研究テーマ　肌色情報に基づいた顔検出と手サイン認識に関する研究

研究者　Jean-Christophe TERRILLON　財団法人ソフトピアジャパン　雇用研究員
渡辺博己　財団法人ソフトピアジャパン　雇用研究員

フェーズⅠ

1 研究の概要

Human skin color is a powerful fundamental cue that can be used in particular, at an early stage, for the important applications of face and hand detection in color images or video sequences, and ultimately, for meaningful human-machine interactions [1]. One important first issue that we addressed during both Phase I and Phase II of the HOIP project is the selection of an efficient chrominance (color) space, because the performance of face and hand detection depends critically on the performance of the initial steps of skin pixel detection and of image segmentation, which in turn ultimately depends on the chrominance space that is used [2], [3]. Secondly, a suitable image segmentation model is applied depending on the selection of the color space [2]. The extraction of binary global facial feature information is then performed by use of fully translation-, scale- and in-plane rotation-invariant Fourier-Mellin moments, in order to ensure invariant face detection [4]. For subsequent binary face/non-face classification in complex scenes, we applied two different statistical learning techniques: a multiplayer perceptron Neural Network (NN) [4], or alternately, Support Vector Machines (SVM), that may [5], [6] or may not [7] rely on color as a first cue. Finally, hand posture recognition of the Japanese Sign Language (JSL) is performed by use of the Phase-Only correlation Filter (POF) [1], that yields an efficient discrimination both between different hand postures and between hand postures and scene background areas that have incorrectly been detected as skin.

This report presents, for Phase I of the HOIP project, a detailed overview and experimental results of the skin color analysis and of the face detection and hand posture recognition system based on skin color (for convenience, later results that were obtained during Phase II are included in this section). The part of the report devoted to Phase II involves the integration of the system based on skin color into a more global, real-time system that is capable to simultaneously detect or track multiple faces as well as recognize hand postures of the JSL in color video sequences.

2 研究の目標

The analysis of the distribution of human skin color for a large set of two-dimensional (2-D) chrominance spaces derived from the standard three-dimensional (3-D) 24-bit RGB color space aims to select the most efficient chrominance space(s) for skin pixel detection and skin color-based image segmentation, in terms of a sufficient number of representative criteria. Subsequently, by applying such spaces to the analysis of color images or of video sequences, our goal is to implement a robust face detection and hand posture recognition system based on skin color, that can be the front end of more complex systems capable of various face recognition tasks, face tracking, dynamic gesture recognition, and ultimately of meaningful human-machine interactions. The two main fields of applications of such interactions, which have a significant impact on society and on its activities, are 1) welfare improvement, and 2) the security of people and of information systems.

3 実施内容 - Skin Chrominance Analysis

Skin pixel detection, or skin color-based image segmentation is relatively robust to changes in illumination, in viewpoint, in scale, to shading, partial occlusions and to cluttered backgrounds as compared to the segmentation of gray-level images. Robustness of segmentation is generally achieved by separating the 2-D chrominance from the luminance channel in the original RGB color images, and then by using only the chrominance for segmentation. This separation implies a dimensionality reduction by a suitable, linear or non-linear transformation from the 3-D RGB color space into a 2-D chrominance space. The selection of an efficient chrominance space motivates an analysis of human skin color for different chrominance spaces. For a given set of skin sample images or of sample pixels that is collected for calibration before thresholding and segmentation of test images, the space that is selected determines the compactness and the shape of the skin chrominance distribution, which in turn determines the complexity of the skin chrominance model that is required in order to obtain a high quality of segmentation. The skin chrominance distribution also depends on the various skin groups that are considered (Asians, Caucasians, dark skin group), the illumination conditions under which the color images were recorded, and on the camera system that is used to record the images. Finally, an important criterion that ultimately limits the quality of skin pixel detection and of image segmentation is the degree of overlap, or of discrimination, between the skin distribution and a distribution of "non-skin" pixels in a given chrominance space, which depends to some extent on the number of skin and non-skin pixels that are collected for calibration.

In this sub-section, we perform an in-depth comparative analysis of the distribution of human skin for a large set of 25 different color spaces (41 chrominance spaces), for facial skin images recorded with two different camera systems, and in terms of seven different criteria. The intrinsic geometrical properties of each space are also briefly discussed. The color spaces considered here that result from a linear transformation from the RGB space are the $I_1I_2I_3$ (Ohta's optimized color features [8]), $h_1h_2h_3$ (Wesolkowski's color space [9]), YCb_1Cr_1 (using the CIE standard illuminant C) and YCb_2Cr_2 (using the CIE standard illuminant D65) [10] [11], YES (a standard space developed by the Xerox company), YIQ and YUV spaces. The color spaces that result from a non-linear transformation form a second group, that can be divided into 4 sub-groups: the normalized color spaces (r-g-b [12] [13], CIE-xyz [12] [13] for both standard C and D65 illuminants, and TSL [14]), the perceptually plausible color spaces (CIE-DSH [12], HSV and HSL [15]), the perceptually uniform color spaces (CIE-L*u*v*[12], CIE-L*a*b*[12], and Farnsworth's F-uv space [12] [16], for both standard C and D65 illuminants), and other color spaces ($C_1C_2C_3$, $l_1l_2l_3$, and $l_1'l_2'l_3'$ proposed as color invariants and used for viewpoint-invariant image retrieval and for color-based object recognition by Smeulders and Gevers [17] [18], r.g and rg.b log-opponent space applied to color image indexing by Berens and Finlayson [19], a'-b' space applied to the extraction of skin color areas in facial images by Kawato and Ohya [20], mod-rgb space proposed by Tominaga [21], P_1-P_2 space used for the construction of the Fourier spectrum of the chromaticity by Vertan *et al.*[22], and (R/G=r/g, R/B=r/b, G/B=g/b) and Yuv spaces). The conversions from the RGB space for both groups are shown in Tables 1-5 (all Figures and Tables are located at the end of the report). These Tables also show the boundaries of each space, as well as the dimensions used to calculate the discrete skin chrominance histogram for each space. For all chrominance spaces considered in this paper, the histogram dimensions are selected such that the histogram resolution is the same for all spaces, in order to ensure a valid comparative study.

Two separate sets of sample images used for the skin chrominance analysis are recorded with an inexpensive SGI camera, and with a high-quality SONY DXC-9000 camera system respectively. The seven criteria used for the analysis for each space are: 1) the robustness of the skin chrominance distribution with respect to the intrinsic variability of skin color (to three different skin groups), 2) its compactness, 3) its shape, 4) the degree of discrimination (or the overlap) between the skin and non-skin distributions, 5) the robustness (or "portability") of the skin distribution to a change of camera system, 6) the relative robustness of the skin distribution to changes in illumination conditions, and finally, 7) the computational cost of the transformation from the 24-bit RGB (NTSC) space into a given chrominance space.

3.1 Parameters used for the Skin Chrominance Analysis

We first define four different parameters that we use to perform a quantitative analysis of the skin chrominance.

1) The compactness of the skin distribution can be calculated as the area of the distribution A_s relative to the area of the gamut of all the possible colors in a given space A_G:

$$C_S = A_S / A_G = (Nb)_S / (Nb)_G \tag{1}$$

Where Nb is the number of bins in the discrete histograms, and where S and G refer to skin and to the gamut respectively.

2) The Kullback-Leibler Divergence [3] (KLD) is selected to estimate the goodness of fit of the skin chrominance distribution to a simple, single elliptical Gaussian. It is defined in the discrete case as:

$$\mathrm{KLD} = \sum_{j=1}^{N}\sum_{i=1}^{M} S'_{ij} \ln\left(\frac{S'_{ij}}{G'_{ij}}\right) + \sum_{j=1}^{N}\sum_{i=1}^{M} G'_{ij} \ln\left(\frac{G'_{ij}}{S'_{ij}}\right) \tag{2}$$

Where S'_{ij} is considered as the "true" distribution (the normalized skin histogram observed in a discrete chrominance space with M x N bins) and G'_{ij} as the "estimated" or "model" distribution (the normalized ideal Gaussian histogram calculated from the mean vector and from the covariance matrix of the skin distribution in the same discrete space). The KLD has the following properties: i) $\mathrm{KLD} \geq 0$ and ii) if KLD=0, then $S'_{ij} = G'_{ij}$. Hence, the lower the value of the KLD, the higher the goodness of fit to the single Gaussian model.

3) The Normalized Histogram Intersection (HIN) is a measure of the overlap between two different distributions, such as the skin and the non-skin distributions. In the discrete case, it is defined as:

$$\mathrm{HIN} = \sum_{j=1}^{N}\sum_{i=1}^{M} \min(S'_{ij}, NS'_{ij}) \tag{3}$$

Where $NS'_{ij} = NS_{ij} / \sum_{j=1}^{N}\sum_{i=1}^{M} NS_{ij}$ is the normalized non-skin histogram calculated in the same discrete chrominance space as S'_{ij}, with M x N bins. The lower the value of the HIN, the higher the degree of discrimination between the two distributions.

4) Finally, the global shift S of a distribution can be calculated as:

$$S = \sqrt{(m_{x1} - m_{x2})^2 + (m_{y1} - m_{y2})^2} \tag{4}$$

Where $\mathbf{m}_x = (m_{x1}, m_{x2})$ and $\mathbf{m}_y = (m_{y1}, m_{y2})$ are the mean vectors for the skin distributions in a given chrominance space (x, y) for cameras 1 and 2.

4 結果

4.1 Experimental Set-up

Facial images of Asian and Caucasian subjects, and of subjects with dark skin color, were recorded under slowly varying illumination conditions (using halogen lamps at 3,200 degrees Kelvin) in the "percept-room" of the HOIP laboratory (under controlled, semi-constrained scene conditions) with both the SONY and the SGI camera systems (using a white balance for both cameras). From the images obtained with the SONY camera, 65, 51 and 10 skin sample images of Asian, Caucasian, and dark skin-colored subjects respectively, were manually selected, yielding a total of 2.115x10E+05, 1.630x10E+05, and 2.580x10E+04 skin pixels for each respective skin group. When using the SGI camera, 111 skin sample images of both Asian and Caucasian subjects were manually selected, for a total of 1.515x10E+05 skin pixels. Also, 80 "non-skin" images were selected from various sources, mainly from the World Wide Web, producing a total of 2.6606x10E+06 non-skin pixels. For each image, the 24-bit RGB values are scaled between 0.0 and 1.0. Achromatic pixels (including black) were assigned suitable values adapted to the particular color space that is considered, as shown in Tables 1-5. For each space yielding negative chrominance values, a shift was performed so that all values are positive, without any influence on the results of the chrominance analysis. Generally, the discrete, cumulative skin and non-skin histograms are calculated over an entire space, except for the CIE-L*u*v and CIE-L*a*b* spaces, whose boundaries are curved, and for the log-opponent and $R_1R_2R_3$ spaces, where the range (hence the histogram dimensions) is determined empirically, by observing the skin and non-skin distributions (we used a range of [-1.0; 2.0] along both the x and y axes for the log-opponent space, and of [0.0;2.0] for the $R_1R_2R_3$ space). The resolution of the skin and non-skin histograms is 0.01 unit in both the x and y directions for all spaces, except for the CIE-L*u*v* and CIE-L*a*b* spaces, where the resolution is 1.0 unit, thus yielding matrices of dimensions 100 x 100 bins for most spaces, as can be seen in Tables 1-5.

Table 1. Linear transformations from RGB color space

COLOR SPACE	CONVERSION EQUATIONS	SPACE BOUNDARIES	HISTOGRAM No OF BINS
$I_1I_2I_3$	$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1 & 0 & -1 \\ -1/2 & 1 & -1/2 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$I_1 \in [0; 1.0]$, $I_2 \in [-1.0; 1.0]$, $I_3 \in [-1.0; 1.0]$	200 x 200
$h_1h_2h_3$	$\begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$h_i \in [-1.0; 1.0]$, $i = 1, 2, 3$ $h_1+h_2+h_3 = 0$	200 x 200
YCb_1Cr_1	$Y = 0.2989R+0.5866G+0.1145B$ $Cb_1 = \dfrac{0.8855B-0.2989R-0.5866G}{1.771}$ $Cr_1 = \dfrac{0.7011R-0.5866G-0.1145B}{1.4022}$	$\lvert Cb_1 \rvert \le 1/2$, $\lvert Cr_1 \rvert \le 1/2$	100 x 100
YCb_2Cr_2	$Y = 0.2126R+0.7152G+0.072B$ $Cb_2 = \dfrac{0.9278B-0.2126R-0.7152G}{1.8556}$ $Cr_2 = \dfrac{0.7874R-0.7152G-0.0722B}{1.5748}$	$\lvert Cb_2 \rvert \le 1/2$, $\lvert Cr_2 \rvert \le 1/2$	100 x 100
YES	$\begin{pmatrix} Y \\ E \\ S \end{pmatrix} = \begin{pmatrix} 0.253 & 0.684 & 0.063 \\ 0.500 & -0.500 & 0.000 \\ 0.250 & 0.250 & -0.500 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$\lvert E \rvert \le 1/2$, $\lvert S \rvert \le 1/2$	100 x 100
YIQ	$\begin{pmatrix} Y \\ I \\ Q \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.275 & -0.321 \\ 0.212 & -0.523 & 0.311 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$\lvert I \rvert \le 0.596$, $\lvert Q \rvert \le 0.523$	120 x 120
YUV	$\begin{pmatrix} Y \\ U \\ V \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ -0.147 & -0.289 & 0.436 \\ 0.615 & -0.515 & -0.100 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$	$\lvert U \rvert \le 0.436$, $\lvert V \rvert \le 0.615$	125 x 125

Table 2. Non-linear conversions from RGB color space into normalized color spaces

COLOR SPACE	CONVERSION EQUATIONS	SPACE BOUNDARIES	HISTOGRAM No OF BINS
rgb	$\mathbf{r}=\frac{\mathbf{R}}{\mathbf{R+G+B}},\ \mathbf{g}=\frac{\mathbf{G}}{\mathbf{R+G+B}},$ $\mathbf{b}=\frac{\mathbf{B}}{\mathbf{R+G+B}}$ If R=G=B=0, set r=g=b=1/3	$(r,g,b)\in[0;1.0]$ $\mathbf{r}+\mathbf{g}+\mathbf{b}=1$	**100 x 100**
CIE-xyz (C Ill.)	$\begin{pmatrix}\mathbf{X}\\\mathbf{Y}\\\mathbf{Z}\end{pmatrix}=\begin{pmatrix}0.607 & 0.174 & 0.201\\0.299 & 0.587 & 0.114\\0.000 & 0.066 & 1.117\end{pmatrix}\begin{pmatrix}\mathbf{R}\\\mathbf{G}\\\mathbf{B}\end{pmatrix}$		**100 x 100**
CIE-xyz (D65 Ill.)	$\begin{pmatrix}X\\Y\\Z\end{pmatrix}=\begin{pmatrix}0.430574 & 0.341550 & 0.178325\\0.222015 & 0.706655 & 0.071330\\0.020183 & 0.129553 & 0.939180\end{pmatrix}\begin{pmatrix}R\\G\\B\end{pmatrix}$		
	$\mathbf{x}=\frac{\mathbf{X}}{\mathbf{X+Y+Z}},\ \mathbf{y}=\frac{\mathbf{Y}}{\mathbf{X+Y+Z}},$ $\mathbf{z}=\frac{\mathbf{Z}}{\mathbf{X+Y+Z}}$ If X=Y=Z=0, set x=y=z=1/3	$(x,y,z)\in[0;1.0]$ $\mathbf{x}+\mathbf{y}+\mathbf{z}=1$	
TSL	$\mathbf{g'}=(\mathbf{g}-1/3),\ \mathbf{r'}=(\mathbf{r}-1/3)$ $\mathbf{T}=\begin{cases}\frac{1}{2\pi}\mathbf{arctan}\left(\frac{\mathbf{r'}}{\mathbf{g'}}\right)+\frac{1}{4} & ,\mathbf{g'}>0\ \mathbf{and}\ \mathbf{r'}\neq 0\\\frac{1}{2\pi}\mathbf{arctan}\left(\frac{\mathbf{r'}}{\mathbf{g'}}\right)+\frac{3}{4} & ,\mathbf{g'}<0\ \mathbf{and}\ \mathbf{r'}\neq 0\\\frac{1}{4} & ,\mathbf{g'}>0\ \mathbf{and}\ \mathbf{r'}=0\\\frac{3}{4} & ,\mathbf{g'}<0\ \mathbf{and}\ \mathbf{r'}=0\\\frac{1}{2} & ,\mathbf{g'}=0\ \mathbf{and}\ \mathbf{r'}>0\\0 & ,\mathbf{g'}=0\ \mathbf{and}\ \mathbf{r'}\leq 0\end{cases}$ $\mathbf{S}=\left[\frac{9}{5}\left(\mathbf{r'}^2+\mathbf{g'}^2\right)\right]^{1/2}$ $\mathbf{L}=0.299\mathbf{R}+0.587\mathbf{G}+0.114\mathbf{B}$	$(T,S,L)\in[0;1.0]$	**100 x 100**

Table 3. Non-linear conversions from RGB color space into perceptually plausible color spaces

COLOR SPACE	CONVERSION EQUATIONS	SPACE BOUNDARIES	HISTOGRAM No OF BINS
CIE-DSH	$D=\frac{1}{3}(R+G+B)$ $S=1-\frac{3[\min(R,G,B)]}{(R+G+B)}$ $H=\begin{cases}\frac{1}{2\pi}\arccos\left(\frac{1/2[\lvert R-G\rvert+\lvert R-B\rvert]}{[\lvert R-G\rvert^{2}+\lvert R-B\rvert\lvert G-B\rvert]^{1/2}}\right)+\frac{1}{2}, & G>B\\ \frac{1}{2}, & G=B\\ \frac{1}{2}-\frac{1}{2\pi}\arccos\left(\frac{1/2[\lvert R-G\rvert+\lvert R-B\rvert]}{[\lvert R-G\rvert^{2}+\lvert R-B\rvert\lvert G-B\rvert]^{1/2}}\right), & G<B\end{cases}$ if R=G=B, S=0, set H=0. If R=G=B=0, set S=H=0	$(D,S,H)\in[0;1.0]$	**100 x 100**
HSV	$H=\begin{cases}\frac{G-B}{\max(R,G,B)-\min(R,G,B)}, & R=\max(R,G,B)\\ 2+\frac{B-R}{\max(R,G,B)-\min(R,G,B)}, & G=\max(R,G,B)\\ 4+\frac{R-G}{\max(R,G,B)-\min(R,G,B)}, & B=\max(R,G,B)\end{cases}$ Normalize by setting $H=H/6$. If $H<0$, set $H=H+1$ $S=\frac{\max(R,G,B)-\min(R,G,B)}{\max(R,G,B)}, \quad \max(R,G,B)\neq 0$ $V=\max(R,G,B)$ If max(R,G,B)=min(R,G,B), then R=G=B, S=0, set H=0 If max(R,G,B)=0, then R=G=B=0, set S=H=0	$(H,S,V)\in[0;1.0]$	**100 x 100**
HSL	$H=\{\text{same as for HSV}$ $S=\begin{cases}\frac{\max(R,G,B)-\min(R,G,B)}{\max(R,G,B)+\min(R,G,B)}, & L\leq 0.5\\ \frac{\max(R,G,B)-\min(R,G,B)}{2-[\max(R,G,B)+\min(R,G,B)]}, & L>0.5\end{cases}$ $L=\frac{\max(R,G,B)+\min(R,G,B)}{2}$ Same limiting conditions as for HSV +if R=G=B=1, set S=H=0	$(H,S,L)\in[0;1.0]$	**100 x 100**

Table 4. Non-linear conversions from RGB color space into perceptually uniform color spaces

