

Thus, our initialization method can be summed up as follows.

INITIALIZATION Algorithm:

1) Apply the fuzzy C-means to obtain the elements, covariance matrix and mean of each component.
2) Apply the MM for each component j to obtain the vector of parameters $\vec{\alpha_j}$.
3) Assign the data to clusters, assuming that the current model is correct.
4) Update the $P(j)$ using the following:

$$P(j) = \frac{\text{Number of elements in class j}}{N}. \tag{44}$$

5) If the current model and the new model are sufficiently close to each other, terminate, else go to 2.

We can readily note that this initialization algorithm takes the distribution into account. In contrast to the *classic* initialization methods which use only algorithms, such as the K means to obtain the initialization parameters, we have introduced the MM with an iterative scheme to refine the results. By using the MM, we suppose from the outset that we have a Dirichlet mixture. This initialization method is designed to work on large databases. When working on small data sets, applying the fuzzy C-means and the MM only once is a feasible option. With this initialization method in hand, our algorithm for estimating a Dirichlet mixture can be summarized as follows.

DIRICHLET MIXTURE ESTIMATION Algorithm:

1) INPUT: $\dim$-dimensional data X_i, $i = 1, \ldots, N$ and an over-specified number of clusters M.
2) INITIALIZATION Algorithm.
3) Update the $\vec{\alpha_j}$ using (33), $j = 1, \ldots, M$.
4) Update the $P(j)$ using (21), $j = 1, \ldots, M$.
5) If $p(j) < \epsilon$ discard component j, go to 3.
6) If the convergence test is passed, terminate, else go to 3.

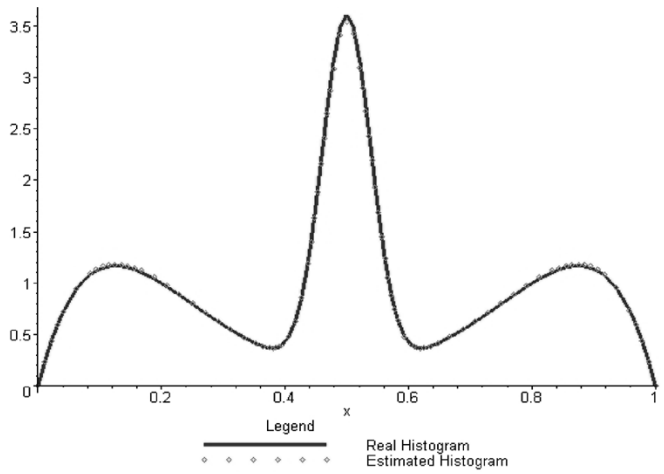


Fig. 3. First artificial histogram.

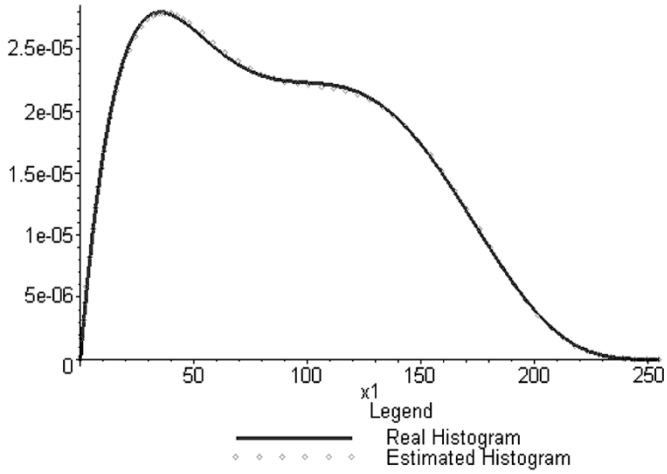


Fig. 4. Second artificial histogram.

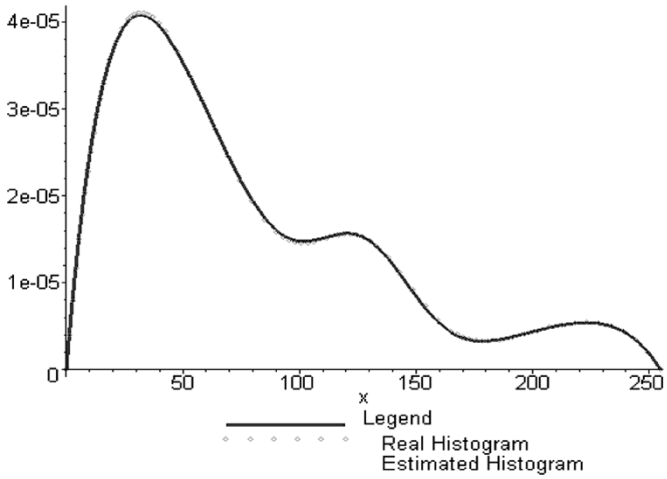


Fig. 5. Third artificial histogram.

