Robust Face Detection and Japanese Sign Language Hand Posture Recognition for Human-Computer Interaction in an "Intelligent" Room

†Jean-Christophe Terrillon, †Arnaud Pilpré,
†Yoshinori Niwa and ‡Kazuhiko Yamamoto
†Office of Regional Intensive Research Project (HOIP), Softopia Japan Foundation,
4-1-7 Kagano, Ogaki-City, Gifu 503-8569, Japan
{terrillon, pilpre, niwa}@softopia.pref.gifu.jp
‡Faculty of Engineering, Gifu University,
1-1 Yanagido, Gifu-City, Gifu 501-1193, Japan
yamamoto@info.gifu-u.ac.jp

Abstract

A system for the detection of human faces and for the classification of hand postures of the Japanese Sign Language in color images inside an "intelligent" room is presented. We first propose to apply a combination of a skin chrominance-based image segmentation with a color vector gradient-based edge detection [1] [2] to efficiently detect faces and hands. Within the framework of a general approach, a statistical model for face detection based on invariant moments [3] [4] is used to discriminate between faces and hands in the segmented images. A novel approach to hand posture recognition based on phase-only correlation [5] is then adopted to classify a subset of static hand postures of the Japanese Sign Language, each posture representing a given phoneme, and also to discriminate between hand postures and the image scene background. Experiments show that the additional use of the color vector gradient significantly improves the correct rate of face detection, and that the phase-only correlation filter yields a high rate of discrimination between different static hand postures as well as between hand postures and the scene background. Ultimately, the system is to contribute to the implementation of meaningful human-machine interactions in a room that we are in the process of establishing, the "percept-room", mainly for welfare applications.

1. Introduction

Recently, so-called "intelligent" environments, in which a range of human activities can be automatically sensed, analyzed and "understood" by use of various computer vision technologies that are the least conspicuously embedded in the environment but are ubiquitous, have been developed [6] [7] [8] [9]. Meaningful human-machine interactions in an "intelligent" room such as the "percept-room", which we are in the process of implementing by use of multiple cameras [10], require as a first step the automatic detection of human faces, as well as of hands, for higher-level face and gesture recognition tasks. In particular, such human-machine interfaces allow a human user to control a variety of devices without any physical contact with remote controls, keyboards, etc... Various applications have been suggested, such as the contact-less control of home appliances (for example, of a television set, as in [11] [12]) for welfare improvement.

A fundamental issue to address is the level of complexity of the scene background that is to be expected in an "intelligent" room, 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. Much work has focused on the robust detection of faces only (for example, [3] [4]), or of hands only [13], or on the robust recognition of (static) hand postures only [14] [15]. Also, it is often implicitly assumed that a face or a hand is present in a scene image. Finally, the "background" may also be considered to include the clothes that a person is wearing, other body parts than faces or hands (such as the neck and arms), and facial attributes such as glasses, hair and hairstyle, etc...

In this paper, we propose a system for the robust simultaneous detection of human faces and classification of hand postures of the Japanese Sign Language (JSL) in color images. The system 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. 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 [5] 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, 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 scene background. An overview of the system is shown in Figure 1.

The paper is organized as follows: in section 2, we describe the face detection module. The hand posture recognition module is presented in section 3, with an emphasis on the phase-only correlation filter and its properties for object recognition. In section 4, we briefly describe the experimental setup that we have used inside the percept-room. Experimental results are presented and discussed in section 5, and conclusions are drawn in section 6, with suggestions for future work.

2. Face Detection Module

As shown in Figure 1, 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 [1] [2]. To increase the robustness of face detection, fully translation-, scale- and in-plane rotation-invariant moments are then calculated for each significant blob that results from the segmentation [3] [4], and the resulting feature vectors are 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 (the segmented faces are approximately elliptical, with holes at the location of the eyes and of the mouth in the segmented images), 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.

