Paper

Human Face Detection In Visual Scenes Using Neural Networks

Student member Stephen Karungaru (University of Tokushima)
Member Minoru Fukumi (University of Tokushima)
Non-member Norio Akamatsu (University of Tokushima)

This paper presents a neural network based face detection system. Our objective is to design a system that can detect human faces in visual scenes at high searching speed and accuracy. We used a neural network with a simple structure but trained using face and non-face samples preprocessed by several methods (position normalization, histogram equalization etc) to attain high accuracy, then pruned the size of the neural network so that it could run faster and reduced the total search area of a target visual scene using skin color detector. Skin color detection assumes that faces reside only in skin color regions. The system designed, is made up of two parts: the face detecting system (FDS) that detects the faces (made up of the face locator, the down sampler, and the merger), and the searching speed improving system (SIS). Speed improvement is achieved by reduction of the size of the face locator (FL) network using structural learning with knowledge and by reducing the face search area using skin color detection system (SCD). Faster training of the neural networks was also achieved using variable step sizes.

Key words: Back propagation, Self- organizing maps, Down-sampling, Merging overlapping detections

1. Introduction

Recently, there is a big tread towards offloading most of the normal tasks performed by people to computers, for example face detection and recognition. However, face detection is quite difficult because no prior knowledge about the lighting, facial expression, scale, orientation, occlusions, pose or size of the faces to be located, is usually available. Accuracy and speed are very important for a face detecting system to be considered useful. This paper explains how, human face detection in a visual scene can be achieved using primarily a neural network based system.

Earlier research has been done in this area using shape and gray scale parameters [(8)], hierarchical knowledge based [(9)], neural networks [(3)(5)], template matching, feature rule generation etc. A comprehensive survey of face detection methods can be found in the references (11)(12). These conventional methods however have not treated visual scenes (images) including many persons (e.g. Over 70 people per image). In the reference (3), the neural networks used have a different structure and are larger than in our model. Therefore it is more time consuming compared to ours. Their method also is designed to work only with grayscale images. Our method trains fast because of its simple structure and also offers face searching at higher speeds because of the various speed improvement systems used.

We have further improved the detection accuracy and speed of our earlier work [(1)]. Higher speed gains were possible primarily because of training the face locator using structural learning with knowledge. "Knowledge" here means using the features of the face that are known to influence the accuracy of face detection, for example, eye pair, nose etc, to help train the network faster and reduce its size considerably. This method made the face locator's size small enough to run at high speed by pruning-off most of the redundant weights. Skin color detection is fast to process and is also orientation invariant. Detecting only the skin color areas and hence the face candidates means that we run our face locator on a smaller area thereby further improving the overall speed.

At present, the face locator can handle face images roughly facing the camera in an almost vertical orientation. The location of the face in the image can be anywhere, but the size should not be less than 20x20 pixels, which is the default training size. Larger faces can be detected by down-sampling an image to reduce its size until the faces are about 20x20 pixels. Note that ideally, the face locator should be re-applied again after every down-sampling step. Since this can be very time consuming, the down-sampling value and the number of times an image is down-sampled is calculated using the size of the face candidates (sec. 3.4). Overlapping detections are merged using the merger.

2. Face detection system (FDS)

2.1 Face Locator The face locator is the part of this system that will do the actual face detection. We chose it to be a three layered back propagation trained neural network. The size of the training samples was set at 20x20 pixels because experiment showed that the facial features, for this size, were still not so much distorted as to affect the total detection accuracy of the face locator.

The face and non-face images used for training were gathered from, scanning printed photographs, newspapers and books, picture taken using a digital camera, etc. These images contained faces of various orientations, positions, and lighting intensities. Since this system is designed to detect faces that are a minimum of 20x20pixels, from these collected images, training samples were normalized using the following steps.

1 The image is re-sampled until the sizes of the faces in it are about 20x20 pixels in size, then it is rotated until the eyes are horizontal.

2 In the 20x20 pixels region, a face is manually extracted such that the distance above the eyebrows (hair above the eye sockets) and that on the left and

right of the edges of the eyes are 1 pixel, respectively. A distance of about 3 pixels was allowed below the center of the lower lip.