If the sample is sufficiently large, the test of convergence can be done using a statistical method [33]. The method uses a quadratic form of the gradient vector. Consider the statistics given by (45). This statistical test can be shown to be distributed as a Chi-square random variable with $\dim + 1$ degrees of freedom. The iteration is continued until S falls below $\chi^2_{\dim+1}(\nu)$ for a fixed ν. Other convergence tests could involve testing the stabilization of the $\vec{\beta_j}$ or the value of the maximum likelihood function

$$S = \left(\frac{\partial}{\partial \hat{\beta}_{j1}}\Phi \quad \cdots \quad \frac{\partial}{\partial \hat{\beta}_{j\dim+1}}\Phi \right) \times \begin{pmatrix} \mathrm{Var}(\hat{\beta}_{j1}) & \cdots & \mathrm{Cov}(\hat{\beta}_{j1}, \hat{\beta}_{j\dim+1}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{\beta}_{j\dim+1}, \hat{\beta}_{j1}) & \cdots & \mathrm{Var}(\hat{\beta}_{j\dim+1}) \end{pmatrix} \times \begin{pmatrix} \frac{\partial}{\partial \hat{\beta}_{j1}}\Phi \\ \vdots \\ \frac{\partial}{\partial \hat{\beta}_{j\dim+1}}\Phi \end{pmatrix}. \quad (45)$$

V. Experimental Results

In this section, we validate the Dirichlet mixture, using contextual and noncontextual evaluation [34] to test the performance of our method. The noncontextual evaluation concerns the estimation of artificial histograms while the contextual evaluation is based on two computer vision applications. Note that we have verified that all the pairwise correlations in a given vector are negative. We begin with the noncontextual evaluation. For this purpose, we generated artificial histograms from artificial Dirichlet mixture models using the following:

$$H(X_i) = \sum_{j=1}^{M} P(j) \frac{\Gamma(\alpha_{j1} + \alpha_{j2})}{\Gamma(\alpha_{j1})\Gamma(\alpha_{j2})} X_i^{\alpha_{j1}-1}(1-X_i)^{\alpha_{j2}-1} \quad (46)$$

where $i = 1 \ldots N$, N is the number of data used to generate the histogram. After that, we tried to estimate the parameters of these artificial histograms. Fig. 3–5 are examples of these histograms. The first histogram presents a Dirichlet mixture of three well-separated components. The second and third histograms present overlapped Dirichlet components. The real and estimated parameters of each histogram are specified in Tables I–III. For these examples, we found the exact number of components by using our method by taking $M = 5$ and the three criteria we employed (Akaike, Schwarz, and minimum description length).

The second validation is contextual and concerns the summarization of image databases. In fact, interactions between users and multimedia databases can involve queries like "retrieve images that are similar to this image." A number of techniques have been developed to handle pictorial queries, e.g, QBIC [35], Photobook [36], Blobworld [37], and VisualSeek [38]. Summarizing the database is very important because it simplifies the task of retrieval by restricting the search for similar images to a smaller domain of the database [39]. Summarization is also very efficient for browsing [40]. Knowing the categories of images in a given database allows the user to find the images he is looking for more quickly. Using mixture decomposition, we can find natural groupings of images and represent each group by the most representative image in the group. In other words, after appropriate features are extracted from the images, it allows us to partition the feature space into regions that are relatively homogeneous with respect to the chosen set of features.

TABLE I
ESTIMATION OF THE PARAMETERS OF THE FIRST ARTIFICIAL HISTOGRAM

	Real parameters	Estimated parameters
	$P(1)$=0.33	$P(1)$=0.333
Mode 1	α_{11} = 8	α_{11} = 8.116
	α_{12} = 2	α_{12} = 2.024
	$P(2)$=0.34	$P(2)$=0.335
Mode 2	α_{21} = 80	α_{21} = 79.567
	α_{22} = 80	α_{22} =79.567
	$P(3)$=0.33	$P(3)$=0.332
Mode 3	α_{31} = 2	α_{31} = 2.024
	α_{32} = 8	α_{32} = 8.116

TABLE II
ESTIMATION OF THE PARAMETERS OF THE SECOND ARTIFICIAL HISTOGRAM

	Real parameters	Estimated parameters
	$P(1)$=0.5	$P(1)$=0.545
Mode 1	α_{11} = 2	α_{11} = 1.942
	α_{12} = 8	α_{12} = 7.141
	$P(2)$=0.5	$P(2)$=0.455
Mode 2	α_{21} = 5	α_{21} = 5.466
	α_{22} = 5	α_{22} =5.108

