HTML Paper

HAND AND FACE SEGMENTATION USING MOTION AND COLOR CUES IN DIGITAL IMAGE SEQUENCES

Nariman Habili and Cheng-Chew Lim

Department of Electrical & Electronic Engineering
Adelaide University, Adelaide, SA 5005, Australia
{nhabili,cclim}@eleceng.adelaide.edu.au

Alireza Moini

VLSI Group Leader, Intelligent Pixels Inc.
R&D Center, Sanori House, 126 Grand Blvd.
Joondalup, WA 6027, Australia
am@intelligent-pixels.com

ABSTRACT

In this paper, we present a hand and face segmentation algorithm using motion and color cues. The algorithm is proposed for the content based representation of sign language image sequences, where the hands and face constitute a video object. Our hand and face segmentation algorithm consists of three stages, namely color segmentation, temporal segmentation, and video object plane generation. In color segmentation, we model the skin color as a normal distribution and classify each pixel as skin or non-skin based on its Mahalanobis distance. The aim of temporal segmentation is to localize moving objects in image sequences. A statistical variance test is employed to detect object motion between two consecutive images. Finally, the results from color and temporal segmentation are analyzed to yield a change detection mask. The performance of the algorithm is illustrated by simulation carried out on the *silent* test sequence.

1. INTRODUCTION

There is a growing trend towards content-based representation in image and video processing applications, as shown by the recent MPEG-4 and 7 standardization efforts. Content-based representation requires the decomposition of an image or video sequence into specific objects, known as video objects (VOs). In this context, a VO may represent a moving person, a fixed background or audio. The instances of VOs at a given time (i.e. frame) are called video object planes (VOPs). A frame can be decomposed into VOPs by means of segmentation.

In sign language communication, or simply signing, the hands and face are perceptually important and thus constitute a VO. The main objective of our research is to devise an algorithm for the segmentation of VOPs in sign language sequences. A comprehensive study on the segmentation of the hands and face, and the coding of sign language sequences was presented in [1]. The author modeled the skin color distribution as a normal mixture in the L*a*b colorspace and used the Bayesian classifier to classify image pixels as skin or non-skin. The algorithm required a separate skin location algorithm to identify skin pixels for distribution training. Due to hand and face motion during signing, motion serves as an important cue for VOP segmentation. The author did not take advantage of motion information to enhance the segmentation results.

Our hand and face segmentation algorithm is composed of three stages. In the first stage, image pixels are classified as skin or non-skin to yield a skin detection mask (SDM). The skin color distribution is modeled as a bivariate normal distribution and the image pixels are classified based on their Mahalanobis distance. In the second stage, the statistical variance test is employed to localize moving objects in the image sequence and yield a change detection mask (CDM). The third stage involves the fusion of the SDM and the CDM to generate the VOP. To distinguish between the hands and face, a face identification method is proposed, employing shape features.

This paper is organized as follows. The color segmentation and temporal segmentation techniques are presented in sections 2 and 3, respectively. In section 4, we present the VOP generation method, and experimental results are presented in section 5. The paper is concluded in section 6.

2. COLOR SEGMENTATION

We employ color information to locate skin regions in each image. The YCbCr color space is considered since it is typically used in video coding and provides an effective use of chrominance information for modeling the human skin color. Experimental results indicate that the skin-color distribution in the CbCr plane remains constant regardless of any variation in the luminance information of an image [2, 3]. Moreover, the CbCr component of the skin pixels of people from European, African and Asian descent occupy the same region in the CbCr plane.

2.1. Pixel Classification

This section describes the classification method employed to classify pixels as skin or non-skin. The method is analogous to the single hypothesis classifier described in [4]. Single hypothesis schemes have been proposed to solve problems in which one class is well defined while others are not. It is assumed that the skin class is well defined, while the non-skin class, which may include a wide variety of different colors, is not.

2.1.1. The Skin-Color Model

Let $\mathbf{x}$ denote the feature vector formed by the Cb and Cr components of a pixel, and $\mathbf{x}$ is in a 2-dimensional Euclidean space $\mathbf{R^2}$, called the feature space. The skin and non-skin classes are denoted by ω_S and $\omega_{\bar{S}}$, respectively. The skin color distribution in the CbCr plane is modeled as a bivariate normal distribution:

$$p(\mathbf{x}|\omega_S) = \frac{1}{2\pi|\mathbf{\Sigma}_S|^{\frac{1}{2}}} exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_S)^T \mathbf{\Sigma}_S^{-1}(\mathbf{x}-\boldsymbol{\mu}_S)], \quad (1)$$

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE
0-7695-1198-8/01/$10.00 (C) 2001 IEEE

where $\boldsymbol{\mu}_S$ and $\boldsymbol{\Sigma}_S$ are the mean vector and covariance matrix of the distribution, respectively. Normal distributions are widely used in the pattern recognition community because of their many desirable properties [4]. The parameters, $\boldsymbol{\mu}_S$ and $\boldsymbol{\Sigma}_S$, are estimated from the skin training pixels. The training pixels were obtained by manually segmenting training images that included people of European, African and Asian descent.