3 The shape selected for training this system is a square, that is 20x20 pixels. Unfortunately, the shape of the face is not square. Therefore, to avoid complications due to different face backgrounds at the areas of the square that are not fully covered by the face, a mask has been used at the bottom to further normalize the input face. This is achieved by setting all the pixels in a triangular selected region to zero intensity. This region is 4 pixels from the bottom corners horizontally, and 5 pixels from the same corners vertically. The mask is as shown in Fig. 1.

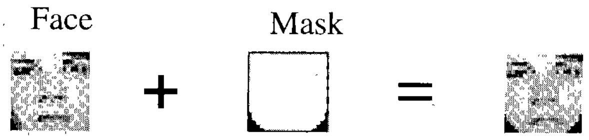


Fig 1: Masking process

4 Since most faces are symmetrical along the line between the eyes, the image is then flipped to produce its mirror image. The flipping is done as if a mirror was placed on the left side of the image.

5 Finally, the image's dynamic range and contrast are modified to produce an image with a fairly normal histogram. This process is also called histogram equalization.

The size of the face locator will depend on the color system used (10). For example, in case of RGB color system, each pixel will be represented by three values each. In other color systems like (Y)IQ, (Y)CrCb, HS(I) or (Y)UV, the brightness, (Y),(I), is separated from the color, (IQ), (CrCb), (HS), (UV). We therefore took advantage of this and represented the color using an addition of the two color components.

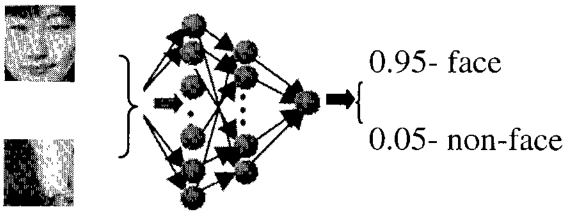


Fig. 2. The Face locator. For YIQ, the number of input units=800, hidden units=20, output units=1. Size= 16041 weights.

We ended up with two instead of three inputs. Therefore, the face locator would have 1200 input units for RGB system and 800 for the YIQ system.

The error back propagation[7] method is used to train the neural network. The system is trained to produce an output of 0.95 for a face and 0.05 for a non-face. In this project, to overcome the problem in selecting the values for both the learning rate η, and the momentum rate α, variable step size method is used.

The total initial number of face example used is 504 and the non-face examples are 370. Since it is very difficult to collect a good representative data for the non-faces, 1500 more non-face examples were added using the "bootstrap" method that was adapted from[5].

2.2 Face locator size reduction This size of the face locator is too large (YIQ 16041, RGB 36061 weights) and therefore renders it slow during searching. Face search is done by the application of the face locator at every pixel in a visual scene. Therefore, the larger the network, the longer it takes to search a given area. Hence, a small sized network is preferable. We investigated the weight distribution after training the face locator and found out that the absolute value of the weights representing the eyes, nose and lips areas had significantly larger values than the rest of the areas. This meant that the face locator depended more on the weights in these regions more than other areas of the face.

It is therefore possible to selectively prune some of the weights from the less important areas from the network without affecting the accuracy of the face locator. Structural learning with knowledge is used to do this. The weight change equation of the back propagation method can be modified to include a forgetting term ε, as shown in equation (1) below. This is an adaptation of the structural learning method [6].

$$\Delta\omega_{ji}(n+1)=\eta(\delta_j o_i)+\alpha\Delta\omega_{ji}(n)-\varepsilon\,\mathrm{sgn}(\omega_{ji}) \qquad (1)$$

where: $\Delta\omega_{ji}(n+1)$ and $\Delta\omega_{ji}(n)$ are the weight change in the (n+1)th and (n)th steps respectively, η is the learning rate, o_i is the nodes output, δ_j is rate of change of error with weights, $(\delta_j o_i)$ represent the current weight change, α is the momentum rate[4] and $\mathrm{sgn}(\omega_{ji})$ is the value of the weight before with the appropriate sign (plus or minus). The forgetting term reduces a weight by ε per training cycle.

Our idea of pruning our network is to let the absolute value of weights in regions A in Fig. 3, grow and trim the weights in other areas as the training progress.

