

# Automatic Detection of Human Faces in Natural Scene Images by Use of a Skin Color Model and of Invariant Moments

Jean-Christophe Terrillon, Martin David and Shigeru Akamatsu  
ATR Human Information Processing Laboratories  
2-2 Hikaridai, Seika-cho,  
Soraku-gun, Kyoto 619-02, Japan  
terril@hip.atr.co.jp

## Abstract

*We use a skin color model based on the Mahalanobis metric and a shape analysis based on invariant moments to automatically detect and locate human faces in two-dimensional natural scene images. First, color segmentation of an input image is performed by thresholding in a perceptually plausible hue-saturation color space where the effects of the variability of human skin color and the dependency of chrominance on changes in illumination are reduced. We then group regions of the resulting binary image which have been classified as face candidates into clusters of connected pixels. Performing median filtering on the image and discarding the smallest remaining clusters ensures that only a small number of clusters will be used for further analysis. Fully translation-, scale- and in-plane rotation-invariant moments are calculated for each remaining cluster. Finally, in order to distinguish faces from distractors, a multilayer perceptron neural network is used with the invariant moments as the input vector. Supervised learning of the network is implemented with the backpropagation algorithm, at first for frontal views of faces. Preliminary results show the efficiency of the combination of color segmentation and of invariant moments in detecting faces with a large variety of poses and against relatively complex backgrounds.*

## 1. Introduction

Automatic detection and localisation of human faces in two-dimensional natural, complex scene images is a difficult task that has been relatively unexplored until recently [5]. On the other hand, a large body of research has addressed the problem of higher-level face recognition tasks such as personal identification, sex/race determination and the understanding of facial expressions, by implicitly assuming that a face has been previously segmented from an arbitrary

background [1]. The implementation of an efficient automatic face recognition system implies that, at the lowest level, the techniques used to detect and locate a face in a scene are robust and amenable to a subsequent higher-level analysis of facial features. Color is a powerful fundamental cue that can be used as a first step in the face detection process because color image segmentation is computationally fast while being relatively robust to changes in illumination, in viewpoint, in scale, to shading and to complex (cluttered) backgrounds. Robustness is achieved if a color space efficiently separating the chrominance from the luminance in the original color image and a plausible model of the human skin chrominance distribution are used for thresholding. However, shape analysis of the segmented scene is also necessary in order to discriminate face candidates from the remaining distractors (such as false positives or other body parts correctly classified as skin). The robustness of face detection and localisation is increased if translation-, scale- and in-plane rotation-invariant quantities characterizing the shape of a face in the segmented scene are used in the shape analysis.

We propose to combine skin-color based image segmentation with a shape analysis using invariant moments as an input vector to a multilayer perceptron neural network (NN). In section 2, the chrominance distribution of human skin in several different color spaces is analyzed. The analysis is mainly based on an anthropometric chromatic scale that assigns an equal statistical weight to each component of the scale under the same illumination conditions. We then define a color space with the perceptual attributes of hue and saturation where the effects of the variability of skin color and the dependency of chrominance on changes in illumination are reduced. A skin chrominance model is then proposed for that particular space. Finally, we determine a threshold value for the Mahalanobis metric inherent to the model. The main scope of section 3 is to present the invariant moments that are used to characterize the clusters

of connected pixels resulting from a connected-component analysis of the segmented images. The architecture of the NN is also briefly described. Experimental results are presented in section 4, and in section 5 we discuss the limits of the performance of the present face detection prototype and summarize the main issues that we plan to address in future research.

## 2. Color segmentation

### 2.1. Selection of a color space

The efficiency of the color segmentation of a human face depends on the color space that is selected, because the color distribution of human skin depends on the color space. For a first analysis, the distribution in a given space is based on an unbiased chromatic scale that is representative of a sufficiently large number of skin colors.

