Comparative Performance of Different Chrominance Spaces for Color Segmentation and Detection of Human Faces in Complex Scene Images

Jean-Christophe Terrillon and Shigeru Akamatsu
ATR Human Information Processing Laboratories
2-2 Hikaridai, Seika-cho,
Soraku-gun, Kyoto 619-0288, Japan
{terril, akamatsu}@hip.atr.co.jp

Abstract

Color is a powerful fundamental cue that can be used at an early stage to detect objects in complex scene images. This paper presents an analysis of the performance of nine different chrominance spaces in the specific problem of automatically detecting and locating human faces in two-dimensional still scene images. For each space, we use a skin color model based on the Mahalanobis metric to segment faces from the scene background by thresholding. We perform feature extraction on the segmented images by use of fully translation-, scale- and in-plane rotation-invariant moments that are derived from the Fourier-Mellin transform, and apply a multilayer perceptron neural network with the invariant moments as the input vector to distinguish faces from distractors. We show that for each chrominance space, the detection efficiency is critically dependent on the goodness of fit of the skin chrominance distribution to the proposed model, and to a lesser extent on the discriminability between skin and "non-skin" distributions. Also, normalized color spaces are shown to yield the best segmentation results, and subsequently the highest rate of detection of faces with a large variety of poses and against relatively complex backgrounds.

1. Introduction

Automatic detection and localisation of human faces in two-dimensional natural, complex scene images is a difficult task that has been relatively unexplored until recently [7] [6] [1]. Face detection has important applications, as a first step in higher-level face recognition tasks such as personal identification for security purposes, the determination of sex and race, the understanding of facial expressions, or in the field of multimedia, for portrait retrieval in a large database of images or for interactive human-machine interfaces. In recent years, an increasing body of research has addressed the specific problem of automatic face detection based on skin color [8] [18] [17] [2] [21]. Color is a powerful fundamental cue that can be used as a first step in the process of face detection in complex scene images because color image segmentation is computationally fast while being relatively robust to changes in illumination, in viewpoint, in scale, to shading and to complex (cluttered) backgrounds as compared to the segmentation of grey-level images. Robustness is achieved if a color space efficiently separating the chrominance from the luminance in the original color image and a plausible model of the chrominance distribution of human skin are used for thresholding. In general, dimensionality reduction is first achieved by a suitable (linear or nonlinear) transformation from a 3-D RGB color space into a 2-D chrominance space (and into a separate luminance component). Normalized r-g chrominance space has often been used for face detection [8] [17] [21] [4] [15] because it reduces the sensitivity of the segmentation to changes in illumination. Other chrominance-luminance spaces that have been commonly used are the perceptually plausible HSV (or HSI) space [16] [9] [14] or the hardware-oriented YIQ or YES spaces [20] [3] [13]. In [10], the comparative efficiency of three different color spaces (HSI, CIE-L*u*v* and Karhunen-Loeve) in discriminating between skin (for Asian subjects only) and different facial features (mouth, eyes and eyebrows) has been analyzed, after a face has been segmented from a background. However, to our knowledge, no analysis of the comparative efficiency of several different chrominance spaces has been performed until now in the general problem of face detection. In effect, the efficiency of the color segmentation of a human face depends on the chrominance space that is selected, because the skin chrominance distribution depends on the chrominance space. Therefore, the selection of an appropriate color space is an important task.

We propose to compare the efficiency of nine different color spaces for face segmentation and detection against complex backgrounds. In section 2, we examine the chrominance

distribution of human skin in the following color spaces : normalized r-g and CIE-xy, a normalized perceptually plausible tint-saturation-luminance space TSL, CIE-DSH, HSV, YIQ, YES, and the perceptually uniform CIE-L*u*v* and CIE-L*a*b* spaces. At first, we describe skin color by use of an unbiased anthropometric chromatic scale that assigns an equal statistical weight to each component of the scale under the same illumination conditions [18] and analyse qualitatively the chrominance distribution of all the components of the scale in the different color spaces. For comparison, the same qualitative analysis is then performed in the general case where illumination conditions vary widely and where different camera systems are used, by use of a large number of skin sample images selected from various sources. Based on these two analyses, preliminary conclusions are drawn on the relative suitability of the different spaces for skin-color based segmentation. Finally, we propose a skin chrominance model based on a quantitative analysis of a set of skin sample images of faces of Asian and Caucasian subjects that were recorded under slowly varying illumination conditions with a single camera and manually selected for our experiments. The model assumes that the chrominance of the skin of Asians and Caucasians is described by a unimodal elliptical Gaussian joint probability density function (pdf). The Mahalanobis metric is inherent to the Gaussian pdf model and is used to determine a threshold value in each chrominance space that would efficiently discriminate between human skin and other objects (or "non-skin" objects). The goodness of fit of the skin distribution to the model and the discriminability between skin and "non-skin" distributions are analyzed for each space. In section 3, we briefly describe a shape analysis based on fully translation-, scale- and in-plane rotation-invariant moments that are generated from the Fourier-Mellin transform and that are calculated for each cluster representing a face candidate in the segmented binary images. In order to discriminate face candidates from distractors (such as false positives or other body parts correctly classified as skin), a multilayer perceptron neural network (NN) is used with the invariant moments as the input vector. The architecture of the NN is also briefly described. Experimental results of face detection for the nine different chrominance spaces are presented in section 4, and in section 5 we summarize the main issues that we plan to address in future research.

