Gaussian Mixture Model for Human Skin Color and Its Applications in Image and Video Databases

Ming-Hsuan Yang and Narendra Ahuja

Beckman Institute and Department of Electrical and Computer Engineering

University of Illinois at Urbana-Champaign, Urbana, IL 61801

ABSTRACT

This paper is concerned with estimating a probability density function of human skin color using a finite Gaussian mixture model whose parameters are estimated through the EM algorithm. Hawkins' statistical test on the normality and homoscedasticity (common covariance matrix) of the estimated Gaussian mixture models is performed and McLachlan's bootstrap method is used to test the number of components in a mixture. Experimental results show that the estimated Gaussian mixture model fits skin images from a large database. Applications of the estimated density function in image and video databases are presented.

Keywords: Gaussian Mixture Model, EM algorithm, Skin Color, Skin Detection

1. INTRODUCTION

Human skin color has been used and proved to be an effective feature in many applications from human face detection to hand tracking. However, most studies use either simple thresholding or a single Gaussian distribution to characterize the properties of skin color. Although skin colors of different races fall into a small cluster in normalied RGB or HSV color space, we find that a single Gaussian distribution is neither sufficient to model human skin color nor effective in general applications. Further, previous approaches use small collections of images to estimate the density function but do not validate the models by verifying the statistical fit of the chosen model to the data. This paper is aimed at estimating the properties of human skin color using the Michigan face database (http://www.engin.umich.edu/faces/) which consists of 2,447 images of human faces from different ethnic groups. More than 9.5 million skin color pixels are used to build a skin color model.

Within the framework of the Gaussian mixture model, two fundamental questions can be asked about the validity of the model. First, on the question of normality and homoscedasticity of the specified components, a test proposed by Hawkins[2] can be used to test both properties simultaneously. The second question is about of the number of Gaussian components used in the mixture. A test for the smallest number of components compatible with the data can be formulated in terms of the likelihood ratio criterion, although unfortunately it does not have its usual asymptotic distribution under the null hypothesis. Basford and McLachlan proposed the use of several measures using the bootstrap method[1] to quantify the strength of the obtained mixture model. In this paper, these two tests are carried out to formally support the estimated model.

This paper is organized as follows. We give a brief description of Gaussian mixture model and parameter estimation in Section 2. Two statistical tests on the normality and the number of are reviewed in Section 3. Section 4 provides the experimental results and applications on the obtained model are presented in Section 5. We conclude this paper with comments in 6.

2. PROPOSED MIXTURE MODEL

In this section, we give a brief description of a finite Gaussian mixture model and then use it to model the distribution of skin color pixels.

Correspondence: E-mail: mhyang@stereo.ai.uiuc.edu, ahuja@stereo.ai.uiuc.edu, WWW: http://uirvli.ai.uiuc.edu/lab/

2.1. Gaussian Mixture Model

Under the finite mixture models to be fitted in this paper, each skin color pixel $\mathbf{x}$ can be viewed as arising from a super population G which is a mixture of a finite number, g, of populations $G_1, \ldots, G_g$ in some proportions $\pi_1, \ldots, \pi_g$, respectively, where

$$\sum_{i=1}^{g} \pi_i = 1 \text{ and } \pi_i \geq 0$$

The probability density function (p.d.f) of an observation $\mathbf{x}$ (of dimensionality d) in the finite mixture form is

$$p(\mathbf{x};\phi) = \sum_{i=1}^{g} \pi_i \cdot p_i(\mathbf{x};\theta) = \sum_{i=1}^{g} \pi_i \cdot p(\mathbf{x}|i;\theta) = \sum_{i=1}^{g} \pi_i \cdot \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_i|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T(\mathbf{\Sigma}_i)^{-1}(\mathbf{x}-\mu_i)}$$

where π_i is the mixing parameter, $p_i(\mathbf{x};\theta)$ is the p.d.f. corresponding to G_i, and θ denotes the vector of all unknown parameters associated with the parametric forms adopted for these g component densities. For the case of multivariate Gaussian components, θ consists of the elements of the mean vectors μ_i, and the distinct elements of the covariance matrices $\mathbf{\Sigma}_i$ for $i = 1, \ldots, g$. The vector

$$\phi = (\pi^T, \theta^T)^T$$

of all unknown parameters belongs to some parameter space Ω and is estimated using the EM algorithm[3] in Section 2.2.

