Humanskin color detection plays an important role in the applications of skin segmentation, face recognition, and tracking. To build a robust human skin color classifier is an essential step. This paper presents a classifier based on beta mixture models (BMM), which uses the pixel values in RGB space as the features. We propose a Bayesian estimation method based on the variational inference framework to approximate the posterior distribution of the parameters in the BMM and take the posterior mean as a point estimate of the parameters. The well-known Compaq image database is used to evaluate the performance of our BMM based classifier. Compared to some other skin color detection methods, our BMM based classifier shows a better recognition performance.

1. INTRODUCTION

Human skin color detection plays an important and effective role in the applications of skin segmentation, face recognition, and tracking problems [1]. Many heuristic and pattern recognition methods have been proposed in the past decades. Different color spaces and methods have been evaluated and compared [2, 3, 4]. From the feature point of view, several color spaces could be used for skin detection such as Red-Green-Blue (RGB) [5, 2, 6], CIE-xy, YIQ, YCbCr [7], Tint-Saturation-Luminance (TSL) [8], and Hue-Saturation-Value/Intensity (HSV/HSD) [9]. From the standpoint of classification methods, the explicitly defined threshold (e.g. normalized R/G ratio [2]), non-parametric probabilistic model [2, 6], and parametric probabilistic model [6, 7] are the most efficient methods applied in the area of skin color detection.

As the RGB space is the mostly used space, we take the pixel value in RGB space as features. Since it has been shown [10, 11] that beta mixture models (BMM) can model data with compact range better than Gaussian mixture models and the pixel value is in [0, 255], we apply BMM to model the skin pixel distribution in RGB space. With the principles of the variational inference (VI) framework [12, 13, 14, 15], we propose a Bayesian estimation algorithm to estimate the parameter distributions. By applying a set of non-linear approximations, the posterior distribution of the parameters in the BMM is obtained. The posterior mean is considered as the point estimate of the parameters.

The BMM-based classifier is trained and tested with the well-known Compaq image database [6]. The Receiver Operating Characteristic (ROC) [16] is used to evaluate the performance of the BMM classifier. This paper presents a method based on the VI framework for the Bayesian estimation of the parameters in the BMM and shows the skin/non-skin color detection results with the BMM classifier.

2. SKIN COLOR MODEL

In the past decades, different color models were proposed for various properties. The color models used nowadays are mostly oriented to applications. The RGB color space originated from the CRT display and describes the color in terms of three primary color channels: red (R), green (G), and blue (B). This is the mostly used space for image display and storage. Fig. 1 shows a 24-bit RGB color image with a human face and the corresponding (normalized) histogram for each color channel. The pixel values of each color channel are linearly compressed to [0, 1] by \( x_c := x_c/255, c \in \{r, g, b\} \).

Although the RGB space is sensitive to the luminance, and the images of one object taken in different environments have diverse characteristics on the RGB space [2], the mixture of probabilistic models (e.g. skin probability map in [2, 6], Gaussian Mixture Models (GMM) in [6]) can still model the distribution of the pixel values in RGB space efficiently. By building the three-dimensional probabilistic skin and non-skin models in RGB space, the distribution of both skin and non-skin pixels can be represented in a probabilistic way. A pixel could be classified as a skin pixel or a non-skin pixel in an optimized way by utilizing the Bayesian classifier [17]. Several studies used parametric or non-parametric techniques to model the pixel value distribution. However, non-parametric techniques, such as histogram based models, need a large amount of training data and have high computational cost. With the parametric technique, we can obtain the parameters of the model, and this is more convenient in practical problems.

From Fig. 1 we can observe that the pixel values in the RGB space are in a compact range for each color channel (in the 24-bit RGB image, the range is [0, 255] and can be linearly compressed to the range [0, 1]). Furthermore, the distribution of the skin color pixel is skewed in each channel. Indeed the conventional GMM can model any distribution shape with a proper number of components. Since it was shown that the BMM outperforms GMM with the same number of components in gray image classification [11], and the color image in RGB space can be considered as the composition...
of three gray images, one from each color channel, the BMM is a more reasonable choice to model the distribution of skin color. The correlations among the three channels can be represented by mixture models.