To our knowledge, the earliest quantitative classification of skin color was attempted in the field of physical anthropology : von Luschan's chromatic scale [8] can be used to match 36 different skin colors, from an unsaturated light color to a saturated dark brown color, as shown in figure 1<sup>1</sup>. Although this scale is crude, it can be used as a reference because it assigns an equal statistical weight (in number of pixels in the distribution analysis) to each component of the scale. All the components in Figure 1 are recorded under the same illumination conditions. They are mapped in different color spaces. Figure 2 shows the color distribution in normalized r-g space ( $r=R/(R+G+B)$  and  $g=G/(R+G+B)$ ) at first separately for light colors matching white Caucasians, for intermediate colors matching Asians, and for dark colors matching dark-skinned people of African descent and Indo-caucasians, and suntanned skins. While the distributions of the first two classes are concentrated near the equal-energy point ( $r=g=1/3$ ), the distribution of the class representative of dark skin covers a significantly larger surface area, with larger values of r and lower values of g. Figure 2d shows the cumulative distribution of all the components on a logarithmic scale. The distribution consists essentially of two clusters, a white Caucasian/Asian complex with a continuous transition to the cluster characteristic of dark skins. Because the dark-skin cluster is more diffuse, it is concealed if a logarithmic scale is not used. Figure 3 shows the same cumulative distribution on a logarithmic scale in the normalized CIE-xy space (with  $x=X/(X+Y+Z)$  and  $y=Y/(X+Y+Z)$ ) and in chrominance (Hue H and Saturation S) in HSV space.

<sup>1</sup>A color image of von Luschan's chromatic scale can be found at [http://www.unifi.it/unifi/msn/antrop/route/stru\\_eng.htm](http://www.unifi.it/unifi/msn/antrop/route/stru_eng.htm).

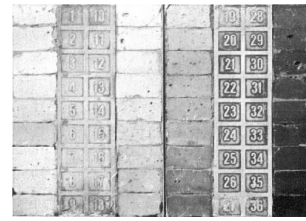


Figure 1. An image of von Luschan's chromatic scale for the classification of skin color.

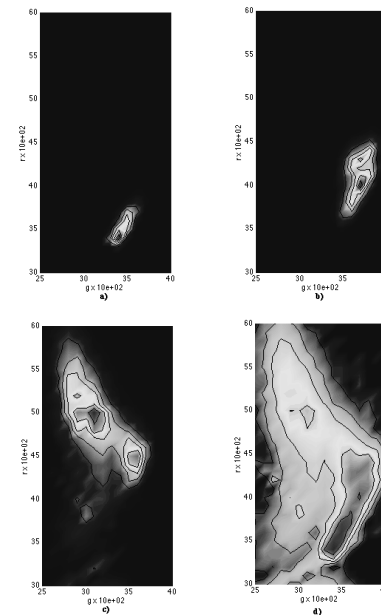
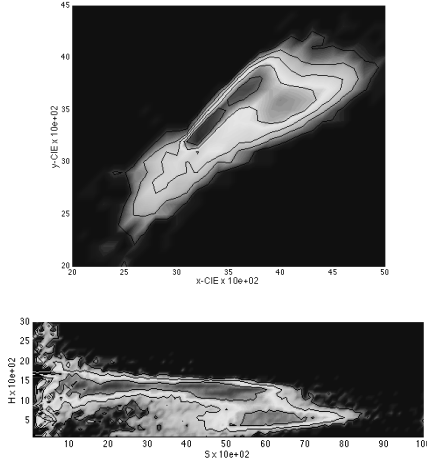


Figure 2. False-color top view of the 2-D histograms in normalized r-g space of the components of von Luschan's chromatic scale : a) for colors matching white Caucasians, b) for colors matching Asians, and c) for colors matching dark-skinned people of African descent and Indo-caucasians; d) cumulative histogram of all the components on a logarithmic scale.