TABLE III
ESTIMATION OF THE PARAMETERS OF THE THIRD ARTIFICIAL HISTOGRAM

	Real parameters	Estimated parameters
	$P(1)$=0.75	$P(1)$=0.746
Mode 1	α_{11} = 2	α_{11} = 2.043
	α_{12} = 8	α_{12} = 8.307
	$P(2)$=0.15	$P(2)$=0.156
Mode 2	α_{21} = 20	α_{21} = 19.185
	α_{22} = 20	α_{22} =19.086
	$P(3)$=0.1	$P(3)$=0.098
Mode 3	α_{31} = 8	α_{31} = 8.220
	α_{32} = 2	α_{32} = 2.035

By identifying the homogeneous regions in the feature space, the task of summarization is accomplished. We used a database containing 600 images of size 128 $\times$ 96 and took color as a feature for categorizing the images. In order to determine the vector of characteristics for each image, pixels were projected onto the $3D$ HSI (H = hue, S = daturation, and I = intensity) space. We, thus, obtained a $3D$ color histogram for each image. Based on the work of Kherfi *et al.* [41], we obtained an $8D$ vector from this histogram. Their method consists of partitioning the space by subdividing each of the axes H, S, and I into n equal intervals. This gives n^3 subspaces. The sum of the elements in each subspace is computed and the result is placed in the corresponding cell of the feature vector. In our application, we chose $n = 2$, so each image was represented by a $2^3 = 8D$ feature vector. We also asked a human subject to determine the number of groups, and he found five categories (see Fig. 7), which is the number of classes found by our algorithm and by the three criteria we used (see Fig. 6). The Dirichlet mixture algorithm was applied to the feature vectors, where each vector represents an image. The classification was performed using the Bayesian decision rule [29] after the class-conditional densities were estimated

$$x \mapsto r(x) = \text{argmax}_j \left\{P(j)p\left(\vec{x_i}|j\right)\right\}. \tag{47}$$

The two classifications (the one generated by the human subject and the one given by our algorithm) were compared by counting the number of misclassified images, yielding confusion matrix (see Table IV). In this confusion matrix, the cell (classi classj) represents the number of images from classi which are classified as classj. The number of images misclassified was small: 40 in all, which represents an accuracy of 93.34%. Table V show the confusion matrix for the Gaussian mixture.

After the database was summarized, we conducted another experiment designed to retrieve images similar to a query. First, we defined a measure to determine the closest component to the query vector. Next, another distance measure was used to determine the similarity between the query vector and the feature vectors in the closest component. The *a posteriori* probabilities were used in order to choose the component nearest to the query. After selecting the closest component, the two-norm was applied to find the images most similar to the query. To measure the retrieval rates, each image was used as a query and the number of relevant images among those that were retrieved was noted. Precision, which is the measure most commonly used by the information retrieval community, was then computed using (48). This measure was then averaged over all the queries. The results are given in Table VI

$$\text{precision} = \frac{\text{number of images retrieved and relevant}}{\text{total number of retrieved images}}. \tag{48}$$

The third application concerns modeling for human skin color using the Dirichlet mixture. In fact, human skin color has been used and has proven to be an effective feature in many applications including teleconferencing [42], face recognition [43], and gesture recognition [44]. The motivation for using a Dirichlet mixture is based on the observation that the color histogram for the skin of people with different ethnic backgrounds does not form a unimodal distribution, but rather a multimodal distribution. Although different people appear to have different colored skin, several studies have shown that major difference lies in intensity rather than in the color itself [45]. Thus, the common *RGB* (red, green, blue) representation of color images is not suitable for characterizing skin color because in the *RGB* space, the triple components rgb represent not only color, but also luminance. To build a skin color model, we can use *CIE LUV* or the chromatic color spaces and discard the luminance value. In our case, we have used chromatic colors (also known as *pure* colors in the absence of luminance), defined by a normalization process as follows[46]:

$$r1 = \frac{r}{r+g+b} \tag{49}$$

$$g1 = \frac{g}{r+g+b} \tag{50}$$

$$b1 = \frac{b}{r+g+b}. \tag{51}$$

In order to train for skin color, we used color images containing human faces and extracted the skin regions in these images manually. Our training set contained more than six hundred images containing human skins of different races. The total number of pixels analyzed was 10 867 932 (skin color pixels), where each sample consists of three values ($r1$, $g1$, $b1$). Fig. 8 shows a part of the data collected in the $r1g1$ space. It is clear that a single