3. Beta Mixture Models

The beta distribution is a family of continuous probability distributions defined on the interval [0, 1] with two positive real parameters. The probability density function of the beta distribution is

$$\text{Beta}(x; u, v) = \frac{1}{\text{Beta}(u, v)} x^{u-1} (1 - x)^{v-1},$$

where \( \text{Beta}(u, v) \) is the beta function

$$\text{Beta}(u, v) = \frac{\Gamma(u) \Gamma(v)}{\Gamma(u + v)}$$

and \( \Gamma(\cdot) \) is the gamma function defined as \( \Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \).

Multivariate data are in most cases statistically dependent. However, for any random vector \( x \) consisting of \( L \) elements, the dependencies among the elements \( x_1, \ldots, x_L \) can be represented by a mixture model, even if each specific mixture component can only generate vectors with statistically independent elements. Therefore, we define the multivariate BMM as

$$f(x; \Pi, U, V) = \sum_{i=1}^L \pi_i \text{Beta}(x; u_i, v_i),$$

where \( \Pi = \{ \pi_1, \ldots, \pi_L \} \), \( U = \{ u_1, \ldots, u_L \} \), and \( V = \{ v_1, \ldots, v_L \} \) denote the parameter vectors of the \( i \)th mixture component and \( u_i, v_i \) are the (scalar) parameters of the beta distribution for element \( x_i \). For representing the skin color distribution, each observation \( x \) is a three-dimensional vector with \( L = 3 \) and each element \( x_i \) is in the range [0, 1]. In the following sections, we will use observed pixel data \( X = \{ x_1, \ldots, x_N \} \) approximate the posterior distribution \( f(U, V, \Pi | X) \) via the variational inference (VI) framework in section 3.3 and the algorithm for Bayesian estimation will be listed in section 3.4.

3.1 Bayesian estimation and conjugate prior

We make a Bayesian estimate of the parameters in the BMM. An important step in the Bayesian estimation is to find the conjugate prior \( f(P) \) such that the posterior distribution \( f(X|P) \) has the same form as \( f(P) \). The conjugate prior distribution to the beta distribution in (1) is

$$f(u, v) = \frac{1}{C} \left[ \frac{\Gamma(u + v)}{\Gamma(u) \Gamma(v)} \right] e^{-\alpha_0 (u-1)} e^{-\beta_0 (v-1)},$$

where \( \alpha_0, \beta_0, \nu_0 \) are free positive parameters and \( C \) is a normalization factor (a function of \( \alpha_0, \beta_0, \nu_0 \)) such that

$$\int_0^\infty \int_0^\infty f(u, v) du dv = 1.$$

Then we obtain the posterior distribution of \( u, v \) as (with \( N \) i.i.d. scalar observations \( x = \{ x_1, \ldots, x_N \} \))

$$f(u, v|x) = \frac{f(x|u, v) f(u, v)}{\int_0^\infty \int_0^\infty f(x|u, v) f(u, v) du dv} = \frac{\Gamma(u + v)}{\Gamma(u) \Gamma(v)} e^{-\alpha_N (u-1)} e^{-\beta_N (v-1)}$$

$$= \int_0^\infty \int_0^\infty \frac{\Gamma(u + v)}{\Gamma(u) \Gamma(v)} e^{-\alpha_N (u-1)} e^{-\beta_N (v-1)} du dv,$$

where

$$\alpha_N = \alpha_0 - \sum_{n=1}^N \ln x_n,$$

$$\beta_N = \beta_0 - \sum_{n=1}^N \ln (1 - x_n),$$

$$\nu_N = \nu_0 + N.$$

3.2 Variational inference and factorized approximation

Analytically, we cannot find a closed-form expression for the posterior distribution in (5) due to the computationally intractable integration expression in the denominator. Some stochastic techniques (e.g. Gibbs sampling [10]) can be used to calculate the posterior distribution numerically. We propose a method based on the VI framework [12] in this paper. According to the VI framework, the posterior density function of variable \( Z \) given the observation \( X \) (i.e. \( f(Z|X) \)) is approximated by \( g(Z) \). We decompose the log likelihood \( \ln f(X) \) as

$$\ln f(X) = \mathcal{L}(g) + KL(g \parallel f),$$

where

$$\mathcal{L}(g) = \int g(Z) \ln \frac{f(X, Z)}{g(Z)} dZ$$

and \( KL(g \parallel f) \) is the Kullback-Leibler (KL) divergence defined as

$$KL(g \parallel f) = - \int g(Z) \ln \frac{f(X|Z)}{g(Z)} dZ.$$

Since the KL divergence is a non-negative measurement, to maximize the lower bound \( \mathcal{L}(g) \) is equivalent to minimize the KL divergence. Especially, when \( g(Z) \) is equal to \( f(X|Z) \), the KL divergence vanishes and the lower bound reaches the true log likelihood \( \ln f(X) \). If the target distribution is analytically intractable, some approximations can be used to achieve tractability with the factorized approximation (FA) method [12, 18].