COLOR SPACE	CONVERSION EQUATIONS	SPACE BOUNDARIES	HISTOGRAM No OF BINS
CIE-L*u*v*	$L^* = \begin{cases} 116\left(\frac{Y}{Y_n}\right)^{1/3} - 16, & \frac{Y}{Y_n} > 0.008856 \\ 903.3\left(\frac{Y}{Y_n}\right), & \frac{Y}{Y_n} \leq 0.008856 \end{cases}$ $u^* = 13\,L^*\,(u' - u'_n),\quad v^* = 13\,L^*\,(v' - v'_n)$ $u' = \frac{4X}{X + 15Y + 3Z},\quad v' = \frac{9Y}{X + 15Y + 3Z}$ $u'_n = \frac{4X_n}{X_n + 15Y_n + 3Z_n},\quad v'_n = \frac{9Y_n}{X_n + 15Y_n + 3Z_n}$ (X_n, Y_n, Z_n) are the CIE-(X, Y, Z) components of a reference illuminant, e.g. the C or D 65 illuminants $\text{C ill.}: \begin{pmatrix} X_n \\ Y_n \\ Z_n \end{pmatrix} = \begin{pmatrix} 0.982 \\ 1.000 \\ 1.183 \end{pmatrix}\quad \text{D65 ill.}: \begin{pmatrix} X_n \\ Y_n \\ Z_n \end{pmatrix} = \begin{pmatrix} 0.950 \\ 1.000 \\ 1.089 \end{pmatrix}$ Achromatic point : $u^* = v^* = 0$;	$0 \leq L^* \leq 100$ Boundaries curved in u^*-v^* space	200 x 200 (skin) or 370 x 370 (non-skin)
CIE-L*a*b*	$L^* = \{$ same as in CIE-$L^*u^*v^*$ space $a^* = 500\left[f\left(\frac{X}{X_0}\right) - f\left(\frac{Y}{Y_0}\right)\right],$ $b^* = 200\left[f\left(\frac{X}{X_0}\right) - f\left(\frac{Y}{Y_0}\right)\right],$ $f(t) = \begin{cases} t^{1/3}, & t \leq 0.008856 \\ 7.787t + \frac{16}{116}, & t > 0.008856 \end{cases}$ $(X_0, Y_0, Z_0) = (X_n, Y_n, Z_n)$	$0 \leq L^* \leq 100$ Boundaries curved in a^*-b^* space	200 x 200 (skin) or 370 x 370 (non-skin)
F-uv	$U = \frac{2X}{3},\quad V = Y,\quad W = \frac{-X+3Y+Z}{2}$ $u_f = \frac{U}{U+V+W} = \frac{4X}{X+15Y+3Z} = \frac{4x}{-2x+12y+3}$ $v_f = \frac{V}{U+V+W} = \frac{6Y}{X+15Y+3Z} = \frac{6y}{-2x+12y+3}$	$u_f \in [0;4.0]$ $v_f \in [0;0.4]$	100 x 100

Table 5. Non-linear conversions from RGB color space into other color spaces

COLOR SPACE	CONVERSION EQUATIONS	SPACE BOUNDARIES	HISTOGRAM No OF BINS
$C_1C_2C_3$	$C_1 = \frac{2}{\pi}\arctan\left(\frac{R}{\max\{G, B\}}\right)$, $C_2 = \frac{2}{\pi}\arctan\left(\frac{G}{\max\{R, B\}}\right)$, $C_3 = \frac{2}{\pi}\arctan\left(\frac{B}{\max\{R, G\}}\right)$ If $R=G=B$, $C_i = 1/2$. If $R=G=B=0$, set $C_i = 0$	$C_i \in [0;\ 1.0]$ $i = 1, 2, 3$	100 x 100
$l_1l_2l_3$	$l_1 = \frac{\lvert R-G\rvert}{\lvert R-G\rvert+\lvert R-B\rvert+\lvert G-B\rvert}$ $l_2 = \frac{\lvert R-B\rvert}{\lvert R-G\rvert+\lvert R-B\rvert+\lvert G-B\rvert}$ $l_3 = \frac{\lvert G-B\rvert}{\lvert R-G\rvert+\lvert R-B\rvert+\lvert G-B\rvert}$ If $R=G=B$ or $R=G=B=0$, set $l_i = 1/3$	$l_i \in [0;\ 1/2]$ $i = 1, 2, 3$ $l_1 + l_2 + l_3 = 1$	100 x 100
$l'_1l'_2l'_3$	$l_1' = \frac{(R-G)^2}{(R-G)^2+(R-B)^2+(G-B)^2}$ $l_2' = \frac{(R-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2}$ $l_3' = \frac{(G-B)^2}{(R-G)^2+(R-B)^2+(G-B)^2}$ If $R=G=B$ or $R=G=B=0$, set $l'_i=1/3$	$l_i' \in [0;\ 2/3]$ $i = 1, 2, 3$ $l_1' + l_2' + l_3' = 1$	100 x 100
Ln-Chroma r_g, rg_b	$r\_g = \ln\left(\frac{R}{G}\right) = \ln R - \ln G = R' - G'$ $rg\_b = \ln\left(\frac{RG}{B^2}\right) = \ln R + \ln G - 2\ln B$ $= R' + G' - 2B'$	$r\_g \in\ ]-\infty;+\infty[$ $rg\_b \in\ ]-\infty;+\infty[$	300 x 300*
a' - b'	$a' = r + \frac{g}{2}$, $b' = \frac{\sqrt{3}}{2}g$	$a' \in [0;\ 1.0]$ $b' \in [0;\ \frac{\sqrt{3}}{2}]$	100 x 100
mod-rgb	Let $I = \sqrt{R^2+G^2+B^2}$, then $m\_r = R/I$, $m\_g = G/I$, $m\_b = B/I$ $(m\_r)^2 + (m\_g)^2 + (m\_b)^2 = 1$ If $R=G=B=0$, set $m\_r=m\_g=m\_b=\sqrt{3}/3$	$m\_r \in [0;1.0]$ $m\_g \in [0;1.0]$ $m\_b \in [0;1.0]$	100 x 100
P_1P_2	$P_1 = \frac{1}{\sqrt{2}}\frac{G-R}{R+G+B}$ $P_2 = \frac{1}{\sqrt{6}}\frac{2B-R-G}{R+G+B}$ If $R=G=B=0$, set $P_1=P_2=0$	$\lvert P_1\rvert \le 1/\sqrt{2}$ $-1/\sqrt{6} \le P_2 \le 2/\sqrt{6}$	150 x 150
$R_1R_2R_3$	$R_1 = G/R$, $R_2 = B/R$, $R_3 = B/G$ If $R=G=B=0$, set $R_1=R_2=R_3=0$	$R_i \in [0;\ \infty[$, $i = 1, 2, 3$	200 x 200*
Yuv	$u = \frac{U}{Y} = \frac{-0.147R-0.289G+0.436B}{0.299R+0.587G+0.114B}$ $= \frac{-0.583r-0.725g+0.436}{0.185r+0.473g+0.114}$ $v = \frac{V}{Y} = \frac{0.615R-0.515G-0.100B}{0.299R+0.587G+0.114B}$ $= \frac{0.715r-0.415g-0.100}{0.185r+0.473g+0.114}$	$u \in [\frac{-289}{587};\ \frac{436}{114}]$ $v \in [\frac{-515}{587};\ \frac{615}{299}]$	100 x 100 (skin) or 440 x 440 (non-skin)

4.2 Robustness to the Intrinsic Variability of Skin Color and Compactness of the Skin Distribution

As an example, Figure 1 shows the skin distribution separately for each of the three different groups of subjects for several representative chrominance spaces, for skin sample images recorded with the SONY camera. Visually, the skin distribution appears more elongated in spaces such as the h1-h3, uv (Yuv) and H-S (HSV) spaces, while it appears more compact in the r-g, CIE-xy, m-g-m-b and P1-P2 spaces, particularly for the Asian subjects. Figure 2 shows the non-skin distribution for all the 41 chrominance spaces, on a logarithmic scale in order to show all non-empty histogram bins. Since the non-skin distribution can include any color, Figure 2 indicates that the gamut in all spaces with rectangular boundaries, except in the CIE-DSH, HSV and HSL spaces, fills only a part of the entire space defined by the space boundaries, and its geometry depends on the space that is considered. Table 6 shows that the area of the gamut can be computed analytically for only a few spaces. Since the area in number of non-empty bins that is observed for those spaces is generally larger than 95% of the total computed area, we assume that the number of non-empty bins observed for the other spaces, as shown in Table 7, is very near the true area. Owing to the particular boundaries of the CIE-L*u*v*, CIE-L*a*b*, r-g-rg-b and R1R2R3 spaces, the area for these spaces is not considered here.

Table 8 shows the KLD and the HIN for the three skin groups for all spaces, when using the SONY camera. Table 9 shows, for each relevant space and for both cameras, the area of the skin distribution relative to the area of the gamut of possible colors. As the blue areas in both tables show, the normalized r-g and CIE-xy spaces, as well as the a'-b', m-r-m-g, m-g-m-b, P1-P2 and R1-R2 spaces yield the most robust distributions with respect to the intrinsic variability of skin color, because: 1) the KLD is consistently lower across the three skin groups than for the other spaces (mainly across the first two skin groups), and 2) the overlap between the skin groups varies typically within a relatively narrow range, between 45% and 64% for most spaces. The l1l2l3, l1'l2'l3' and C1-C2 spaces yield the highest overlap between skin groups, indicative of a higher robustness (relevant blue areas), but this advantage is offset by the large overlap between the skin and non-skin distributions in those spaces, as shown by the relevant pink areas, and also in Subsection 4.4. In almost all chrominance spaces, the distribution for the Asian subjects, who have an intermediate skin color, is the most compact, in terms of the relative area of the distribution, in particular in the r-g, CIE-xy, C1-C2, a'-b', mod-rgb and P1-P2 spaces. This result is confirmed visually by Figure 1.

4.3 Shape of the Skin Distribution

A few representative examples of the cumulative skin chrominance distribution for the Asian + Caucasian subjects, obtained with both cameras, are shown in Figure 3. Visually, the skin distribution in the normalized spaces fits well to the single Gaussian model, whereas in the un-normalized spaces, its shape is generally complex and cannot be described well by a simple model. Table 9 shows the KLD for all the chrominance spaces and for both cameras. Since the KLD is consistently lowest for the normalized r-g and CIE-xy spaces, together with the C_2-C_3, a'-b', mod-rgb and P_1P_2 spaces, the skin distribution in those spaces can be modeled by a single Gaussian.

4.4 Discrimination between Skin and Non-skin Distributions

Table 10 also shows the HIN for Asian + Caucasian subjects, for all the chrominance spaces and for both camera systems. For both camera systems, the overlap between the skin and non-skin distributions is lowest for the r-g, CIE-xy, TSL, CIE-DSH, HSV, CIE-L*u*v*, CIE-L*a*b*, C_2-C_3, C_1-C_3, r_g-rg_b (ln-chroma), a'-b', mod-rgb, P_1P_2 and $R_1R_2R_3$ spaces. Hence, the discrimination capabilities between skin pixels and non-skin pixels are highest in these spaces. The low overlap observed in the CIE-DSH and HSV spaces may be due to the fact that the gamut in those spaces fills the entire space defined by the space boundaries. The lowest discrimination is found for the $l_1l_2l_3$ and l_1'l_2'l_3' spaces, as we already mentioned in Subsection 4.2.

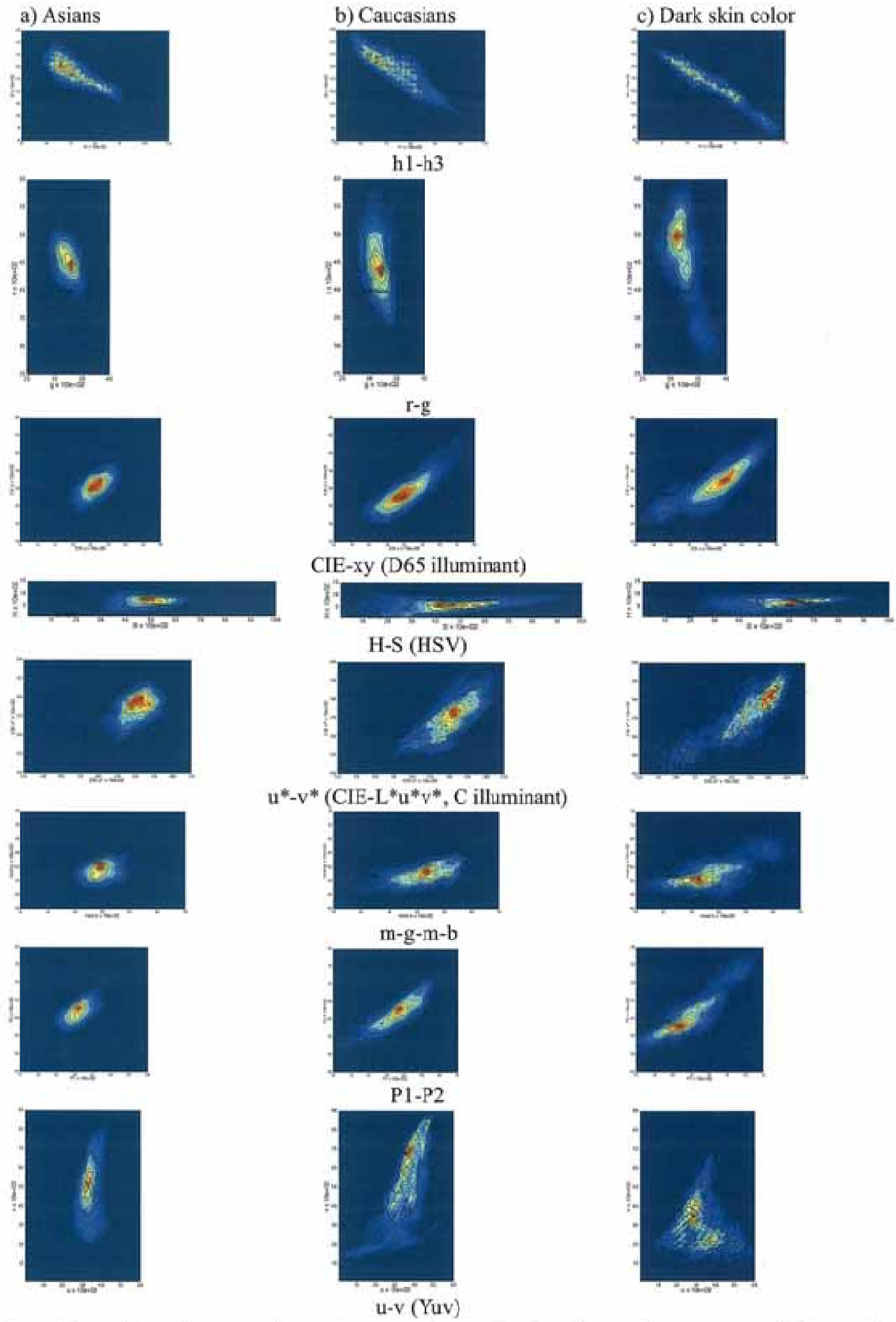
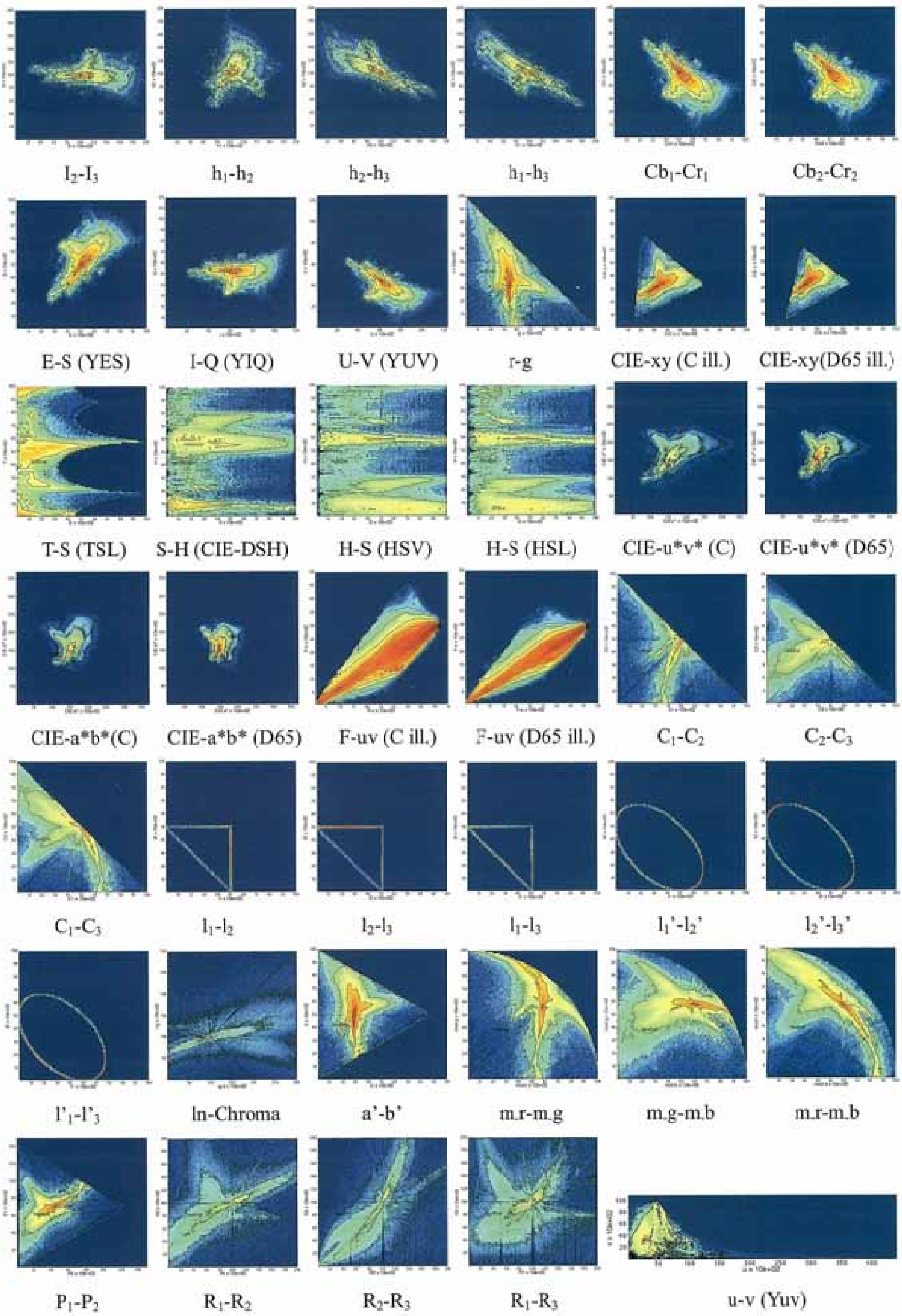


Figure 1. 2-D top view of the cumulative histograms in several selected chrominance spaces of skin sample images of a) Asian, b) Caucasian subjects, and c) of subjects with dark skin color, recorded with the SONY camera. Here, only the relevant part of the histograms is shown.



I_2-I_3 h_1-h_2 h_2-h_3 h_1-h_3 Cb_1-Cr_1 Cb_2-Cr_2

E-S (YES) I-Q (YIQ) U-V (YUV) r-g CIE-xy (C ill.) CIE-xy(D65 ill.)

T-S (TSL) S-H (CIE-DSH) H-S (HSV) H-S (HSL) CIE-u*v* (C) CIE-u*v* (D65)

CIE-a*b*(C) CIE-a*b* (D65) F-uv (C ill.) F-uv (D65 ill.) C_1-C_2 C_2-C_3

C_1-C_3 l_1-l_2 l_2-l_3 l_1-l_3 l_1'-l_2' l_2'-l_3'

l'_1-l'_3 ln-Chroma a'-b' m.r-m.g m.g-m.b m.r-m.b

P_1-P_2 R_1-R_2 R_2-R_3 R_1-R_3 u-v (Yuv)

Figure 2. 2-D top view of the cumulative histograms in 41 different chrominance spaces of 80 non-skin sample images collected from various sources, on a logarithmic scale in order to show all non-empty bins.

Table 6. Area of the gamut of all possible colors in a given chrominance space computed analytically, and corresponding number of non-empty bins in the discrete histograms, for 10 different color spaces (16 chrominance spaces).

COLOR SPACE	COMPUTED	No OF BINS
r-g	1/2	5,000
CIE-DSH	1.0	10,000
HSV	1.0	10,000
HSL	1.0	10,000
$C_1C_2C_3$	1/2	5,000
r_g-rg_b	∞	∞
a' - b'	$\sqrt{3}/4$	4,330
mod-rgb	$\pi/4$	7,854
P_1-P_2	$\sqrt{3}/2$	8,660
$R_1R_2R_3$	∞	∞

Table 7. Area of the gamut of all possible colors in a given chrominance space expressed as the number of non-empty bins in the discrete histograms, for 15 different color spaces (21 chrominance spaces).