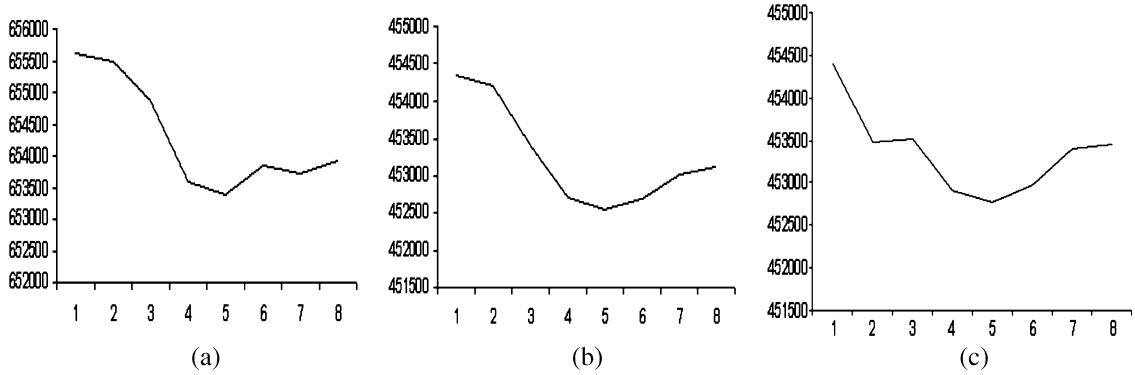


Fig. 6. Number of classes found by the three criteria. (a) Akaike. (b) MDL. (C) Schwartz.

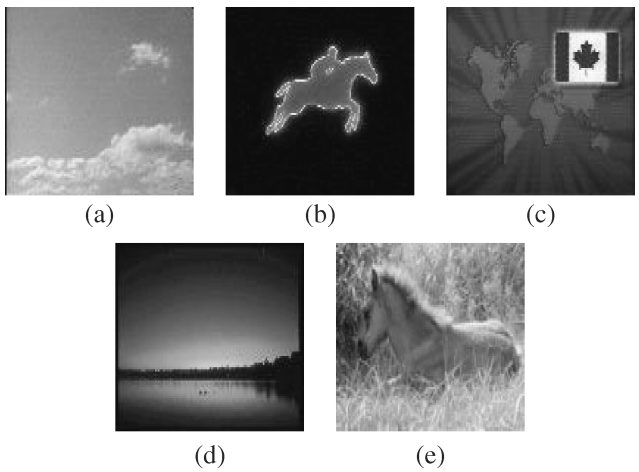


Fig. 7. Sample images from each group. (a) Class1. (b) Class2. (c) Class3. (d) Class4. (e) Class5.

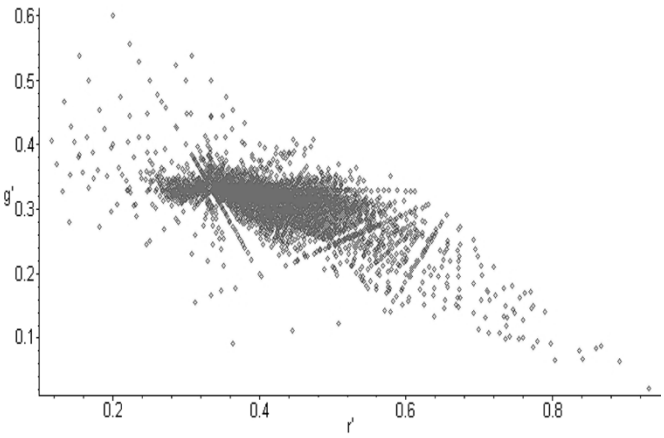


Fig. 8. Part of the data collected for human skin color modeling.

TABLE IV
CONFUSION MATRIX FOR IMAGE CLASSIFICATION BY A DIRICHLET MIXTURE

	Class1	Class2	Class3	Class4	Class5
Class1	101	0	10	0	0
Class2	0	120	0	13	0
Class3	0	0	108	0	0
Class4	0	6	0	104	4
Class5	0	2	0	5	127

TABLE V
CONFUSION MATRIX FOR IMAGE CLASSIFICATION BY A GAUSSIAN MIXTURE

	Class1	Class2	Class3	Class4	Class5
Class1	101	0	10	0	0
Class2	0	112	0	21	0
Class3	2	0	106	0	0
Class4	0	10	0	96	8
Class5	0	6	0	9	119

TABLE VI
RETRIEVAL PRECISION FOR THE CONTENT-BASED IMAGE RETRIEVAL APPLICATION

Algorithm	Number of retrieved images		
	32	64	86
Gaussian	0.812	0.625	0.581
Dirichlet	0.906	0.781	0.697

Dirichlet density function is not sufficient to model the distribution of skin color. We used our algorithm to estimate the parameters of the Dirichlet mixture and we found two components. The estimated density function is shown in Fig. 9. Given an image, a segmentation was performed to obtain homogenous regions. Each pixel was classified as skin color if its probability measure was above a threshold. Each region was then recognized as a skin area if most of the pixels in the region had a high probability of being skin color. Fig. 10 shows the results of skin detection in different cases. Skin color alone is usually not sufficient in detecting human faces or hands. However, a good estimated mixture is very useful in simplifying the task of skin area detection. Using skin color and area information, human faces can be detected robustly [43]. In our application, if more than 75% of the pixels in a region are classified to be skin color, then the region is recognized as a skin area. Fig. 11 shows an image sequence and Fig. 12 shows the results of skin detection.

