DIRICHLET-BASED PROBABILITY MODEL APPLIED TO HUMAN SKIN DETECTION

Nizar Bouguila and Djemel Ziou

DMI, Faculté des sciences
Université de Sherbrooke
Sherbrooke, Qc, Canada J1K 2R1
{bouguila,ziou}@dmi.usherb.ca

ABSTRACT

The performance of a statistical signal processing system depends in large part on the accuracy of the probabilistic model used. This paper presents a robust probabilistic mixture model based on a generalization of the Dirichlet distribution. An unsupervised algorithm for learning this mixture is given, too. The proposed approach for estimating the parameters of a Dirichlet mixture is based on the Maximum Likelihood (ML) and Fisher scoring methods. Experimental results involve human skin color modeling and its application to skin detection in images.

1. INTRODUCTION

Many signal processing techniques employ mixture models. The performance of such approach depends in large part on the accuracy of the model- that is, which density is used to model the data. 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 Decomposition (GM) a popular technique. 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 [1]. In this paper we present a generalization of the Dirichlet distribution which can be a very good choice to overcome the disadvantages of the Gaussian. The Dirichlet distribution is the multivariate generalization of the Beta distribution, which offers considerable flexibility and ease of use [2]. In fact, it was used in many image processing applications such as the identification of the blur and the restoration of noisy images [3] and image database categorization [4]. Although its flexibility the Dirichlet distribution has a very restrictive covariance structure as we will show in the next section. In this paper, we will present a method to estimate the parameters of a generalized Dirichlet mixture and test it with real image processing application.

The completion of this research was made possible thanks to Bell Canada's support through its Bell University Laboratories R&D program.

The paper is organized as follows. The next section describes the generalized Dirichlet mixture in details. In section 3, we propose a method for estimating the parameters of the mixture. In section 4, we give the complete estimation algorithm. Section 5 is devoted to experimental results. We end the paper with some concluding remarks.

2. THE GENERALIZED DIRICHLET MIXTURE

If the random vector $\vec{X} = (X_1, \dots, X_{dim})$ follows a Dirichlet distribution, the joint density function is given by:

$$p(X_1, \dots, X_{dim}) = \frac{\Gamma(\sum_{i=1}^{dim+1} \alpha_i)}{\prod_{i=1}^{dim+1} \Gamma(\alpha_i)} \prod_{i=1}^{dim+1} X_i^{\alpha_i - 1} \quad (1)$$

Where: $\alpha_i > 0 \, \forall i = 1 \dots dim{+}1$, $\sum_{i=1}^{dim} X_i < 1$, $X_{dim+1} = 1 - \sum_{i=1}^{dim} X_i$ and $0 < X_i < 1 \, \forall i = 1 \dots dim$. The covariance of the Dirichlet distribution are given by:

$$Cov(X_i, X_j) = \frac{-\alpha_i \alpha_j}{(\sum_{j=1}^{dim+1} \alpha_j)^2 (\sum_{j=1}^{dim+1} \alpha_j + 1)} \quad (2)$$

Thus, any two random variable in $\vec{X}$ are negatively correlated. In some practical cases, two random variables may be positively correlated, and hance the Dirichlet distribution will not be a reasonable choice to be a prior distribution in mixture analysis. Connor and Mosimann [5] used the concept of complete neutrality to generalize the Dirichlet distribution. Random vector $\vec{X}$ is said to be completely neutral if $(X_1, \dots, X_j)$ is independant of $(X_{j+1}, \dots, X_{dim})/V_j$ for all $j < dim$, where $V_j = 1 - X_1 - X_2 - \dots - X_j$. Let $Z_1 = X_1$ and let $Z_j = X_j / V_{j-1}$ for $j = 2, 3 \dots, dim$. When the Z_j are independant, then $\vec{X}$ is also completely neutral. Connor and Mosimann supposed that each Z_j has a beta distribution with parameters α_j and β_j, and derived the density function for the generalized Dirichlet distribution as follows:

$$p(X_1, \dots, X_{dim}) = \prod_{i=1}^{dim} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} X_i^{\alpha_i - 1} (1 - \sum_{j=1}^{i} X_j)^{\gamma_i} \quad (3)$$