The FA method partitions the variable \( Z \) into disjoint parts \( \{ Z_m \}, \ m = 1, \ldots, M \) and decomposes the distribution as

$$g(Z) = \prod_{m=1}^M g_m(Z_m).$$

Amongst all the distributions having the form in (9), we need to seek a distribution \( g(Z) \) that drives the lower bound \( \mathcal{L}(g) \) to be largest. By substituting (9) into (7) and denoting \( g_m(Z_m) \) by \( g_m \) simply, we obtain

$$\mathcal{L}(g) = \int g_m \ln f(X, Z) - \sum_{m=1}^M \ln g_m \; dZ_m$$

$$= \int g_m \ln f(X, Z) \prod_{m \neq m} g_m dZ_m \; dZ_m$$

$$- \int g_m \ln g_m dZ_m + \text{const}. \equiv g_m \ln f(X, Z) dZ_m - \int g_m \ln g_m dZ_m + \text{const}.$$

1To prevent the infinity quantity in the practical implementation, we assign \( \varepsilon_1 \) to \( x_n \) when \( x_n = 0 \) and \( 1 - \varepsilon_2 \) to \( x_n \) when \( x_n = 1 \). Both \( \varepsilon_1 \) and \( \varepsilon_2 \) are slightly positive real numbers.
Furthermore, by introducing the Dirichlet distribution as the prior
and
where
For each observation
is nonnegative variables, we assign the
So that the KL divergence in (8) is almost equal to

By recognizing that the first two integrals in the final line of (10) is
a negative KL divergence between
and
, we can maximize
with respect to any possible form of
, while keeping
fixed, by minimizing the KL divergence. The optimal value occurs when
, which gives us the optimal solution to
as

3.3 Factorized approximation for BMM
The prior of the beta distribution is analytically intractable. With the
principles of FA, the conjugate prior in (4) can be approximated as

Since both
and
are nonnegative variables, we assign the

where

For each observation
the corresponding
is an indication vector with one element equals to
and the rest equal to
, where
. The
ith observation is generated from the
th component in the BMM. The conditional distribution of
is
and
given latent variables
is

The logarithm of the joint distribution function of
and
is

\[ \mathcal{L}(X, Z) = \ln f(X, Z) \]

\[ = \sum_{n=1}^{N} \sum_{i=1}^{L} \left\{ \ln \pi_i + \ln \left( \frac{\Gamma(u_i + v_i)}{\Gamma(u_i) \Gamma(v_i)} \right) \right\} \]

\[ + \sum_{n=1}^{N} \left\{ \ln c_i - 1 \right\} \ln x_{in} + \left( v_i - 1 \right) \ln \left( 1 - x_{in} \right) \]

3.4 Algorithm of Bayesian estimation
The latent variables we have now are
and
with the hyper-
parameters
and
. The optimal distribution for
and
are obtained by taking the expected value of
as (element-wise)

Obviously, the expectations in the RHS of (18) could not lead to
a closed-form expression. The second-order Taylor expansion of
in terms of
and
can be proven to be a lower bound [19]. The expectation of this lower bound can yield the
optimal solution to
asymptotically. The update equations for the hyper-parameters of
and
are listed as follows (element-wise):

\[ \alpha_i^* = \alpha_i + \sum_{n=1}^{N} E \{ z_{in} \} \]

\[ \beta_i^* = \beta_i + \sum_{n=1}^{N} E \{ z_{in} \} \ln x_{in} \]

\[ \alpha_i^* = \alpha_i + \sum_{n=1}^{N} E \{ z_{in} \} \ln x_{in} \]

\[ \beta_i^* = \beta_i + \sum_{n=1}^{N} E \{ z_{in} \} \ln \left( 1 - x_{in} \right) \]

where (the estimation of
is in (20) in the next page)

\[ \bar{\pi} = \frac{\mu}{\alpha}, \quad \bar{\nu} = \frac{\nu}{\beta} \]

\[ E \{ z_{in} \} = \frac{\rho_{ni}}{\sum_{k=1}^{K} \rho_{nk}} \]

\[ E_u \{ \ln u \} = \psi(\mu) - \ln \alpha \]

\[ E_v \{ \ln v \} = \psi(\nu) - \ln \beta \]

\[ \left( \ln u - \ln \bar{\pi} \right)^2 = \psi(\mu) - \ln \alpha \]