VI. CONCLUSION

In this paper, we have introduced a new mixture based on the Dirichlet distribution. The Dirichlet distribution has the advantage that by varying its parameters, it permits multiple modes and asymmetry and can thus approximate a wide variety of shapes. We estimated the parameters of this mixture using the maximum likelihood and Fisher scoring methods. An interesting interpretation, based on statistical geometric information, was given. In order to make our method less sensitive to initialization, we proposed an initialization algorithm based on the moments, which takes the distribution into account from the outset. Contextual and noncontextual evaluations were

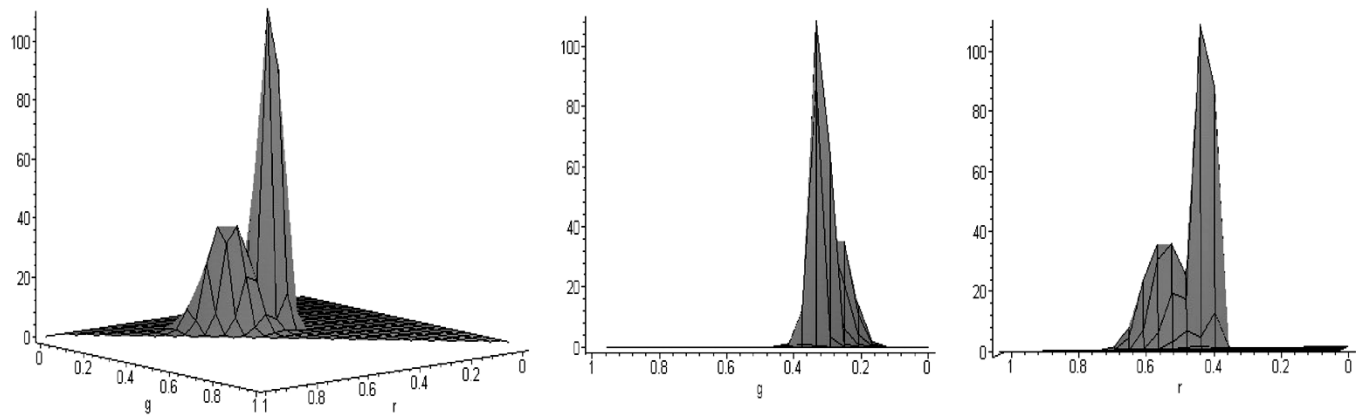


Fig. 9. Estimated density function viewed from different angles.

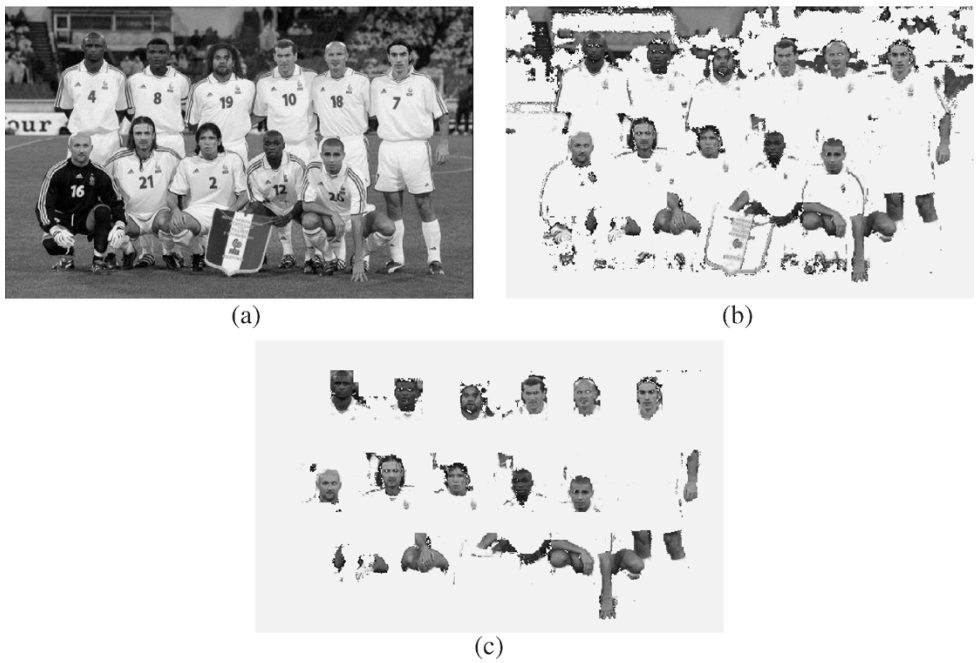


Fig. 10. Original image and results of skin detection in different cases. (a) Original image. (b) Skin regions extracted by a mixture of Gaussian distributions. (c) Skin regions extracted by a mixture of Dirichlet distributions.