2.2. Estimating Parameters Using EM Algorithm

Various procedures have been developed to determine the parameters of a Gaussian mixture model from a set of data. Here we briefly describe the EM algorithm for parameter estimation. A review on parameter estimation using maximum likelihood techniques can be found in a paper by Render and Walker.[3]

For the case of Gaussian components, the mixture density contains the following adjustable parameters: π_i, μ_i and $\mathbf{\Sigma}_i$ (where $i = 1, \ldots, g$). The negative log-likelihood for the data set is given by

$$E = -\ln \mathcal{L} = -\sum_{j=1}^{n} \ln p(\mathbf{x}_j) = -\sum_{j=1}^{n}(\sum_{i=1}^{g} \pi_i p(\mathbf{x}_j|i)) \tag{1}$$

which can be regarded as an error function. Maximizing the likelihood $\mathcal{L}$ is then equivalent to minimizing E.

The EM algorithm begins by making some initial guess for the parameters of the Gaussian mixture model, which we call the "old" parameter values. We can then evaluate the new parameters using the following equations. This will give a revise estimate for the parameters, which we call the "new" parameter values, for which we might hope the value of error function is smaller. These parameter values then become the "old" values, and the process is repeated.

We write the change in error function when we replace the old parameter values by the new values in the form

$$\Delta^{t+1} = E^{t+1} - E^t = -\sum_{j}^{n} \ln(\frac{p^{t+1}(\mathbf{x}_j)}{p^t(\mathbf{x}_j)}) \tag{2}$$

where $p^{t+1}(\mathbf{x})$ denotes the probability density evaluated using the new values for the parameters, while $p^t(\mathbf{x})$ represents the density evaluated using the old parameter values.

By setting the derivatives of Δ^{t+1} to zero (see Render and Walker's review[3] for details), we obtain the following update equations for the parameters of the mixture model:

$$\mu_i^{t+1} = \frac{\sum_{j=1}^{n} p^t(i|\mathbf{x}_j)\mathbf{x}_j}{\sum_{j=1}^{n} p^t(i|\mathbf{x}_j)} \tag{3}$$

$$\mathbf{\Sigma}_i^{t+1} = \frac{\sum_{j=1}^{n} p^t(i|\mathbf{x}_j)||\mathbf{x}_j - \mu_i^t||^2}{\sum_{j=1}^{n} p^t(i|\mathbf{x}_j)} \tag{4}$$

$$\pi_i^{t+1} = \frac{1}{n}\sum_{j=1}^{n} p^t(i|\mathbf{x}_j) \tag{5}$$

where

$$p^t(i|\mathbf{x}_j) = \frac{p^t(\mathbf{x}_j|i)\pi^t(i)}{p^t(\mathbf{x}_j)} \tag{6}$$

3. STATISTICAL TESTS

In order to determine the goodness-of-fit of a Gaussian mixture, we use Hawkins' method to test the normality and homoscedaticity of the mixture models and the bootstrap method to test the number of components required in the model. These methods are briefly described in the following sections. For details, see the treatments in Titterington[4] and McLachlan.[1]

3.1. Hawkins' Test for Normality and Homoscedasticity

Under the normality assumption with equal covariance matrices (homoscedasticity) for the component distribution in a Gaussian mixture, we have a null hypothesis

$$H_0 : \mathbf{x} \sim N(\mu_i, \mathbf{\Sigma}) \text{ in } G_i \ (i = 1, \ldots, g) \tag{7}$$

on the basis of the classified data $\mathbf{y}_{ij}$ $(1 = 1, \ldots, g; j = 1, \ldots, m_i)$. Let

$$\bar{\mathbf{y}}_i = \sum_{j=1}^{m_i} \mathbf{y}_{ij}/m_i \tag{8}$$

and

$$\mathbf{S}_i = \sum_{j=1}^{m_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T/(m_i - 1) \tag{9}$$

for $i = 1, \ldots, , g$ and let

$$\mathbf{S} = \sum_{i=1}^{g} (m_i - 1)\mathbf{S}_i/(m - g) \text{ where } m = \textstyle\sum_{i=1}^{g} m_i \tag{10}$$