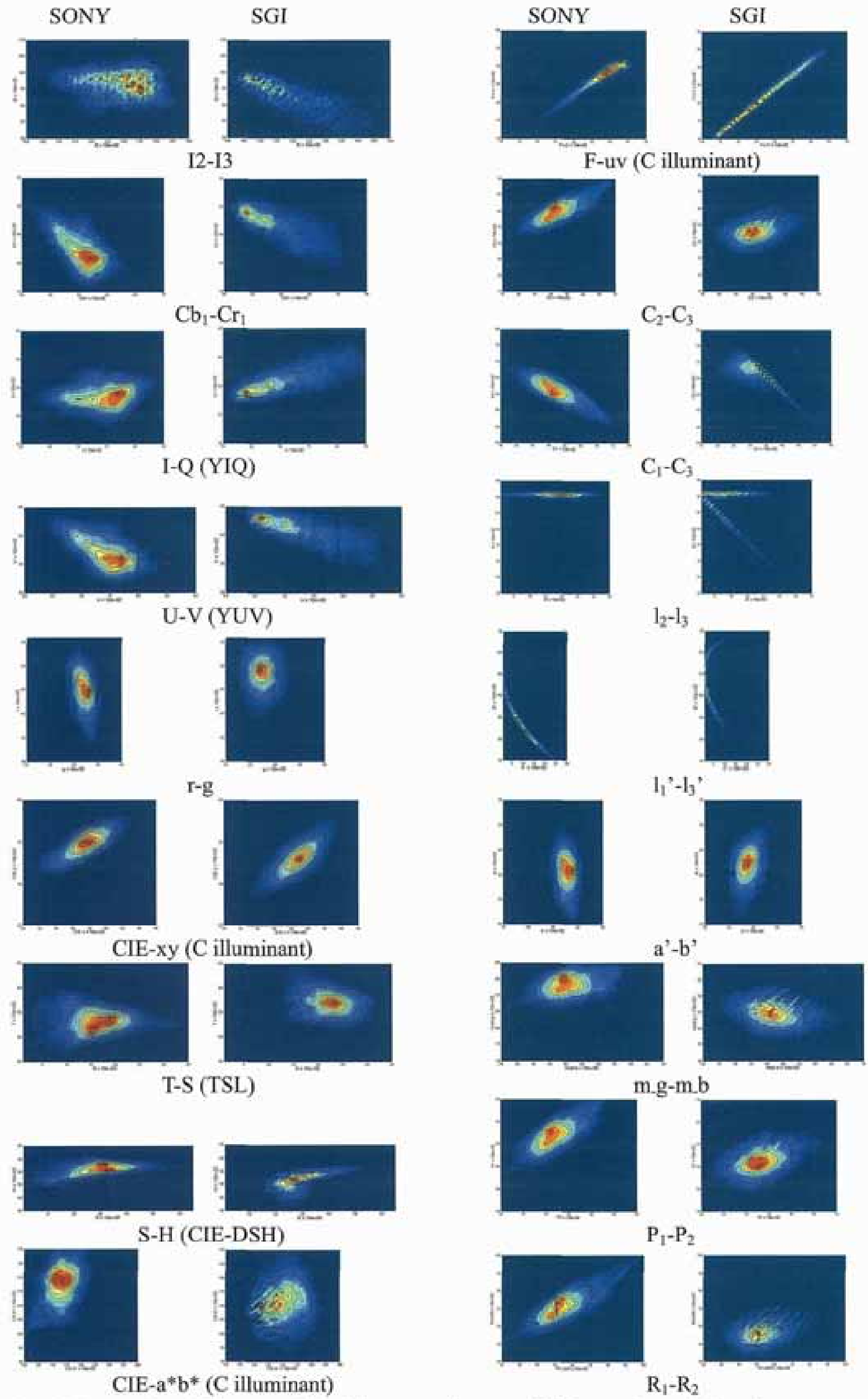
COLOR SPACE	No OF BINS
I_2-I_3	9,540
$h_1h_2h_3$	9,533
Cb_1-Cr_1	2,616
Cb_2-Cr_2	2,659
E-S	2,760
I-Q	2,793
U-V	2,782
CIE-xy (C III.)	1,692
CIE-xy (D65 III.)	1,239
TSL	5,317
F-uv (C III.)	919
F-uv (D65 III.)	745
$I_1I_2I_3$	310
$I'_1I'_2I'_3$	265
Yuv	10,291

Table 8. Fit of the skin distribution to a single Gaussian (KLD) and overlap of skin/skin and skin/non-skin distributions (HIN) for Asians (A), Caucasians (C) and subjects with dark skin color (D), for 41 chrominance spaces and for skin sample images recorded with the SONY camera.

	KULLBACK-LEIBLER DIVERGENCE			OVERLAP (HIN) BETWEEN DIFFERENT SKIN GROUPS			OVERLAP (HIN) BETWEEN SKIN GROUPS AND NON-SKIN		
COLOR SPACE	A	C	D	A / C	A / D	C / D	A / NS	C / NS	D / NS
l_2-l_3	1.1771	0.9658	2.2285	0.5919	0.6259	0.5349	0.1444	0.1992	0.2460
h_1-h_2	1.3536	0.9798	2.2138	0.5934	0.6255	0.5363	0.1463	0.2000	0.2489
h_2-h_3	1.1237	0.9539	2.4292	0.5943	0.6244	0.5395	0.1461	0.2009	0.2503
h_1-h_3	1.1950	0.8987	2.7134	0.5935	0.6285	0.5355	0.1471	0.2004	0.2505
Cb_1-Cr_1	1.0569	0.7582	2.3620	0.6034	0.6464	0.5499	0.1520	0.2083	0.2632
Cb_2-Cr_2	1.0990	0.7198	2.1675	0.6012	0.6489	0.5511	0.1509	0.2063	0.2645
E-S	1.1008	0.7440	1.6099	0.6025	0.6475	0.5528	0.1506	0.2095	0.2609
I-Q	1.1613	0.6412	1.8226	0.6030	0.6415	0.5620	0.1529	0.2088	0.2662
U-V	1.0696	0.7073	2.1329	0.6034	0.6495	0.5561	0.1536	0.2093	0.2674
r-g	0.4878	0.6690	1.7057	0.5010	0.4739	0.5399	0.1047	0.1894	0.2558
CIE-xy (C Ill.)	0.3771	0.5269	0.9205	0.5031	0.4882	0.5661	0.1132	0.2014	0.2637
CIE-xy (D65 Ill.)	0.4323	0.4906	0.8467	0.5103	0.5066	0.5692	0.1151	0.2017	0.2686
TSL	10.0736	4.6134	17.2312	0.4909	0.4681	0.5289	0.1112	0.1826	0.2432
CIE-DSH	0.8066	8.7930	18.7132	0.4929	0.4688	0.5284	0.1026	0.1804	0.2305
HSV	2.7220	20.0670	17.7311	0.4922	0.4675	0.5271	0.1020	0.1803	0.2227
HSL	1.7272	18.6020	17.7355	0.4780	0.5537	0.6204	0.1405	0.2121	0.2200
CIE-L*u*v* (C Ill.)	0.2653	0.4121	3.3201	0.5535	0.5456	0.5905	0.1124	0.1935	0.2388
CIE-L*u*v* (D65 Ill.)	0.2828	0.2832	2.8929	0.5944	0.5275	0.6503	0.1147	0.1840	0.2283
CIE-L*a*b* (C Ill.)	0.1979	0.6845	3.4265	0.5283	0.5308	0.5742	0.1096	0.1975	0.2404
CIE-L*a*b* (D65 Ill.)	0.2137	0.3792	2.4141	0.5679	0.5111	0.6395	0.1103	0.1914	0.2471
F-uv (C Ill.)	1.6057	4.8038	1.6153	0.5322	0.4002	0.3969	0.1830	0.2367	0.2567
F-uv (D65 Ill.)	1.6348	4.0665	1.0664	0.5572	0.4156	0.3936	0.2140	0.2487	0.2743
c_1-c_2	23.6305	26.1568	29.4019	0.7683	0.5471	0.6909	0.1671	0.2637	0.2931
c_2-c_3	0.2057	2.0942	5.8982	0.4947	0.4690	0.5355	0.1029	0.1862	0.2593
c_1-c_3	0.6431	12.8966	16.1455	0.4947	0.4692	0.5366	0.1029	0.2026	0.2484
l_1-l_2	1.9433	20.1158	28.7567	0.6756	0.8200	0.7159	0.3900	0.4103	0.4870
l_2-l_3	1.9335	19.9077	28.7633	0.6757	0.8189	0.7188	0.3920	0.4117	0.4846
l_1-l_3	28.4289	21.3616	29.0135	0.6808	0.8330	0.7172	0.3872	0.4093	0.4827
l'_1-l'_2	20.6039	26.6590	30.5420	0.6823	0.8217	0.7145	0.3902	0.4123	0.4807
l'_2-l'_3	22.4415	28.3576	31.0432	0.6825	0.8331	0.7200	0.3928	0.4138	0.4825
l'_1-l'_3	20.8104	26.2118	29.8854	0.6821	0.8212	0.7175	0.3889	0.4087	0.4819
r_g-rg_b	0.4316	4.3017	10.6025	0.4764	0.4285	0.4766	0.0933	0.1696	0.1726
a' - b'	0.4282	0.6836	1.5669	0.4995	0.4820	0.5444	0.1046	0.1873	0.2533
m_r-m_g	0.5026	0.5638	3.6269	0.4988	0.4709	0.5334	0.1036	0.1852	0.2455
m_g-m_b	0.1916	0.6016	1.4755	0.4890	0.4658	0.5246	0.1014	0.1818	0.2435
m_r-m_b	0.3913	1.8047	5.6624	0.4946	0.4714	0.5352	0.1023	0.1848	0.2450
P_1-P_2	0.1613	0.6085	1.6328	0.4928	0.4698	0.5331	0.1031	0.1851	0.2492
R_1-R_2	0.2508	0.6462	7.9108	0.4853	0.4575	0.5121	0.0995	0.1768	0.2084
R_2-R_3	0.3881	2.1389	11.7300	0.4850	0.4590	0.5115	0.0988	0.1766	0.2114
R_1-R_3	0.3101	1.2686	9.8428	0.4841	0.4535	0.5020	0.0971	0.1734	0.2022
Yuv	0.9517	3.0968	3.7347	0.5109	0.3532	0.3559	0.1651	0.2608	0.2546

Table 9. Area of the skin distribution for Asians (A), Caucasians (C), and subjects with dark skin color (D) relative to the area of the gamut of all possible colors in a given chrominance space, for 33 of 41 chrominance spaces and for both the SONY and SGI camera systems.

	AREA OF SKIN DISTRIBUTIONS (% OF ALL SPACE)					
	SONY DXC-9000					SGI
COLOR SPACE	A	C	D	A+ C	A+C+D	A+ C
I_2-I_3	6.279	8.396	5.975	9.518	10.388	18.040
h_1-h_2	6.273	8.465	5.916	9.535	10.469	17.707
h_2-h_3	6.346	8.528	5.948	9.682	10.563	17.644
h_1-h_3	6.294	8.486	5.811	9.556	10.416	17.749
Cb_1-Cr_1	6.499	8.945	6.499	10.092	11.124	23.012
Cb_2-Cr_2	6.544	8.988	6.581	10.041	10.982	22.828
E-S	6.558	8.804	6.522	9.783	10.725	22.971
I-Q	6.481	8.844	6.447	9.882	10.813	23.022
U-V	6.505	9.166	6.505	10.209	11.107	23.077
r-g	3.260	8.900	10.000	9.080	13.140	12.000
CIE-xy (C Ill.)	4.078	12.234	12.707	12.411	17.317	14.894
CIE-xy (D65 Ill.)	4.278	12.752	13.156	12.833	17.918	17.111
TSL	6.188	15.103	18.450	15.648	24.074	16.005
CIE-DSH	5.380	12.600	15.420	13.120	20.390	14.050
HSV	6.470	14.730	18.500	15.420	25.160	15.430
HSL	7.240	15.380	15.370	15.910	23.100	20.020
F-uv (C Ill.)	12.622	19.804	20.131	20.131	26.551	26.659
F-uv (D65 Ill.)	12.349	19.329	20.269	19.597	25.638	28.993
C_1-C_2	0.760	3.260	7.160	3.280	8.540	11.260
C_2-C_3	6.680	13.760	12.940	15.600	18.040	20.620
C_1-C_3	6.400	15.160	16.440	14.180	20.720	10.800
I_1-I_2	27.419	53.226	56.129	53.226	58.710	69.355
I_2-I_3	27.419	54.839	57.097	54.839	60.645	61.290
I_1-I_3	25.807	53.548	55.484	53.871	60.000	65.807
I'_1-I'_2	35.094	72.453	75.472	72.453	84.906	89.434
I'_2-I'_3	30.566	67.547	73.585	68.302	82.264	84.151
I'_1-I'_3	39.623	71.698	76.981	72.830	83.774	88.302
a' - b'	3.349	9.261	10.370	9.423	13.764	12.494
m_r-m_g	3.489	6.685	10.008	7.550	11.752	11.879
m_g-m_b	5.895	14.311	14.337	14.680	19.430	18.118
m_r-m_b	4.469	10.122	9.740	12.401	12.270	15.623
P_1-P_2	2.991	7.979	8.718	8.153	11.571	10.347
Yuv	12.283	19.901	15.927	20.581	24.507	30.833



I2-I3

Cb$_1$-Cr$_1$

I-Q (YIQ)

U-V (YUV)

r-g

CIE-xy (C illuminant)

T-S (TSL)

S-H (CIE-DSH)

CIE-a*b* (C illuminant)

F-uv (C illuminant)

C$_2$-C$_3$

C$_1$-C$_3$

l$_2$-l$_3$

l$_1$'-l$_3$'

a'-b'

m.g-m.b

P$_1$-P$_2$

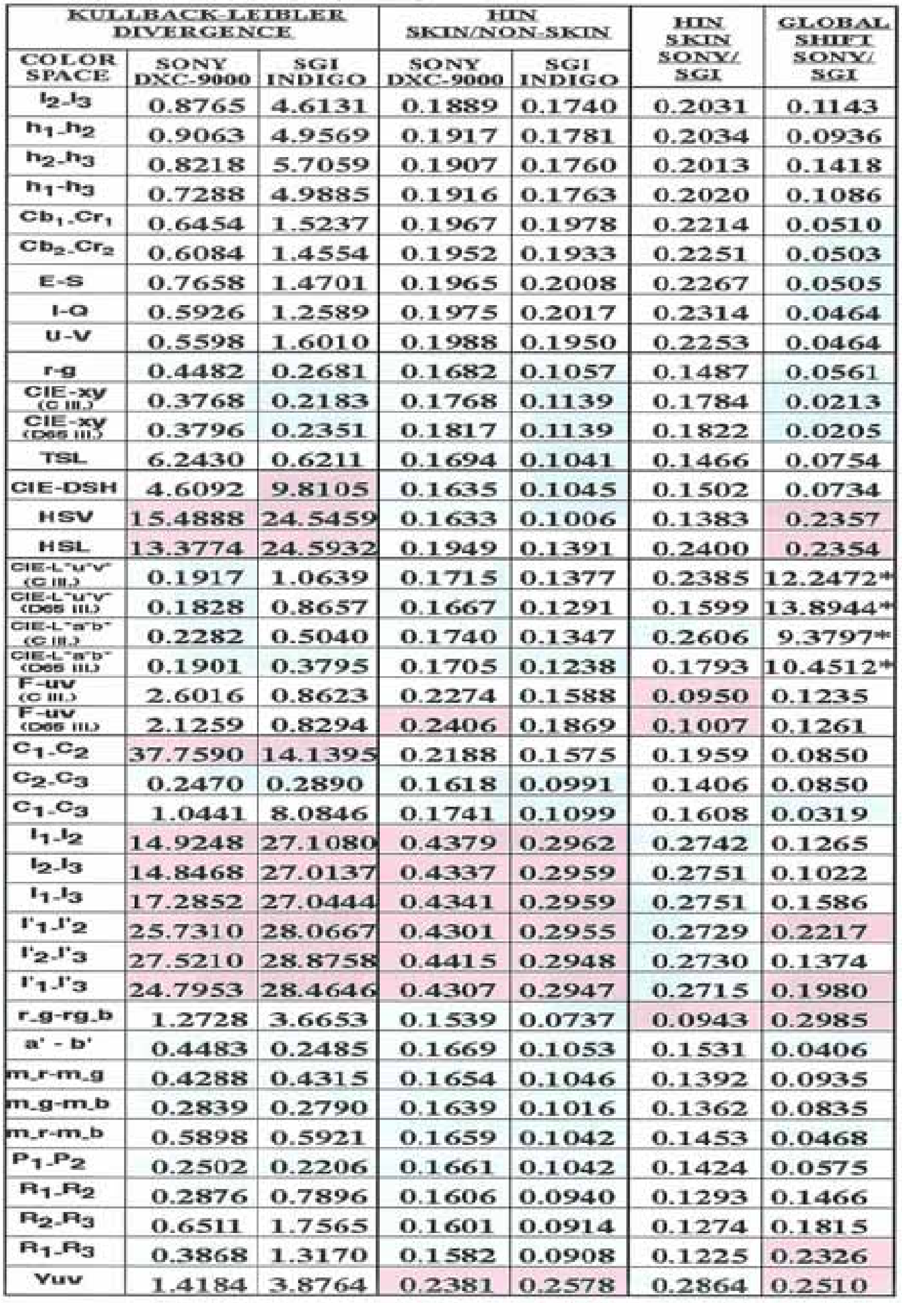
R$_1$-R$_2$

Figure 3.　2-D top view of the cumulative histograms in several different selected chrominance spaces of skin sample images of Asian + Caucasian subjects recorded with the SONY camera (left) and with the SGI camera (right). Here, only the relevant part of the histogram is shown.

4.5 Robustness to a Change of Camera System

The robustness to a change of camera system can be measured as the change in the KLD, in the HIN and the global shift S of the distribution. As seen from Table 10, the change in the KLD is lowest for the r-g, CIE-xy, C_1-C_3, a'-b', m.r-m.b and P_1-P_2 spaces, while the overlap of the skin distributions between the two camera systems (HIN skin SONY/SGI) is intermediate to low for those spaces. The highest overlap is found for the $l_1l_2l_3$, and $l_1'l_2'l_3'$ but, as for the overlap between the three different skin groups, for those spaces this advantage is offset by a significant overlap between the skin and non-skin distributions, and also by very large values of the KLD for both cameras. The global shift S of the distribution is low to lowest for the above-mentioned 6 chrominance spaces, and is also low for some of the spaces resulting from a linear transformation from the RGB space. The low global shift for the r-g, CIE-xy, a'-b' and P_1-P_2 spaces is confirmed visually by Figure 3.

Table 10. Fit of the skin distribution to a single Gaussian (KLD), overlap of skin/non-skin and skin/skin distributions (HIN), and global shift of the skin distribution for Asian + Caucasian subjects, for 41 chrominance spaces and for skin sample images recorded with both the SONY and SGI cameras.