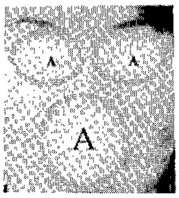


Fig. 3. Important areas (A, low ε), Rest of areas are not vital(high ε)}

The face locator is first trained up to an acceptable error level with forgetting factors of zero. Then two forgetting factors are selected, ε_A for areas A in Fig. 3 and ε_0 for the rest of the areas, and added or subtracted to the weights during back propagation, depending on if the current weight is negative or positive. Weights that are less than the forgetting factor ε, are then removed. Since the face locator depends more on the areas around the eyes, nose and lips, the forgetting factors are therefore, set such that starting with ε_A =0.01, the value of ε_0 is 0.1 larger than that of ε_A. That is if ε_A=0.01, then ε_0=0.11. This means that if a weight in the eye region has the same value with one in e.g. the cheeks region, say 0.02, then the cheeks region weight is removed while the other weight remains. The forgetting factors are then increased and the above process repeated. This process is repeated until the size of the network reduces enough without affecting the overall face detection accuracy. Structural learning drastically reduces the size of this network and improves its accuracy. Care however, should be taken during the pruning of the weights so as not to prune the bias weights regardless of their values or positions. We found out that it was not possible to continue training the neural network without them.

2.3 Testing Threshold Setting During training of the face locator, outputs of 0.05 and 0.95 were used to denote a non-face and face respectively. This means that, when detecting faces using the face locator, outputs of 0.5 and below should represent a non-face, and above 0.5 the presence a face in that position.
However it was found that the criterion above needed to be adjusted so as to improve the ability of the system to detect faces. Starting at a threshold of 0.1 up to 0.9 the system was tested on a set of images and the results were as shown in Fig. 4.

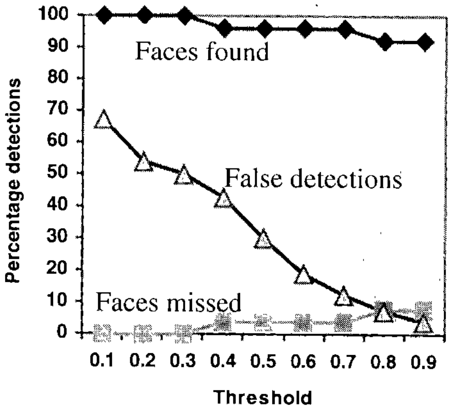


Fig. 4: Threshold results: "Faces found" means percentage of correctly detected, "Faces missed" is for percentage not detected and "False detections" is the percentage of positions detected as faces but actually are not. Faces found plus faces missed equals 100%.

At a threshold off 0.9 the ratio of the correct detections to the missed faces and the false detections was the best. Therefore, 0.9 was adapted as the operating threshold for this system. That is, if the output of a window is 0.9 or above, the window contains a face. Otherwise a non-face is contained.

2.4 Down Sampler Although the default training size of the samples is 20x20 pixels, this system has the ability to detect faces that are larger than this size. This is achieved using the process of down sampling. In this method, the image is down-sampled by a factor decided on depending on the size of the face candidate (section 3.4 below). Face candidates are skin regions whose length and width are bigger than 10 pixels. They also contain skin area over half of their total area. Large faces are thus reduced in size at every down-sampling stage until they are about 20x20 pixels. The face locator can then be used to successfully detect them.

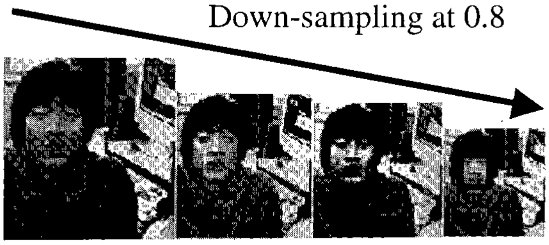


Fig 5: Down-sampling performed at rate of 0.8 until size of image is about 20x20 pixels.

As shown in Fig. 5, the size of the face on the left is larger than the default size and is thus not detected on the first run. Consequent down sampling yields the face location as Fig. 5 shows.

2.5 Merger Around the areas with faces, multiple face detection are likely to occur as shown in Fig. 6 left. To be able to merge these areas, we have used a system in which the centroids (the x and y values of the center of the detected face regions) of these areas are compared with each other consecutively. Simple rules are used to decide which areas qualify for merging.