In both spaces, the overall distribution is similar to that observed in r-g space. However, while the distribution is confined in both r-g and CIE-xy spaces, it covers almost all the possible range of saturation S for a limited range in hue H in H-S space. Therefore, normalization of RGB values by  $R+G+B$  or of CIE-XYZ values by  $X+Y+Z$  yields color spaces that are more efficient for skin color segmentation than HSV space where the normalization is not performed,



**Figure 3. Cumulative histograms in normalized CIE-xy space and in H-S space of all the components of von Luschan's chromatic scale (logarithmic scale).**

because the sensitivity to the variation of skin color is reduced. This conclusion also applies to the more general case where illumination conditions vary widely and where different camera systems are used, since the normalization theoretically renders the chrominance independent of any identical changes in R, G and B or X, Y and Z values. In practice such normalized spaces are shown to be relatively robust to changes in measured intensity.

The normalization of RGB values by R+G+B can be followed by a further transformation into a perceptually plausible hue-saturation chrominance space. We selected a modified color space where saturation S, tint T and value V are calculated by use of the following transformations :

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

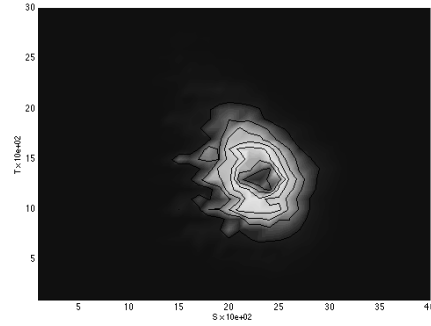
$$T = [\arctan(r'/g')]/\pi + 1/2 \quad (2)$$

$$V = (R + G + B)/3 \quad (3)$$

where  $r'=(r-1/3)$  and  $g'=(g-1/3)$ . The values of S, T and V are normalized in the range [0. ; 1.0].

As a first step, only the white Caucasian/Asian complex was considered for color segmentation. Images of 11 asian and 19 white caucasian subjects were recorded under slowly varying illumination conditions with a single video camera mounted on an SGI computer. 110 skin sample images were manually selected to analyze the color distribution in T-S space and to calibrate the camera for color segmentation. The cumulative distribution of all the samples, shown in

Figure 4, covers a small surface area. This is also observed with von Luschan's scale, in which case the range of S is greatly reduced.



**Figure 4. Cumulative histogram in S-T space of 110 skin sample images of 11 asian and 19 white caucasian subjects used for the calibration of the SGI camera.**

## 2.2. Chrominance model for human skin

For the color space that we have selected and as Figure 4 shows , the skin color distribution for white Caucasians and Asians can be modeled by an elliptical Gaussian chrominance probability density function (pdf). If the vector  $\underline{\mathbf{X}}(i, j) = [\underline{T}(i, j)\underline{S}(i, j)]^T$  represents the random measured values of tint  $\underline{T}$  and of saturation  $\underline{S}$  of a pixel with coordinates  $(i, j)$  in an image, the pdf is given by

$$p[\underline{\mathbf{X}}(i, j)/W_s] = (2\pi)^{-1} |\mathbf{C}_s|^{-1/2} \exp\{-[\lambda_s(i, j)]^2/2\} \quad (4)$$

where  $W_s$  is the class describing skin,  $\mathbf{C}_s$  is the covariance matrix for skin color :

$$\mathbf{C}_s = \begin{bmatrix} \sigma^2_{Ts} & \sigma_{TSs} \\ \sigma_{TSs} & \sigma^2_{Ss} \end{bmatrix} \quad (5)$$

and where  $\lambda_s(i, j)$  is the Mahalanobis distance from the vector  $\underline{\mathbf{X}}(i, j)$  to the mean vector  $\mathbf{m}_s = [m_{Ts} m_{Ss}]^T$  obtained for skin color, defined as

$$[\lambda_s(i, j)]^2 = [\underline{\mathbf{X}}(i, j) - \mathbf{m}_s]^T \mathbf{C}_s^{-1} [\underline{\mathbf{X}}(i, j) - \mathbf{m}_s] \quad (6)$$

Eq. (6) defines elliptical surfaces in chrominance space of scale  $\lambda_s(i, j)$ , centered about  $\mathbf{m}_s$  and whose principal axes are determined by  $\mathbf{C}_s$ . The value of  $\lambda_s(i, j)$  for a pixel with coordinates  $(i, j)$  determines the probability that the pixel belongs to the class  $W_s$  representing human skin, as seen from Eq. (4). The larger  $\lambda_s(i, j)$ , the lower the probability that the pixel  $(i, j)$  belongs to  $W_s$ .