2. Color segmentation

2.1. Chrominance Distribution of Human Skin in Different Color Spaces

For a first analysis, to describe skin color we have used an unbiased anthropometric chromatic scale that is representative of a sufficiently large number of skin colors and that assigns an equal statistical weight (in number of pixels in the distribution analysis) to each component under the same illumination conditions, so that every color component may be equally well represented. Von Luschan's chromatic scale [18] [19] can be used to match 36 different skin colors, from an unsaturated light color to a saturated dark brown color. Figure 1 shows the cumulative distribution of all the components of the scale in the nine different chrominance spaces on a logarithmic scale. For the normalized TSL chrominance-luminance space, we selected the following transformations:

$$S = [9/5(r'^2 + g'^2)]^{1/2}$$

$$T = \begin{cases} arctan(r'/g')/2\pi + 1/4, & g' > 0 \\ arctan(r'/g')/2\pi + 3/4, & g' < 0 \\ 0 & g' = 0 \end{cases}$$

$$L = 0.299R + 0.587G + 0.114B \quad (1)$$

where r'=(r-1/3) and g'=(g-1/3), r=R/(R+G+B) and g=G/(R+G+B), S is the saturation, T is the tint, and where L is the luminance (for Gamma-corrected RGB values). The values of S, T and L are normalized in the range [0. ; 1.0]. The distribution consists essentially of two classes in the normalized r-g, CIE-xy and T-S spaces, as well as in CIE-SH and H-S spaces. The distribution in I-Q and E-S spaces is more complex and difficult to model, while three classes are clearly distinguishable in CIE-u*v* and CIE-a*b* spaces. In r-g space, while the distribution of both light and intermediate skin colors (corresponding to the first 24 components of the scale) is concentrated near the equal-energy or achromatic point (r=g=1/3), the distribution of the class representative of dark skin colors covers a significantly larger surface area, with larger values of r and lower values of g. There is a continuous transition from a cluster representative of light and intermediate skin colors to a cluster describing dark skin colors. Because the dark-skin cluster is spatially more diffuse in the different chrominance spaces, it is concealed if a logarithmic scale is not used. While the distribution is confined in the normalized r-g, CIE-xy and T-S spaces, it covers most of the possible range of saturation S for a limited range in hue H in CIE-SH and H-S spaces. Therefore, a normalization of RGB values by (R+G+B) or of CIE-XYZ values by (X+Y+Z) with or without a further transformation (such as into T-S space) yields chrominance spaces that are more efficient for skin color segmentation than CIE-SH and H-S spaces where such a normalization is not performed, because the sensitivity of the distribution to the variability of skin color is significantly reduced. Finally, the distribution in the normalized spaces may be described by a simpler model (such as a 2-class mixture density) than in I-Q, E-S, CIE-u*v* and CIE-a*b* spaces where the distribution is also confined. The above conclusions also apply to the general case where illumination conditions vary widely and where different camera systems are used.