The first merging rule is: If the centroids of two areas are within 3 pixels distance and 3 lines spacing then merge, otherwise move to next area. To merge, the average values of the x and y values of the two centroids under consideration are found and the result used as the new face position. Results of this first step are shown in Fig. 6, right.

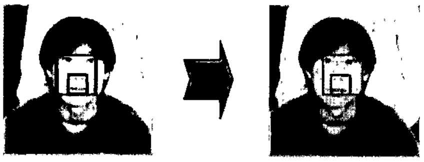


Fig 6: Initial results from the BP, left, Results after first merge process, right.

The second merging rule is: If two unequal areas overlap, choose the larger one. See Fig. 7,left.

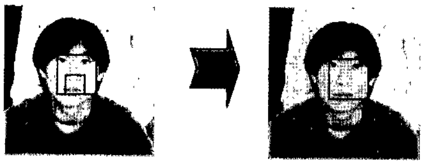


Fig 7: Results after second merge process. The smaller area is ignored, right.

If an overlap still occurs, that is, the spacing between the overlapping detections is more than 3 pixel and 3 lines and less than 20 pixels and 20 lines respectively, as shown in Fig. 8 left, the third merging rule is applied.

The third merging rule is: Choose the detected area that produced a higher output from the face locator and discard the other one.

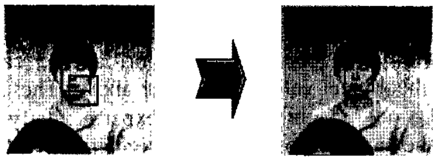


Fig 8: Another kind of overlap, left. Results after third merging process, right.

3. Speed Improvement system (SIS)

The face detection system as described in the sections above, is capable of detecting faces in visual scenes. However, since the FDS is applied pixel by pixel through out the image, it is quite slow and therefore cannot be used in real world situations. To improve the speed at which the system operates, we introduce the use of skin color detectors. We have assumed that human faces can only be found in skin color regions. We at first detect skin color regions at high speeds and then apply the FDS only in the face candidates in the skin color areas found. We considered three methods to detect skin color as explained below. The skin color data used in this section was collected from the face samples used to train the face locator.

3.1 SIS: Skin Color Detector 1: Kohonen Neural Network

This system searches for skin color regions in the visual scene using Kohonen self-organizing neural network [2]. After the training data is collected, the choice of the output grid is then done. The input is one-dimensional skin and non-skin color pixels and the output is a two-dimensional grid. In this project, the grid size chosen was a square of 3x3 nodes. Experiment showed that, this was the smallest size with which separation of the 2 spaces was possible. Note that a larger output grid produces better separation on training but also have more weights and is therefore slower to run. The system is then trained over the chosen output grid. The result of training showed that, minimum at nodes 1,2,4 and 9 will be the result of a skin color pixels. Nodes 5,6,7 and 8 will be non-skin pixels. From the results node 3 was an overlap and is therefore ignored during testing. Fig. 9 shows this result.

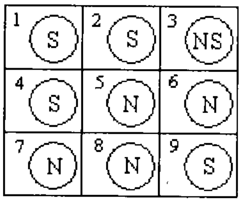


Fig 9. Kohonen network output. S represents skin color nodes, N, non-skin and NS shows an overlap.

3.2 SIS: Skin Color Detector 2: Back Propagation Neural Network

A back propagation trained neural network can be trained to separate the skin color pixels from non-skin pixels. We used a neural network with three layers, that is, input, one hidden layer and output. The number of nodes at the input depends on the color system used. For example, if RGB is used, then 3 inputs are necessary, but if YIQ is used, then we can combine the color components I and Q, and end up with only two inputs. The hidden-layer had five nodes and the output layer one node. The activation function used in the back propagation is the sigmoid [4] with 0 to 1 output characteristics. Teacher signals are skin color and non-skin color pixels with the output set to produce 0.95 for skin color and 0.05 for non-skin color, as we can only approach the theoretical values of 0 and 1 above. These values can produce good recognition results experimentally, which were almost the same for other teacher signals 0.9 to 0.98 for skin color and 0.02 to 0.1 for non-skin color after training. In this paper, however, 0.95 and 0.05 are used as the teacher signals to obtain threshold results using varying threshold as shown in Fig. 4.