The Mahalanobis squared distance between $\mathbf{y}_{ij}$ and $\bar{\mathbf{y}}_i$ with respect to $\mathbf{S}$ is denoted by $D(\mathbf{y}_{ij}, \bar{\mathbf{y}}_i; \mathbf{S})$, so that

$$D(\mathbf{y}_{ij}, \bar{\mathbf{y}}_i; \mathbf{S}) = (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T \mathbf{S}^{-1} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)$$

The notation for the Mahalanobis squared distance between two vectors with respect to some positive definite symmetric matrix is to be used in this paper. For a given population G_i, the test considers the Mahalanobis squared distance between each $\mathbf{y}_{ij} (j = 1, \ldots, m_i)$ and the mean of the sample from G_i, but where each $\mathbf{y}_{ij}$ is deleted from the sample if it severely contaminates the estimates of the mean and covariance matrix of $G_i (i = 1, \ldots, g)$. Accordingly, the Mahalanobis squared distance

$$D(\mathbf{y}_{ij}, \bar{\mathbf{y}}_i; \mathbf{S}_{(ij)}) \tag{11}$$

is computed where $\bar{y}_{i(ij)}$ and $\mathbf{S}_{(ij)}$ denote the resulting values of $\bar{\mathbf{y}}_i$ and $\mathbf{S}$ after the deletion of $\mathbf{y}_{ij}$ from the data. It follows that under H_0,

$$c(m_i, \nu) D(\mathbf{y}_{ij}, \bar{y}_{(ij)}; \mathbf{S}_{(ij)}) \tag{12}$$

is distributed according to an F distribution with p and $\nu = m - g - p$ degrees of freedom, where

$$c(m_i, \nu) = ((m_i - 1)\nu)/((m_i p)(\nu + p - 1))$$

To avoid the recomputation of $\bar{\mathbf{y}}_i$ and $\mathbf{S}$ after the deletion of each $\mathbf{y}_{ij}$ from the data, it was shown that (12) can be computed using the result that

$$c(m_i,\nu)D(\mathbf{y}_{ij},\bar{y}_{i(ij)};\mathbf{S}_{(ij)}) = \frac{(\nu m_i/p)D(\mathbf{y}_{ij},\bar{\mathbf{y}}_i;\mathbf{S})}{(\nu+p)(m_i-1)-m_i D(\mathbf{y}_{ij},\bar{\mathbf{y}}_i;\mathbf{S})} \tag{13}$$

Let a_{ij} denote the area to the right of the observed value of (13) under the $F_{p,\nu}$ distribution, then under H_0 we have that

$$H_{0i} : a_{i1},\ldots,a_{im_i} \overset{iid}{\sim} U(0,1) \ (i=1,\ldots,g) \tag{14}$$

holds approximately, where $U(0,1)$ denotes the uniform distribution on the unit interval. The result of (14) is only an approximate as for a given i, the a_{ij} are only independent exactly as $m_i \to \infty$, due to the presence of the estimates of μ_i and $\mathbf{\Sigma}$ in the formation of (11). Hawkins[2] has reported empirical evidence which suggests that subsequent steps in his test which treat (14) as if it were an exact result should be approximately valid.

A close inspection of the tail areas a_{ij} including $Q-Q$ plots can be used to detect departures from the g hypotheses H_{0i}, and hence from the original hypothesis H_0. In conjunction with this detailed analysis, Hawkins advocated the use of Anderson-Darling statistic for assessing (14), as this statistic is particularly sensitive to fit in the tails of the distribution. The Anderson-Darling statistic can be computed for the sample of m_i values $a_{ij}(j=1,\ldots,m_i)$ by

$$W_i = -m_i - \sum_{j=1}^{m_i}(2j-1)(\log a_{i(j)} + \log(1-a_{i(m_i-j+1)})/m_i \ (i=1,\ldots,g)$$

where for each i, $a_{i(1)} \leq a_{i(2)} \leq \cdots \leq a_{i(m_i)}$ denote the m_i order statistics of the a_{ij}. In the asymptotic resolution of each W_i into standard normal variances W_{ik} according to