2.1 Skin Chrominance-based Image Segmentation

The image segmentation based on the chrominance of human skin is performed with the normalized r-g or CIE-xy chrominance spaces (where r=R/(R+G+B) and g=G/(R+G+B) for a 24-bit RGB image). We recently showed [16] that these two normalized spaces were the most efficient for image segmentation based on skin color among nine different chrominance spaces and in terms of six different criteria, that include in particular the robustness to a change of illumination conditions and to a change of the camera system that is used to record the images. Moreover, in these spaces, a simple, single Gaussian model can be used to estimate the skin chrominance distribution, which obviates the necessity to apply a complex and computationally intensive model in

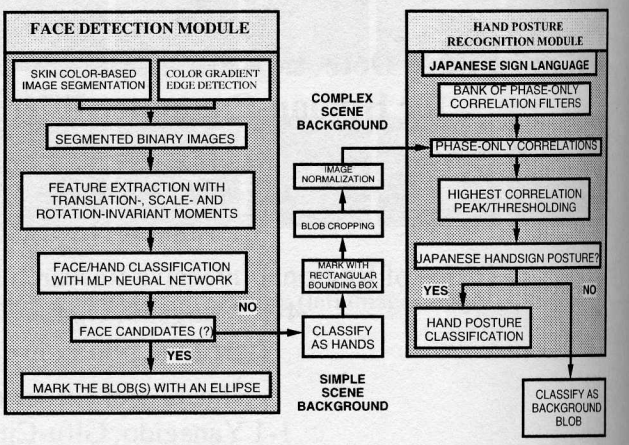

Figure 1. Flowchart of the face detection and hand posture recognition system, taking into account the image scene background.

order to achieve a high quality of segmentation. 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 [17].

2.2 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 [1] and later more thoroughly investigated by Lee and Cok [2]. As shown in [2], 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.

If we let $\vec{C} = [\, R(x,y), G(x,y), B(x,y) \,]^T$ be the color vector that contains the three color components R, G and B as a function of the Cartesian coordinates (x,y), the modulus of the color gradient is then, for a continuous image [2]:

$$\left\| \vec{\nabla} \vec{C} \right\| = \left\{ \frac{1}{2} \left\{ g_{xx} + g_{yy} + \sqrt{ (g_{xx} - g_{yy})^2 + 4 g_{xy}^2 } \right\} \right\}^{1/2} \quad (1)$$

where the quantities g_{xx}, g_{xy} and g_{yy} are given by

$$g_{xx} = \left(\frac{\partial R}{\partial x}\right)^2 + \left(\frac{\partial G}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial x}\right)^2 \quad (2)$$

$$g_{xy} = \left(\frac{\partial R}{\partial x}\right)\left(\frac{\partial R}{\partial y}\right) + \left(\frac{\partial G}{\partial x}\right)\left(\frac{\partial G}{\partial y}\right) + \left(\frac{\partial B}{\partial x}\right)\left(\frac{\partial B}{\partial y}\right) \quad (3)$$

$$g_{yy} = \left(\frac{\partial R}{\partial y}\right)^2 + \left(\frac{\partial G}{\partial y}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2 \quad (4)$$

The numerical computation of the partial derivatives in Eqs. (2)-(4) is performed by use of either the Sobel or Prewitt operators (in our experiments, the edge detection results with both operators are very similar). The edge detection is performed in the original RGB color image. In the color gradient modulus map, the threshold for edge detection is typically set at 10% of the maximum gradient modulus value. The edge image is then subtracted from the converted r-g or CIE-xy chrominance images, before applying the skin chrominance-based image segmentation.

2.3 Feature Extraction and Face-Hand Classification

For feature extraction, after a connected-component analysis of the segmented binary images, a selected number of low-order invariant moments [3] [4] are calculated for each blob in the images. We apply either the Fourier-Mellin moments generalized by Li [18] or the orthogonal Fourier-Mellin moments developed by Sheng and Shen [19]. 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 [4], 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 [3].

3. Hand Posture Recognition Module

As Figure 1 indicates, any face candidate that is not classified as a face may be classified as a hand (eventually connected to an arm), 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 process, and consequently of being detected as a hand (or as a face). In contradistinction to segmented frontal images of faces, 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, representing a letter or a phoneme, is consistent. The hand posture recognition module is linked to the face detection module by the following process: each blob classified as a hand is first 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 [14], 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 [15], 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 [5] to every normalized blob in the segmented images, to simultaneously discriminate between hand postures, and between hand postures and background blobs.