### 2.3. Thresholding of color test images

Both  $\mathbf{m}_s$  and  $\mathbf{C}_s$  are estimated by use of the 110 skin sample images recorded with the SGI camera.  $[\lambda_s(i, j)]^2$  is then calculated for every pixel over all the skin samples and over five large image regions not containing skin and is compared to a parameter  $\lambda_{s,t}^2 > 0$  for thresholding. A "standard" threshold value  $\lambda_{s,t}^2$  is obtained when the proportion of true positives TP over the ensemble of regions of skin becomes equal to the proportion of true negatives TN over the ensemble of regions not containing skin [6], or equivalently, when the proportion of false negatives FN equals the proportion of false positives FP (since TP+FN=1 and TN+FP=1). This initial skin color calibration must be performed for any color camera before color image segmentation. However, under slowly varying illumination conditions the use of a single camera does not require the further application of an adaptive threshold as in [6]. Color segmentation is performed on test images by calculating  $[\lambda_s(i, j)]^2$  for every pixel in the test images and comparing its value to the standard threshold  $\lambda_{s,t}^2$ . A value of 1 is assigned to pixel  $(i, j)$  if  $\lambda_s \leq \lambda_{s,t}$  and a value of 0 if  $\lambda_s > \lambda_{s,t}$ . The result of thresholding is a binary image that is subjected to further (shape) analysis in order to isolate face candidates in a scene.

### 3. Shape analysis

A local median filter with a window of 3x3 or 5x5 pixels is applied to the thresholded binary image in order to obtain clusters of connected pixels (connected component analysis). The clusters of area less than 0.5% of the area of the image in number of pixels are discarded so that only a small number of main clusters are used for further analysis.

#### 3.1. Calculation of Invariant moments

In order to characterize the shape of each cluster, we use the method of invariant moments that were first developed by Hu [2] and recently generalized by Li [3]. Hu's moments are fully translation, scale and in-plane rotation invariant. Although they have been used extensively in low-level pattern-recognition problems [3], to our knowledge they have not until now been used in combination with color segmentation for the detection of human faces. Hu's generalized moments arise as a particular case of the circular Fourier and radial Mellin transform (FMT) [7], defined in polar coordinates  $(r, \theta)$  as

$$M_{s,m} = \int_0^{2\pi} \int_0^\infty r^s f(r, \theta) \exp(-im\theta) r dr d\theta \quad (7)$$

where  $f(r, \theta)$  is the object to be analyzed, the Mellin transform order  $s$  is in general complex and where the circu-

lar harmonic expansion order  $m = 0, \pm 1, \pm 2, \dots$ . When  $s$  is purely imaginary, the FMT is reduced to a two-dimensional Fourier transform in log-polar coordinates. It has been proposed that such a log-polar transform describes the mapping of information from the retina to the visual cortex [7]. When  $s$  is an integer,  $s = 0, 1, 2, \dots$ , the FMT becomes directly related to Hu's moments. First, in order to achieve translation, scale and rotation invariance, the centroid of the object is subtracted in Eq. (7) (without loss of generality, the origin of the polar coordinate system is located at the centroid), Eq. (7) is normalized by the quantity  $M'_{0,0}{}^{(s/2+1)}$ , where  $M'_{0,0}$  is the zero-order geometric moment of the object, and the object is rotated such that  $\theta \rightarrow \theta + \alpha$ , where  $\alpha$  is an arbitrary angle. Then, after a transformation from polar coordinates  $(r, \theta)$  to cartesian coordinates  $(x, y)$ , Eq. (7) becomes