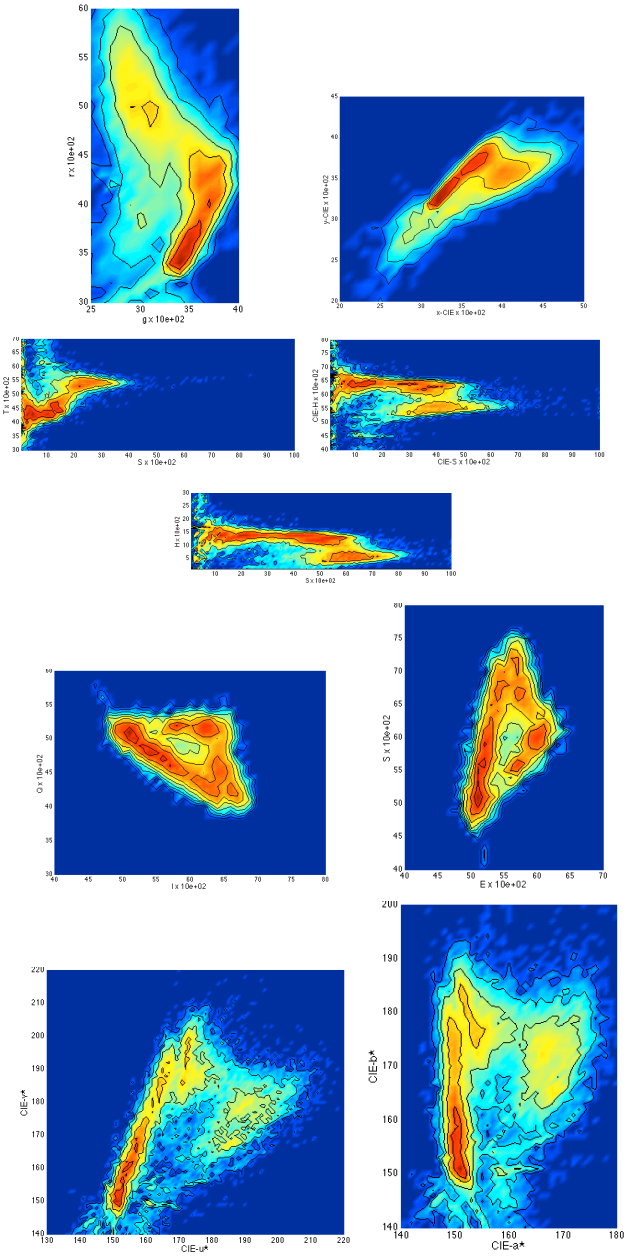

Figure 1. False-color top view of the 2-D cumulative histograms on a logarithmic scale of all the components of von Luschan's chromatic scale in nine different chrominance spaces (N = 5.9 x 10E+04 pixels). From top to bottom and left to right : normalized r-g, CIE-xy and T-S spaces, CIE-DSH, H-S, I-Q, E-S, and perceptually uniform CIE-u*v* and CIE-a*b* spaces (appropriately shifted or corrected for discontinuities, except H-S space). Total histogram dimensions are 100 x 100 bins in all spaces except in CIE-u*v* and CIE-a*b* spaces where the dimensions are 200 x 200 bins.

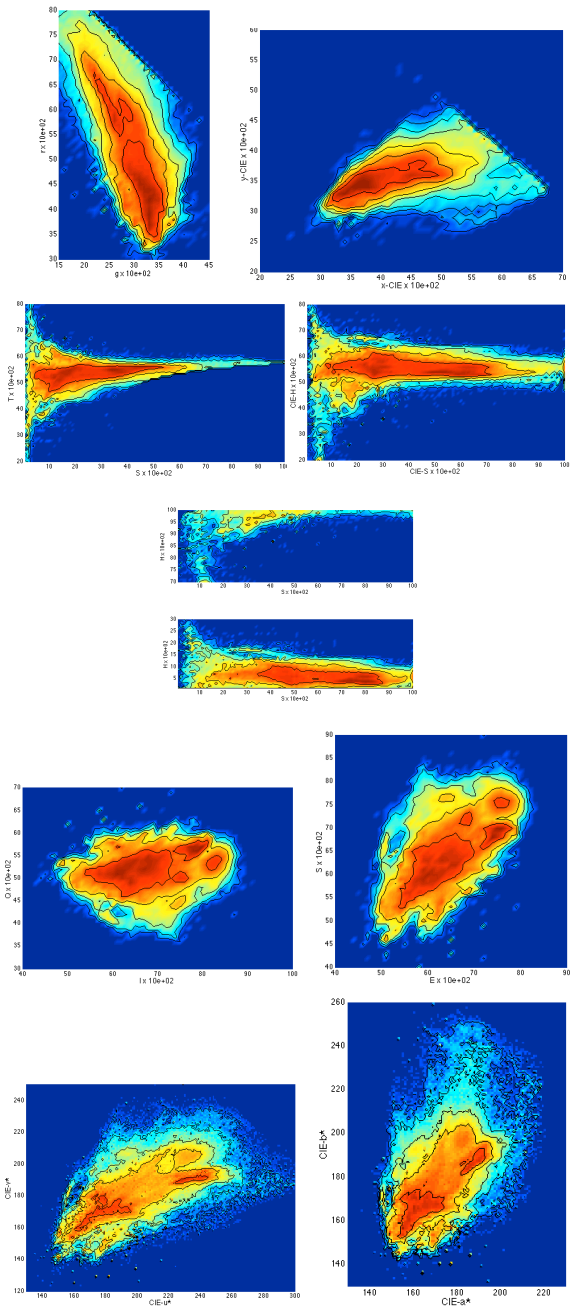

Figure 2. Cumulative histograms on a logarithmic scale in nine different chrominance spaces of 300 skin sample images manually selected from various sources (N = 1.118 x 10E+06 pixels). From top to bottom and left to right : normalized r-g, CIE-xy and T-S spaces, CIE-DSH, H-S, I-Q, E-S, and perceptually uniform CIE-u*v* and CIE-a*b* spaces (shifted or corrected for discontinuities, except H-S space). Total histogram dimensions are 100x100 bins in all spaces except in CIE-u*v* and CIE-a*b* spaces where the dimensions are 370x370 bins.

Figure 2 shows the cumulative distribution in the nine chrominance spaces (on a logarithmic scale) of 300 skin sample images manually selected from various sources and originally recorded in unconstrained conditions. While covering a larger surface area than in the case of the anthropometric scale, the distribution in the normalized spaces is still confined, whereas that in CIE-SH and H-S spaces covers the whole range of S and a significantly larger range of H at low values of S. Again, while still confined, the distribution in I-Q, E-S, CIE-u*v* and CIE-a*b* spaces cannot be described by a simple model. In this general case, the distribution in r-g space appears to be compatible with a unimodal elliptical Gaussian joint pdf model that is assumed in [2] [21] to describe the skin chrominance of the different human races in the same space, whereas the cumulative distribution of all the components of the unbiased anthropometric scale is not compatible with such a model.