$$W_i = \sum_{k=1}^{\infty} W_{ik}^2/(k(k+1)) \ (i=1,\ldots,g)$$

where attention is focused on the first two components

$$W_{i1} = -(3/m_i)^{1/2}\sum_{j=1}^{m_i}(2a_{i(j)}-1)$$

and

$$W_{i2} = -(5/m_i)^{1/2}\sum_{j=1}^{m_i}\frac{1}{2}(3(2a_{i(j)}-1)^2-1)$$

Similarly, the Anderson-Darling statistic W_T and its first two components W_{T1} and W_{T2} can be computed for the single sample where all the a_{ij} are combined.

Some simulations performed by Hawkins[2] suggest that the size of the test will be approximately 0.1 if (14) is rejected if any W_i exceeds 2.5 (the asymptotic 95th percentile) or any W_{ik} (k=1,2) exceeds 2.54 in magnitude.

3.2. Statistical Test for the Number of Components

The "bootstrap" method, first introduced by Efron,[5] is a powerful technique which permits the variability in a random quantity to be assessed using just the data at hand. An estimate $\hat{F}$ of the underlying distribution is formed from the observed sample. Conditional on the later, the sampling distribution of the random quantity of interest with F replaced by $\hat{F}$, defines its so-called bootstrap distribution, which provides an approximation to its true distribution. It is assumed that $\hat{F}$ has been so formed that the stochastic structure of the model has been preserved. Usually, it is impossible to express the bootstrap distribution in simple form, and it must be approximated by Monte Carlo methods whereby pseudo-random samples (bootstrap samples) are drawn from $\hat{F}$. The bootstrap method can be implemented nonparametrically by using the empirical distribution function constructed from the original data. In the following application the bootstrap is applied in a parametric framework in which the bootstrap samples are drawn from the parametric likelihood estimate of the underlying distribution function.

The log likelihood ratio statistic for the test of the null hypothesis $H_0 : g = g_1$ groups versus the alternative $H_1 : g = g_2$ can be bootstrapped as follows. Proceeding under H_0, a bootstrap sample is generated from a mixture of g_1 groups where, in the specified from their densities, unknown parameters are replaced by their likelihood estimates formed under H_0 from the original sample. The value of $-2\log\lambda$ is computed fro the bootstrap sample after fitting mixture models for $g = g_1$ and $g = g_2$ in turn to it. This process is repeated independently a number of times K, and the replicated values of $-2\log\lambda$ formed from the successive bootstrap samples provide an assessment of the bootstrap, and hence of the true, null distribution of $-2\log\lambda$. It enables an approximation to be made to the achieved level of significance P corresponding to the value of $-2\log\lambda$ evaluated from the original sample.

If a very accurate estimate of the P-value were required, then K may have to be very large. Indeed for less complicated models than mixtures, Efron et. al.[5] have shown that whereas 50 to 100 bootstrap replications may be sufficient fro standard error and bias estimation, a larger number, say 350, are needed to give a useful estimate of a percentile or P-value, and many more for a highly accurate assessment. Usually, however, there is no interest in estimating a P-value with high precision. Even with a limited replication number K, the amount of computation involved is still considerable, in particular for values of g_1 and g_2 not close to one.

In the narrower sense where the decision to be made concerns solely the rejection or retention of the null hypothesis at a specified significance level α, Aitkin[6] noted how analogous to the Monte Carlo test produce of Hope,[7] the bootstrap replications can be used to provide a test of approximate size α. The test which rejects H_0 if $-2\log\lambda$ for the original data is greater than the hth smallest of its K bootstrap replications, has size

$$\alpha = 1 - j/(K+1) \tag{15}$$

approximately. For if any difference between the bootstrap and true null distribution of $-2\log\lambda$ is ignored, then the original and subsequent bootstrap values of $-2\log\lambda$ can be treated as the realizations of a random sample of size $K+1$, and the probability that a specified member is greater than j of the others is $1 - j/(K+1)$.