3.3 SIS: Skin Color Detector 3: Color Filtering

In color spaces like YIQ, Y,CrCb etc, the brightness Y is separated from the color (10). The color also remains fairy constant as brightness changes. Using the skin data, we plotted a graph of the values of I(Cr) and Q(Cb), converted from RGB (conversion equations can be found in reference (10)), verses the brightness Y. We found out that the values of I, Q, Cr, and Cb fairly lay between two thresholds as shown in the following equations.

$$YCrCb\ \ 98 \le Cr \le 128\ \ and\ \ 133 \le Cb \le 168 \quad (2)$$

$$YIQ:\ \ 21 \le I \le 54\ \ and\ \ 40 \le Q \le 85 \quad (3)$$

3.4 Face Candidates and Down Sampling Rates

Not all the areas identified as skin regions are automatically searched for faces. The reason is that skin color will also be detected on the hands, legs and other open body parts. A decision needs to be made as to which areas qualify for the searching and which areas don't.

Face candidates are skin regions that have width and breath greater than ten pixels and contain skin color pixels more than half their area. The threshold of ten was arrived at by comparison to the minimum face size of 20x20 pixels, and also from experimenting

There are however, some exceptions to this rule. These are skin regions near the boundaries of the face search area. At these areas, may be only part of face is searched for skin color.

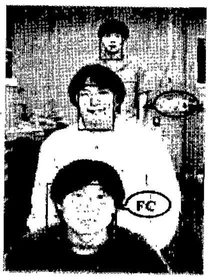


Fig 10: (FC-)Face candidates and (NFC)-Not Face Candidates.

In the Fig. 10, the marked areas show the skin color regions detected by the skin color detector. Of those, only areas larger than 10 pixels in length and width, and with skin area more than half their areas are selected as shown by the marked areas in Fig. 10.

The selection of the face candidates also serves as a means of reducing the time taken by the face detection system through the down sampler. This is because the size of the face candidate is usually an indication of how big the face in that region is. By comparing this size with the default training size of 20x20 pixels, an approximate down sampler rate can be found. For example, if the face candidate size were 45x45 pixels, then the down sampling rate will be about 20/45=0.44. This can then be rounded off to 0.5.

Another advantage of the face candidate is that, during down sampling, the whole image need not be re-sampled. The re-sampling rate found is used to down-sample only that face candidate area. This saves some valuable time taken by the unnecessary re-samplings and researching the whole image.

4. Experimental Results

Our test set contains 13 images with a total of 111 faces. The faces in the set are of various lighting, sizes, and from people of Asian, African and Caucasian origins. The faces, however, are all frontal, without glasses or occlusion. The total face search windows are 493,043. A window is here defined as a 20x20 region in an image.

The skin color detector (SCD) is first applied to the image to determine skin color regions and hence the face candidates. Once found, the face locator (FL) is then used to decide whether or not these regions contain a face.

4.1 Test 1: The Face Detection System (FDS) The face locator was trained using five different color spaces in an effort to find out which one produced the best result. The color systems used were the YIQ, YUV, YCrCb, HSI and RGB. Details about these color systems are available in$^{(10)}$. Table 1 shows the results of the face detection system.

In Table 1, FL size is the initial size, in weights, of the face locator, SR ε, is the final value of the forgetting term, NS/SL is the new size (no. of weights) of the face locator after structural learning, FST is time taken to search for faces in the test set in seconds, and ACC is the percentage of faces found.

Table 1. Face detection system with structural learning.

System	FDS				
	FL Size	SR ε	NS/ SL	FST (secs)	ACC %
YUV	16041	0.9	86	10.07	97.3
YIQ	16041	0.8	94	10.09	96.4
YCrCb	16041	0.8	78	10.05	97.3
HSI	16041	0.9	116	10.2	92.8
RGB	36061	1.3	143	10.4	96.4

False detections (defined as a position whose face locator output is more than the threshold but which does not contain a face) found were acceptably low, in the range of 10 in 493,043 search windows.

4.2 Test 5: FDS with SL (YCrCb) and all Skin Color Detection Systems In Table 2, (Sch time) is the time taken by the skin color detector to search through the test set, (Wins after) is the total face candidate area reduced from the original 493,043 windows. FST and ACC are as in Table 1.