for $\sum_{i=1}^{dim} X_i < 1$ and $0 < X_i < 1$ for $i = 1 \ldots dim$, where $\gamma_i = \beta_i - \alpha_{i+1} - \beta_{i+1}$ for $i = 1 \ldots dim - 1$ and $\gamma_{dim} = \beta_{dim} - 1$. Note that the generalized Dirichlet distribution is reduced to a Dirichlet distribution when $\beta_i = \alpha_{i+1} + \beta_{i+1}$. A generalized Dirichlet mixture with M components is defined as :

$$p(\vec{X}/\Theta) = \sum_{j=1}^{M} p(\vec{X}/j, \Theta_j) P(j) \tag{4}$$

where the $P(j)$ ($0 < P(j) < 1$ and $\sum_{j=1}^{dim} P(j) = 1$) are the mixing proportions and $p(\vec{X}/j, \Theta_j)$ is the generalized 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 $\vec{\alpha_j} = (\alpha_{j1}, \beta_{j1}, \ldots, \alpha_{jdim}, \beta_{jdim})$ is the parameter vector for the j^{th} population. In the following developments, we use the notation $\Theta_j = (\vec{\alpha_j}, P(j))$ for j = 1 …M.

3. MAXIMUM LIKELIHOOD ESTIMATION

With ML estimation, the problem of determining Θ becomes:

$$max_{\Theta} p(\vec{X}/\Theta) \tag{5}$$

with the constraint: $\sum_{j=1}^{M} P(j) = 1$ and $P(j) > 0 \quad \forall j \in [1, M]$. These constraints permit us to take into consideration *a priori* probabilities $P(j)$. Using Lagrange multipliers, we maximize the following function:

$$\begin{aligned} \Phi(\vec{X}, \Theta, \Lambda) &= ln(p(\vec{X}/\Theta)) + \Lambda(1 - \textstyle\sum_{i=1}^{M} P(i)) \\ &+ \mu \textstyle\sum_{j=1}^{M} P(j)^2 \end{aligned} \tag{6}$$

where Λ is the Lagrange multiplier. For convenience, we have replaced the function $p(\vec{X}/\Theta)$ in Eq. 5 by the function $ln(p(\vec{X}/\Theta))$. If we assume that we have N random vector $\vec{X_i}$ which are independent, we can write: $p(\vec{X}/\Theta) = \prod_{i=1}^{N} p(\vec{X_i}/\Theta)$, thus:

$$\begin{aligned} \Phi(\vec{X}, \Theta, \Lambda) &= \textstyle\sum_{i=1}^{N} ln(\sum_{j=1}^{M} p(\vec{X_i}/j, \Theta_j) P(j)) \\ &+ \Lambda(1 - \textstyle\sum_{j=1}^{M} P(j)) + \mu \sum_{j=1}^{M} P^2(j) \end{aligned} \tag{7}$$

In order to automatically find the number of components needed to model the mixture, we use an entropy-based criterion. Thus, the first term in Eq. 7 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. We will now try to resolve this optimization problem. To do this, we must determine the solution to the following equations: $\frac{\partial}{\partial \Theta} \Phi(\vec{X}, \Theta, \Lambda) = 0$ and $\frac{\partial}{\partial \Lambda} \Phi(\vec{X}, \Theta, \Lambda) = 0$. Calculating the derivative with respect to Θ_j, we obtain:

$$\frac{\partial}{\partial \Theta_j} \Phi(\vec{X}, \Theta, \Lambda) = \sum_{i=1}^{N} p(j/\vec{X_i}, \Theta_j) \frac{\partial}{\partial \Theta_j} ln(p(\vec{X_i}/j, \Theta_j)) \tag{8}$$

where $p(j/\vec{X_i}, \Theta_j)$ is the posterior probability.
Since $p(\vec{X_i}/j, \vec{\alpha_j})$ is independent of $P(j)$, straight forward manipulations yield:

$$P(j)^{new} = \frac{\sum_{i=1}^{N} p^{old}(j/\vec{X_i}, \vec{\alpha_j}) + 2\mu (p^2(j))^{old}}{N + 2\mu \sum_{j=1}^{M} (p^2(j))^{old}} \tag{9}$$