The result of (15) applies to the unconditional size of the test and not to its size conditional on the K bootstrap values of $-2\log\lambda$. For a specified significance level α, the values of j and K can be appropriately chosen according to (15). For example, for $\alpha = 0.05$, the smallest value of K needed is 19 with $j = 19$. As cautioned above on the estimation of the P-value for the likelihood ratio test, K needs to be very large to ensure an accurate assessment. For the 0.05 level test of a single normal population versus a mixture of two normal homoscedastic populations, McLachlan[8] performed some simulations to demonstrate the improvement in the power as K increased from 19 through 39 to 99.

4. EXPERIMENTAL RESULTS

Each image from the Michigan face database is segmented using the multiscale transform[9] and the skin color regions are selected. The total number of analyzed pixels is 9,565,862 (skin color pixels) where each sample consists of three values (r,g,b). To reduce the dependence on the lighting condition, each sample is transformed from RGB to CIE LUV color space and then the lightness value is discarded. Figure 1 shows the resulting 2D histogram (downsampled by a factor of 10) of skin color histogram of $\mathbf{x}$ ($\mathbf{x} = (u,v)^T$). It is clear that a single Gaussian density function is not sufficient to model the distribution of skin color.

4.1. Estimated Density Function

We use the EM algorithm to estimate the parameters of the Gaussian mixture. The samples are initially labeled using the k-means clustering where k is equal to g (the number of components in the mixture model). In our experiments, k is set to 2 because the histogram in Figure 1 can be modeled as a mixture model with 2 components (i.e. $g = 2$). The parameters, ϕ, are estimated using the E-step (expectation) and M-step (maximization) iteratively.[3] The estimated density function (Figure 2) perceptually fits the histogram of the samples (Figure 1(a)). It is evident that a finite Gaussian mixture model is more appropriate for estimating the density function of human skin color. To further support the argument, statistical tests on the normality and the number of components are performed.

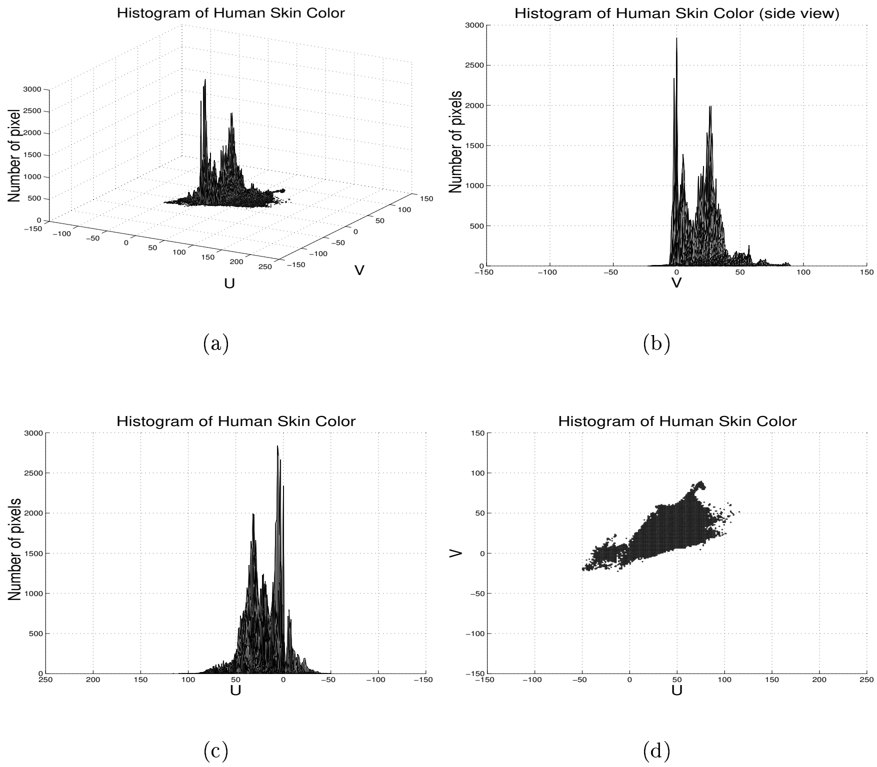

Figure 1. Histogram of skin color (downsampled by a factor of 10) viewed from different angles