The normalization theoretically renders the chrominance independent of any identical changes in R, G and B or X, Y and Z values. In practice such normalized spaces significantly reduce the dependency of skin chrominance on changes in illumination and on the camera system used to record the images. In our experiments, only the cluster corresponding to light and intermediate skin colors in the distribution of von Luschan's chromatic scale was considered for segmentation. Images of 11 Asian and 19 white Caucasian subjects were recorded under slowly varying illumination conditions in an office environment with a single video camera mounted on an SGI computer. 110 skin sample images were manually selected to analyze the color distribution in the nine different chrominance spaces and to calibrate the camera for color segmentation in each space. The cumulative distribution of all the sample images are shown in the left part of Figure 4, in section 4. Although the distribution in such controlled conditions is confined in each space, its area is generally smaller in the normalized spaces.

2.2. Skin Chrominance Model and Thresholding of Color Test Images

On the basis of the histograms shown in the left part of Figure 4, we assume that the skin chrominance distribution for Asians and white Caucasians may be modeled by a unimodal elliptical Gaussian joint pdf given by

$$p[\mathbf{x}(i,j)/W_s] = (2\pi)^{-1}|\mathbf{C}_s|^{-1/2} \exp\left[-\lambda_s^2(i,j)/2\right] \quad (2)$$

where the vector $\underline{\mathbf{x}}(i,j) = [\underline{x}(i,j)\underline{y}(i,j)]^T$ represents the random measured values of the chrominance (x,y) of a pixel with coordinates (i,j) in an image, W_s is the class describing skin, $\mathbf{C_s}$ is the covariance matrix for skin chrominance ,and where $\lambda_s(i,j)$ is the Mahalanobis distance from the vector $\underline{\mathbf{x}}(i,j)$ to the mean vector $\mathbf{m_S} = [m_{Xs} m_{Ys}]^T$ obtained for skin chrominance, defined as

$$[\lambda_s(i,j)]^2 = [\underline{\mathbf{x}}(i,j) - \mathbf{m}_s]^T {\mathbf{C}_s}^{-1} [\underline{\mathbf{x}}(i,j) - \mathbf{m}_s] \quad (3)$$

Eq. 3 defines elliptical surfaces in chrominance space of scale $\lambda_s(i,j)$, centered about $\mathbf{m}_s$ and whose principal axes are determined by $\mathbf{C}_s$. The value of $\lambda_s(i,j)$ for a pixel with coordinates (i,j) determines the probability that the pixel belongs to the class W_s representing human skin, as seen from Eq. 2. The larger $\lambda_s(i,j)$, the lower the probability that the pixel (i,j) belongs to W_s.

Both $\mathbf{m_s}$ and $\mathbf{C_s}$ are estimated in each space by use of the 110 skin sample images recorded with the SGI camera. $[\lambda_s(i,j)]^2$ is then calculated for every pixel over all the skin samples and over five large image regions not containing skin, and is compared to a parameter $\lambda_{s,t}^2 > 0$ for thresholding. As a suitable compromise to discriminate between the skin class and the "non-skin" class, a "standard" threshold value $\lambda_{s,t}^2$ is obtained when the proportion of true positives TP over the ensemble of regions of skin becomes equal to the proportion of true negatives TN over the ensemble of regions not containing skin [13], or equivalently, when the proportion of false negatives FN equals the proportion of false positives FP (since TP+FN=1 and TN+FP=1). This initial skin color calibration must be performed for any color camera before color image segmentation and is semi-automatic. However, under slowly varying illumination conditions the use of a single camera does not require the further application of an "adaptive" threshold as in [13]. Color segmentation is performed on test images by calculating $[\lambda_s(i,j)]^2$ for every pixel in the test images and comparing its value to the standard threshold $\lambda_{s,t}^2$. A value of 1 is assigned to pixel (i,j) if $\lambda_s \leq \lambda_{s,t}$ and a value of 0 if $\lambda_s > \lambda_{s,t}$. The result of thresholding is a binary image that is subjected to further (morphological) analysis to isolate face candidates in a scene.