To start the iterations, the values of
and
are chosen such that the prior distributions are assigned with
non-informative distributions (flat broad distribution). The parameters
for the Dirichlet distribution
and
are assigned with a
small value (i.e., 0.001) to ensure the number of mixture components is
controlled by the data. By updating the hyper-parameters
and
recursively in order, the algorithm will converge
so that the KL divergence in (8) is almost equal to
. Compared to the
conventional expectation maximization (EM) based maximum
likelihood estimation (MLE), this Bayesian estimation can prevent
overfitting and estimate the effective number of mixture components
automatically. The latent variables are all unimodally distributed
and the posterior distributions are highly peaked. Considering the
posterior mean as the point estimate of
and
, we take
and
. More details about the
derivations of this algorithm can be found in [19].
\[
\ln \rho_{nli} \approx E[\ln 0] + \sum_{l=1}^{L} (\ln u_{nli} - 1) \ln x_{lni} + (\ln \tau_{nli} - 1) \ln (1 - x_{lni}) \\
+ \sum_{l=1}^{L} \left\{ \ln \left[ \frac{\psi(\tau_{nli} + \tau_{nli})}{\psi(\tau_{nli})} \right] + \psi(\tau_{nli} + \tau_{nli}) - \psi(\tau_{nli}) \right\} + \sum_{l=1}^{L} \left( \ln \left[ \frac{\psi(\tau_{nli} + \tau_{nli})}{\psi(\tau_{nli})} \right] + \psi(\tau_{nli} + \tau_{nli}) - \psi(\tau_{nli}) \right) \right\}
\]

For each evaluation round, the Compaq database was partitioned randomly into a training sub-database and a test sub-database. Each sub-database consists of skin and non-skin images. We randomly selected a training set from the training sub-database with 500,000 skin color pixels and 1,500,000 non-skin color pixels. Also, a test set with the same size was randomly drawn from the test sub-database. These pixels were selected randomly and labelled. For the training procedure, the labelled pixels were applied to classify the pixels in the test set. The overall performance was good and both the skin area of the black lady and the white couple were detected. The missing part on the man’s left arm and misclassified part of the hair in a relative brighter background indicates that the illumination of the image has influence on the detection result.
Our BMM classifier reported TPR with 90% of randomness, we executed when the TPR is equal to one minus the FPR. To prevent the effect of randomness, we executed 60 rounds of the train-test procedures mentioned above. The mean values of the TPR, the FPR and the accuracy rate are reported.

Some other classifiers based on the pixel probabilistic model were also analyzed with the Compaq database in the previous literature [5, 2, 8, 20, 6, 7]. They reported different classification scores with different methods and with different color spaces (e.g., RGB, YCbCr). The best classification scores from some previous studies and our BMM classifier are listed in table 1. As mentioned in [21], if the transformation from one color space to another is invertible and provided the optimal skin classifier for the color space is used, the differences of the color space does not have influence on the classifier’s performance. Although different methods used different separations of the database and employed different learning strategies, it is still interesting to compare our ROC curve with the results in table 1. In Fig. 3, all the results listed in table 1 are in the southeast side of the ROC curve, which means that our BMM classifier outperforms all the other methods. For some other tasks of classification with data in a compact range, the BMM is probably also a promising model.

6. CONCLUSION

This paper presented a BMM-based classifier for the task of human skin/non-skin color detection. A Bayesian estimation algorithm for the parameters was proposed. With the variational inference framework and a set of non-linear approximations, the posterior distributions for the BMM parameters were approximated and the posterior mean was used as the point estimate of the parameters.

The BMM classifier was applied to the well-known Compaq image database, using the pixel values in the RGB color space as the features. The overall detection performance is good. In comparison with other methods based on pixel probabilistic models, our BMM classifier outperforms the previous results.

REFERENCES