Table 2. Comparison of the Skin color detectors

Skin Detector	Sch Time	Wins After	FST (secs)	ACC %
Kohonen	0.71	121,333	1.02	96.2
Back Propagation	0.55	103,910	1.05	97.3
Color Filtering	0.21	160,321	1.6	97.3

Note that the Color filtering method is the fastest of the skin color detectors. However, it results in the highest number of face candidates and hence when the face detection system is applied, it becomes the slowest at 1.6 seconds.

4.3 Discussion The skin color detection tremendously improves the speed of this system. However, the accuracy of the system could reduce if the face candidate area includes non-skin pixels. These pixels increases the size of the face candidates and hence the re-sampling rate of the down sampler, could be incorrect. The merger steps 1 and 2 work very well but step 3 fails about 10 percent of the times. Even though amongst the skin color detectors, the color filter is the fastest, we however think that the back propagation trained detector is more reliable. The effect of illumination was not investigated but our system worked well with images taken in daylight and florescent lamps illuminating conditions.

Although the training of the face locator with structural learning required repeated retraining and was therefore time consuming, the overall improvement in speed and accuracy was worth the effort. The high threshold of 0.9 adapted for face locator helps reduce the number of false detections. However, in low brightness images some faces are missed and the overall accuracy falls. This could be solved by lighting correction or use of more face samples to train the face locator.

4.4 Other Results The images below, selected from the test image set, show the pictorial results of our system. In Fig. 11-15, (a/b/c/d) represents Faces in image/Faces found/Faces missed/False detects

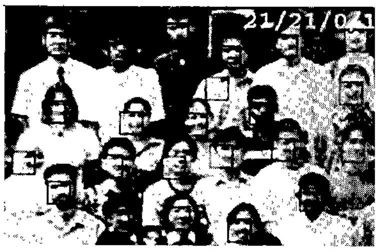


Fig. 11. (21/21/0/1).

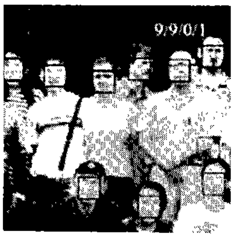


Fig. 12. (9/9/0/1).

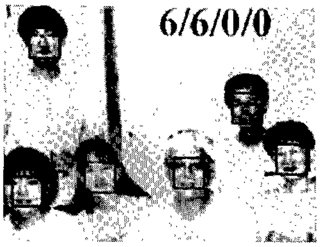


Fig. 13. (6/6/0/0).

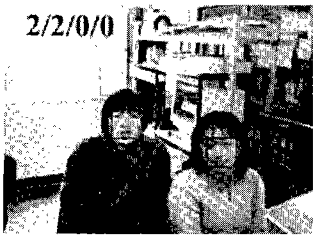


Fig. 14. (2/2/0/0).

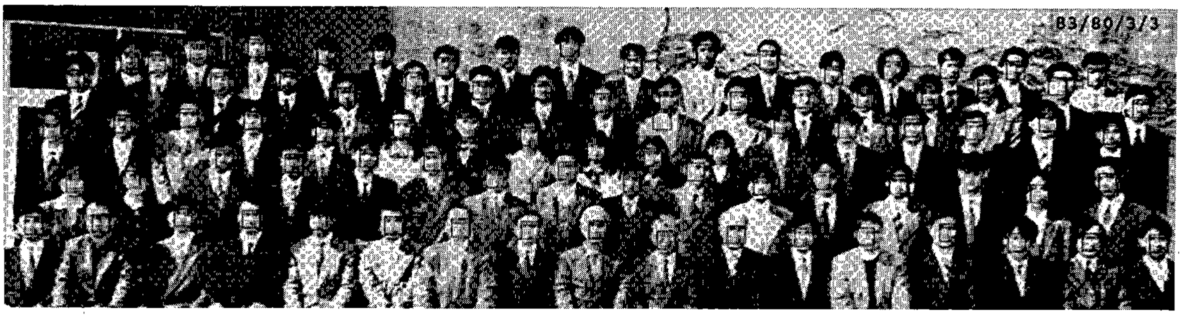


Fig. 15. (83/80/3/3).

5. Summary and Future Research

An accuracy of 97.3% at a speed of 1.6 seconds per about 500,000 windows was achieved for the chosen test set using the YCrCb color system and the back propagation trained skin color detector. Loading the images to memory took about 2.0 seconds. The false detections rate is 10 per 500,000 windows. The project was carried out on a PENTIUM III 750MHz personal computer.