In order to assess the plausibility of the unimodal Gaussian pdf model in each chrominance space, we define a measure of the goodness of fit of the skin distribution shown in the left part of Figure 4 to the model as the Mean Square Deviation between each normalized histogram observed in a discrete space with M x N bins and the corresponding "ideal" discrete unimodal Gaussian pdf calculated from $\mathbf{m}_s$ and $\mathbf{C}_s$:

$$MSDN = \sum_{j=1}^{N} \sum_{i=1}^{M} (G'_i j - S'_i j)^2 \quad (4)$$

where $S'_{ij} = S_{ij} / \sum_{j=1}^{N} \sum_{i=1}^{M} S_{ij}$ is the normalized observed histogram and $G'_{ij} = G_{ij} / \sum_{j=1}^{N} \sum_{i=1}^{M} G_{ij}$ is the normalized "ideal" Gaussian histogram. Also, the discriminability between skin and "non-skin" distributions is measured as the degree of overlap between the distributions, defined as the intersection between the normalized skin and "non-skin" histograms :

$$HIN = \sum_{j=1}^{N} \sum_{i=1}^{M} min(S'_{ij}, NS'_{ij}) \quad (5)$$

where $NS'_{ij} = NS_{ij} / \sum_{j=1}^{N} \sum_{i=1}^{M} NS_{ij}$ is the normal-

ized observed "non-skin" histogram calculated by use of the five large image regions not containing skin. The values of the MSDN,HIN and the results of the skin color calibration for each chrominance space are presented in Figure 3.

color space	MSDN	HIN	Calibration TP=TN (%)
TSL	1.0000	0.2653	83.32
r-g	1.0034	0.2627	81.35
CIE-xy	2.1915	0.2845	73.77
CIE-DSH	3.3466	0.2657	77.73
HSV	3.4951	0.2618	76.43
YIQ	4.9076	0.5801	61.59
YES	3.6299	0.5809	60.69
CIE-L*u*v	3.3806	0.4177	68.36
CIE-L*a*b	3.7381	0.4050	80.10

Figure 3. Goodness of fit to the unimodal Gaussian model (MSDN), overlap between skin and "non-skin" distributions (HIN) and results of the skin color calibration in nine different chrominance spaces for 110 skin sample images recorded with the SGI camera (MSDN = 1.73 x 10E-03 for the normalized TSL color space). The total histogram dimensions in CIE-u*v* and CIE-a*b* spaces have been downsampled to the same dimensions as those of the other spaces to calculate MSDN, for proper comparison.

The normalized spaces, in particular TSL and r-g spaces, yield the best fit to the unimodal Gaussian model, as the MSDN of the skin distribution is smallest in those spaces. Moreover, the discrimination between the skin class and the "non-skin" class is among the highest in the same spaces, as lower values of the HIN show, which leads to the highest values of TP and TN (hence the lowest values of FN and FP). Therefore it is expected that the normalized spaces should produce the best segmentation results and subsequently the highest performance of face detection. As expected, the results of the skin color calibration are generally poor (the values of TP and TN are significantly smaller) for the chrominance spaces for which the MSDN and/or the HIN are larger.

3. Shape Analysis

A local median filter with a window of 3x3 pixels is applied to the thresholded binary images in order to obtain clusters of connected pixels (connected-component analysis). The clusters of area less than 0.5% of the area of the image in number of pixels are discarded so that only a small number of main clusters or face candidates are used for further analysis.

The robustness of face detection and localization is increased if invariant features are extracted from each face candidate in the segmented images. We use the method of invariant moments that were first developed by Hu [5] and recently generalized by Li [11]. Hu's moments are fully translation-, scale- and in-plane rotation invariant. Although they have been used extensively in low-level pattern-recognition problems [11], to our knowledge they have not been used in combination with color segmentation for the detection of human faces. Hu's generalized moments arise as a particular case of the circular Fourier and radial Mellin transform (FMT) [11]. An infinite number or moments can be generated by use of the FMT and are expressed as combinations of scale- and translation-invariant geometric moments. A detailed derivation of the moments can be found in [18] [11].

The computation of the discrete moments for binary images yields theoretically an error-free estimate of the continuous moments as opposed to the computation performed for grey-level images [12], and the moments are then also independent of illumination. However, higher orders amplify the contributions of the peripheral parts of an object to the moments. The contours of segmented face candidates constitute a correlated noise because they are variable, and thus may reduce the degree of invariance of the moments, so that only a small number of the lowest invariant moments should be used in the shape analysis.

We use the 11 lowest-order moments, that include the second to the fourth-order geometric moments. The 11 moments are calculated for each face candidate and are the input units of a feed-forward multilayer perceptron neural network (NN), with one hidden layer containing 6 nodes and with one output unit. All the units take on continuous bipolar-sigmoid activation values in the range [-1.0 ; 1.0]. For training, the backpropagation algorithm is applied to perform gradient descent on a quadratic error function (or total error) using batch processing. A momentum term is included in the learning rule in order to increase the learning rate of the network. When a face is detected by the network, it is marked by an ellipse using the already computed first and second-order geometric moments of the cluster representing the face, as described in [18].

4. Experimental results

The face detection system is implemented on an SGI Indigo 2 Impact 10000 computer. After the color segmentation and the connected-component analysis, N=220 elements (clusters) with an approximate ratio of 1:1 between elements describing faces (of 9 Asians and 20 Caucasians) and those representing objects other than faces were used to train the NN. The training was performed on images segmented by use of the normalized T-S chrominance space, which produces the best results in the chrominance analysis and the color calibration. Considering then that the faces in the training images are well segmented and are assumed to be approxi-

mately elliptical (with holes at the location of the eyes and of the mouth), it is reasoned that the training need not be performed for each chrominance space separately, so that the same NN weights may be applied to test images segmented by use of the other chrominance spaces.