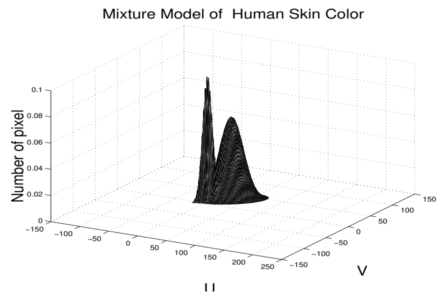

Figure 2. Estimated Density Function

4.2. Hawkins' Test on Normality and Homoscedasticity

The result of Hawkins' test of

$$H_0 : \mathbf{x} \sim \mathcal{N}(\mu_i, \Sigma) \text{ in } G_i \ (i = 1, 2)$$

for normality and homoscedasticity are given in Table 1. They are for the application of Hawkins' test to the data in their known classified form. That is, for each observation $\mathbf{x}_{ij} (i = 1, 2; j = 1, \ldots, n_i)$, the tail area a_{ij} to the right of the Anderson-Darling statistic[2] was computed under the F distribution. The Anderson-Darling statistic and its first two asymptotic $\mathcal{N}(0, 1)$ components for the a_{ij} from G_i and for the totality of the a_{ij} are useful in interpreting qualitatively departures from H_0. From Table 1, the difference in sign and significance of the first components of the Anderson-Darling statistics for the individual populations indicate hetroscedasticity while the nonsignificance of the Anderson-Darling statistic and its components for the totality of the a_{ij} gives a fair indication that G_1 and G_2 are bivariate normal.

Table 1. Results of Hawkins' test for normality and homoscedasticity(applied to data in known classified form)

Source	Anderson-Darling statistic	Components of Anderson-Darling statistic	
		First	Second
G1	3.07	-2.14	-1.50
G2	2.30	1.90	0.97
Total	0.15	0.12	-0.31

4.3. Bootstrap Test on the Number of Components

Now that we know the normality and hetroscedasticity of the data set, it is of interest to see if the likelihood ratio test of

$$H_0 : g = 2 \text{ versus } H_1 : g = 3$$

would lead to the rejection of the null hypothesis of two components.

For the bootstrap test suggested by McLachlan,[1] H_0 is rejected at a nominal α level if $-2 \log \lambda$ evaluated for the original sample exceeds the $(1 - \alpha)(K + 1)$th smallest of K bootstrap values subsequently replicated for this statistic, where $\alpha(K + 1)$ is an integer. Given the amount of computation involved, K was limited to 19. Proceeding under H_0, 19 bootstrap samples were generated from two component Gaussian mixture. For each of the 19 bootstrap samples a mixture of $g = 3$ with unequal covariance matrices was fitted, and the increase in the log likelihood over that for a single population calculated. As the value of 10.6 for $-2 \log \lambda$ from the original samples was less than 6 of 19 replications, H_0 would clearly not be rejected in favor of H_1. The conclusion, based on $K = 19$, is that two component Gaussian mixture hypothesis is not rejected.

5. APPLICATIONS

One application of using the estimated mixture density is to detect human faces in an image database. Given an image, a multiscale segmentation is performed to obtain homogeneous regions. Each pixel is classified to be skin color if its probability measure is above a threshold. Each region is then recognized as a skin area if the most pixels of the regions have high probability to be skin color. Figure 3(a) shows an image and Figures 3(a), 3(b) shows the results after skin detection using simple thresholds. The result using estimated mixture density function is shown in Figure 3(c). It is clear that a mixture model has better performance than a single Gaussian distribution function in detecting skin regions.

Skin color alone is usually not sufficient in detecting human faces or hands. For example, people wear clothes whose colors are similar to skin color as shown in Figure 3. Nevertheless, a good estimated density function of skin color is very useful and effective in simplifying the tasks of skin area detection. Using the skin color and structure information, human faces can be detected robustly.[10]

In our experiments, if more than 70% of the pixels in a region are classified to be skin color, then the region is recognized as a skin area. Figure 4 shows an image sequence and Figure 5 shows the results of skin detection.