The face locator neural network and the skin color detectors were all implemented using all the color systems mentioned. Of these color systems, YCrCb proved to be the best. The major problem faced was the acquisition of the face and non-face examples. More work needs to be done in areas like training samples pre-processing, including brightness, and illumination normalization, adaptive color filtering method so as to make use of its speed and upgrading the system to detect faces in all possible orientations (Note that skin color is orientation invariant).

(Manuscript received April 12, 2001, revised Jan. 4, 2002)

References

(1) S. Karungaru, M. Fukumi and N. Akamatsu, "Improved speed for human face detection in visual scenes using neural networks," *Proc. of ICONIP* 2000, FBP-30, pp 1-6. 2000.

(2) Teuvo Kohonen, "Self-Organizing and associative memory," *Springer-Verlag*, pp 125-160.1984.

(3) Rowley, Baluja and Kanade, " Human face detection in visual scenes," CMU-CS-95-158R,*Carnegie Mellon University*,1995.

(4) Yoh-Han Pao, "Adaptive Pattern Recognition and Neural networks,"*Addison-Wesle* , no.12, pp.113-139, 1989.

(5) Kah-Kay Sung, "Learning and example selection for object and pattern recognition," *PhD Thesis, MIT AI Lab* , 1996.

(6) M. Ishikawa, "Structure learning with forgetting," *neural networks*, pp 509-521, 1993.

(7) Rumelhart D.E, G. E Hinton and R. J Williams, "Learning internal representations by error propagation in PDP," *The MIT press*, pp 318-362, 1986.

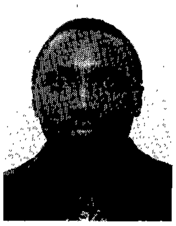
(8) A. Lanitis, C. J. Taylor and T. F. Cootes, "Human face detection in complex background," *Pattern recognation*, vol. 27, no. 1, pp.53-63,1994.

(9) G. Yang and T. Huang, "An automatic face identification system using flexible appearance models," *image and vision computing*, vol. 13, no. 5, pp.393-401,1995.

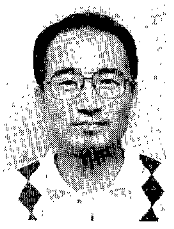
(10) Charles Poynton, "Frequently Asked Questions about Color," *Published electronically, http://www.inforamp.net*, 1999.

(11) S. Akamatsu, "Computer Recognition of Human Faces-A Survey" Trans. of IEEE, Vol.J80-DII, No.8, pp 2031-2046,1997, in Japanese.

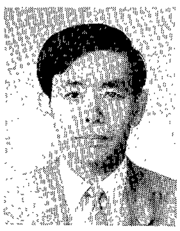
(12) M. Yang, D. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey", to appear in *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2001.



Stephen Karungaru (Student Member) Stephen Karungaru received the BSc degree from Moi University in Kenya in 1992. He then joined the Department of Electrical and Electronics Engineering, Jomo Kenyatta University of Agriculture and Technology, also in Kenya. He received the ME degree in 2001 from Department of Information science and Intelligent Systems, the University of Tokushima and is currently a PhD student in System design Engineering in the same department. His research interests include neural networks, Image processing and computer networks. He is a student member of the IEEE.



Minoru Fukumi (Member) Minoru Fukumi received the B.E and M.E degrees from the University of Tokushima, in 1984 and 1987,and the doctor degree from Kyoto University in 1996. Since 1987, he has been with the Department of Information science and Intelligent Systems, the University of Tokushima. In 1996, he became an associate professor in the same department. He received the best paper award from the SICE in 1995. His research interests include neural networks, evolutionary algorithms and image processing. He is a member of the IEEE, SICE, ISCIE and IEICE.



Norio Akamatsu (Non-member) Norio Akamatsu received the BE in Electrical Engineering from the University of Tokushima in 1966 and the M.S. and PhD degrees in Electrical Engineering from Kyoto University in 1968 and 1974 respectively. From 1968 to 1974 he was an associate lecturer of Electrical Engineering at the University of Tokushima. Currently, he is Professor at the Department of Information science and Intelligent Systems, the University of Tokushima. His research interests are neural networks and non-linear circuits.