The face detection efficiency was investigated for each chrominance space, by use of a test file of 90 images with 133 faces and 59 subjects (27 Asians, 31 Caucasians and one subject of African descent), with a large variety of poses and against different complex backgrounds. 77% of the images in the test file are not part of the training set.

Figure 4 shows an example of the simultaneous detection of the faces of an Asian subject and of a Caucasian subject for the nine different chrominance spaces, while Figure 5 presents the general results of face detection for the different spaces. As seen on the middle part of Figure 4 and in Figure 5, the segmentation based on all the chrominance spaces other than the normalized T-S and r-g spaces produces a significantly larger number of clusters or face candidates as well as an erosion of facial parts, due to larger classification errors of pixels belonging to the background or to a face respectively. Such results can be explained by the shape of the corresponding skin chrominance distribution in the left part of Figure 4 , which determines the value of the MSDN. Hence, the quality of the segmentation depends critically on the goodness of fit of each distribution to the unimodal Gaussian model, as the results of Section 2.2 suggested, and to a certain extent on the degree of overlap between "skin" and "non-skin" histograms. Subsequently, larger proportions of false negatives, of false positives and of face localization errors occur during the face detection process, as can be seen in the right part of Figure 4. A face localization error is defined as the detection of only part of a face and/or of the face together with some other object in the immediate vicinity of the face that has been classified as skin during the color segmentation. A face is properly detected when all the facial features are detected and when the ellipse marking the face follows the contours of the face, with an error on the angle between the major axis of the ellipse and the face axis that does not exceed 15^o . The general results in Figure 5 show a high face detection performance for the normalized T-S space, an acceptable performance for the r-g space, but the performance is generally poor for the other spaces. The large proportion of false negatives is due to the erosion of the face clusters but also to a larger probability of distractors being connected to a face cluster when the total number of clusters is larger, thus modifying significantly the shape of the face cluster. Although the invariant properties of the moments tend to increase their tolerance to false positives, the correct rejection rate remains relatively high for all spaces, because the NN discriminates successfully between clusters having a similar shape to that of a face (approximately elliptical) and distractors with significantly different shapes.

It should be borne in mind that the complexity of the scene images contributes also to reduce the performance of face detection : in the case of the normalized T-S space, for which a 90.8% correct detection rate and a 84.9% correct rejection rate are achieved, faces of Asian and Caucasian subjects are equally well detected, despite a large variety of backgrounds and of poses that include small out-of-plane rotations, as the examples of Figure 6 show. Exposed body parts other than faces that are correctly classified as skin during the segmentation are also well rejected, except when their shape is similar to that of an ellipse, as can be seen in an example of Figure 7. However, as the examples of Figure 7 show (here normalized T-S space has also been used), four main types of errors limit the performance of the system, whatever the color space that is used. First, other body parts (such as the neck) connected to or in contact with a face, as well as false positives due to other objects in contact with a face, may lead to face localization errors. In this particular case, the NN still detects the face because the entire cluster of connected pixels retains a similar shape to that of a face. This type of error can also occur when partial occlusions, by dark glasses or by facial hair for example, erode the face cluster or divide it into two distinct clusters. Second, false positives may occur due to other body parts correctly classified as skin precisely because of the higher tolerance of the invariant moments and of a shape of the skin cluster that is similar to that of an ellipse. Third, false positives due to other objects in the scene result mainly from a poor color segmentation performance, and to some extent from the limited color discrimination of the SGI camera system. Finally, false negatives may be caused by strong shadows and/or strong illumination (specularities) that eliminate the chromatic information in regions of a face and thus reduce the efficiency of the color segmentation, or by partial occlusions that divide the face cluster into two or more distinct clusters, or by other objects in the scene that are connected to the face cluster (including other faces or other body parts) and that modify its shape significantly.

Such problems are magnified significantly when segmenting the images by use of the other chrominance spaces. In addition to significantly lower rates of correct face detection, the face localization error rate (FLER) obtained for the non-normalized spaces varies between 33% and 57%, whereas the FLER is 13% for the normalized T-S space. For well segmented images in T-S space, the mean absolute error on the angle between the face axis and the major axis of the ellipse marking a detected face is 7^o, with a standard deviation of 4^o. Finally, the time required for face detection, without optimization of the algorithms, varies between 0.2 second and 2.5 seconds for images of dimensions 320 x 243 pixels, depending on the chrominance space that is used (non-linear color transformations are the most computer-intensive process in the face detection system, in particular the transformations into CIE-L*u*v* and CIE-L*a*b* spaces). However, the face detection time is not affected by the number of faces that are present in a scene.