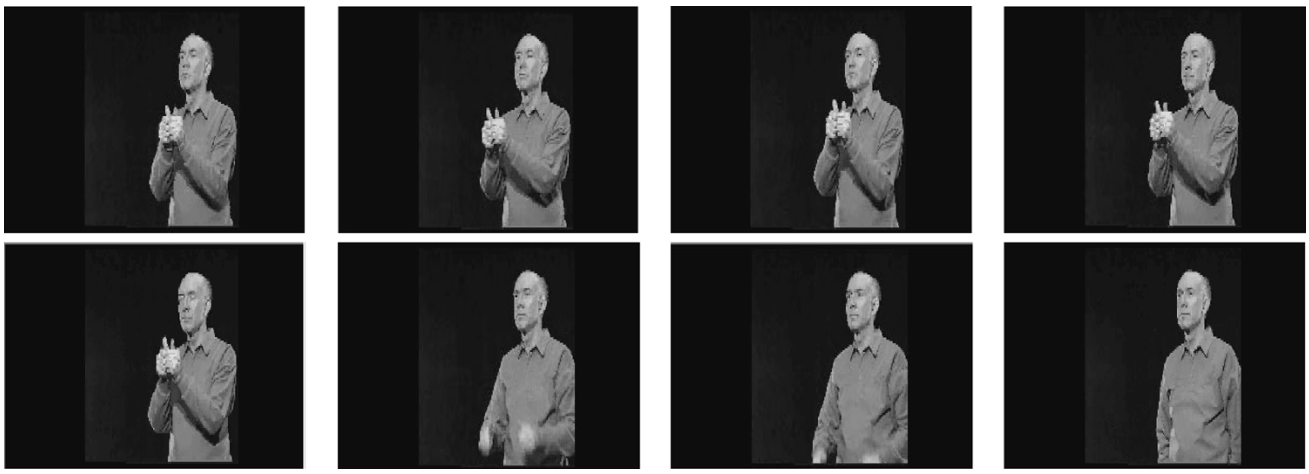


Fig. 11. Image sequence of ASL sign "accompany."

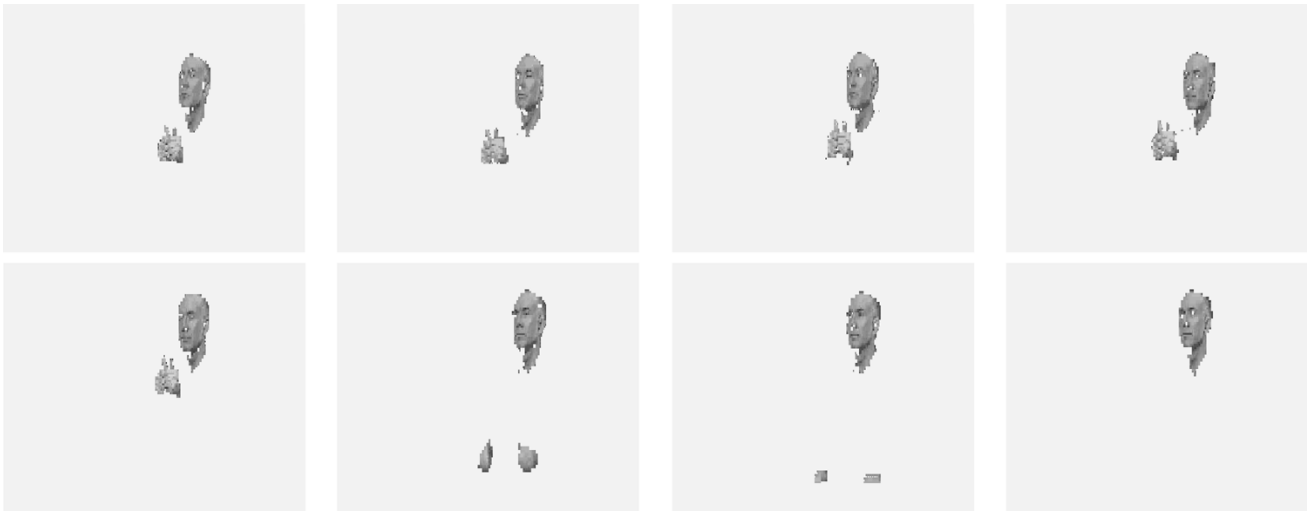


Fig. 12. Skin areas detected in the sequence of Fig. 11.

used to test the performance of our method. The noncontextual evaluation was based on artificial histograms. The contextual test involved the summarization of image databases for efficient retrieval and the detection of human skin color. From the results of these evaluations, we can say that the Dirichlet mixture has good modeling capabilities. Future work will be devoted to generalizing of this distribution in order to improve its covariance structure.