3.1 Phase-Only Correlation Filter for Hand Posture Recognition

It has been found by Oppenheim and Lim [20] 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 [5] 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 $i^2=-1$. As an example, Figure 2 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.

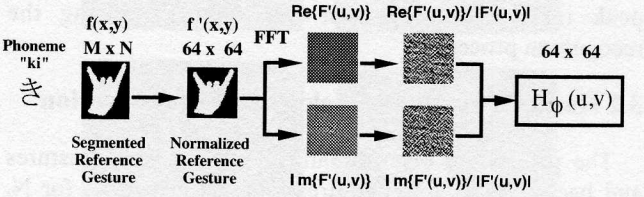

Figure 2. Synthesis of a phase-only correlation filter from the normalized image of a Japanese Sign Language hand posture obtained from a given segmented reference hand posture image, here symbolizing the phoneme "ki".

As Figure 1 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 [5], 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 3, 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 4 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, as it was shown in [21], 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 is able to 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.

3.2 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

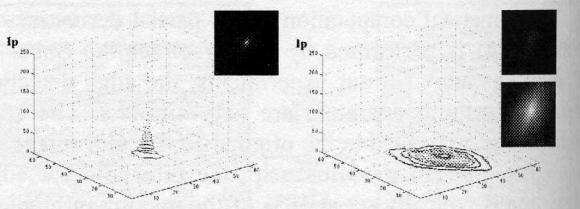

Figure 3. Comparison of the phase-only autocorrelation of the normalized hand posture image representing the phoneme "ki" (left) with the corresponding classical autocorrelation obtained with the classical matched filter (right), in units of intensity. The top view of the autocorrelation at the top right of the figure is shown in relation with the phase-only autocorrelation intensity, whereas the lower top view is scaled between 0 and 255 units.

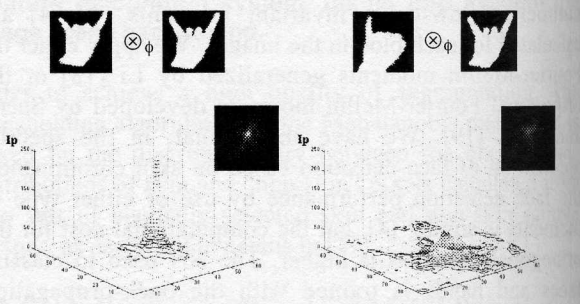

Figure 4. Comparison of the phase-only correlation between a normalized input hand posture image for the phoneme "ki" and the reference hand posture for the same phoneme (left), with the phase-only cross-correlation between a similarly shaped normalized input hand posture image for the phoneme "i" and the same reference hand posture (right), in units of intensity.

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.

4. Experimental Setup

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 with a SONY DXC-9000 camera 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 hand signs of the Japanese Sign Language (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). Outside the database, a number of face and hand images selected from various sources, that include Caucasian persons, are used for preliminary tests and for comparison.

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 79 large non-skin sample images (3x10E+06 non-skin pixels) selected from various sources.

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.

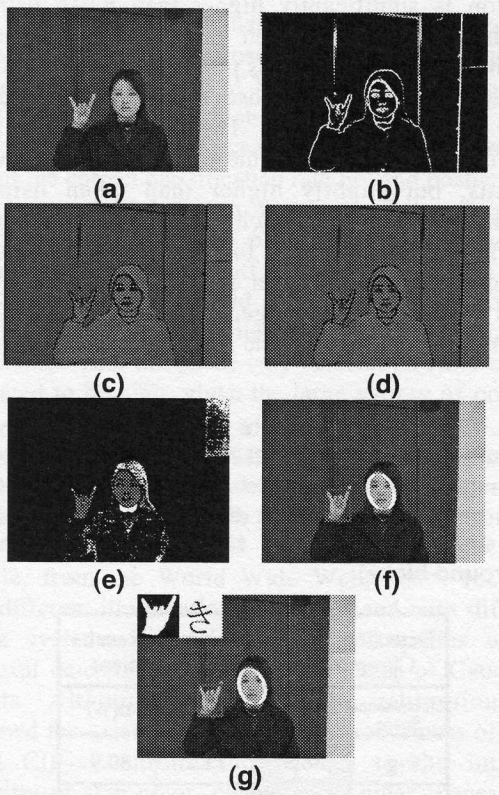

Figure 5. Illustration of the face detection and hand posture recognition process. (a) original input image, (b) results of the color vector gradient-based edge detection, (c) and (d): results of the conversion of the original image into the r-g and CIE-xy chrominance spaces respectively, with subtraction 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".