(a) original image

(b) extracted skin regions by a Gaussian density function with low threshold

(c) extracted skin regions by a Gaussian density function with high threshold

(d) extracted skin regions by a mixture density function

Figure 3. Original image and results of skin detection

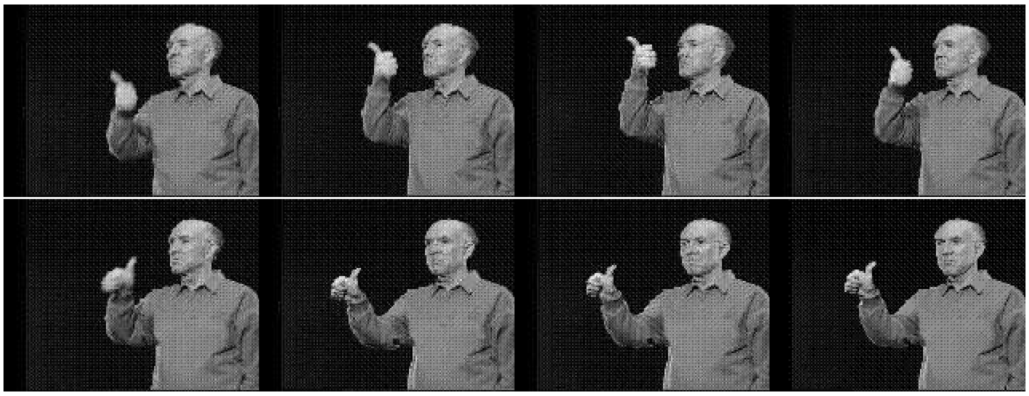

Figure 4. Image sequence of ASL sign "any" (time increases left to right and top to bottom)

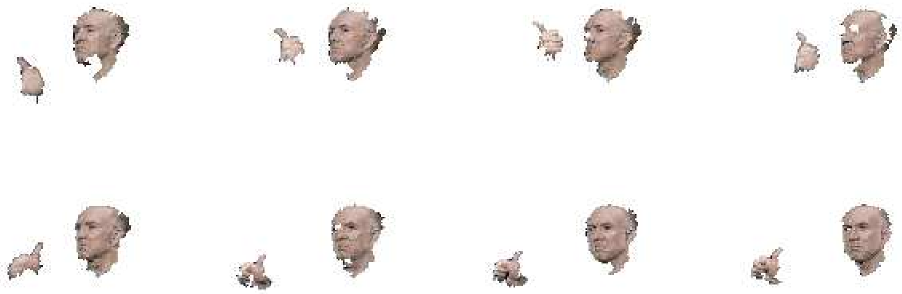

Figure 5. Detected skin areas in the sequence in Figure 4

6. CONCLUSION

In this paper, a Gaussian mixture model for human skin color is introduced. It is evident, both from the histogram of large samples and the estimated density function, that a Gaussian mixture is more appropriate than a single Gaussian function in estimating the distribution of skin color. Statistical tests on the normality and the number of components are performed to justify the hypotheses. Applications using the estimated density function show that the estimated density function of skin color is useful and effective in detecting skin regions.

REFERENCES

1. G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, 1988.
2. D. M. Hawkins, "A new test for multivariate normality and homoscedasticity," *Technometrics* **23**, pp. 105–110, 1981.
3. R. A. Render and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM Review* **26**(2), pp. 195–239, 1984.
4. D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York, 1985.
5. B. Efron, "Bootstrap methods: another look at the jackknife," *Ann. Statist.* **7**, pp. 1–26, 1979.
6. M. Aitkin, D. Anderson, and J. Hinde, "Statistical modeling of data on teaching styles," *Journal Royal Statistics Society* **144**, pp. 419–461, 1981.
7. A. C. A. Hope, "A simplified monte carlo significance test procedure," *Journal of Royal Statistics Society* **30**, pp. 582–598, 1968.
8. G. J. McLachlan, "Assessing the performance of an allocation rule," *Comp. & Maths. with Appls.* **12A**, pp. 261–272, 1986.
9. N. Ahuja, "A transform for multiscale image segmentation by integrated edge and region detection," *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, pp. 1211–1235, December 1996.
10. M.-H. Yang and N. Ahuja, "Detecting human faces in color images," in *Proceedings of the International Conference on Image Processing*, vol. 1, pp. 127–130, 1998.