	KULLBACK-LEIBLER DIVERGENCE		HIN SKIN/NON-SKIN		HIN SKIN SONY/SGI	GLOBAL SHIFT SONY/SGI
COLOR SPACE	SONY DXC-9000	SGI INDIGO	SONY DXC-9000	SGI INDIGO		
l_2-l_3	0.8765	4.6131	0.1889	0.1740	0.2031	0.1143
h_1-h_2	0.9063	4.9569	0.1917	0.1781	0.2034	0.0936
h_2-h_3	0.8218	5.7059	0.1907	0.1760	0.2013	0.1418
h_1-h_3	0.7288	4.9885	0.1916	0.1763	0.2020	0.1086
Cb_1-Cr_1	0.6454	1.5237	0.1967	0.1978	0.2214	0.0510
Cb_2-Cr_2	0.6084	1.4554	0.1952	0.1933	0.2251	0.0503
E-S	0.7658	1.4701	0.1965	0.2008	0.2267	0.0505
I-Q	0.5926	1.2589	0.1975	0.2017	0.2314	0.0464
U-V	0.5598	1.6010	0.1988	0.1950	0.2253	0.0464
r-g	0.4482	0.2681	0.1682	0.1057	0.1487	0.0561
CIE-xy (C Ill.)	0.3768	0.2183	0.1768	0.1139	0.1784	0.0213
CIE-xy (D65 Ill.)	0.3796	0.2351	0.1817	0.1139	0.1822	0.0205
TSL	6.2430	0.6211	0.1694	0.1041	0.1466	0.0754
CIE-DSH	4.6092	9.8105	0.1635	0.1045	0.1502	0.0734
HSV	15.4888	24.5459	0.1633	0.1006	0.1383	0.2357
HSL	13.3774	24.5932	0.1949	0.1391	0.2400	0.2354
CIE-L*u*v* (C Ill.)	0.1917	1.0639	0.1715	0.1377	0.2385	12.2472*
CIE-L*u*v* (D65 Ill.)	0.1828	0.8657	0.1667	0.1291	0.1599	13.8944*
CIE-L*a*b* (C Ill.)	0.2282	0.5040	0.1740	0.1347	0.2606	9.3797*
CIE-L*a*b* (D65 Ill.)	0.1901	0.3795	0.1705	0.1238	0.1793	10.4512*
F-uv (C Ill.)	2.6016	0.8623	0.2274	0.1588	0.0950	0.1235
F-uv (D65 Ill.)	2.1259	0.8294	0.2406	0.1869	0.1007	0.1261
C_1-C_2	37.7590	14.1395	0.2188	0.1575	0.1959	0.0850
C_2-C_3	0.2470	0.2890	0.1618	0.0991	0.1406	0.0850
C_1-C_3	1.0441	8.0846	0.1741	0.1099	0.1608	0.0319
l_1-l_2	14.9248	27.1080	0.4379	0.2962	0.2742	0.1265
l_2-l_3	14.8468	27.0137	0.4337	0.2959	0.2751	0.1022
l_1-l_3	17.2852	27.0444	0.4341	0.2959	0.2751	0.1586
l'_1-l'_2	25.7310	28.0667	0.4301	0.2955	0.2729	0.2217
l'_2-l'_3	27.5210	28.8758	0.4415	0.2948	0.2730	0.1374
l'_1-l'_3	24.7953	28.4646	0.4307	0.2947	0.2715	0.1980
r_g-rg_b	1.2728	3.6653	0.1539	0.0737	0.0943	0.2985
a' - b'	0.4483	0.2485	0.1669	0.1053	0.1531	0.0406
m_r-m_g	0.4288	0.4315	0.1654	0.1046	0.1392	0.0935
m_g-m_b	0.2839	0.2790	0.1639	0.1016	0.1362	0.0835
m_r-m_b	0.5898	0.5921	0.1659	0.1042	0.1453	0.0468
P_1-P_2	0.2502	0.2206	0.1661	0.1042	0.1424	0.0575
R_1-R_2	0.2876	0.7896	0.1606	0.0940	0.1293	0.1466
R_2-R_3	0.6511	1.7565	0.1601	0.0914	0.1274	0.1815
R_1-R_3	0.3868	1.3170	0.1582	0.0908	0.1225	0.2326
Yuv	1.4184	3.8764	0.2381	0.2578	0.2864	0.2510

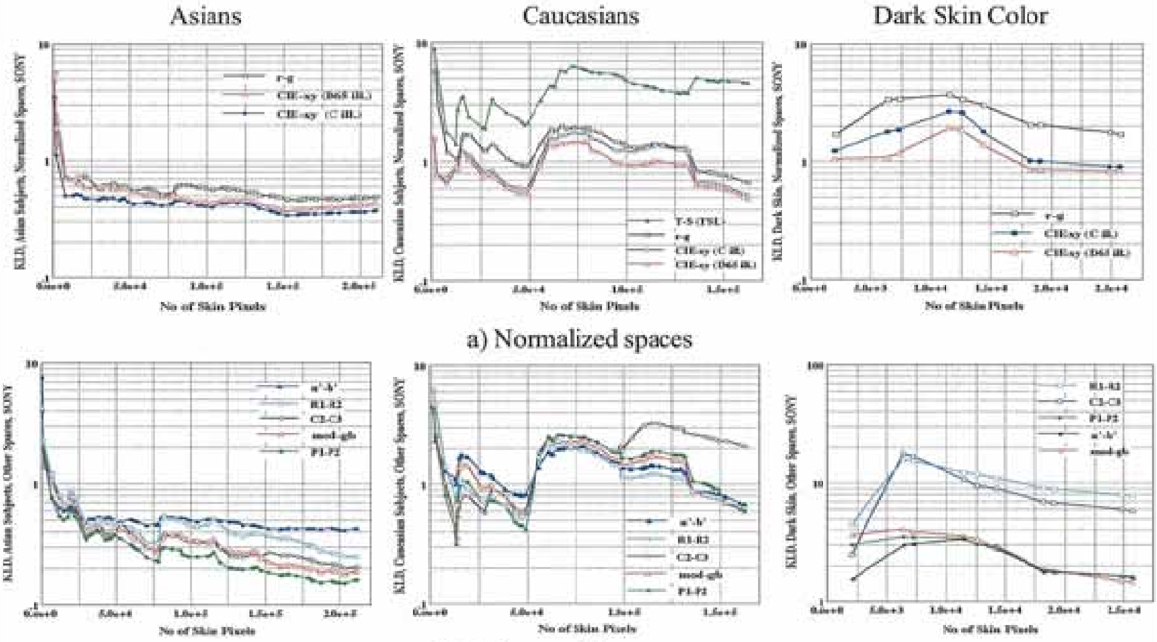
4.6 Robustness to Changes in Illumination Conditions and Computational Cost of the Color Space Transformation

Finally, it is well known that a normalization of RGB values by (R+G+B) or of CIE-XYZ values by (X+Y+Z) reduces the most the sensitivity of the skin distribution to changes in illumination (the normalized spaces are robust to minor changes in illumination conditions), and a linear transformation from the RGB space, or a non-linear conversion into the normalized rgb coordinates and into the CIE-xyz space is not computationally intensive compared to that into other spaces. The mod-rgb space also provides a suitable normalization.

4.7 Influence of the Number of Skin Sample Pixels on the KLD and on the HIN

The results of the skin chrominance analysis have so far been presented for the fixed numbers of skin sample images (or of skin sample pixels) specified in Subsection 4.1. Although these numbers are believed to be sufficiently large to warrant a statistical analysis, the results of the skin chrominance analysis depend on the number of skin sample pixels. Thus, it is necessary to analyze the dependency of the skin chrominance distribution on the number of skin sample pixels (or images) that are collected in order to statistically validate the results presented previously in this report [23]. Most importantly, the variation of the shape of the skin distribution, that determines the complexity of the skin chrominance model that is required in order to obtain a high efficiency of skin pixel detection, and the variation of the degree of discrimination (or the overlap) between the skin and the non-skin distributions, that ultimately limits the performance of skin pixel detection, are to be analyzed. Here we consider a subset of the global set of chrominance spaces, namely the normalized and perceptually plausible spaces, as well as some members of the "other non-linear spaces" sub-group.

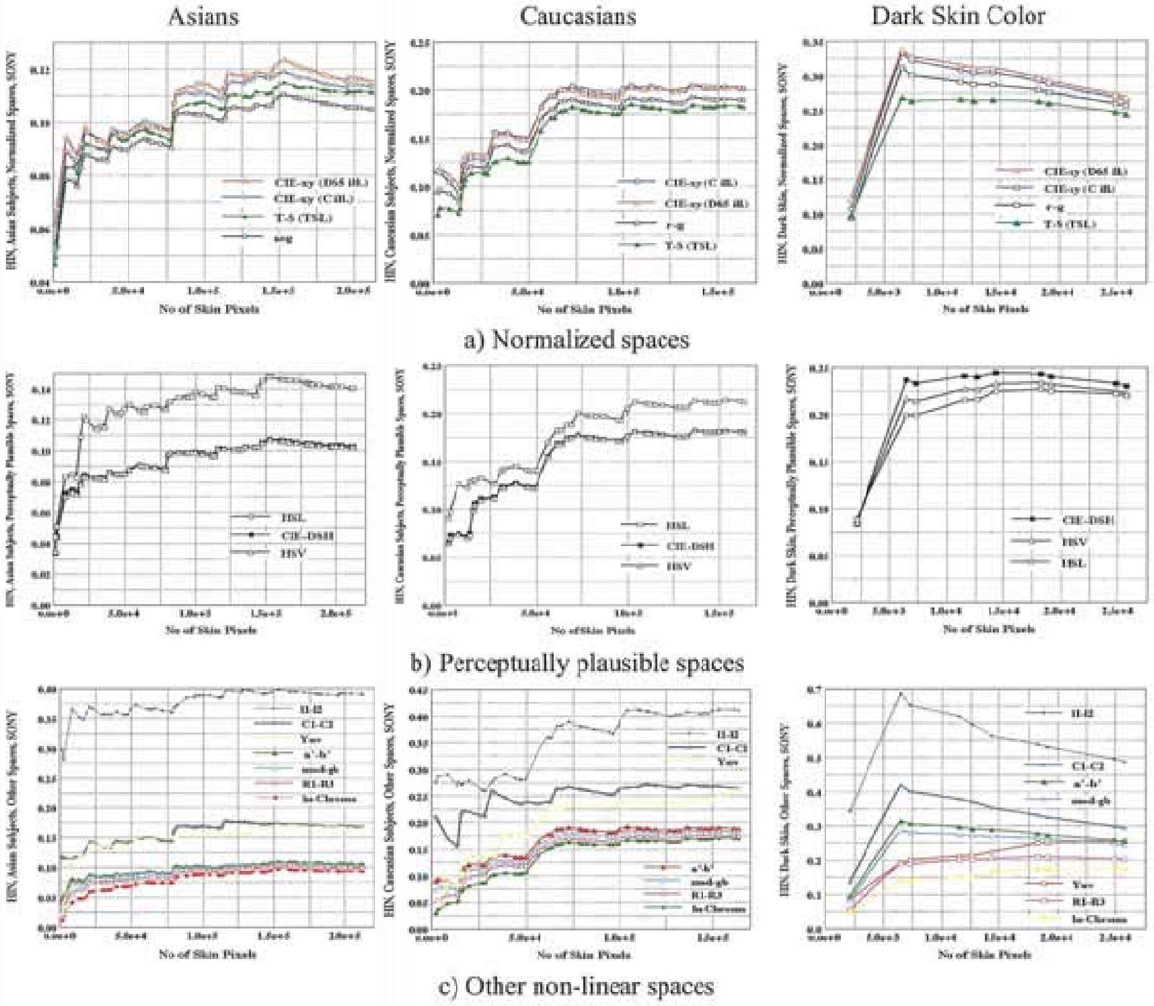
Figure 4 shows graphs of the KLD as a function of the cumulative number of skin sample pixels from the skin sample images collected with the SONY camera, for the three different skin groups separately, for the normalized spaces and for the "other non-linear spaces" sub-group. Figure 5 shows the corresponding graphs of the HIN, for the normalized spaces, the perceptually plausible spaces, and finally for the other non-linear spaces sub-group. The HIN is of course calculated for the usual fixed number of 80 non-skin sample images or equivalently, of 2.6606x10E+06 non-skin sample pixels.



a) Normalized spaces

b) Other non-linear spaces

Figure 4. Graphs of the Kullback-Leibler Divergence (KLD) as a function of the cumulative number of skin pixels from skin sample images collected with the SONY camera, for Asian (left column), Caucasian (middle column), and dark-skin colored subjects (right column), a) for the normalized chrominance spaces, and b) for the "other non-linear spaces" sub-group.



a) Normalized spaces

b) Perceptually plausible spaces

c) Other non-linear spaces

Figure 5. Graphs of the Normalized Histogram Intersection (HIN) between skin and non-skin distributions as a function of the cumulative number of skin pixels from skin sample images collected with the SONY camera, for Asian (left column), Caucasian (middle column), and dark-skin colored subjects (right column), a) for the normalized chrominance spaces, b) for the perceptually plausible spaces, and c) for the "other non-linear spaces" sub-group.

In the case of the KLD, for the Asian subjects, the KLD decreases rapidly when the number of skin sample pixels Ns is small, when it is less than about 1.0 x 10E+04 for the normalized spaces and 3.0 x 10E+04 for the other non-linear spaces sub-group. For larger Ns, it converges rapidly and saturates at low values. The behavior of the KLD is more complex for the Caucasian subjects, but convergence also occurs when Ns is larger than about 7.0 x 10E+04. The KLD for the dark skin group also converges, but always to significantly high values compared to the other two groups. Thus, the fit to a single elliptical Gaussian is generally highest for the Asian subjects, and only a relatively small number of skin sample pixels are required for calibration (with the single Gaussian model, typically by use of the Mahalanobis metric [2] [4]).

In all cases, the HIN increases rapidly when Ns is small (as new skin sample images are added to the histograms), generally, when it is less than about 2.5 x 10E+04 for Asian subjects and 7.0 x 10E+04 for Caucasian subjects, but for larger Ns it converges rapidly and saturates at a stable value. The comparison with the dark skin group is made difficult owing to the comparatively small number of skin sample images from that group that were available, but it can be seen that the HIN for dark skin colors always converges to higher values than that for the Caucasian subjects, and the HIN for the latter group always saturates at higher values than the HIN for the Asian subjects. Hence, the degree of discrimination between skin and non-skin pixels is generally the highest, and significantly so, for the Asian subjects than for the other two skin groups, the degree of discrimination for the Caucasian subjects reaching an intermediate level. Subsequently, the highest performance of skin pixel detection may be achieved for the Asian subjects, once again requiring only a relatively small number of skin sample pixels for calibration. We note that, when Ns is small, hence for a small initial number of skin sample images, the order in which the sample images are added to the cumulative histograms may significantly influence the precise shape of the curves for both the HIN and the KLD, but the convergence, and eventually the saturation of the curves are statistically warranted as Ns increases to large values.

In conclusion, given that, for the total number of skin sample pixels (images) that we collected, the KLD and the HIN lie well into the saturation range, and by generalizing the results presented here to the chrominance spaces that we did not consider in this subsection, the previous results presented in this report may be considered to be statistically valid.

4.8 Example of Skin Pixel Detection

The skin pixel detection is performed by use of the Mahalanobis metric, that is inherent to the single Gaussian chrominance model. A detailed analysis of the skin color calibration, the thresholding and subsequent segmentation algorithm can be found in [2] [4]. Thresholded images are subjected to a connected-component analysis. Figure 6 shows the segmentation (and face detection) results for two images of Asian subjects with a complex scene background, for the r-g space and for the H-S (HSV) space. Due to the complex shape of the skin distribution in the H-S space, in that space a large number of non-skin pixels are misclassified as skin, and parts of the background are connected to faces and to a hand, thus making difficult the task of automatic face or hand detection.

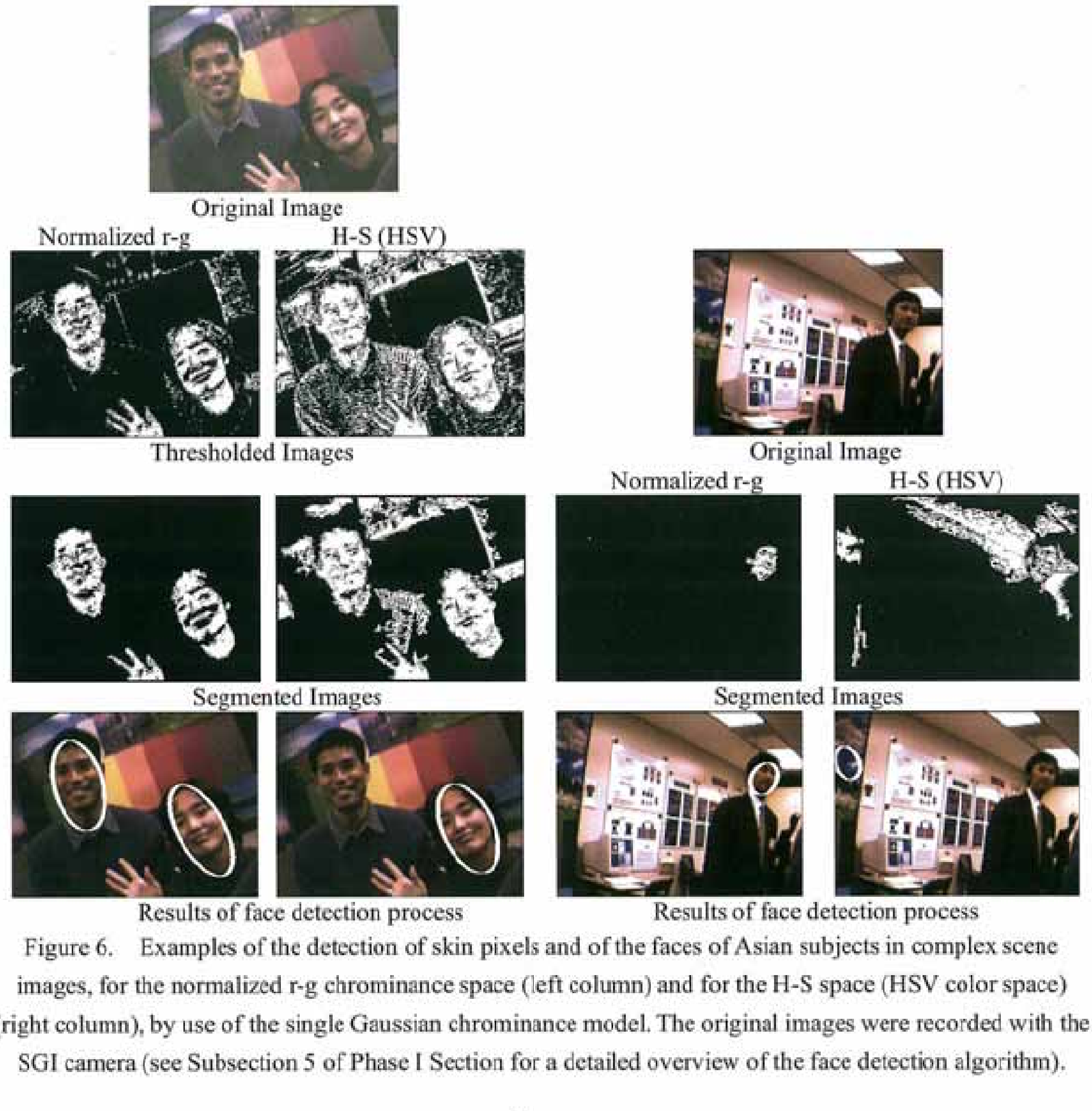


Figure 6. Examples of the detection of skin pixels and of the faces of Asian subjects in complex scene images, for the normalized r-g chrominance space (left column) and for the H-S space (HSV color space) (right column), by use of the single Gaussian chrominance model. The original images were recorded with the SGI camera (see Subsection 5 of Phase I Section for a detailed overview of the face detection algorithm).

4.9 Analysis of the KLD and of the HIN for Skin Sample Images Recorded under Unconstrained Scene Conditions

We now consider the distribution of human skin for skin sample images recorded under unconstrained scene conditions [24]. Typically, unconstrained environments include indoor or outdoor illumination conditions that are highly variable both in space and time, complex scenes with a large variety of objects, different camera systems used to record images, and also, in our analysis, a large spectrum of intrinsic skin colors (physically, the skin spectral reflectance power density and its variabiality along a continuum of possible skin colors). Hence, unconstrained environments imply that scene conditions are generally uncontrolled and/or uncontrollable. The World Wide Web (WWW) may be considered to be a suitable medium to collect a statistically significant number of skin sample pixels under such conditions. We therefore semi-randomly selected and manually collected a large number of skin and non-skin sample images on the WWW: In total, 300 skin sample images and 80 large non-skin sample images were collected (the non-skin sample images being the same ones as those that we have used previously), yielding a total of 1.118237x10E+06 skin pixels and of 2.6606x10E+06 non-skin pixels respectively. Figure 7 shows several examples of skin sample images that reveal the large diversity of observable skin colors, particularly under unconstrained image scene conditions. Examples of the non-skin sample images can be seen in Figure 8. Figure 9 shows the cumulative chrominance distribution for all the skin sample images, for a sub-set of 6 different chrominance spaces. A logarithmic scale is used so that all histogram bins that are not empty can be seen, because the distribution of dark skin colors is very diffuse [4]. Visually, the skin distribution in each space covers a much larger surface area and appears to be significantly more complex-shaped than the distributions in the same spaces for the three different skin color groups (Asians, Caucasians, and a dark skin group) that were shown for controlled image scene conditions. In particular, in the perceptually plausible chrominance spaces, the skin distribution covers the whole range of saturation S and a significantly larger range of hue H (or of tint T) at low values of S, thus requiring for these spaces a complex model for skin pixel detection under general, unconstrained image scene conditions.

We now analyze the variation of the shape of the skin distribution or its fit to a single elliptical Gaussian in terms of the KLD, and the variation of the degree of discrimination between skin pixels and non-skin pixels in terms of the HIN (for the fixed cumulative number of 2.6606x10E+06 non-skin pixels), as a function of the cumulative number N_s of collected skin sample pixels. Figures 10 and 11 show graphs of the KLD and of the HIN respectively, for 6 different chrominance space sub-groups: the spaces resulting from a linear transformation from the RGB color space, the normalized, perceptually plausible and perceptually uniform spaces, and finally for the other non-linear spaces sub-group, divided further into two smaller subsets for visual clarity.

For the large majority of chrominance spaces, the KLD first decreases rapidly and then fluctuates at low values of N_s, as new skin sample images are added to the histograms, typically when N_s is less than about 1.0-2.0 x 10E+05 pixels. For larger N_s the KLD converges rapidly, except for the "linear spaces" sub-group where fluctuations still occur, and it saturates at a stable value when N_s is larger than 1.0 x 10E+06 pixels. However, the saturation occurs very rapidly (N_s < 1.0 x 10E+05 pixels) and at high values of the KLD (20.0 < KLD < 30.0 units) for the $l_1l_2l_3$, $l_1'l_2'l_3'$ and for the C_1-C_2 spaces. The fit to a single elliptical Gaussian appears to be generally highest for the linear spaces, the normalized r-g and CIE-xy spaces, and for the C_2-C_3, a'-b', mod-rb, mod-gb, P_1-P_2, R_1-R_2, R_1-R_3 and Yuv spaces, where the KLD saturates in the vicinity of 1.0 unit. Thus the single Gaussian model can in principle be applied to these spaces for skin pixel detection and subsequent skin color-based image segmentation (typically by use of the Mahalanobis metric [2] [4]), whereas the $l_1l_2l_3$, $l_1'l_2'l_3'$ and C_1-C_2 spaces require a more complex model, independently of the number of skin sample pixels that are used for initial color calibration and image thresholding. Also, in accordance with a visual inspection of the histograms shown in Figure 9, the KLD for the perceptually plausible chrominance spaces is relatively high.