$$I_{s,m} = \frac{\exp(im\alpha)}{M'_{0,0}{}^{(s/2+1)}} \int_{-\infty}^\infty \int_{-\infty}^\infty (x^2 + y^2)^{(s-m)/2} \times (x - iy)^m f(x, y) dx dy \quad (8)$$

where  $i = \sqrt{-1}$ . Hu's invariant moments can be generated from Eq. (8) by appropriate combinations of the  $I_{s,m}$  that cancel the phasor  $\exp(im\alpha)$ . Since the geometric moments of order  $p, q = 0, 1, 2, \dots$ , are defined as

$$M'_{p,q} = \int_{-\infty}^\infty \int_{-\infty}^\infty x^p y^q f(x, y) dx dy \quad (9)$$

and that translation and scale-invariant moments are obtained from Eq. (9) by use of the equation

$$\mu_{p,q} = \frac{1}{M'_{0,0}{}^{[(p+q)/2+1]}} \int_{-\infty}^\infty \int_{-\infty}^\infty (x - x_0)^p (y - y_0)^q \times f(x, y) dx dy \quad (10)$$

where  $(x_0, y_0) = (M'_{1,0}/M'_{0,0}, M'_{0,1}/M'_{0,0})$  is the centroid, Hu's moments can be calculated from Eqs (8) and (10), provided that  $p+q=s$  and that  $s = |m|+2, |m|+4, \dots$ . The 11 lowest-order invariant moments, that include the second to the fourth-order geometric moments, are

$$\begin{aligned} H_1 &= \mu_{2,0} + \mu_{0,2} = I_{2,0} \\ H_2 &= (\mu_{2,0} - \mu_{0,2})^2 + 4\mu_{1,1}^2 = |I_{2,2}|^2 \\ H_3 &= (\mu_{2,0} - \mu_{0,2})[(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] \\ &\quad + 4\mu_{1,1}(\mu_{3,0} + \mu_{1,2})(\mu_{2,1} + \mu_{0,3}) = |I_{3,3}|^2 \\ H_4 &= (\mu_{3,0} + \mu_{1,2})^2 + (\mu_{2,1} + \mu_{0,3})^2 = |I_{3,1}|^2 \\ H_5 &= (\mu_{3,0} - 3\mu_{1,2})^2 + (3\mu_{2,1} - \mu_{0,3})^2 \\ &= \text{Re}\{I_{3,3}(I_{3,-1})^3\} \\ H_6 &= (\mu_{3,0} + \mu_{1,2})(\mu_{3,0} - 3\mu_{1,2}) \\ &\quad \times [(\mu_{3,0} + \mu_{1,2})^2 - 3(\mu_{2,1} + \mu_{0,3})^2] \\ &\quad + (\mu_{2,1} + \mu_{0,3})(3\mu_{2,1} - \mu_{0,3}) \\ &\quad \times [3(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] \\ &= \text{Re}\{I_{2,2}(I_{2,-1})^2\} \end{aligned}$$

$$\begin{aligned}
H_7 &= (\mu_{3,0} + \mu_{1,2})(3\mu_{2,1} - \mu_{0,3}) \\
&\times [(\mu_{3,0} + \mu_{1,2})^2 - 3(\mu_{2,1} + \mu_{0,3})^2] \\
&- (\mu_{2,1} + \mu_{0,3})(\mu_{3,0} - 3\mu_{1,2}) \\
&\times [3(\mu_{3,0} + \mu_{1,2})^2 - (\mu_{2,1} + \mu_{0,3})^2] \\
&= Re\{I_{3,3}(I_{3,-1})^3\} \\
H_8 &= \mu_{4,0} + 2\mu_{2,2} + \mu_{0,4} = I_{4,0} \\
H_9 &= (\mu_{4,0} - \mu_{0,4})^2 + 4(\mu_{3,1} + \mu_{1,3})^2 = |I_{4,2}|^2 \\
H_{10} &= (\mu_{4,0} - 6\mu_{2,2} + \mu_{0,4})^2 + 16(\mu_{3,1} - \mu_{1,3})^2 = |I_{4,4}|^2 \\
H_{11} &= (\mu_{4,0} - 6\mu_{2,2} + \mu_{0,4}) \\
&\times [(\mu_{4,0} - \mu_{0,4})^2 - 4(\mu_{3,1} + \mu_{1,3})^2] \\
&+ 16(\mu_{4,0} - \mu_{0,4})(\mu_{3,1} + \mu_{1,3})(\mu_{3,1} - \mu_{1,3}) \\
&= Re\{I_{4,4}(I_{4,-2})^2\} \quad (11)
\end{aligned}$$