REFERENCES

[1] S. J. Roberts and L. Rezek, "Bayesian approach to Gaussian mixture modeling," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 1133–1142, Nov. 1998.

[2] M. A. T. Figeiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 4–37, Mar. 2002.

[3] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4–37, Jan. 2000.

[4] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*. London, U.K.: Chapman & Hall, 1981.

[5] S. Medasani and R. Krishnapuram, "A comparison of Gaussian and pearson mixture modeling for pattern recognition and computer vision applications," *Pattern Recognit. Lett.*, vol. 20, pp. 305–313, 1999.

[6] A. E. Raftery and J. D. Banfield, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 49, pp. 803–821, 1993.

[7] T Kato, S. Omachi, and H. Aso, "Asymmetric Gaussian and its application to pattern recognition," in *Proc. Joint IAPR Int. Workshops SSPR 2002 and SPR 2002*, Aug. 2002, pp. 404–413.

[8] A. El Zaart and D. Ziou, "Statistical modeling of multimodal SAR images," *Int. J. Remote Sens.*, to be published.

[9] R. J. Connor and J. E. Mosimann, "Concepts of independence for proportions with a generalization of the Dirichlet distribution," *J. Amer. Stat. Assoc.*, vol. 39, pp. 1–38, 1977.

[10] A. Narayanan, "A note on parameter estimation in the multivariate Beta distribution," *Comput. Math. Applicat.*, vol. 24, no. 10, pp. 11–17, 1992.

[11] G. Ronning, "Maximum likelihood estimation of Dirichlet distributions," *J. Stat. Comput. Simul.*, vol. 32, pp. 215–221, 1989.

[12] B. D. Fielitz and B. L. Myers, "Estimation of parameters in the Beta distribution," *Decision Sci.*, vol. 6, pp. 1–13, 1975.

[13] G. H. Weis and M. Dishon, "Small sample comparison of estimation methods for the Beta distribution," *J. Stat. Comput. Simul.*, vol. 11, pp. 1–11, 1980.

[14] K. Samuel, K. W. Ng, and K. Fang, *Symmetric Multivariate and Related Distributions*. London, U.K.: Chapman & Hall, 1990.

[15] K. Samuel, K. Balakrishman, and J. Norman, *Continous Multivariate Distributions*. New York: Wiley, 2000, vol. 1.

[16] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and EM algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, Apr. 1984.

[17] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.

[18] T. J. Santner and D. E. Duffy, *The Statistical Analysis of Discrete Data*. New York: Springer-Verlag, 1989.

[19] P. Rao, "Asymtotic theory of statistical inference," in *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley, 1987.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B*, vol. 39, pp. 1–38, 1977.

[21] S. Ikeda, "Acceleration of the EM algorithm," *Syst. Comput. Jpn.*, vol. 31, no. 2, pp. 10–18, Feb. 2000.

[22] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, June 1974.

[23] J. Rissanen, "Modeling by shortest data description," *Biometrika*, vol. 14, pp. 465–471, 1978.

[24] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[25] Z. Rong and M. Zwolinski, "Mutual information theory for adaptive mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 4, pp. 713–724, Apr. 2001.

[26] S. Medasani and R. Krishnapuram, "Categorization of image databases for efficient retrieval using robust mixture decomposition," *Comput. Vis. Image Understanding*, vol. 83, pp. 216–235, 2001.

[27] N. Bouguila, D. Ziou, and J. Vaillancourt, "A maximum likelihood estimation of the generalized Dirichlet mixture," Dept. Mathématiques et d'informatique, Univ. Sherbrooke, Shrebrooke, QC, Canada, 2002.

[28] C. R. Rao, *Advanced Statistical Methods in Biomedical Research*. New York: Wiley, 1952.

[29] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[30] S. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao, *Differential Geometry in Statistical Inference*. Hayward, CA: Inst. Math. Stat., 1987, vol. 10.

[31] S. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, pp. 251–276, 1998.

[32] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[33] S. C. Choi and R. Wette, "Maximum likelihood estimation of parameters of the gamma distribution and their bias," *Technometrics*, vol. 11, no. 4, pp. 683–690, Nov. 1969.

[34] T. B. Nguyen and D. Ziou, "Contextual and noncontextual performance evaluation of edge detectors," *Pattern Recognit. Lett.*, no. 21, pp. 805–816, 2000.