Let us note that we observed similar trends under controlled image scene conditions [23], where the KLD saturates generally below 1.0 unit for the r-g, CIE-xy, C_2-C_3, a'-b', mod-rgb, P_1-P_2 and $R_1R_2R_3$ spaces, for Asian and Caucasian subjects (when N_s > 1.0-3.0 x 10E+04 pixels and N_s > 7.0 x 10E +04 pixels respectively). As can be expected, owing to the large variability of observable skin colors, under unconstrained image scene conditions a significantly larger cumulative number of collected skin sample pixels (or images) are required for the KLD to saturate, and at higher values, than under controlled scene conditions.

In the case of the HIN, for all spaces, the HIN increases rapidly when the number of skin sample pixels N_s is small (as new skin sample images are added to the histograms), generally when N_s is less than about 1.3 x 10E+05 pixels. Fluctuations of the HIN then typically occur until N_s reaches about 5.0 x 10E+05 pixels. For larger values of N_s, the HIN converges rapidly to a narrow range of values, between 27% and 38%, and saturates for most spaces between 30% and 35%. As for the KLD, the behavior of the $l_1l_2l_3$ and $l_1'l_2'l_3'$ spaces (and to a lesser extent that of the C_1-C_2 space) is atypical, in that the HIN for those spaces saturates at significantly high values, near 55%. Hence, under unconstrained image scene conditions, and when a sufficiently large number of skin sample pixels are collected for color calibration, the degree of discrimination between skin pixels and non-skin pixels does not generally vary significantly depending on which chrominance space is used, although some spaces, such as the "linear" spaces or the F-uv space, may be slightly better suited for skin pixel detection when taking this particular criterion into account, as seen in Figure 11. The $l_1l_2l_3$ and $l_1'l_2'l_3'$ spaces are the least efficient for discriminating between skin pixels and non-skin pixels, probably because of the particular geometric characteristics of these spaces, which we discussed briefly in Subsection 4.2. This plausible explanation also applies to the results that are obtained for the same spaces in terms of the KLD.

For the HIN, the trends under controlled image scene conditions [23] are quite similar to those that prevail here under unconstrained conditions. In that case also, for the same reason as was given for the KLD, the overlap between the skin and non-skin distributions is lower and saturates more rapidly (typically, for N_s < 1.0 x 10E+05 pixels) than under unconstrained conditions (including for the ln-chroma space, for which the results are not shown here).

Taking into account both the KLD and the HIN, it is important to note that: 1) once again, when N_s is small, hence for a small initial number of skin sample images, the order in which the sample images are added to the cumulative histograms may significantly influence the precise shape of the curves for both the HIN and the KLD, whatever the chrominance space that is considered, but the convergence, and eventually the saturation of the curves are statistically warranted as N_s increases to large values; 2) the fluctuations of the KLD and of the HIN observed at low to intermediate values of N_s are probably caused by larger random effects of a smaller number of skin sample images on the shape and on the area of the cumulative skin chrominance distribution, and owing to their random nature, they would hence occur irrespectively of the order in which the sample images are added to the cumulative histograms; 3) finally, the experimental results presented in this subsection seem to indicate that at least nearly a million skin sample pixels are required for calibration under unconstrained image scene conditions in order to obtain reproducible skin pixel detection results on given input scene images.

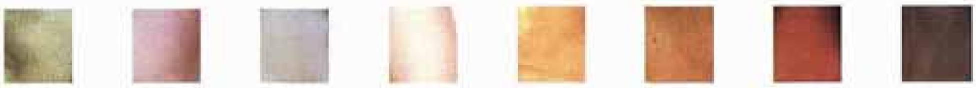


Figure 7. Examples of skin sample images semi-randomly selected and manually collected on the World Wide Web.

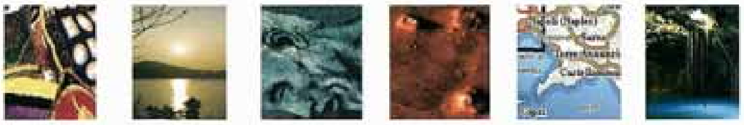
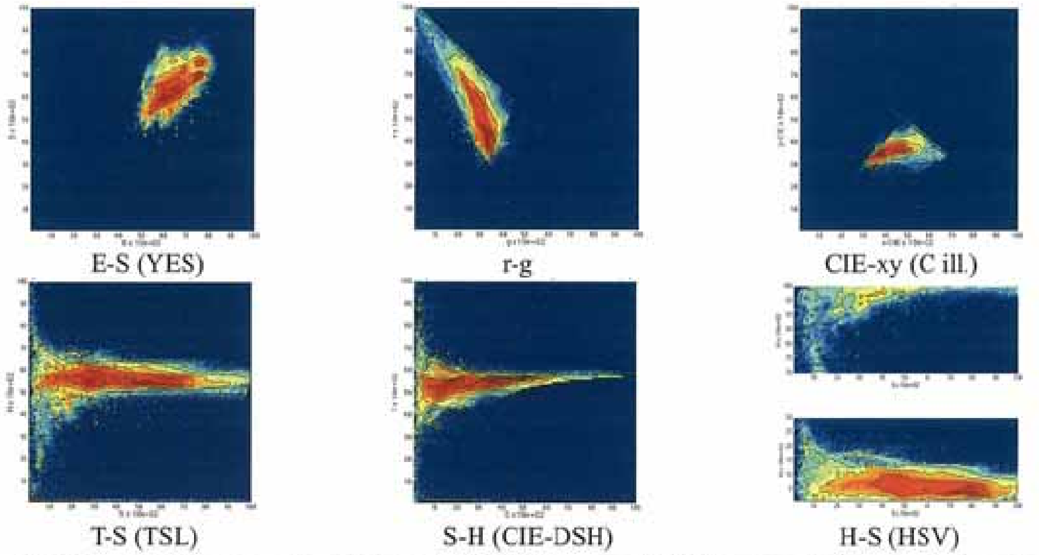


Figure 8. Examples of non-skin sample images semi-randomly selected and manually collected on the World Wide Web.



E-S (YES)

r-g

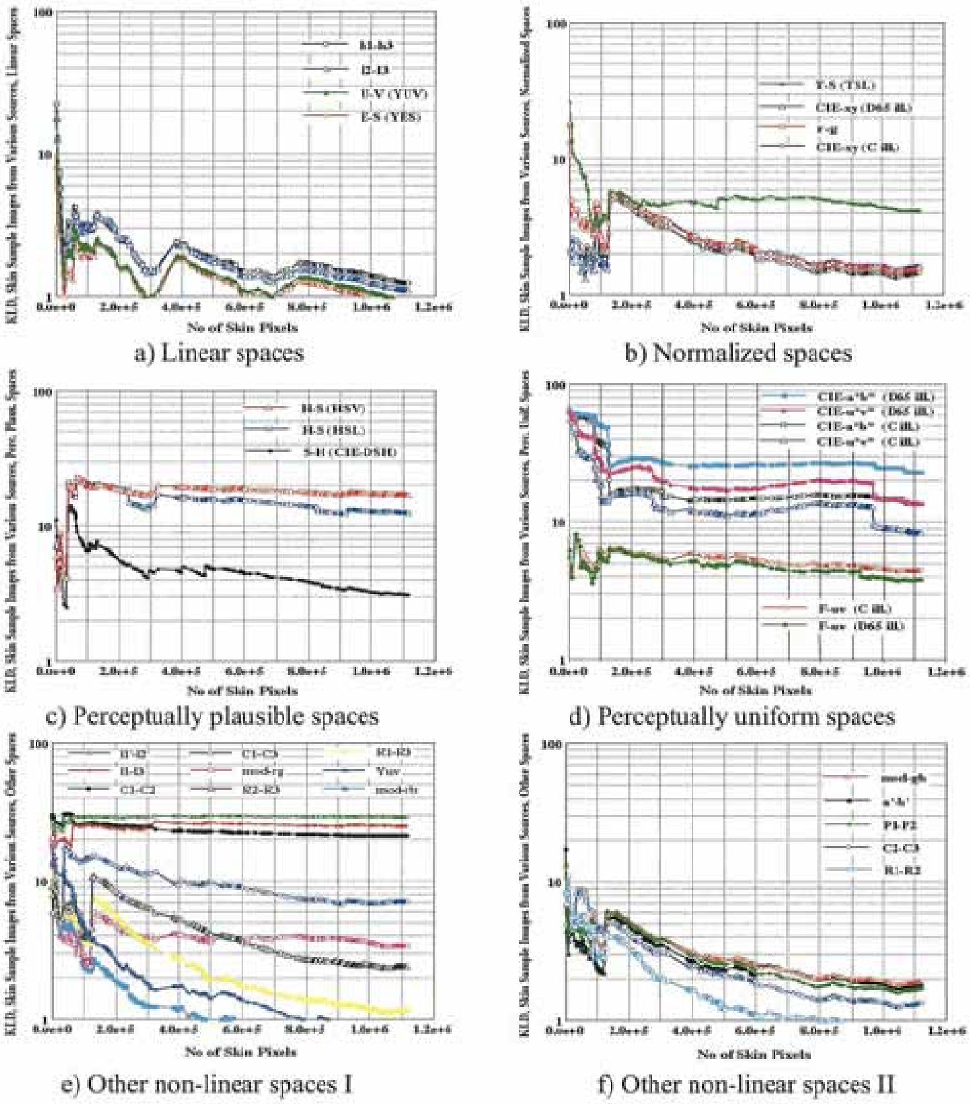
CIE-xy (C ill.)

T-S (TSL)

S-H (CIE-DSH)

H-S (HSV)

Figure 9. Representative examples of the cumulative histograms of 300 skin sample images semi-randomly selected and manually collected on the World Wide Web, for 6 different chrominance spaces (top view, logarithmic scale).



a) Linear spaces

b) Normalized spaces

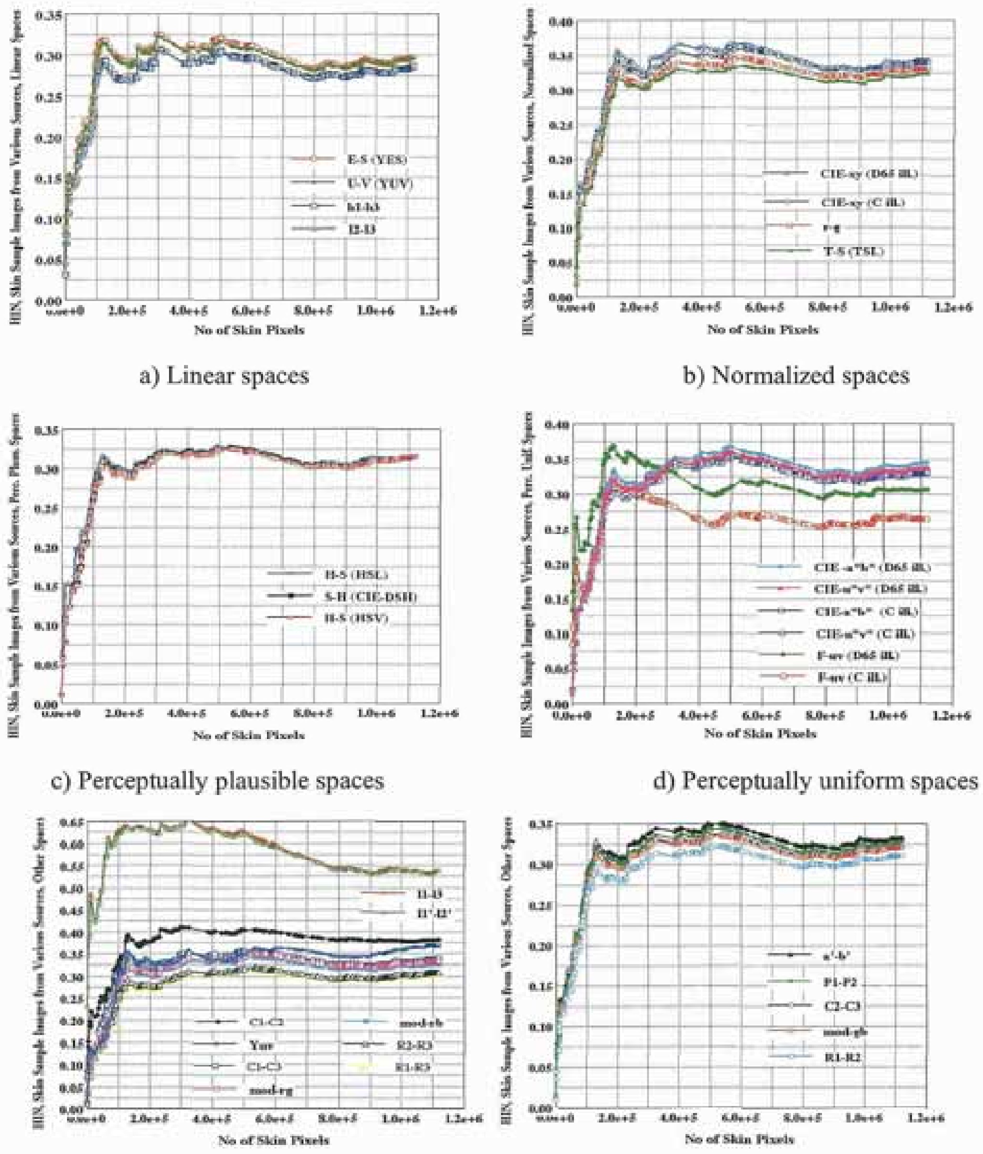
c) Perceptually plausible spaces

d) Perceptually uniform spaces

e) Other non-linear spaces I

f) Other non-linear spaces II

Figure 10. Graphs of the Kullback-Leibler Divergence (KLD) or fit to a single Gaussian distribution as a function of the cumulative number of skin pixels from skin sample images collected on the World Wide Web, for 6 different chrominance space sub-groups.



a) Linear spaces

b) Normalized spaces

c) Perceptually plausible spaces

d) Perceptually uniform spaces

e) Other non-linear spaces I

f) Other non-linear spaces II

Figure 11. Graphs of the Normalized Histogram Intersection (HIN) between skin and non-skin distributions (or of the discrimination between skin and non-skin pixels) as a function of the cumulative number of skin pixels from skin sample images collected on the World Wide Web, for 6 different chrominance space subgroups.

4.10 Conclusions

In conclusion, overall, in terms of seven different criteria, the normalized r-g and CIE-xy chrominance spaces, or spaces such as the a'-b' and P_1-P_2 spaces that are constructed as a linear combination of normalized r, g and b values, offer the best tradeoff and appear consistently to be the most efficient for skin color-based image segmentation. In particular, the use of these normalized spaces obviates the necessity to apply a complex and computationally intensive skin chrominance model in order to obtain a high quality of segmentation, as is the case with most un-normalized spaces, in which the skin distribution is complex-shaped. The C_2-C_3 space, the mod-rgb space that also results from a suitable normalization, and to a lesser extent the $R_1R_2R_3$ space which can be expressed as ratios of normalized r, g and b values, are also good candidates. Owing to their particular geometry, the $l_1l_2l_3$ and $l_1'l_2'l_3'$ spaces are the least efficient for the specific problems of skin pixel detection and of image segmentation based on skin color.

5 実施内容 - Face Detection and Hand Posture Recognition System

A flowchart of the face detection and static hand posture recognition system is shown in Figure 12. In implementing the system, a fundamental issue to address is the level of complexity of the scene background at the location where the system is to be applied, because the robustness of the simultaneous detection and discrimination of faces and of hands (or recognition of hand postures) against complex scene backgrounds is a difficult problem which, to our knowledge, has not yet received much attention. The "background" also includes the clothes that a person is wearing, other body parts, and facial attributes such as glasses, hair and hairstyle, etc…

The system shown in Figure 12 can adapt to varying degrees of scene background complexity in indoor environments (office, home), to slowly varying illumination conditions, and it does not imply any a priori assumption about the presence of a face (or of more than one face) or of a hand (posture) simultaneously in an image (it is often implicitly assumed that a face is present in a scene when constructing a face detection system). The system first uses a statistical skin color model to segment images and a statistical regularity-based shape model to detect faces. We then apply, to our knowledge for the first time, phase-only correlation [25] to classify a subset of static hand postures of the JSL, each posture representing a given phoneme, and to discriminate between hand postures and the image scene background. In effect, as can be seen from Figure 12, we decompose a 3-class problem, that involves the "face", "hand (posture)" and "scene background" classes, into two binary classification problems, that involve, in succession, the classification of faces and of hands, and the classification of hand postures and of the background.

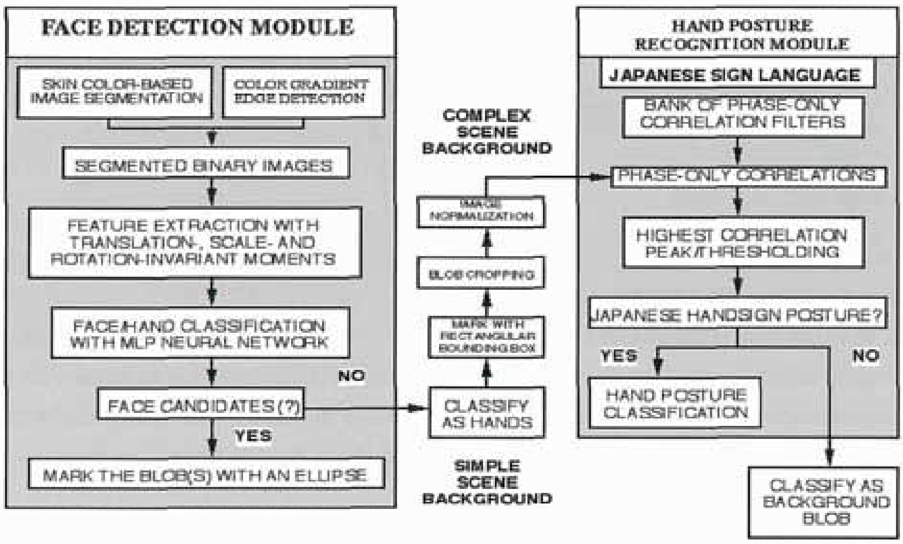

FACE DETECTION MODULE
SKIN COLOR-BASED IMAGE SEGMENTATION
COLOR GRADIENT EDGE DETECTION
SEGMENTED BINARY IMAGES
FEATURE EXTRACTION WITH TRANSLATION-, SCALE- AND ROTATION-INVARIANT MOMENTS
FACE/HAND CLASSIFICATION WITH MLP NEURAL NETWORK
NO
FACE CANDIDATES (?)
YES
MARK THE BLOB(S) WITH AN ELLIPSE
COMPLEX SCENE BACKGROUND
IMAGE NORMALIZATION
BLOB CROPPING
MARK WITH RECTANGULAR BOUNDING BOX
CLASSIFY AS HANDS
SIMPLE SCENE BACKGROUND
HAND POSTURE RECOGNITION MODULE
JAPANESE SIGN LANGUAGE
BANK OF PHASE-ONLY CORRELATION FILTERS
PHASE-ONLY CORRELATIONS
HIGHEST CORRELATION PEAK/THRESHOLDING
JAPANESE HANDSIGN POSTURE?
YES
NO
HAND POSTURE CLASSIFICATION
CLASSIFY AS BACKGROUND BLOB

Figure 12. Flowchart of the face detection and hand posture recognition system, taking into account the image scene background.

5.1 Face Detection Module

As shown in Figure 12, in the face detection module, the segmentation of an image input to the system is performed at each pixel by use of both the chrominance of human skin and of an edge detection based on the color vector gradient [26] [27]. To increase the robustness of face detection, fully translation-, scale- and in-plane rotation-invariant Fourier-Mellin moments are then calculated for each significant blob that results from the segmentation [4] [6], and the resulting feature vectors are here input to a multi-layer perceptron Neural Network (NN) for the classification of faces and hands. The NN is trained to detect frontal views of faces, which possess a statistical regularity, whereas hands can have a large variety of shapes. Hence, at this stage, it is assumed that any blob that is not detected as a face is a hand.

5.1. 1 Skin Chrominance-based Image Segmentation

In accordance with the results of the skin chrominance analysis presented in Section 4, the image segmentation is performed here with the normalized r-g or CIE-xy chrominance spaces. In these spaces, a simple, single Gaussian model is used to estimate the skin chrominance distribution. The thresholding algorithm uses the Mahalanobis metric and is based on the discriminability between the chrominance *distributions of skin and "non-skin" pixels calculated from two sets of manually selected skin and non-skin* sample images, as described in details in [2].

5.1. 2 Color Vector Gradient-based Edge Detection

In order to complement the image segmentation based on the skin chrominance, we apply an edge detection algorithm based on the gradient of the three RGB channel image field or the color vector gradient, that was first proposed by Di Zenzo [26], and later more thoroughly investigated by Lee and Cok [27]. As shown in [27], because the three R, G and B channels of a color image are generally correlated, the color vector gradient is less sensitive to noise than the scalar gradient computed from each channel separately or for gray-level images. The numerical computation of the modulus of the color gradient is performed by use of either the Sobel or Prewitt operators. The edge detection is performed in the original RGB color image. The edge image is then subtracted from the converted r-g or CIE-xy chrominance images, before applying the skin chrominance-based image segmentation.