The computation of the discrete moments for binary images yields theoretically an error-free estimate of the continuous moments as opposed to the computation performed for grey-level images [4], and the moments are then also independent of illumination. Higher orders  $s$  (or  $p$  and  $q$ ) amplify the contributions of the peripheral parts of an object to the moments. The contours of segmented face candidates constitute a correlated noise because they are variable, so that only a small number of the lowest invariant moments should be used in the shape analysis.

### 3.2. Neural network applied to invariant moments

Hu's 11 lowest-order moments are the input units of a feed-forward multilayer perceptron, with one hidden layer containing 6 nodes and with one output unit. All the units take on continuous bipolar-sigmoid activation values in the range [-1.0 ; 1.0]. For training, the backpropagation algorithm is applied to perform gradient descent on a quadratic error function (or total error) using batch processing. A momentum term is included in the learning rule in order to increase the learning rate of the network. When a face is detected by the network, it is marked by an ellipse using the already computed first and second-order geometric moments of the cluster representing the face, according to the equations [7] :

$$(a, b) = \left[ \frac{\nu_{2,0} + \nu_{0,2} \pm \sqrt{(\nu_{2,0} - \nu_{0,2})^2 + 4\nu_{1,1}^2}}{\nu_{0,0}/2} \right]^{1/2} \quad (12)$$

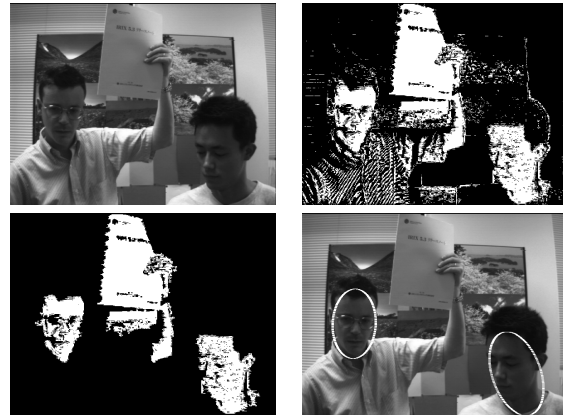
$$\theta = \frac{1}{2} \arctan\left(\frac{2\nu_{1,1}}{\nu_{2,0} - \nu_{0,2}}\right) \quad (13)$$

where  $\nu_{p,q} = M_{0,0}^{[(p+q)/2+1]} \mu_{p,q}$ ,  $a$  and  $b$  are respectively the semi-major and semi-minor axes of the ellipse, and where  $\theta$  defines its orientation.

## 4. Experimental results

The face detection system is implemented on an SGI Indigo 2 Impact 10000 computer. The Mahalanobis thresh-

old distance is found to be  $\lambda_{s,t}^2 = 1.62$  for TP=TN=83.3% (hence FN=FP=16.7%). After the color segmentation and the connected-component analysis, N=220 elements (clusters) with an approximate ratio of 1:1 between elements describing faces and those representing objects other than faces were used to train the neural network. The efficiency of the network was investigated using a test file of 100 elements, with again a 1:1 ratio between faces and other objects. Preliminary results for the test images used indicate that 85% of faces were correctly detected while 82% of clusters not representing faces were correctly rejected. Figure 5 illustrates the procedure used to detect faces in a scene : two faces are successfully detected by the NN while another object that was classified as skin is correctly rejected. In Figure 6, other examples show that faces of asian and white caucasian subjects (including also an indo-caucasian subject) are equally well detected against relatively complex backgrounds, and demonstrate the invariant detection capabilities of the NN as well as the capability of the network to discriminate between faces and other body parts. The time required for face detection, without optimization of the algorithms, is 0.2 second for images of dimensions 320 x 243 pixels which are typically used on the SGI computer.