The quantity d in

$$d^2 = (\mathbf{x} - \boldsymbol{\mu}_S)^T \boldsymbol{\Sigma}_S^{-1} (\mathbf{x} - \boldsymbol{\mu}_S) \tag{2}$$

is known as the Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}_S$. Pixels can be classified as skin or non-skin based on their Mahalanobis distance. The value of d is related to the probability that a given pixel belongs to class ω_S. A small value of d indicates a high skin pixel probability and vice-versa.

2.1.2. Test of Normality

An effective test to check the assumption of bivariate normality is the chi-square test [5]. Equation (2) can be expressed as:

$$d^2 = (\mathbf{x} - \boldsymbol{\mu}_S)^T {\boldsymbol{\Sigma}_S}^{-1} (\mathbf{x} - \boldsymbol{\mu}_S) = \mathbf{z}^\mathbf{T} \mathbf{z} = \sum_{j=1}^{n} z_j^2 \tag{3}$$

where $\mathbf{z} = A^T(\mathbf{x} - \boldsymbol{\mu}_S)$ and A is the whitening transformation [4]. Since the mean vector and covariance matrix of $\mathbf{z}$ are $[0\ 0]^T$ and the identity matrix respectively, the z_i's are independent random variables with zero mean and unity variance. If $\mathbf{x}$ is indeed normal, then $\sum_{j=1}^{n} z_j^2$ in equation (3) is a chi-square (χ_n^2) random variable with $n = 2$ degrees of freedom. Therefore, the test for bivariate normality is to compare the goodness of fit of the Mahalanobis distances

$$d_i^2 = (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_S)^T \hat{\boldsymbol{\Sigma}}_S^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_S) \tag{4}$$

to χ_2^2, where $\hat{\boldsymbol{\mu}}_S$ and $\hat{\boldsymbol{\Sigma}}_S$ are estimated from the skin training pixels. The procedure is as follows:

1. The squared distances in equation (4) are ordered in ascending order as $d_{(1)}^2 \leq d_{(2)}^2 \leq \ldots \leq d_{(N_S)}^2$, where N_S is the number of skin training pixels. Note that $d_{(i)}^2$ is the ith smallest squared distance, whereas d_i^2 is the squared distance associated with the chrominance vector for the ith skin training pixel.
2. $d_{(i)}^2$ is plotted against $\chi_2^2[(i - 0.5)/N_S]$, where $\chi_2^2[(i - 0.5)/N_S]$ is the $100(i - 0.5)/N_S$ percentile of the chi-square distribution with 2 degrees of freedom (the factor of 0.5 is added as a correction for continuity).

The plot should follow a straight line. The chi-square plot of the ordered distances shown in figure 1 does not show any significant deviation from a straight line. It can therefore be asserted that the skin class pixels in the CbCr plane follow a bivariate normal distribution.

2.1.3. The Segmentation Threshold

The skin detection mask (SDM) is defined as:

$$SDM(m,n) = \begin{cases} 1, & \text{if } d_{m,n} < \tau \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

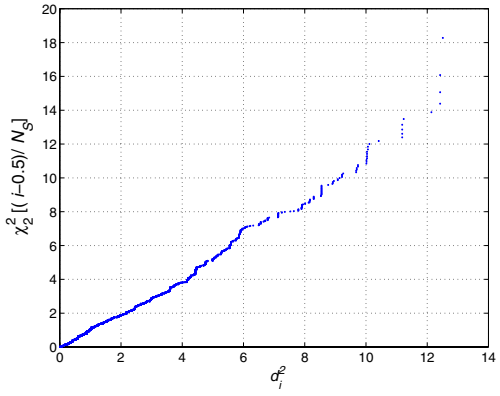

Figure 1: The chi-square plot of the ordered distances.

where τ is the segmentation threshold, and $d_{m,n}$ is the Mahalanobis distance for the pixel at location (m,n). τ is derived by examining the probability of classification error, P_{error}.