5.1. 3 Face-Hand Classification

For feature extraction, a selected number of low-order invariant moments [4] [6] are calculated for each blob in the binary segmented images. We apply either the Fourier-Mellin moments generalized by Li [28] or the orthogonal Fourier-Mellin moments developed by Sheng and Shen [29]. We have shown that, in the specific application of face detection based on skin chrominance, the face detection performance by use of either type of moments is similar [6], but the computational cost for the orthogonal moments is higher. The NN used to classify faces and hands is trained with the back-propagation algorithm. When a face is detected, it is marked by an ellipse, as described in details in [4].

5.2 Hand Posture Recognition Module

As Figure 12 indicates, any face candidate that is not classified as a face may be classified as a hand, as long as the scene background is simple. Increasing the background complexity increases the probability of background regions being misclassified as skin during the segmentation, and consequently of being detected as a hand (or as a face). Both hands and background blobs incorrectly detected as skin may have a large variety of shapes, so that the discrimination between the two classes by use of the invariant moments is poor. One of the most important uses of hands in human-machine interactions is gesture recognition, for which the shape distribution of a given segmented hand posture is consistent. The hand posture recognition module is linked to the face detection module as follows: each blob classified as a hand is marked by a rectangular bounding box, and then cropped from the image. A size normalization is then applied to each cropped image to ensure robust recognition with respect to scale changes. There are several different hand posture recognition algorithms that can be used. For example, in [30], an elastic graph matching and Gabor wavelets are applied to hand postures of varying sizes and shapes against complex scene backgrounds in gray-level images, with a correct classification rate of 86.2% for 10 different hand postures. In [31], 25 hand postures of the International Sign Language are classified in gray-level images by use of only one pair of moment-based size functions as features and of a NN for subsequent recognition. A correct recognition rate of over 85% is achieved, but the scene background is almost uniform.

We propose to apply phase-only correlation [25] to every normalized blob in the segmented images, to simultaneously discriminate between hand postures, and between hand postures and background blobs.

5.2. 1 Phase-Only Correlation Filter for Hand Posture Recognition

It has been found by Oppenheim and Lim [32] that the phase information in the Fourier domain of an image is considerably more important than the amplitude information in preserving the features of the image. Horner and Gianino [25] used this result to construct a novel matched spatial filter that can be used for optical pattern recognition, and derived the phase-only correlation filter: given an object to be recognized f(x,y), where f(x,y) usually represents gray levels at Cartesian coordinates (x,y) in a monochrome image, we construct in the corresponding Fourier domain of f(x,y), F(u,v), where (u,v) are the spatial frequencies corresponding to (x,y), a filter with transfer function $H_\phi(u,v$ such that

$$H_\phi(u,v) = \frac{F^*(u,v)}{|F(u,v)|} = e^{-i\phi(u,v} \tag{5}$$

Where F*(u,v) is the complex conjugate of the Fourier transform of f(x,y), |F(u,v)| and $\phi(u,v$ are respectively the modulus and the phase of F(u,v), and where $i^2=-1$. Figure 13 describes the synthesis of the phase-only correlation filter from a well-segmented, normalized reference (or template) hand posture for the Japanese phoneme "ki". The Fast Fourier Transform (FFT) of the normalized hand posture f'(x,y) is calculated for image dimensions of 64 x 64 pixels, Re{F'(u,v)} is the real part of the FFT of f'(x,y) and Im{F'(u,v)} is its imaginary part.

As Figure 12 indicates, we construct off-line a phase-only correlation filter for each well-segmented and normalized "reference" hand posture image belonging to a set of N_r static hand postures of the JSL. After size normalization, any input hand posture image is correlated on-line with the resulting bank of N_r phase-only filters, resulting in N_r correlation images, each of dimensions 64 x 64 pixels. By using the FFT, the computational load is thus $O((N_r + 1)\,M(\log_2 M) + N_r M^2)$ for each input hand posture, with M=64. Since M is small, hand postures can be recognized in real time.

The main advantage of the phase-only correlation filter over the classical matched filter (classical correlation) is that it yields much higher and sharper correlation peaks [25], because it behaves as a high-pass filter, and thus enhances the contributions of the contours of objects. This property is illustrated by the example of Figure 14, in units of intensity, for the normalized reference hand posture representing the phoneme "ki". Also, the phase-only filter has very good discrimination capabilities between different objects with similar shapes. As an example, Figure 15 shows, in units of intensity, the phase-only correlations of normalized input hand postures for the phonemes "ki" and "i" with the reference hand posture for the phoneme "ki". Despite the similarity between the two hand postures, the phase-only correlation peak (maximum) intensity for the phoneme "ki" is 3.4 times higher than the corresponding phase-only cross-correlation peak intensity for the phoneme "i". Finally, the application of the phase-only filter is a simple technique that neither requires a manual initialization, nor the tuning of any parameter. However, the phase-only filter is much more sensitive than the classical matched filter to distortions of objects to be recognized, and it is not rotation-invariant.

When the scene background is not taken into account, and when one addresses the specific problem of the discrimination between different hand postures, the correlation with the highest peak intensity among the N_r correlations is selected to recognize a given hand posture (and phoneme). Despite its sensitivity to the distortions of objects, the phase-only filter can discriminate between quite similar hand postures, as long as the distortion of a hand posture to be recognized is not too large, because only the relative values of the correlation peak intensities are taken into account during the recognition process.

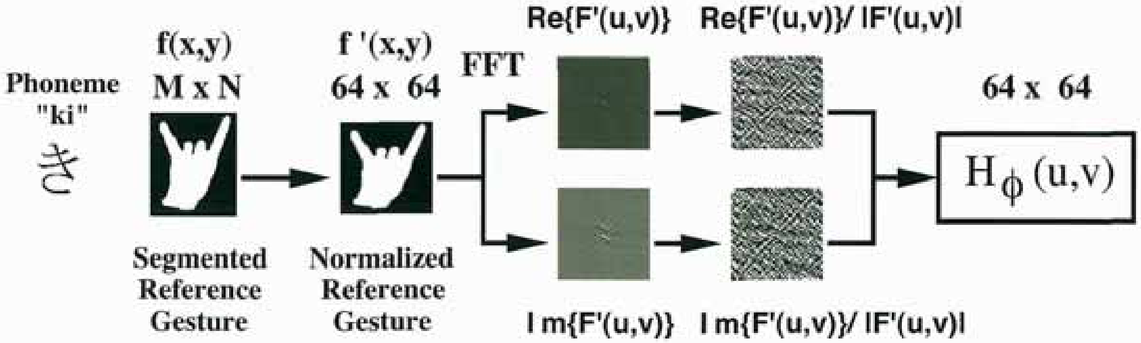


Figure 13. Synthesis of a phase-only correlation filter from the normalized, segmented reference image of a Japanese Sign Language (JSL) hand posture, here symbolizing the phoneme "ki".

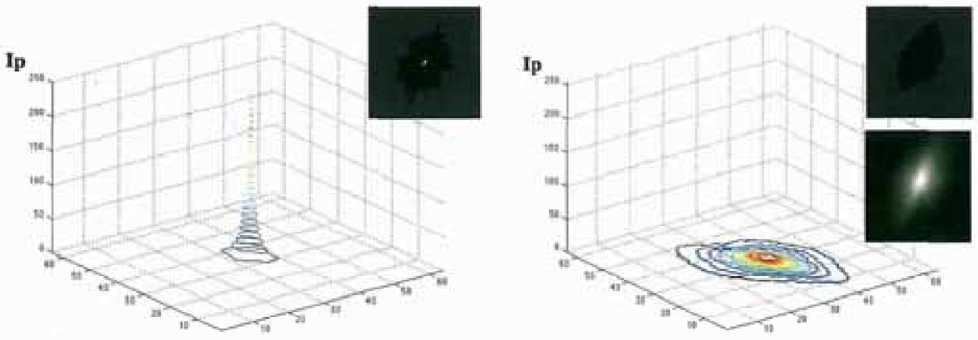


Figure 14. Phase-only autocorrelation of the normalized hand posture image for the phoneme "ki" (left) and corresponding classical autocorrelation obtained with the classical matched filter (right). The higher top view of the classical autocorrelation is shown in relation with the phase-only autocorrelation, whereas the lower top view is scaled between 0 and 255 units.

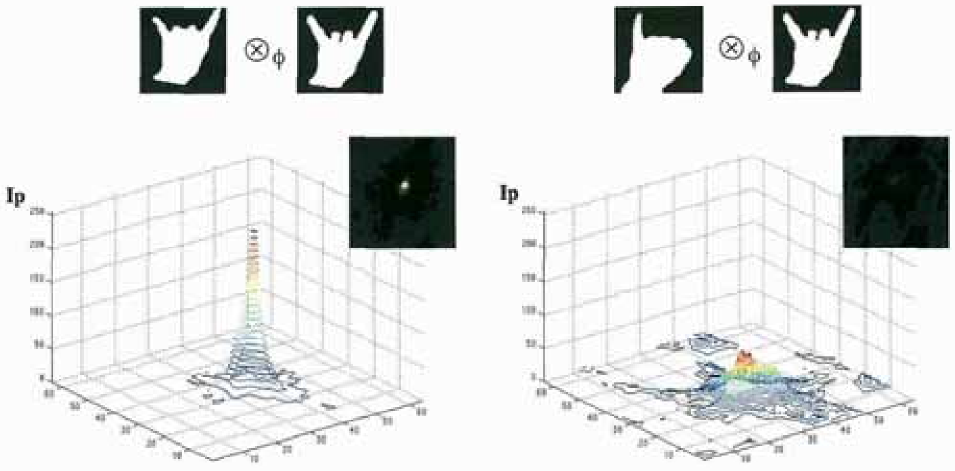


Figure 15. Phase-only correlations of the normalized input hand posture images for the phoneme "ki" (left) and for the similarly shaped phoneme "i" (right) with the reference hand posture for the phoneme "ki".

5.2. 2 Hand Posture and Background Classification

The process of discriminating between hand postures and background blobs requires that one examine, for N_r different hand postures to be recognized (or reference hand postures), the N_r phase-only correlations obtained for each of a set of N_G normalized input hand posture blobs and N_B normalized background blobs. If we assign a threshold intensity (or amplitude modulus) to the correlations, such a threshold should be high in order to reject background blobs, and conversely, low in order to detect hand postures with a high probability. We then analyze the percentage of True Positives (or the correct classification rate) for hand postures $CR_G = \sum_{i=1}^{N_G} TP_G/N_G$ and the percentage of True Positives for background blobs $CR_B = \sum_{j=1}^{N_B} TP_B/N_B$ as a function of the correlation peak threshold intensity or amplitude modulus. A threshold value yielding the best trade-off discrimination between hand posture blobs and background blobs can then be found at the intersection of the curves for CR_G and CR_B.

6 結果

The face and hand posture image database consists of 516 frame sequences of 258 Japanese subjects. Each sequence contains 30 frames, recorded in the "percept-room" of the HOIP laboratory (using halogen lamps at 3,200 degrees Kelvin as a source of illumination) with a SONY DXC-9000 camera (using a white balance) that zooms on each subject in each frame sequence, for a total of 15,480 static images of faces with a variety of poses, scales, in-plane rotations and facial attributes, and of hand postures. The hand postures represent 45 static hand signs of the JSL (there are 11 frame sequences for each hand sign). Each image contains only one face and one hand posture. The training and test sets both consist of 541 static images semi-randomly selected from the 516 frame sequences (no test image is part of the training set). The skin chrominance calibration for the r-g and CIE-xy spaces is performed by use of 901 skin sample images (1.9x10E+06 skin pixels) manually selected from the image database, and of the 80 non-skin sample images (2.6606x10E+06 non-skin pixels) selected from various sources that were used for the skin chrominance analysis. The face detection and hand posture recognition system is implemented on a PC Pentium-III, 1 GHz. The dimensions of the input images are 640 x 480 pixels.

The complete process of face detection and hand posture recognition is illustrated in Figure 16. The results of the skin chrominance-based image segmentation are generally very similar for both the r-g and CIE-xy chrominance spaces. We note that the color gradient-based edge detection efficiently separates the neck of the subject from her face, but that it also tends to separate the fingertips from the hand.

Before presenting the general results of face detection and hand posture recognition, it is instructive to examine some particular examples. We first focus on the detection of faces and of hands. Figure 17 shows an example of the successful detection and discrimination of the face and the hand (as well as an arm) of a Japanese subject at three different scales. The segmentation results vary significantly as the camera zooms on the subject. In this particular case, a binary face and hand classification problem is valid, since the scene background (which is almost uniform) and the clothes, as well as the hair, have been correctly classified as "non-skin" during the segmentation. However, in the example of figure 18, the clothes of a subject have been misclassified as skin, and the color gradient-based edge detection successfully separates the subject's face from her clothes. Consequently, the face is successfully detected. Figure 19 show various misclassification errors, such as a hair region of a subject detected as a hand, a hand misclassified as a face, and the double detection of a face owing to the presence of glasses. Other types of errors include face localization errors, a face or part of a face misclassified as a hand, or parts of clothes detected as either hands or faces. The misclassification of hands or the detection of background blobs as faces are due to the invariant properties of the moments used for feature extraction, and also to the resemblance of the shape of the misclassified blobs to segmented frontal views of faces. Faces misclassified as hands typically are connected to the neck, as the color gradient-based edge detection does not always completely separate the neck from the face, and the misclassification may also occur because the shape of the segmented face blobs varies significantly as the camera is zooming on the subjects.

In order to evaluate the general performance of the face and hand detection and discrimination sub-system, without taking background blobs into account, we first define the rate of correct face detection as:

$$CD_F = \sum_{i=1}^{N_F} TP_F / N_F \qquad (6)$$

Where TP_F is a true positive for faces (a face that is correctly detected), and where N_F is the total number of faces in the test set, which also includes the false negatives for faces FN_F (faces that are not detected, or misclassified as hands). Hence,

$$\sum_{i=1}^{N_F} TP_F + \sum_{i=1}^{N_F} FN_F = N_F \qquad (7)$$

Similarly, we define the rate of correct detection of hands CD_H as

$$CD_H = \sum_{j=1}^{N_H} TP_H / N_H \qquad (8)$$

Where TP_H is a true positive for hands and N_H is the total number of hands in the test set. As for faces, we have the following relation:

$$\sum_{j=1}^{N_H} TP_H + \sum_{j=1}^{N_H} FN_H = N_H \qquad (9)$$

Where FN_H is a false negative for hands (a hand that is misclassified as a face). Finally, we define the rate of discrimination between faces and hands as

$$D = \left(\sum_{i=1}^{N_F} TP_F + \sum_{j=1}^{N_H} TP_H \right) / \left(N_F + N_H \right) \tag{10}$$

Since in our experiments, $N_F = N_H = 541$, in this particular case, $D = \left(CD_F + CD_H \right) / 2$, or D is the average of the two detection rates.

Table 11 presents the general results of face and hand detection and discrimination when both the color gradient-based edge detection and the skin chrominance-based image segmentation are applied. The performance of face detection is significantly higher than when using the chrominance only (in which case $CD_F = 70\%$ for both chrominance spaces), and it is practically the same for both chrominance spaces, because both spaces yield very similar segmentation results. The correct detection rate of hands is lower, because of the tolerance of the invariant moments, but slightly higher than when using the chrominance alone (in which case $CD_H = 72\%$ for both spaces). The time required for face and hand detection, that depends on the number and size of the blobs that are present in a segmented image, was on average 270 [ms] on the Pentium-III PC and for the input image dimensions that we used, before being significantly reduced during Phase II of the project.

We now focus on the recognition of hand postures of the JSL, before presenting general results of the discrimination between hand postures and background blobs. Figure 20 illustrates the recognition of three different hand postures representing the phonemes “ni”, “ma”, and “wo” respectively, among a set of 12 different (reference) hand postures, thus requiring the computation of 144 phase-only correlations of dimensions 64 x 64 pixels. All hand postures are correctly classified, except the hand posture for the phoneme “wo”, which is confused with the posture for the phoneme “ho”. This particular misclassification example illustrates a problem that is bound to occur in realistic situations, namely, the classification of a hand posture where an exposed forearm or arm is connected to the hand, and thus changes the shape of the hand posture significantly. Table 12 presents general results of the recognition of a subset of 8 hand postures of the JSL, with 94 input hand postures, each corresponding to one of the 8 reference phoneme hand postures, thus requiring the analysis of 752 phase-only correlations. In this experiment (which does not take the scene background into account), the phase-only filter achieves a correct classification rate of over 95%. However, it is expected that the classification rate decreases as the number of hand postures to be recognized and the number of input hand postures increase.

We analyze the discrimination between hand postures and background blobs by use of 15 different reference hand postures, and of a set of 236 input hand posture blobs and of 261 background blobs. Hence, a total of 7,455 phase-only correlations are examined. Figure 21 shows the graph of the correct classification rates of hand postures CR_G and of background blobs CR_B as a function of the threshold amplitude modulus of the phase-only correlation, for convenience. The best tradeoff discrimination rate is found to be 86.1%, for a threshold amplitude of 669 units (or 4.476 x 10E+05 units of intensity). This discrimination rate obtained with the phase-only correlation filter can be considered to be high, given the large variety of possible shapes of both hand postures and background blobs.

Finally, because of the relatively general approach that we adopt to detect faces and to recognize hand postures, our system can be applied to color images with complex scene backgrounds selected from various sources, for example, from the World Wide Web, hence recorded under different illumination conditions and with different camera systems. Figure 22 shows examples of the successful simultaneous detection of faces and of hands of Caucasian subjects with different skin colors. Although the skin color calibration was performed here for Asian subjects only, the robustness of the r-g and CIE-xy chrominance spaces to the intrinsic variability of skin color, as compared to other spaces, leads to the correct detection of a relatively large range of skin colors. We note in particular that several faces can be detected simultaneously (although the hands in the last image could not be discriminated against the background).

In conclusion, the biggest challenge for the problem of the detection of faces and of hands in color images lies with the background regions that have been detected as skin during the color segmentation. We nevertheless could achieve a face detection rate of over 88%, a correct hand posture classification rate of over 95%, and a best tradeoff discrimination rate between hand postures and background blobs of 86.1% for three different sets of test images. The phase-only correlation filter is a simple and promising technique for the recognition of static hand postures in binary segmented images, although it may not be suitable for all real situations. We finally suggest that the performance of face detection (as well as the correct classification of hands) can be further improved by searching for facial features in each face candidate blob. The quality of segmentation can also be improved by considering correlations between neighboring pixels.

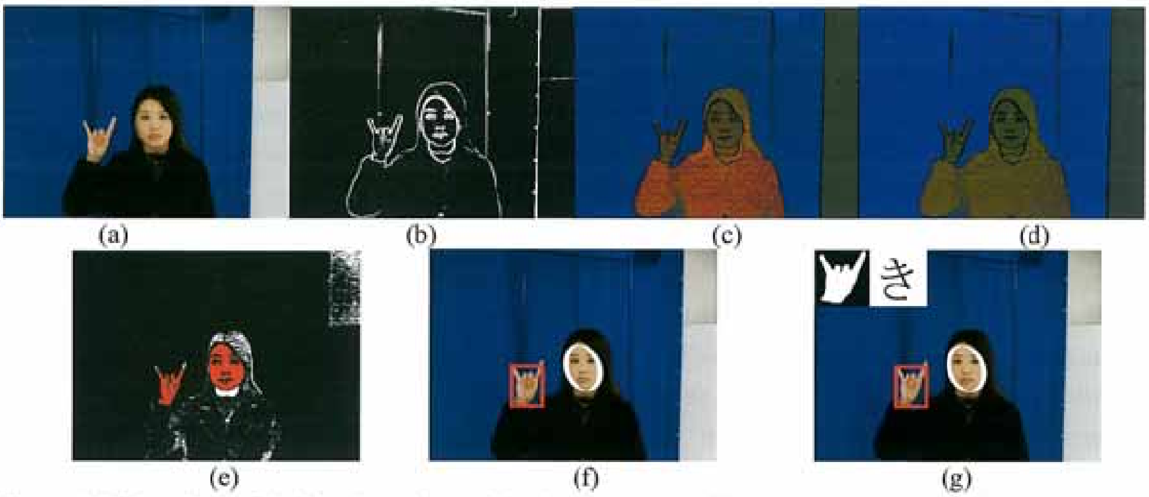


Figure 16. Illustration of the face detection and hand posture recognition process. (a) original input image, (b) results of the color gradient-based edge detection, (c) and (d): results of the conversion of the original image into the r-g and CIE-xy chrominance spaces respectively, with subtraction of the color edges, (e) skin chrominance-based segmented image, with two blobs (in red) used for feature extraction, (f) results of the face/hand classification, and (g) final results of hand posture recognition, with the recognition of the phoneme "ki".