5. Experimental Results and Discussion

The complete process of face detection and hand posture recognition is illustrated in Figure 5. 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 6 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 and the clothes, as well as the hair, have been correctly classified as "non-skin" during the segmentation. However, in the example of figure 7, the clothes of a subject have been misclassified as skin, and the color gradient-based edge detection is required in order to separate the subject's face from her clothes and to detect her face. The examples of figure 8 show various 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 due 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.

Figure 6. 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.

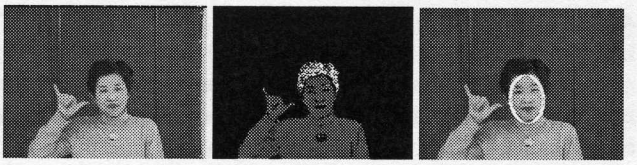

Figure 7. Example of the successful detection of the face of a Japanese subject when combining the skin chrominance-based image segmentation with the color gradient-based edge detection.

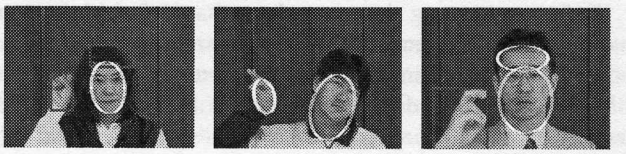

Figure 8. Example of problems occurring with the present system. From left to right: detection of a hair region as a hand, misclassification of a hand as a face, and double detection of a face due to the presence of glasses.

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:

$$CD_F = \sum_{i=1}^{N_F} TP_F / N_F \tag{6}$$

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,

$$\sum_{i=1}^{N_F} TP_F + \sum_{i=1}^{N_F} FN_F = N_F \tag{7}$$

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

$$CD_H = \sum_{j=1}^{N_H} TP_H / N_H \tag{8}$$

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:

$$\sum_{j=1}^{N_H} TP_H + \sum_{j=1}^{N_H} FN_H = N_H \tag{9}$$

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

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

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

Table 1 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, which depends on the number and size of the blobs that are present in a segmented image, is on average 270 [ms] on the Pentium-III PC and for the input image dimensions that we use.

Table 1. General results of the detection and discrimination of faces and hands inside the percept-room 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

We now focus on the recognition of hand postures of the Japanese Sign Language, before presenting preliminary general results of the discrimination between hand postures and background blobs.

Figure 9 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 2 presents preliminary 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, 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 10 shows the graph of the correct classification rate of hand postures and of background blobs 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 11 shows examples of the successful detection of faces and of hands of Caucasian subjects. Although the skin color calibration was performed for Asian subjects only, the robustness of the r-g and CIE-xy chrominance spaces to the intrinsic variability of skin color, compared to other spaces [16], leads to the correct classification of a 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).

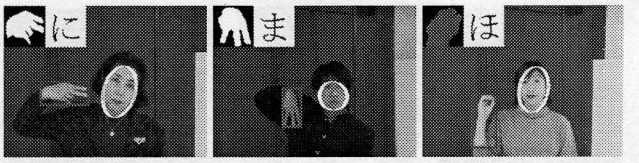

Figure 9. Example of the recognition of three different hand postures representing the phonemes "ni", "ma" and "wo" respectively. In the rightmost image, the posture for "wo" was misclassified as the posture for "ho", because the forearm of the subject is exposed and connected to her hand.

Table 2. 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).

ki to na ni nu ha mu wo

きとなにぬはむを

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

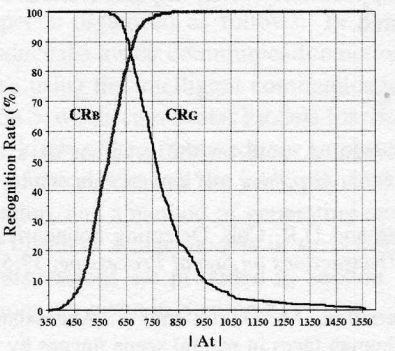

Figure 10. Graph of the correct classification rate of hand postures CR_G and of background blobs CR_B as a function of the value of the threshold amplitude modulus of the phase-only correlation.