Let $\mathcal{R}_S$ denote the region in the feature space where the classifier decides ω_S and likewise for $\mathcal{R}_{\bar{S}}$ and $\omega_{\bar{S}}$. There are two ways in which a classification error can occur; either an observation $\mathbf{x}$ falls in $\mathcal{R}_S$ and the true class is $\omega_{\bar{S}}$, or $\mathbf{x}$ falls in $\mathcal{R}_{\bar{S}}$ and the true class is ω_S. Since these events are mutually exclusive and collectively exhaustive, the probability of classification error is

$$\begin{aligned} P_{error} &= P(\mathbf{x} \in \mathcal{R}_{\bar{S}}, \omega_S) + P(\mathbf{x} \in \mathcal{R}_S, \omega_{\bar{S}}) \\ &= P(\mathbf{x} \in \mathcal{R}_{\bar{S}} | \omega_S) P(\omega_S) + P(\mathbf{x} \in \mathcal{R}_S | \omega_{\bar{S}}) P(\omega_{\bar{S}}) \end{aligned} \tag{6}$$

where $P(\omega_S)$ and $P(\omega_{\bar{S}})$ denote the *a priori* probabilities of the skin and non-skin classes, respectively. For the remainder of this paper, the following notations, borrowed from radar terminology, will be used

$$\begin{aligned} P_F &= P(\mathbf{x} \in \mathcal{R}_S | \omega_{\bar{S}}) \\ P_D &= P(\mathbf{x} \in \mathcal{R}_S | \omega_S) \\ P_M &= P(\mathbf{x} \in \mathcal{R}_{\bar{S}} | \omega_S). \end{aligned} \tag{7}$$

P_F, P_D and P_M are the probabilities of false alarm, detection and miss, respectively. Note that $P_M = 1 - P_D$.

Using the above notations, the probability of classification error can now be expressed as

$$P_{error} = P_M(\theta) P(\omega_S) + P_F(\theta) P(\omega_{\bar{S}}), \tag{8}$$

where θ is a threshold. Therefore, the probability of error is a function of θ and the *a priori* probabilities. P_D and P_F are evaluated for the set of training images I_k, $k = 1, \ldots, K$,

$$P_D(\theta) = \frac{1}{N_S} \sum_{k=1}^{K} \sum_{\mathbf{x} \in I_k} \alpha(\mathbf{x}, k, \theta), \tag{9}$$

and

$$P_F(\theta) = \frac{1}{N_{\bar{S}}} \sum_{k=1}^{K} \sum_{\mathbf{x} \in I_k} \beta(\mathbf{x}, k, \theta), \tag{10}$$

where $\alpha(\mathbf{x}, k, \theta)$ and $\beta(\mathbf{x}, k, \theta)$ are defined as:

$$\alpha(\mathbf{x}, k, \theta) = \begin{cases} 1, & \text{if } \mathbf{x}_k \in \omega_S \text{ and } d_{\mathbf{x}_k} \leq \theta \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE
0-7695-1198-8/01/$10.00 (C) 2001 IEEE

and

$$\beta(\mathbf{x}, k, \theta) = \begin{cases} 1, & \text{if } \mathbf{x}_k \in \omega_{\bar{S}} \text{ and } d_{\mathbf{x}_k} \leq \theta \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

$\mathbf{x}_k$ denotes the feature vector of a pixel in training image k, $k = 1, ..., K$. The *a priori* probabilities can be either estimated or assumed. The θ that minimizes equation (8) is then designated as the segmentation threshold.

3. TEMPORAL SEGMENTATION

In this section, a temporal segmentation method is developed based on the variance statistical test. The motion of a moving object from one image to the next generates intensity variations that can be represented in the form of a difference image. However, intensity variations can also occur due to camera or quantization noise. The noise is usually modeled as a zero-mean normal distribution [6]. The objective of temporal segmentation is to distinguish between temporal variations caused by noise and those caused by object motion. We refer to intensity variations caused by motion as foreground and those caused by noise as background.

Let σ_B^2 denote the variance of the background population, and W a sliding observation window. We use the statistical variance test to detect background and foreground regions in the difference image. The statistical variance test can be formally stated as:

$$H_0 : \sigma^2 = \sigma_B^2$$
$$H_1 : \sigma^2 > \sigma_B^2. \quad (13)$$

The null hypothesis, H_0, implies that the set of difference pixels in W is drawn from a normal population with variance σ_B^2. The hypothesis is rejected if the variance of the difference pixels in W is significantly greater than σ_B^2. The intensity variation induced by a moving object is greater than that of the background because of a higher intensity gradient at the edge and inside of a moving object. W is set to 3×3 samples (i.e. $n = 9$ samples) and the significance level, α, is set to 1%. If the hypothesis is true, then

$$Y = \frac{(n-1)S^2}{\sigma_B^2} \quad (14)$$

has a χ^2 distribution with $n - 1 = 8$ degrees of freedom. S^2 is the sample variance. For a significance level of 1%, the critical value of Y is 20.1. Therefore if $Y > 20.1$, we would reject the hypothesis.