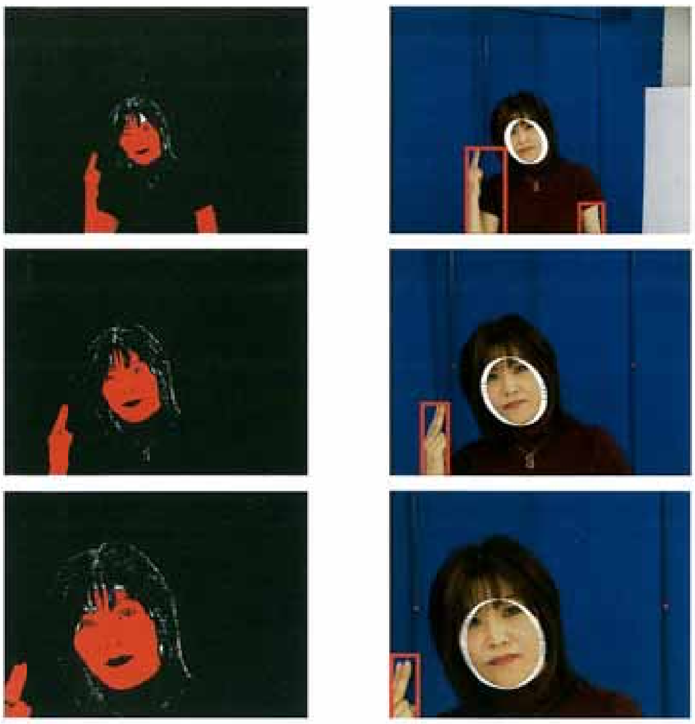


Figure 17. Examples of the successful detection and discrimination, at different scales, of the face and of the hand of a Japanese subject. The segmented images are in the left column of the figure.

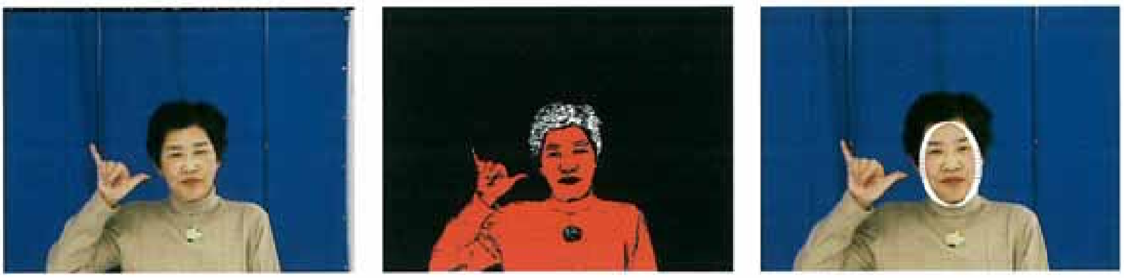


Figure 18. Example of the successful detection of the face of a Japanese subject when combining the skin chrominance-based image segmentation with the color gradient-based edge detection.

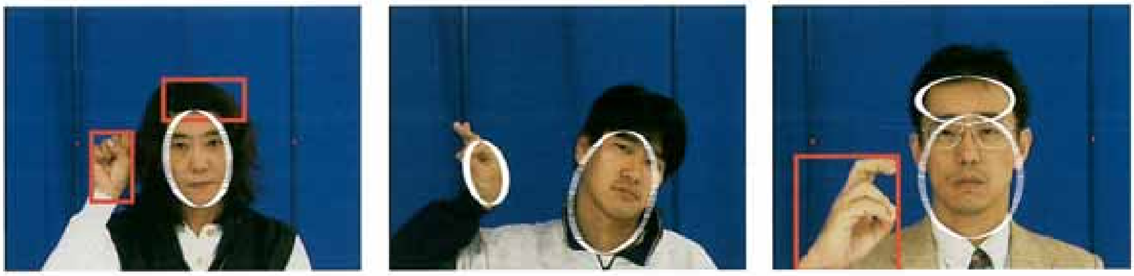


Figure 19. Example of problems occurring with the present face and hand detection system. From left to right: detection of a hair region as a hand, misclassification of a hand as a face, and double detection of a face due to the presence of glasses.

Table 11. General results of the correct detection and discrimination of faces and of hands when combining the skin chrominance-based image segmentation with the color gradient-based edge detection (without taking into account the background blobs).

Detection and Discrimination Results Chrominance and Color Gradient			
Chrominance Space	CD_F (%)	CD_H (%)	D(%)
r-g	88.5	73.2	80.9
CIE-xy	88.3	72.9	80.6

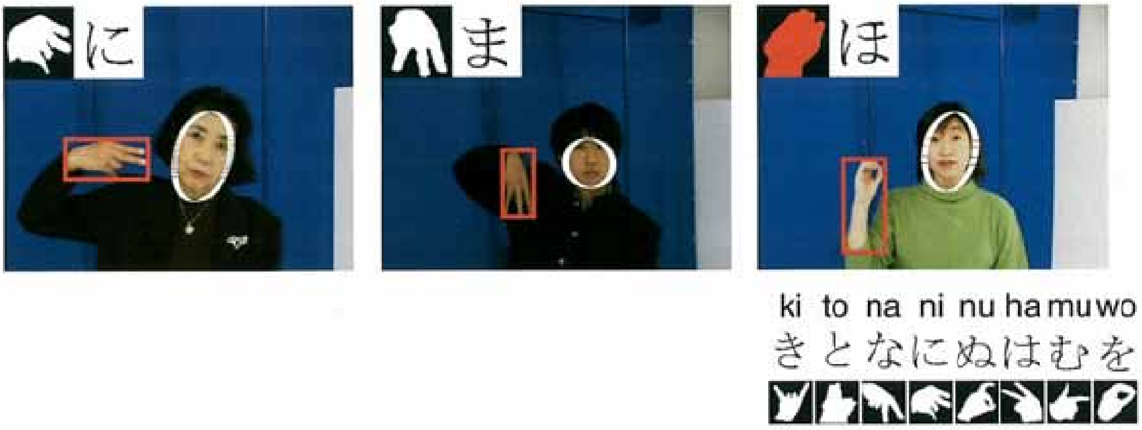


Figure 20. Example of the recognition of three different hand postures representing the phonemes "ni", "ma" and "wo" respectively. In the rightmost image, the posture for "wo" was misclassified as the posture for "ho", because the forearm of the subject is exposed and connected to her hand.

Table 12. Results of the correct classification of 8 different Japanese phoneme hand postures for a total of 94 input hand postures, each corresponding to one of the 8 reference phoneme hand postures (without taking into account the image scene background).

Phonemes	Number of Test Gestures	Recognition Rate (%)
ki	6	100.00
to	12	83.33
na	17	100.00
ni	10	100.00
nu	9	100.00
ha	6	83.33
mu	15	100.00
wo	19	94.74
Total	94	95.74

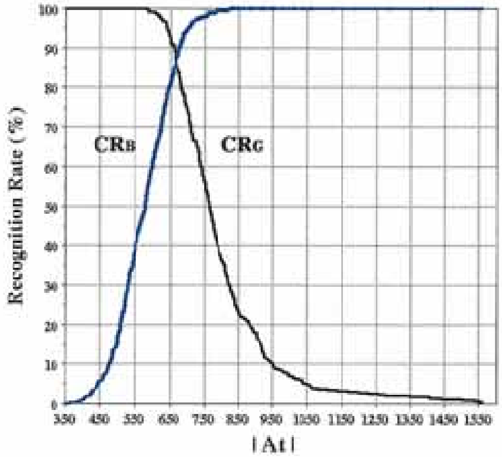


Figure 21. Graph of the correct classification rate of hand postures CR_G and of background blobs CR_B as a function of the value of the threshold amplitude modulus of the phase-only correlation.

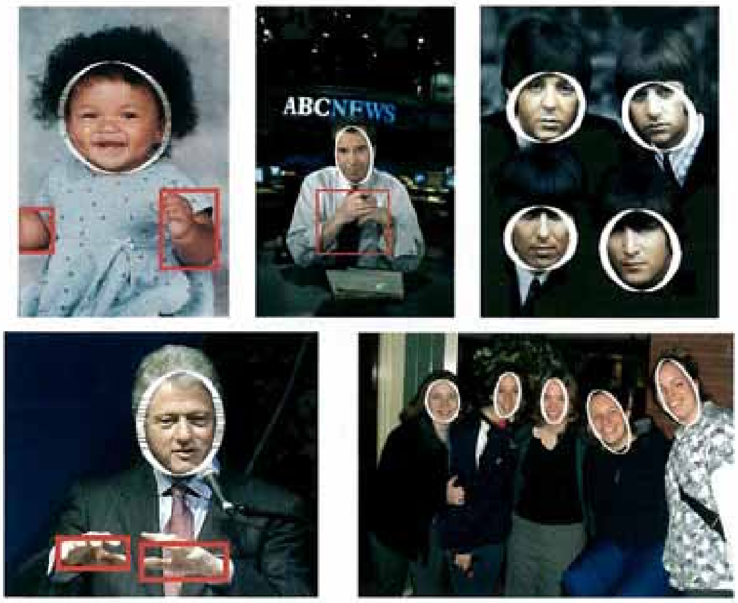


Figure 22. Examples of the successful detection of faces and of hands of Caucasian subjects in images selected from various sources, mostly from the World Wide Web.

フェーズII

1 研究の概要

During Phase II of the HOIP project, the face detection and hand posture recognition system based on skin color was integrated into a more global, real-time system that relies on the three fundamental cues of color, shape and motion: "DRUIDE", which stands for "Detection, Recognition, Unification, Interpretation, Decision, Evolution", consists of three mutually complementary and independent sub-systems, in order to simultaneously detect or track multiple faces as well as recognize hand postures of the JSL in color video sequences .In addition to the module based on skin color, a second module detects faces concurrently in gray-level images by use of SVMs , while a third module performs face tracking, so that the overall robustness of detection and recognition of the system is significantly increased. This report presents, for Phase II of the HOIP project, an overview and experimental results of the capabilities of the DRUIDE system.

2 研究の目標

The implementation of the DRUIDE system is designed to increase the robustness of the initial skin color-based face detection and hand posture recognition system, by adding SVM-based face detection that uses luminance information, and also by tracking faces in color video sequences.

3 実施内容 - Architecture of the DRUIDE System

DRUIDE was developed to offer high-level functions covering many aspects in computer vision-based human-machine interaction, enabling shortened development time and increased application-oriented productivity. It consists of a library that includes algorithms for detecting and tracking objects, more specifically faces [1], recognizing static hand postures [1], and also functions related to face recognition and to the detection of facial features. Basic routines cover hardware optimization, image processing and scene understanding. DRUIDE seamlessly integrates and builds upon APIs such as DirectShow, DirectX, and Intel's OpenCV [33]. The flowchart of the system is shown in Fig. 23.

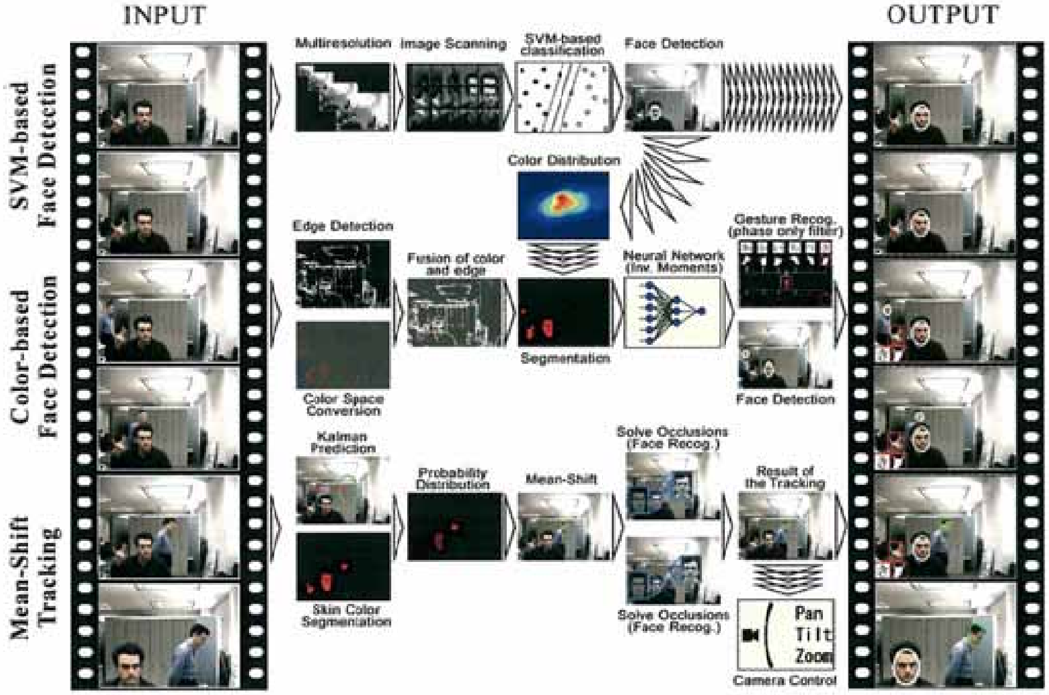


Figure 23. Flowchart of the DRUIDE face detection, tracking and hand posture recognition system, for a given video sequence.

3.1 Skin Color-based Face Detection and Hand Posture Recognition

The core sub-system of DRUIDE is the color-based face detection and static hand posture recognition module, located in the middle section of the flowchart. The system first detects faces in an input 24-bit RGB color video stream acquired with a Pan-Tilt-Zoom (PTZ) camera. The skin color-based image segmentation is implemented as a first step at each video frame pixel. As we showed in the Phase I section of this report, suitably normalized chrominance spaces are the most efficient for image segmentation based on skin color. Here, the normalized r-g space was selected. To insure a more robust segmentation, we also have the option to apply the color (RGB) vector gradient-based edge detection, so that chrominance and luminance information are fused. In this module, a multi-layer perceptron NN is used for the classification of faces and "non-faces".

3.2 SVM-based Face Detection

In order to more robustly avoid confusions between faces and "non-face" blobs with similar shapes (hand palms for example), the image region enclosing a given blob is normalized and analyzed by use of a Support Vector Machine (SVM) classifier, as shown in the top section of Figure 22. Our approach is similar to the one described in [34], and our implementation is based on the LIBSVM library. The μ-support vector binary classifier has the capability to discriminate between faces and non-faces (the parameter μ lets the user control the number of support vectors, as described in [35]). The training phase for the detection of faces involved 12,000 face images and 16,000 non-face images from various sources, mainly from the World Wide Web. After optimizing the SVM-based face detection algorithm, we use this part of the DRUIDE system as an independent face detector that runs concurrently with the skin color-based face detector, as shown in Figure 23, in order to increase the robustness of the face detection process. This approach also allows to regularly re-initialize parameters related to the skin color distribution in order to cope with changes of illumination.

3.3 Mean-Shift Face Tracking

In order to robustly track faces in a video stream, we use a statistical approach based on the mean-shift algorithm [36] [37], which consists of a gradient ascent search over the skin color distribution. Although the algorithm is limited to the tracking of elliptical objects, it is somewhat robust to changes in illumination conditions. The system can track multiple moving objects simultaneously and is robust to partial occlusions and to camera motion.

3.4 Camera Motion

Fixed video cameras, with the exception of omni-directional sensing devices and fisheye lens-mounted cameras, can sense only a very limited part of their environment. Moreover, their resolution is often too limited to precisely analyze distant objects. By mounting the video camera on mechanical devices, it is possible to overcome such restrictions: while pan-tilt capabilities, provided by two rotational engines, can compensate for the narrow field of view of cameras, zooming allows the observation of remote objects. When linked to computer vision algorithms and when precisely controlled, pan-tilt-zoom (PTZ) cameras can help solve complex tasks related to vision-based human-machine interaction. We use two SONY EVID-100 PTZ cameras that are controlled through the serial port of a PC (software and hardware extensions to up to eight cameras and control through the VISCA protocol are supported). The two cameras were optically and mechanically calibrated [38] [39] [40]. Both the face tracking and the camera control are illustrated in the bottom section of Figure 23.

4 結果

The DRUIDE system is implemented on a single PC computer with a 1.0 GHz Pentium-III processor. The dimensions of the input video frames are 320 x 240 pixels. Preliminary experiments, carried out on 540 frames semi-randomly selected from 510 video sequences, yield a correct face detection rate of about 90%. When only hand postures are considered, for a sub-set of 8 reference hand postures of the JSL and a set of 100 input hand postures, the POCF achieves a correct classification rate of over 95%. When the scene background is also taken into account, the POCF yields a best tradeoff discrimination rate of over 86% between a set of 236 input hand posture blobs and another set of 261 background blobs. The performance of the DRUIDE system is illustrated in Figures 24 to 28, for faces with different skin colors and with a variety of poses and scales. The color video sequences are recorded in an office environment, under slowly varying illumination conditions, but with varying degrees of scene background complexity. Despite the obvious robustness to dynamic scene conditions, errors typically include face localization errors, a hand or other body part, or a background blob misclassified as a face, or a single detection or tracking of two connected faces or a double detection of a single face.

Finally, in order to achieve real-time performance (the system runs at 30 frames/sec), we took special care of reducing the computation time of each separate process. Critical processes have been rewritten in Assembly language, and look-up tables are used to speed up time-consuming processes. Average computing times (in milliseconds) are shown in Table 13.

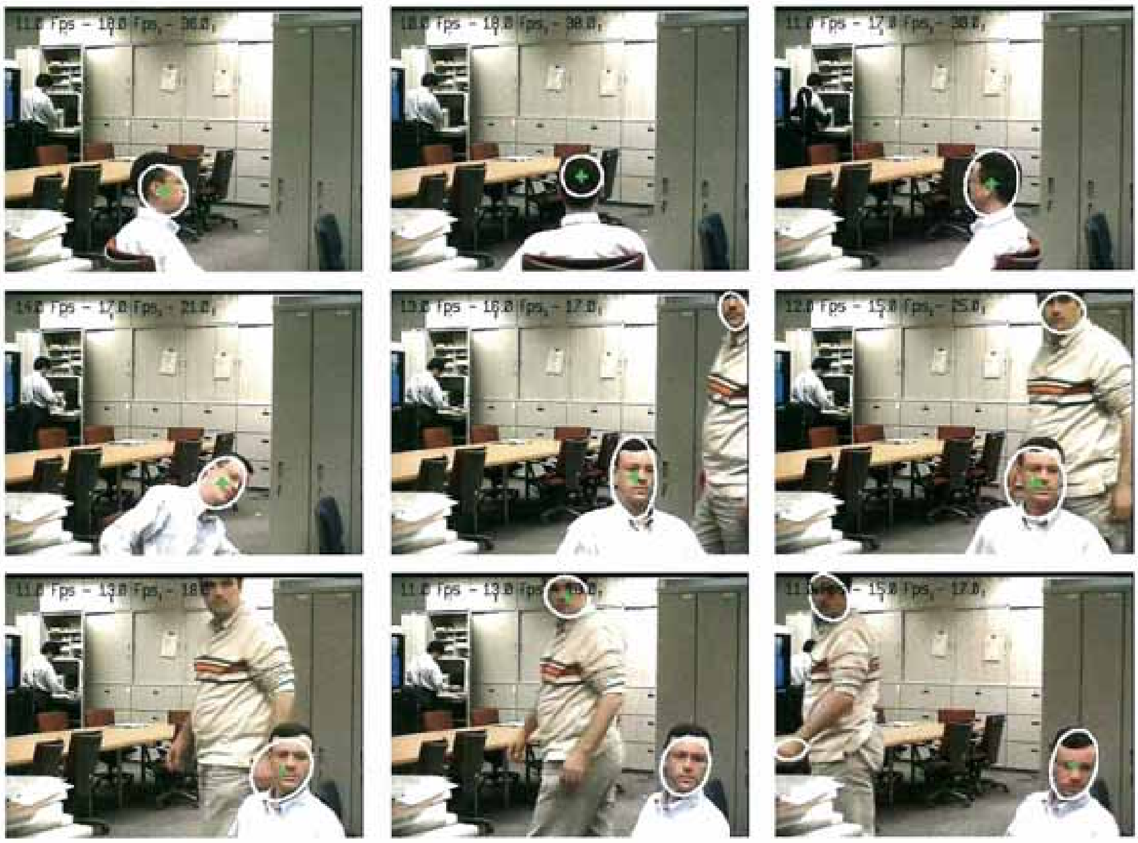


Figure 24. Simultaneous detection of the faces of two Caucasian subjects in a given color video sequence.