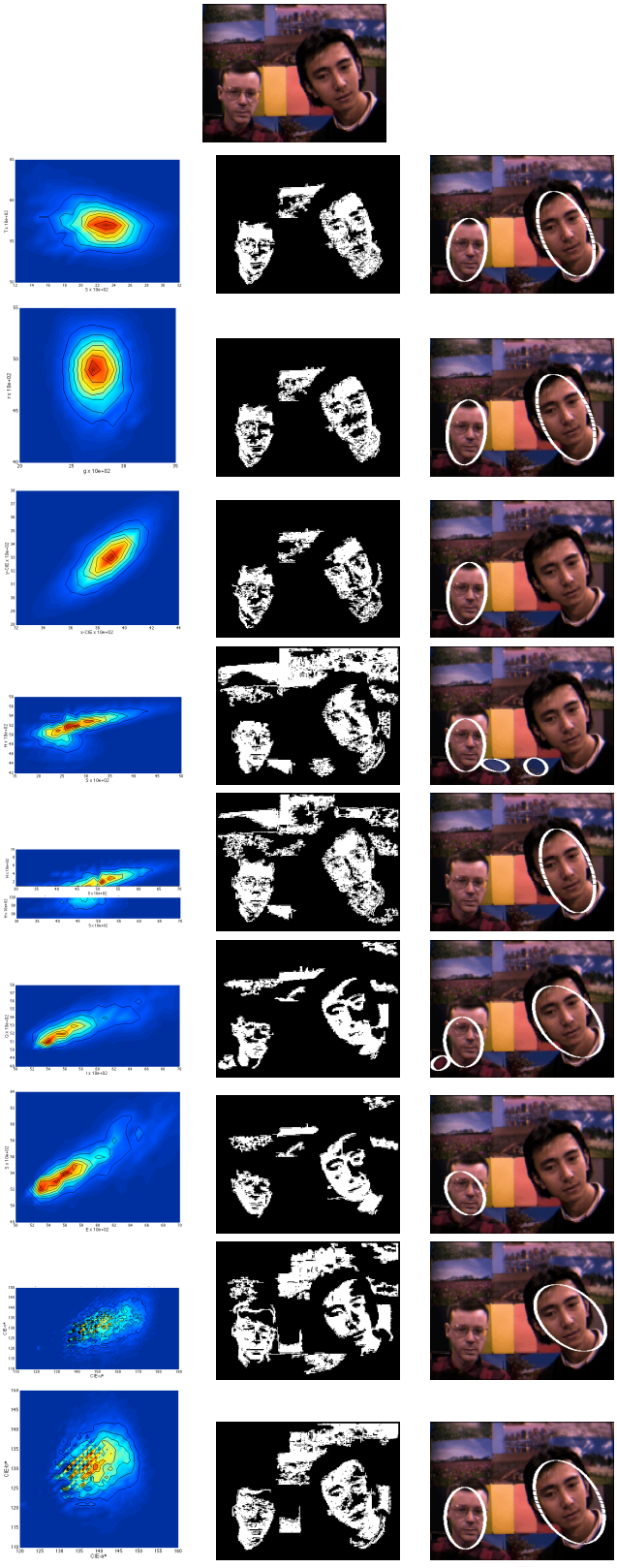

Figure 4. Example of the simultaneous detection of the face of and Asian and of a Caucasian subject. Top : original scene. From top to bottom, left : 2-D cumulative histograms in nine different chrominance spaces of 110 skin sample images (1.5 x 10E+05 pixels) of 11 Asian and 19 white Caucasian subjects used for calibrating the SGI camera : normalized T-S, r-g and CIE-xy spaces, CIE-SH, H-S, I-Q, E-S, CIE-u*-v* and CIE-a*b* spaces (appropriately shifted or corrected for discontinuities, except H-S space). Total histogram dimensions are 100 x 100 bins in all spaces except in CIE-u*-v* and CIE-a*-b* spaces where the dimensions are 200 x 200 bins; middle : Scene image after color segmentation in each space and after a connected-component analysis; right : results of the face detection process.

color space	Total # of Elements	CD(%)	CR(%)
TSL	258	90.8	84.9
r-g	328	74.6	80.3
CIE-xy	388	56.6	83.5
CIE-DSH	318	60.9	75.0
HSV	408	55.7	84.7
YIQ	471	47.3	79.8
YES	494	41.6	80.3
CIE-L*u*v	418	24.1	79.0
CIE-L*a*b	399	38.4	83.6

Figure 5. General results of face detection in nine different chrominance spaces. The total number of elements is the cumulative number of clusters (face candidates) remaining after the connected-component analysis, CD is the rate of correct face detection and CR is the rate of correct rejection of distractors.

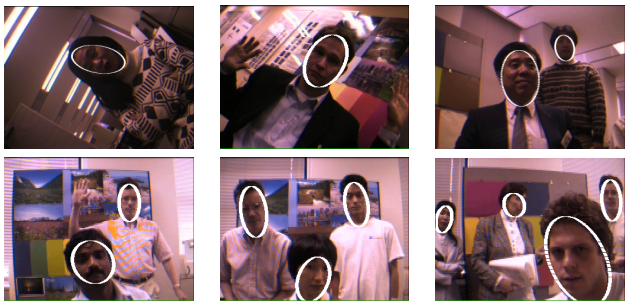

Figure 6. Examples of the detection of faces with different poses and different skin colors against various complex backgrounds. All images are segmented by use of the normalized T-S space.