The foreground and background regions in the difference image are represented in the form of a binary map, called the change detection mask (CDM). If the null hypothesis is rejected, a binary 1 is allocated to the center pixel in W, otherwise a binary 0 is allocated. The parameter σ_B^2 can be estimated by the histogram fitting technique described in [7] or the least median of squares technique described in [8].

4. VIDEO OBJECT PLANE GENERATION

This section describes the VOP generation method. Firstly, connected components analysis on both the SDM and the CDM is performed to remove all connected components of 50 or less pixels (with 8-neighborhood connectivity). These regions can be generally attributed to false alarms. After connected components analysis, holes in the remaining connected components are filled. This was performed to promote the formation of semantic objects and improve the accuracy of VOP generation.

4.1. SDM and CDM Analysis

Due to face and hand motion during signing, the CDM can be utilized to identify the hands and face in the SDM. First, the SDM is superposed on top of the CDM. When 80% or more of a connected component in the SDM is covered by a foreground region in the CDM, the connected component is declared as either a face or a hand.

4.2. Face Identification

It may sometimes be necessary to discriminate between the face and the hands. One method is to compare the areas of the connected components in the VOP. Intuitively, the face would have the largest area, however if a subject has part of an arm exposed, the arm may have a greater area than the face and thus result in inaccurate identification. An effective method to distinguish between the face and the hands is to model the face as a rigid object, and the hands as non-rigid objects, due to wrist and finger motion. Such a model would allow the use of shape features to differentiate between the face and hands. We have devised three tests to make the differentiation.

It is a well known fact that the shape of the face can be approximated by an ellipse [9]. The best-fit-ellipse of a connected component, $\mathcal{C}$, is defined by its center $(\bar{m}, \bar{n})$, its orientation ϕ, and the length of its major (a) and minor (b) axes [10]. The center of gravity of $\mathcal{C}$ gives the center of the ellipse:

$$\bar{m} = \frac{1}{N} \sum_{(m,n) \in \mathcal{C}} m, \quad (15)$$

and

$$\bar{n} = \frac{1}{N} \sum_{(m,n) \in \mathcal{C}} n, \quad (16)$$

where N denotes the number of pixels in $\mathcal{C}$. Orientation is defined as the angle of axis of the least moment of inertia. It can be computed by utilizing the central moments $\mu_{p,q}$ of the connected component:

$$\phi = \frac{1}{2} \tan^{-1} \left[\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}} \right]. \quad (17)$$

The first test is the orientation test. We have observed that during signing, the head can tilt in the range $(-40^\circ, +40^\circ)$ from vertical. Therefore, if the orientation of a connected component is not within this range, it cannot be the face.

The second test deals with the aspect ratio (a/b) of $\mathcal{C}$. We have observed that the aspect ratio of the face can range from 1.4 to 1.8. Therefore, any connected component outside of this range, cannot represent the face. a and b are determined by computing the moments of inertia of $\mathcal{C}$. The least and greatest moments of inertia for an ellipse are

$$I_{min} = \frac{\pi}{4} a b^3, \quad (18)$$

and

$$I_{max} = \frac{\pi}{4} a^3 b. \quad (19)$$

For a given ϕ, the above moments can be calculated as

$$I'_{min} = \sum_{(m,n) \in \mathcal{C}} \left[(n - \bar{n}) cos\phi - (m - \bar{m}) sin\phi \right]^2, \quad (20)$$

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE
0-7695-1198-8/01/$10.00 (C) 2001 IEEE

and

$$I'_{max} = \sum_{(m,n)\in\mathcal{C}} \left[(n-\bar{n})sin\phi - (m-\bar{m})cos\phi\right]^2 . \qquad (21)$$

The requirements for a best fit ellipse are $I_{min} = I'_{min}$ and $I_{max} = I'_{max}$, which gives the lengths of a and b, respectively:

$$a = \left(\frac{4}{\pi}\right)^{\frac{1}{4}} \left[\frac{(I'_{max})^3}{I'_{min}}\right]^{\frac{1}{8}}, \qquad (22)$$

and

$$b = \left(\frac{4}{\pi}\right)^{\frac{1}{4}} \left[\frac{(I'_{min})^3}{I'_{max}}\right]^{\frac{1}{8}}. \qquad (23)$$

The final test is to assess the similarity between a connected component and its best fit ellipse. This is accomplished by computing the difference between the area of $\mathcal{C}$ inside and outside the ellipse. The difference is then divided by the area of the ellipse. We have found that the above similarity measure should be 0.8 or higher for facial regions.