In order to estimate the $\vec{\alpha}$ parameters we will use the fact that each Z_i (see section 2) has a beta distribution with parameters α_i and β_i. Then, the problem of estimating the parameters of a generalized Dirichlet mixture can be reduced to the estimation of the parameters of dim beta mixtures. For this, we must maximize this equation for every dimension:

$$\Phi_Z(Z, \theta_d) = \textstyle\sum_{i=1}^{N} ln(\sum_{j=1}^{M} p_{beta}(Z_{id}/j, \theta_{jd}) P(j)) \tag{10}$$

where $0 < d < dim$, p_{beta} is the Beta distribution, $\theta_d = (\alpha_{1d}, \beta_{1d}, \ldots, \alpha_{Md}, \beta_{Md})$, $\theta_{jd} = (\alpha_{jd}, \beta_{jd})$ and $P(j)$ are the mixing parameters founded by Eq. 9. This approach which consist of considering a multidimensional density as individual 1-dimensional densities was previously used in [6] in the case of Gaussian mixture. In order to estimate the θ_d parameters we will use Fisher's scoring method. The scoring method is based on the first, second and mixed derivatives of $\Phi_Z(Z, \theta_d)$ function. During iterations, the α_{jd} and β_{jd} can become negative. In order to overcome this problem, we reparametrize, setting $\alpha^{'}_{jd} = e^{\alpha_{jd}}$ and $\beta^{'}_{jd} = e^{\beta_{jd}}$, where $\alpha^{'}_{jd}$ and $\beta^{'}_{jd}$ are unconstrained real numbers. Given a set of initial estimates, Fisher's scoring method can now be used. The iterative scheme of the Fisher method is given by the following equation:

$$\begin{pmatrix} \hat{\alpha}^{'}_{jd} \\ \hat{\beta}^{'}_{jd} \end{pmatrix}^{new} = \begin{pmatrix} \hat{\alpha}^{'}_{jd} \\ \hat{\beta}^{'}_{jd} \end{pmatrix}^{old} + V^{old} \times \begin{pmatrix} \frac{\partial \Phi_Z}{\partial \hat{\alpha}^{'}_{jd}} \\ \frac{\partial \Phi_Z}{\partial \hat{\beta}^{'}_{jd}} \end{pmatrix}^{old} \tag{11}$$

where j is the class number: $1 \leq j \leq M$ and d is the current dimension: $1 \leq d \leq dim$.
The variance-covariance matrix V is obtained as the inverse of the Fisher's information matrix $\mathbf{I}$. The information matrix

I is:

$$\mathbf{I} = \begin{pmatrix} -E[\frac{\partial^2}{\partial^2 \alpha'_{jd}} \Phi_Z(Z,\theta_d)] & -E[\frac{\partial^2}{\alpha'_{jd}\beta'_{jd}} \Phi_Z(Z,\theta_d)] \\ -E[\frac{\partial^2}{\beta'_{jd}\alpha'_{jd}} \Phi_Z(Z,\theta_d)] & -E[\frac{\partial^2}{\partial^2 \beta'_{jd}} \Phi_Z(Z,\theta_d)]] \end{pmatrix} \quad (12)$$

4. ALGORITHM

In order to make our algorithm less sensitive to local maxima, we have used some initialization schemes including the Fuzzy C-means and the method of moments (MM) [2] . Thus, our initialization method can be resumed as follows: INITIALIZATION Algorithm

1. INPUT: dim-dimensional data X_i, $i = 1, \ldots, N$ and the number of clusters M.
2. Apply the Fuzzy C-means to obtain the elements, covariance matrix and mean of each component.
3. Compute the Z_i. $Z_1 = X_1$ and $Z_j = X_j/V_{j-1}$ for $j = 2, 3 \ldots, dim$, where $V_j = 1 - X_1 - X_2 - \ldots - X_j$.
4. Apply the MM for each component j and for each dimension d to obtain the vector of parameters $\vec{\theta}_{jd}$.
5. Assign the data to clusters, assuming that the current model is correct.
6. If the current model and the new model are sufficiently close to each other, terminate, else go to 4.

With this initialization method in hand, our algorithm for estimating of generalized Dirichlet mixture can be summarized as follows:
GENERALIZED DIRICHLET MIXTURE ESTIMATION Algorithm

1. INPUT: dim-dimensional data X_i, $i = 1, \ldots, N$ and an over-specified number of clusters M.
2. Apply the INITIALIZATION Algorithm.
3. Update the $\vec{\theta_{jd}}$ using Eq. 11, $j = 1, \ldots, M$. and $d = 1, \ldots, dim$.
4. Update the $P(j)$ using Eq. 9, $j = 1, \ldots, M$.
5. If $p(j) < \epsilon$ discard component j, go to 3.
6. If the convergence test is passed, terminate, else go to 3.