Figure 7. Examples of errors occurring with the present face detection system.

5. Conclusion

The most important conclusion of the skin-color based segmentation and of the face detection analyses is that normalized color spaces, in particular the normalized T-S space that we use in this paper and that we proposed in [18], yield the best segmentation results and the most robust face detection system. We hypothesize that such a conclusion may hold not only in the specific application of skin-color based face detection, but also in the general problem of the detection of any object that is based on color. Based on the discussion in Section 4, three main issues should be addressed in future research : first, if normalized color spaces are selected for segmentation and detection, the efficiency of the segmentation needs further improvement in order to increase bost the robustness to larger variations of illumination and the portability between different camera systems. this would require a color correction (or color constancy) algorithm as a preprocessing step. Also, dark skin colors pose a problem of dicrimination against "non-skin" colors because low values of R, G and B lead to large fluctuations in their relative values, hence to noise. Secondly, the luminance L is an important information that is to complement the color segmentation. Finally, the NN should be trained to detect also side views of faces, and its generalization ability should be increased. Other discrimination methods than the application of a NN may be considered for face detection, for example a template matching performed on the segmented images that could use the invariant moments. The performance of this approach to face detection could be compared to that of the NN.

References

[1] M. Bichsel, editor. *Proc. of the International Workshop on Automatic Face- and Gesture-Recognition*, Zurich, 1995.

[2] Q. Chen, H. Wu, and M. Yachida. Face detection by fuzzy pattern matching. In *Proc. of the 5th International Conference on Computer Vision*, MIT, Boston, 1995. pp. 591-596.

[3] Y. Dai and Y. Nakano. Extraction of facial images from complex background using color information and sgld matrices. In *Proc. of the International Workshop on Automatic Face and Gesture Recognition*, Zurich, 1995. pp. 238-242.

[4] H. P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan. Multi-modal system for locating heads and faces. In *Proc. of the Second International Conf. on Automatic Face and Gesture Recog.*, Killington, Vermont, 1996. pp. 88-93.

[5] M. K. Hu. Visual pattern recognition by moment invariants. *IEEE Trans. Inf. Theory*, IT-8:179–187, 1962.

[6] I.C.S. Press. *Proc. of the Second Intern. Conf. on Automatic Face and Gesture Recog.*, Killington, Vermont, 1996.

[7] I.C.S. Press. *Proc. of the Third International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998.

[8] S.-H. Kim, N.-K. Kim, S. C. Ahn, and H.-G. Kim. Object oriented face detection using range and color information. In *Proc. of the Third International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998. pp. 76-81.

[9] R. Kjeldsen and J. Kender. Finding skin in color images. In *Proc. of the Second Intern. Conf. on Automatic Face and Gesture Recog.*, Killington, Vermont, 1996. pp. 312-317.

[10] C. H. Lee, J. S. Kim, and K. H. Park. Automatic face location in a complex background using motion and color information. *Pattern Recognition*, 29(11):1877–1889, 1996.

[11] Y. Li. Reforming the theory of invariant moments for pattern recognition. *Pattern Recognition*, 25(7):723–730, 1992.

[12] S. X. Liao and M. Pawlak. On image analysis by moments. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-18(3):254–266, 1996.

[13] E. Saber, A. M. Tekalp, R. Eschbach, and K. Knox. Automatic image annotation using adaptive color classification. *Graph. Models and Image Proc.*, 58(2):115–126, 1996.

[14] D. Saxe and R. Foulds. Toward robust skin identification in video images. In *Proc. of the Second International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, 1996. pp. 379-384.

[15] B. Schiele and A. Waibel. Gaze tracking based on face-color. In *Proc. of the International Workshop on Automatic Face and Gesture Recognition*, Zurich, 1995. pp. 344-349.

[16] K. Sobottka and I. Pitas. Segmentation and tracking of faces in color images. In *Proc. of the Second International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, 1996. pp. 236-241.

[17] Q. B. Sun, W. M. Huang, and J. K. Wu. Face detection based on color and local symmetry information. In *Proc. of the Third International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998. pp. 130-135.

[18] J.-C. Terrillon, M. David, and S. Akamatsu. Automatic detection of human faces in natural scene images by use of a skin color model and of invariant moments. In *Proc. of the Third International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998. pp. 112-117.

[19] F. von Luschan. *Voelker, Rassen, Sprachen : Anthropologische Betrachtungen*. Deutsche Buchgemeinshaft, Berlin, 1927. 382 pp.

[20] M. Yamada, K. Ebihara, and J. Ohya. A new robust real-time method for extracting human silhouettes from color images. In *Proc. of the Third International Conf. on Automatic Face and Gesture Recognition*, Nara, Japan, 1998. pp. 528-533.

[21] J. Yang and A. Waibel. Tracking human faces in real time. Technical Report CMU-CS-95-210, C.M.U., 1995.