5. EXPERIMENTAL RESULTS

For simulation we used the *silent* test sequence. Results for frames 11 and 12 are shown in figure 2. The SDMs are shown in figure 2(b). The false alarms present in the SDMs are due to similar skin and background color characteristics. The CDMs, shown in figure 2(c), also contain false alarms. The false alarms to the subject's right are due to shadow, induced by hand motion. The false alarms are largely eliminated after connected components analysis. The face and hands of the subject have been segmented quite effectively, as shown in figure 2(d).

6. CONCLUSIONS

A new hand and face segmentation algorithm has been presented in this paper. The algorithm consists of three steps, namely color segmentation, temporal segmentation and VOP generation. In color segmentation, the aim is to segment skin regions in an image. Meanwhile, in temporal segmentation, moving objects in the image sequence are localized. The color and motion information is then used to generate the VOP. Experimental results indicate that the technique is capable of segmenting the hands and face quite effectively. The algorithm allows the flexibility of incorporating additional techniques to enhance the results. Work is currently under way to incorporate a tracking technique to track the hands and face throughout the sequence.

7. REFERENCES

[1] R. P. Schumeyer, *A Video Coder Based on Scene Content and Visual Perception*, PhD thesis, University of Delaware, 1998.

[2] D. Chai and K. N. Ngan, "Face Segmentation Using Skin-Color Map in Videophone Applications", *IEEE Trans. Circuits Sys. Video Tech.*, vol. 9, no. 4, pp. 551–564, June 1999.

[3] H. Wang and S.-F. Chang, "A Highly Efficient System for Automatic Face Region Detection in MPEG Video", *IEEE Trans. Circuits Sys. Video Tech.*, vol. 7, no. 4, pp. 615–628, Aug. 1997.

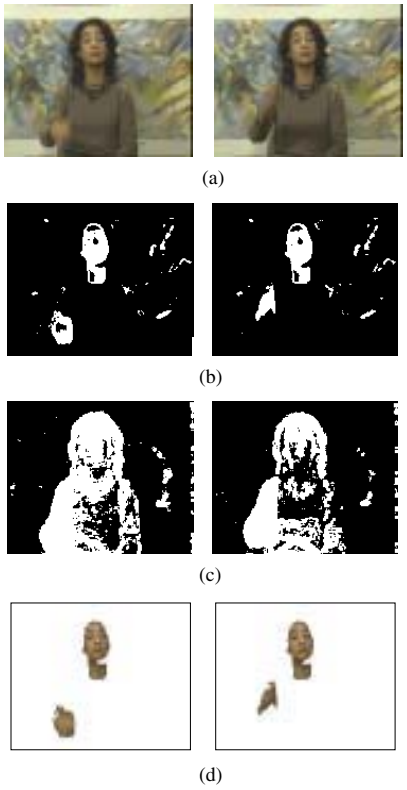

Figure 2: (a): Frames 11 and 12 of the *silent* sequence. (b): SDMs. (c): CDMs. (d): VOPs.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Boston, 1990.

[5] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, N.J, 1982.

[6] M. Kim, J. G. Choi, D. Kim, H. Lee, M. H. Lee, C. Ahn, and Y.-S. Ho, "A VOP Generation Tool: Automatic Segmentation of Moving Objects in Image Sequences Based on Spatio-Temporal Information", *IEEE Trans. Circuits Sys. Video Tech.*, vol. 9, no. 8, pp. 1216–1226, Dec. 1999.

[7] N. Habili, A. R. Moini, and N. Burgess, "Histogram Based Temporal Object Segmentation for VOP Extraction in MPEG-4", in *Proc. The Fist IEEE Pacific-Rim Conference on Multimedia*, Sydney, Australia, Dec. 2000, pp. 310–313.

[8] P. L. Rosin, "Thresholding for Change Detection", Tech. Rep. ISTR-97-01, Brunel University, UK, June 1997.

[9] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking", *Signal Processing: Image Communication*, vol. 12, no. 3, pp. 263–281, June 1998.

[10] A. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, N.J, 1989.

2001 IEEE International Conference on Multimedia and Expo
ISBN 0-7695-1198-8/01 $17.00 © 2001 IEEE