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 Eq. 7 of each iteration. Convergence tests could involve testing the stabilization of the $\vec{\alpha_j}$ or the value of the maximum likelihood function.

	α_1	β_1	α_2	β_2	$P(j)$
Class1	55.39	64.84	32.50	25.65	0.443
Class2	63.96	102.53	124.36	115.34	0.557

Table 1. Mixture of Dirichlet Skin Color Model

	α_1	β_1	α_2	β_2	$P(j)$
Class1	26.45	25.34	4.29	3.15	0.120
Class2	16.13	42.74	8.22	3.95	0.560
Class3	31.97	18.07	0.36	0.05	0.087
Class4	124.16	107.83	19.80	6.55	0.133

Table 2. Mixture of Dirichlet non-skin Color Model

5. EXPERIMENTAL RESULTS

The 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, face recognition, and gesture recognition. 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 [7]. Thus, the common RGB 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 shown here: $r1 = \frac{r}{r+g+b}$, $g1 = \frac{g}{r+g+b}$ and $b1 = \frac{b}{r+g+b}$. 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 10867932 (skin color pixels), where each sample consists of two values $(r1, g1)$. We used our algorithm to estimate the parameters of the Dirichlet mixture and we found 2 components (see Table 1). The estimated density function is shown in Fig. 1. To improve our application, we train for non-skin model, too. The training set contains 23256874 non-skin color pixels. Table 2 shows the estimated parameters of this dataset and we obtain 4 classes.

Given an image, a segmentation was performed to obtain homogenous regions. Each pixel was classified as skin color if its probability measure to be a skin was above a threshold and its probability measure to be a non-skin was below another threshold. Each region was then recognized

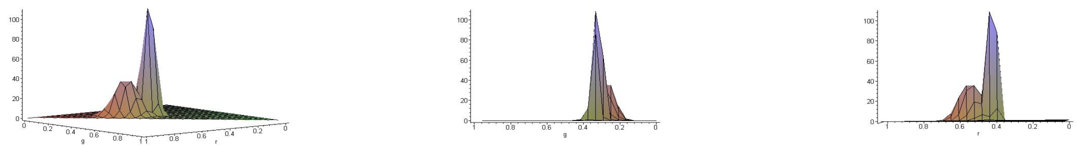

Fig. 1. Estimated Density Function for the skin model viewed from different angles.

(a) original image (b) Extraction by a mixture of Gaussian (c) Extraction by a mixture of Dirichlet

Fig. 2. Original image and results of skin detection in different cases.

as a skin area if most of the pixels in the region had a high probability of being skin color. Figure 2 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 [7]. 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.

6. CONCLUSION

In this paper, we have introduced a new mixture, based on a generalization of the Dirichlet distribution. The generalized 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. The experiments involved the detection of human skin color. From the results, we can say that the Generalized Dirichlet mixture has good modeling capabilities.

7. REFERENCES

[1] Raftery, A. E. and Banfield, J. D., "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, vol. 49, pp. 803–821, 1993.

[2] N. Bouguila, D. Ziou and J. Vaillancourt, "Novel Mixtures Based on the Dirichlet Distribution: Application to Data and Image Classification," in *Proceedings of MLDM*, Leipzig, Germany, 2003, pp. 172–181, Lecture Notes in Computer Science 2734 Springer.

[3] R. Molina, A. K. Katsaggelos and J. Abad and J. Mateos, "A Bayesian Approach to Blind Deconvolution Based on Dirichlet Distributions," in *Proceedings of ICASSP*, Munich, Germany, 1997, pp. 2809–28012.

[4] N. Bouguila, D. Ziou and J. Vaillancourt, "A Probabilistic Multimedia Summarizing Based on the Dirichlet Mixture," in *Supplementary Proceedings of ICANN/ICONIP*, Istanbul, Turkey, 2003, pp. 150–154.

[5] R. J. Connor and J. E. Mosimann, "Concepts of Independence for Proportions With a Generalization of the Dirichlet Distribution," *Journal of the American Statistical Association*, vol. 39, pp. 1–38, 1977.

[6] K. Popat and R. W. Picard, "Cluster-Based Probability Model Applied To Image Restoration and Compression," in *Proceedings of ICASSP*, Aldelaide, Australia, 1994, pp. 381–384.

[7] M. Yang, D. J. Kriegman and N . Ahuja, "Detecting Faces in Images: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, January 2002.