6. Conclusions and Future Work

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

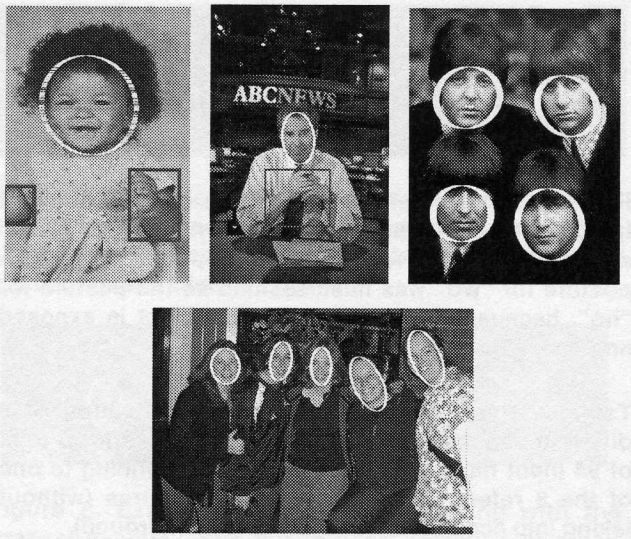

Figure 11. Examples of the successful detection of faces and hands of Caucasian subjects in images selected from various sources.

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, and consequently a higher recognition rate of hand postures, can be achieved by searching for facial features in each face candidate resulting from the image segmentation. The quality of segmentation can also be improved by considering correlations between neighboring pixels. These two approaches are an important motivation for our future research.

References

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

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

[3] 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*, Nara, Japan, 1998. pp. 112-117.

[4] J.-C. Terrillon, D. McReynolds, M. Sadek, Y. Sheng, and S. Akamatsu. Invariant neural network-based face detection with orthogonal Fourier-Mellin moments. In *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, Spain, September 2000. Vol.2, pp. 993-2000.

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

[6] M. Weiser. The computer for the 21st century. *Scientific American*, 256(3):66-76, 1991.

[7] M. Torrance. Advances in human-computer interaction: the intelligent room. *CHI'95 Research Symposium*, 1995.

[8] A. Pentland. Smart rooms, smart clothes. *Scientific American*, 274(4):68-76, 1996.

[9] A. Pentland. Looking at people: sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-22(1):107-119, 2000.

[10] M. Yasumoto, H. Hongo, H. Watanabe, and K. Yamamoto. Face direction estimation using multiple cameras for human computer interaction. In *Proceedings of the Third International Conference on Advances in Multi-modal Interfaces*, Beijing, China, October 2000. pp. 222-229.

[11] W. T. Freeman and C. D. Weissman. Television control by hand gestures. In *Proceedings of the International Workshop on Automatic Face- and Gesture- Recognition*, Zurich, Switzerland, June 1995. pp. 179-183.

[12] S. Ito, K. Yamamoto, H. Hongo, K. Kato, and Y. Niwa. The ubiquitous interface system supports bedridden people (proposal of the system and basic experiment). In *Proceedings of the 5th Symposium on Intelligent Information Media*, Tokyo, Japan, December 1999. pp. 111-116. [in Japanese].

[13] X. Zhu, J. Yang, and A. Waibel. Segmenting hands of arbitrary color. In *Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, March 2000. pp. 446-453.

[14] 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*, Killington, Vermont, October 1996. pp. 170-175.

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

[16] J.-C. Terrillon, Y. Niwa, and K. Yamamoto. On the selection of an efficient chrominance space for skin color-based image segmentation with an application to face detection. In *Proceedings of the International Conference on Quality Control by Artificial Vision*, Le Creusot, France, May 2001. Vol. 2, pp. 409-414.

[17] 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 4th International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, March 2000. pp. 54-61.

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

[19] 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.

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

[21] P. D. Gianino and J. L. Horner. Additional properties of the phase-only correlation filter. *Optical Engineering*, 23(6):695-697, 1984.