**Figure 5. Details of the automatic detection of human faces in a scene. From left to right and from top to bottom are the original color image containing two faces, the thresholded image, the results of the connected-component analysis and the successfully detected faces, each marked by an ellipse.**

## 5. Discussion

While preliminary results are very encouraging, four main types of errors limit the performance of the system,



**Figure 6. Examples of the detection of faces with different poses and different skin colors.**



**Figure 7. Examples of problems encountered with the present face detection prototype.**

as the examples of Figure 7 show.

First, other body parts (such as the neck) connected to or in contact with a face may lead to face localization errors. In this particular case, the NN still detects the face because the entire cluster of connected pixels retains a similar shape to that of a face (with holes at the location of the eyes and of the mouth). This type of error can also occur when partial occlusions, by dark glasses or by facial hair for example, erode the face cluster or divide it into two distinct clusters. Second, false positives may occur due to other body parts correctly classified as skin. While significantly increasing the range of face poses that can be detected, the invariant moments also tend to increase the frequency of detection of other regions of skin as faces precisely because of their higher tolerance. Third, false positives due to other objects in the scene result from the sensitivity of the color segmentation to the value of the threshold  $\lambda_{s,t}$  and to the limited color discrimination of the camera system. Finally, false negatives may be caused by strong shadows and/or strong illumination that eliminate the chromatic information in regions of a face and thus reduce the efficiency of the color segmentation, or by partial occlusions that divide the face cluster into two or more distinct clusters, or by other objects in the scene that are connected to the face cluster and modify its shape significantly.

Based on the above discussion, three main issues are the focus of future research : first, the efficiency of the color segmentation is to be improved by including an adaptive thresholding algorithm that would increase both the robustness to larger variations of illumination and the portability between different camera systems. Also, a color model for dark skins is to be combined with the present model. Second,

the luminance information (value V) is to complement the color segmentation by an analysis of regions of skin where the chrominance information is eliminated and in order to separate faces from other connected body parts or objects. Finally, we plan to train the NN to detect also side views of faces, and to increase its generalization capability by use of a larger set of both training and test images.

#### Acknowledgements

We are grateful to Aude Oliva, Michael Lyons and Nicolas Schweighofer for helpful discussions.

#### References

- [1] M. Bichsel, editor. *Proc. of the International Workshop on Automatic Face- and Gesture-Recognition*, Zurich, 1995.
- [2] M. K. Hu. Visual pattern recognition by moment invariants. *IEEE Trans. Inf. Theory*, IT-8:179–187, 1962.
- [3] Y. Li. Reforming the theory of invariant moments for pattern recognition. *Pattern Recognition*, 25(7):723–730, 1992.
- [4] S. X. Liao and M. Pawlak. On image analysis by moments. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-18(3):254–266, 1996.
- [5] I. C. S. Press, editor. *Proc. of the Second International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, 1996.
- [6] E. Saber, A. M. Tekalp, R. Eschbach, and K. Knox. Automatic image annotation using adaptive color classification. *Graphical Models and Image Processing*, 58(2):115–126, 1996.
- [7] Y. Sheng and C. Lejeune. Invariant pattern recognition using fourier-mellin transforms and neural networks. *J. Optics (Paris)*, 22(5):223–228, 1991.
- [8] F. von Luschan. *Voelker, Rassen, Sprachen : Anthropologische Betrachtungen*. Deutsche Buchgemeinschaft, Berlin, 1927. 382 pp.