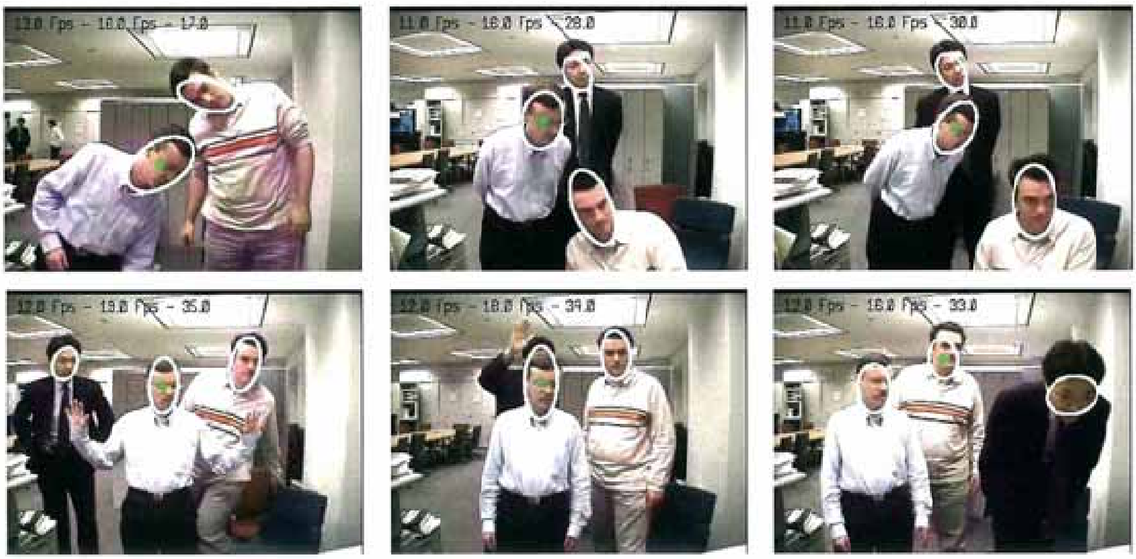


Figure 25. Simultaneous detection of the faces of two Caucasian subjects and of one Asian subject.

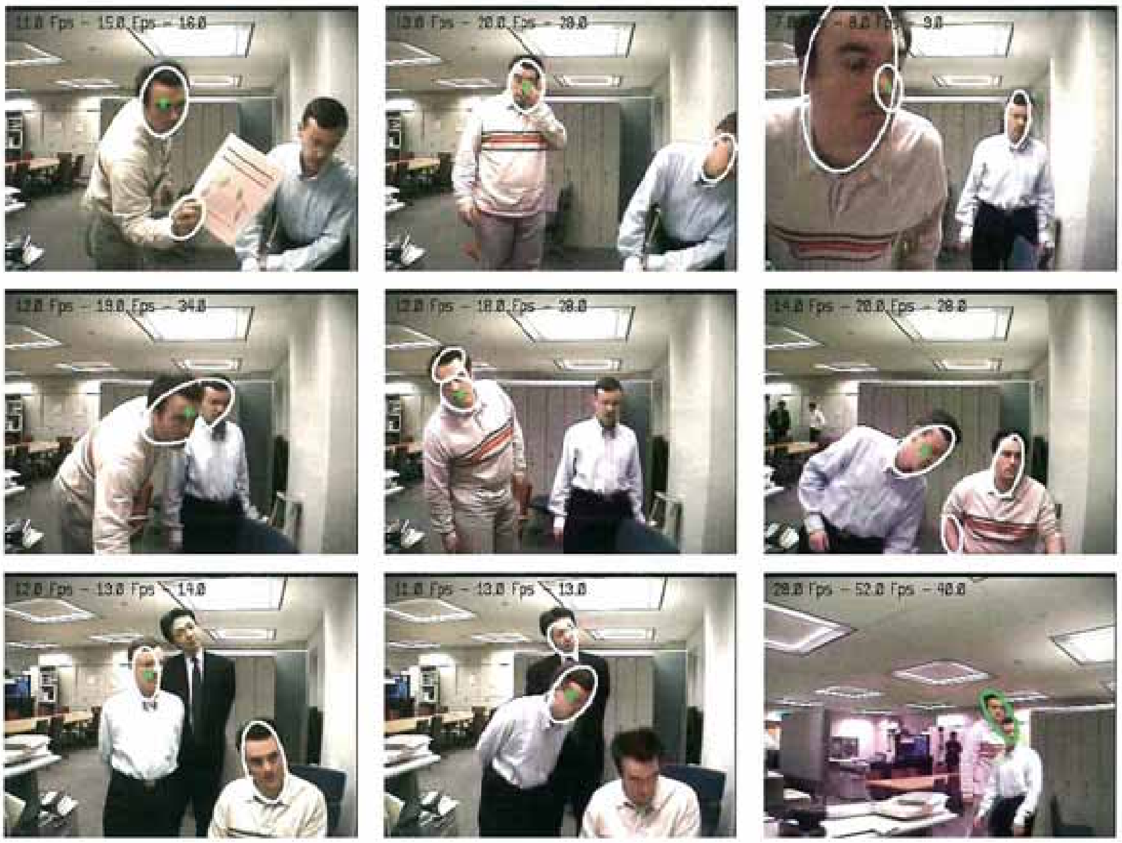

11.0 Fps - 15.0 Fps - 16.0
13.0 Fps - 20.0 Fps - 20.0
- 8.0 Fps - 9.0
12.0 Fps - 19.0 Fps - 34.0
12.0 Fps - 18.0 Fps - 20.0
14.0 Fps - 20.0 Fps - 20.0
12.0 Fps - 13.0 Fps - 14.0
11.0 Fps - 13.0 Fps - 13.0
20.0 Fps - 52.0 Fps - 40.0

Figure 26. Examples of errors occurring with the present face detection and tracking system.

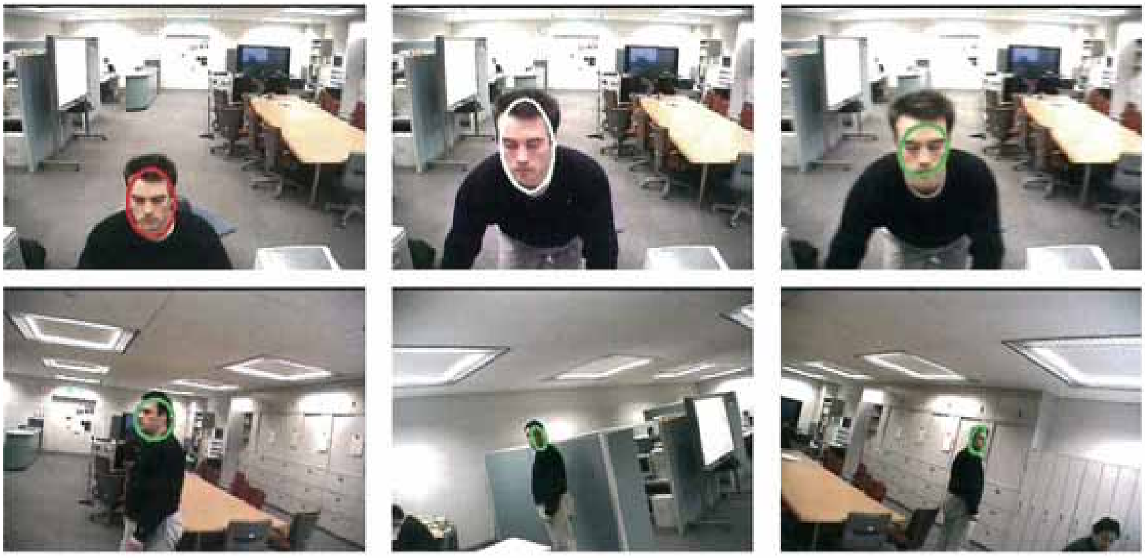


Figure 27. Successive SVM-based detection, skin color-based detection, and camera-controlled tracking of the face of a Caucasian subject (active zoom is not performed here).

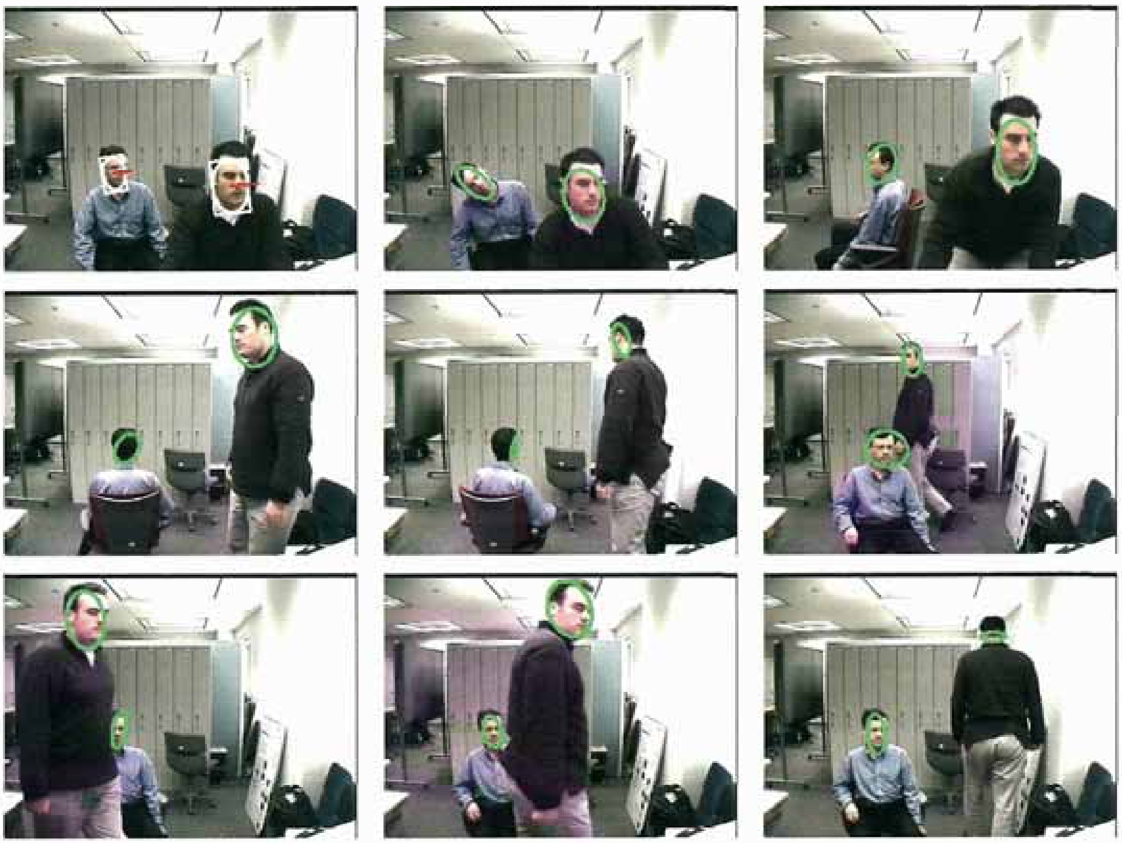


Figure 28. Simultaneous tracking of the faces of two Caucasian subjects (without camera control).

PROCESS	COMPUTATION TIME (1.0 GHz Pentium PC, 320 x 240 Pixel Images)
Edge Detection	
Monochrome Image	10 ms
Color Gradient	30 ms
Face Detection	
Color Space Conversion	3 ms
Segmentation + Blobbing	8 ms < X < 15 ms
Face Detection (per Blob)	(FMMs) 3 ms
Blob Size Normalization (per Blob)	< 3 ms
Color-based Face Detection	< 20 ms
Face Detection with SVMs	20 ms
Face Verification with SVMs	< 20 ms
Tracking	
Mean-Shift Tracking	< 20 ms
Feature Tracking	< 30 ms
Face Recognition (Eigenface-based) (depends on the Database)	< 20 ms
Camera Control	< 1 ms
Hand Posture Recognition (with the Phase-Only Correlation Filter) (depends on the Database)	< 15 ms

Table 13. Computational cost of the different processes implemented in the DRUIDE system, in milliseconds.

フェーズIII

今後の取り組み

Our current work consists in widening the set of basic functionalities of the DRUIDE system, by integrating tasks such as face direction estimation, gaze detection, and a combination of tracking with stereo information to increase the robustness to occlusions and the volume of sensed space. More generally, the fusion of information extracted from videos and the use of evolutionary techniques in order to increase the robustness of the system as a whole, as well as platform interoperability, are important aspects of future research. Within this framework, rapid advances in computing power are to enable DRUIDE to run in real-time while integrating algorithms of significantly increased complexity. Finally, although the primary focus of DRUIDE is human faces and hand postures, its ultimate goal extends beyond human-computer interactions, to also encompass in future work the adaptive detection, tracking and recognition of various objects under unconstrained scene conditions.

References

[1] J.-C. Terrillon, A. Pilpré, Y. Niwa, and K. Yamamoto. Robust face detection and hand posture recognition in color images for human-machine interaction. In *Proceedings of the 16th International Conference on Pattern Recognition* (ICPR), Québec City, Canada, August 2002. Volume 1, pp. 204-209.

[2] J.-C. Terrillon, M. N. Shirazi, H. Fukamachi, and S. Akamatsu. Comparative Performance of Different Skin Chrominance Models and Chrominance Spaces for the Automatic Detection of Human Faces in Color Images. In *Proceedings of the Fourth International Conference on Face and Gesture Recognition* (ICFGR), Grenoble, France, March 2000. pp 54-61.

[3] J.-C. Terrillon, A. Pilpré, Y. Niwa and K. Yamamoto. Analysis of a Large Set of Color Spaces for Skin Pixel Detection in Color Images. In *Proceedings of the 6th International Conference on Quality Control by Artificial Vision* (QCAV), Gatlinburg, Tennessee, U.S.A., May 2003. SPIE Vol. 5132, pp. 433-446.

[4] 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 *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition* (ICFGR), Nara, Japan, April 1998. pp. 112-117.

[5] J.-C Terrillon, M. N. Shirazi, M. Sadek, H. Fukamachi and S. Akamatsu. Invariant face detection with support vector machines. In *Proceedings of the 15th International Conference on Pattern Recognition* (ICPR), Barcelona, Spain, September 2000. Volume 4, pp. 210-217.

[6] J.-C. Terrillon, M. N. Shirazi, D. MacReynolds, M. Sadek, Y. Sheng, S. Akamatsu and K. Yamamoto. Invariant face detection in color images using orthogonal Fourier-Mellin moments and support vector machines. In *Proceedings of the 2nd International Conference on Advances in Pattern recognition* (ICAPR), Rio de Janeiro, Brazil, March 2001. Lecture Notes in Computer Science, Springer Verlag. pp. 83-92.

[7] J.-C. Terrillon, A. Pilpré, Y. Niwa, and K. Yamamoto. DRUIDE: a real-time system for multiple face detection, tracking and hand posture recognition in color video sequences. Accepted for publication at the 17th International Conference on Pattern Recognition (ICPR), to be held in Cambridge, U.K., on 23-26 August 2004.

[8] Y.-I. Ohta, T. Kanade, and T. Sakai. Color information for region segmentation. *Computer Graphics and Image Processing*, 13(3):222-241, July 1980.

[9] S. Wesolkowski, M. E. Jernigan, and R. D. Dony. Comparison of color image edge detectors in multiple color spaces. In *Proceedings of the International Conference on Image Processing* (ICIP), Vancouver, Canada, September 2000.

[10] CIE Colorimetry, *Official recommendations of the International Commission on illumination*, Publication CIE No. 15.2, Second Edition, Central Bureau of the Commission Internationale de L'Éclairage, Vienna, Austria, 1986.

[11] ITU-R Recommendation BT.709, *Basic Parameter Values for the HDTV Standard for the Studio and for International Programme Exchange* , [formerly CCIR Rec. 709], ITU, 1211 Geneva 20, Switzerland, 1990.

[12] G. Wyszecki, and W. S. Styles. *Color science: concepts and methods, quantitative data and formulae*, 2nd edition, John Wiley, New York, 1982.

[13] R. W. G. Hunt. *Measuring Colour*, 2nd edition, Ellis Horwood, 1991.

[14] J.-C. Terrillon, M. David, and S. Akamatsu. Detection of human faces in complex scene images by use of a skin color model and of invariant Fourier-Mellin moments. In *Proceedings of the 14th International*

Conference on Pattern Recognition (ICPR), Brisbane, Australia, August 1998. Volume 2, pp. 1350-1355.

[15] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Second Edition in C Computer Graphics Principles and Practice*, Addison-Wesley, New York, 1996.

[16] H. Wu, Q. Chen, and M. Yachida. Face detection from color images using a fuzzy pattern matching method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-21(6):557-563, 1999.

[17] T. Gevers and A. W. M. Smeulders. Color-based object recognition. *Pattern Recognition*, 32(3):453-464, 1999.

[18] T. Gevers and A. W. M. Smeulders. Content-based image retrieval by viewpoint-invariant color indexing. *Image and Vision computing*, 17:475-488, 1999.

[19] J. Berens and G. D. Finlayson. Log-opponent chromaticity coding of colour space. In *Proceedings of the 15th International Conference on Pattern Recognition* (ICPR), Barcelona, Spain, September 2000. pp. 206-211.

[20] S. Kawato and J. Ohya. Real-time detection of nodding and head shaking by directly detecting and tracking the between eyes. In *Proceedings of the 4th International Conference on Face and Gesture Recognition* (ICFGR), Grenoble, France, March 2000. pp. 40-45.

[21] S. Tominaga. Illuminant estimation of natural scenes from color images. In *Proceedings of the International Conference on Color in Graphics and Image Processing* (ICCGIP), Cépaduès Éditions, Saint-Etienne, France, 2000. pp. 35-40.

[22] C. Vertan, M. Ciuc, and N. Boujemaa. On the introduction of a chrominance spectrum and its applications. In *Proceedings of the International Conference on Color in Graphics and Image Processing* (ICCGIP), Cépaduès Éditions, Saint-Etienne, France, 2000. pp. 214-218.

[23] J.-C. Terrillon, A. Pilpré, Y. Niwa and K. Yamamoto. Analysis of the Influence of the Number of Skin Sample Pixels on the Chrominance Distribution of Human Skin. In *Proceedings of the Japanese Industrial Applications Society Conference* (JIASC'2003), Tokyo University of Technology, 26-28 August 2003. Vol. 2, pp.291-296.

[24] J.-C. Terrillon, A. Pilpré, Y. Niwa and K. Yamamoto. Analysis of the Chrominance Distribution of Human Skin under Unconstrained Image Scene. In *Proceedings of the Joint Technical Meeting on Information Processing and Information-Oriented Industrial Systems*, Arima Onsen, 16 January 2004. Technical report of the Institute of Electrical Engineers of Japan (IEEJ) IP-04-1. pp. 1-6.

[25] J. L. Horner and P. D. Gianino. Phase-only matched filtering. *Applied Optics*, 23(6): 812-816, 1984.

[26] S. Di Zenzo. A note on the gradient of a multi-image. *Computer Vision, Graphics and Image Processing*, 33:116-125, 1986.

[27] H.-C Lee and D. R. Cok. Detecting boundaries in a vector field. *IEEE Transactions on Signal Processing*, 39(5):1181-1194, 1991

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

[29] Y. Sheng and L. Shen. Orthogonal fourier-mellin moments for invariant pattern recognition. *Journal of the Optical Society of America-A*, 11(6):1748-1757, 1994.

[30] J. Triesch and C. von der Malsburg. Robust Classification of hand postures against complex backgrounds. In *Proceedings of the Second International Conference on Face and Gesture Recognition* (ICFGR), Killington, Vermont, October 1996. pp. 170-175.

[31] M. Handouyahia, D. Ziou, and S. Wang. Sign language recognition using moment-based size functions. In *Proceedings of the 12th Conference on Vision Interface* (VI), Trois-Rivières, Canada, May 1999. pp. 210-216.

[32] A. V. Oppenheim and J. S. Lim. The importance of phase in signals. In *Proceedings of the IEEE*, 69:529-541, 1981.

[33] http://www.intel.com/research/mrl/research/opencv/

[34] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *Proceedings of the CVPR*, Puerto Rico, June 1997. pp. 130-136.

[35] B. Schoelkopf, A. Smola, R. C. Williamson, and P. L. Bartlett. New Support Vector Algorithms. *Neural Computation*, 12:1207-1245, 2000.

[36] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proceedings of the CVPR*, Hilton Head Island, S.C., U.S.A., 2000. Vol. 2, pp. 142-149.

[37] G. Bradsky. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, 2nd Quarter 1998.

[38] E. Hayman, I. D. Reid, and D. W. Murray. Zooming while tracking using affine transfer. In *Proceedings of the 7th British Machine Vision Conference*, BMVA Press, 1996. pp. 395-404.

[39] D. W. Murray, K. J. Bradshaw, P. F. McLaughlan, I. D. Reid, and P. M. Sharkey. Driving saccade to pursuit using image motion. *International Journal of Computer Vision*, 16(3):205-228, 1995.

[40] L. de Agapito, E. Hayman, and I. Reid. Self-calibration of rotating and zooming cameras. Technical Report OUEL 0225/00, Department of Engineering Science, University of Oxford.