[35] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. H. Glasman, D. Yanker, P. Faloutsos, and G. Taubin, "The QBIC project: Querying images by content using color, texture and shape," in *Proc. SPIE Conf. Storage and Retrieval for Images and Video Databases*, 1993, pp. 173–187.

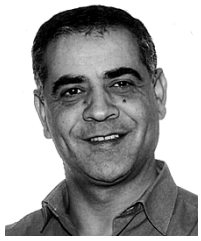
[36] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: tools for content-based manipulation of image databases," in *Proc. SPIE Conf. Storage and Retrieval for Images and Video Databases II*, Feb. 1994.

[37] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A system for region-based image indexing and retrieval," in *Proc. 3rd Int. Conf. Visual Information Systems*, 1999.

[38] J. R. Smith and S. F. Chang, "VisualSEEK: A fully automated content-based image query system," in *Proc. ACM Int. Conf. Multimedia*, Boston, Nov. 1996, pp. 87–98.
[39] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H. Zhang, "Image classification for content-based indexing," *IEEE Trans. Image Processing*, vol. 10, pp. 117–130, Jan. 2001.
[40] S. Newsman, B. Sumengen, and B. S. Manjunath, "Category-based image retrieval," in *Proc. IEEE Int. Conf. Image Processing Special Session on Multimedia Indexing, Browsing and Retrieval*, Thessaloniki, Greece, Sept. 2001.
[41] M. L. Kherfi, D. Ziou, and A. Bernardi, "Combining positive and negative examples in relevance feedback for content-based image retrieval," *J. Vis. Commun. Image Representation*, vol. 14, pp. 428–457, 2003.
[42] H. P. Graf, E. Cosatto, D. Gibbon, M. Kocheison, and E. Petajan, "Multimodal system for location heads and faces," in *Proc. 2nd IEEE Int. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 88–93.
[43] M. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 34–58, Jan. 2002.
[44] M. Yang and N. Ahuja, "Extracting gestural motion trajectory," in *Proc. 3rd Int. Conf. Automatic Face and Gesture Recognition*, 1998, pp. 10–15.
[45] J. Yang, W. Lu, and A. Waibel, "Skin-color modeling and adaption," in *Proc. Asian Conf. Computer Vision*, 1998, pp. 687–694.
[46] G. Wyszecki and W. S. Styles, *Color Science: Concepts and Methods, Quantative Data and Formulae*, 2nd ed. New York: Wiley, 1982.

Nizar Bouguila received the B.Eng. degree in computer Science from the Faculté des Sciences, Tunis, Tunisia, in 2000 and the M.Sc. degree in computer science from the Université de Sherbrooke, Sherbrooke, QC, Canada, in 2002. He is currently pursuing the Ph.D. degree at the Université de Sherbrooke.

His research interests include image processing, information retrieval, computer vision, and pattern recognition.



Djemel Ziou received the B.Eng. degree in computer science from the University of Annaba, Algeria, in 1984 and the Ph.D. degree in computer science from the Institut National Polytechnique de Lorraine (INPL), Lorraine, France, in 1991.

From 1987 to 1993, he served as a Lecturer at several universities in France. During the same time period, he was a Researcher in the Centre de Recherche en Informatique de Nancy (CRIN), Nancy, France, and the Institut National de Recherche en Informatique et Automatique (INRIA), France. Presently, he is a Full Professor at the Department of Computer Science, Université de Sherbrooke, Sherbrooke, QC, Canada. He has served on numerous conference committees as member or chair. He heads the laboratory MOIVRE and the consortium CoRIMedia, which he founded. His research interests include image processing, information retrieval, computer vision, and pattern recognition.

Jean Vaillancourt received the B.S. degree from the Université Laval, Canada, in 1981 and the Doctor of Sciences degree from Carleton University, Ottawa, ON, Canada, in 1987.

He has been the Dean of Research at the Université du Québec en Outaouais, Hull, QC, Canada, since December 2001. After starting his career as Junior Methodologist at Statistics Canada, he became Professor of mathematics in 1986 and was Associate Dean of Sciences from 1997 to 2001 at the Université de Sherbrooke, Sherbrooke, QC. During this period, he served as member of the Board of Directors of the Montreal Mathematics Research Center (CRM), the Management Committee of the Quebec Mathematical Sciences Institute (ISM), and the Board of Directors of the Statistical Society of Canada. He is currently a member of the prestigious Committee on Research Grants of the Natural Sciences and Engineering Research Council of Canada, as well as Vice-President of the Board of Directors of the Quebec Institute for the sustainable management of deciduous forests (IQAFF). He remains a productive researcher in the field of statistical estimation in the context of the numerical treatment of images. He is a founding member of the multi-university consortium CoRIMedia for automated content-based image searches, a research network funded by several Canadian peer-review agencies.