Facial Feature Detection and Head Orientation based Gaze Tracking

Abstract: - In this paper, we propose a robust, fast and economical scheme for locating the eyes, lip-corners, and nostrils. A head pose estimation scheme for eye-head controlled human computer interface under an unconstrained background is included. The method we propose uses geometric similarity matching from thresholded images. and the two objects that have the greatest similarity are selected as eyes, also, we located mouth and nostrils in turn using eye location information and size information. Two kindes of gaze tracking methods are supposed in this paper. There are template matching method and neural network approach. It has been tested on several sequential facial images with different illuminating conditions and varied head poses, It returned quite a satisfactory performance in both speed and accuracy.

Key-Words: - HCI, face detection, gaze tracking, head pose, neural network

1 Introduction

Many applications in Human-Computer Interaction require tracking a human face. The area has witnessed intense research activity in recent times. Image containing faces are essential to intelligent vision-based human comuter interaction.[1] Skin-color model-based method [2], Template Matching Method [3], Feature-based approach [4], and Neural network approach[6] provide some examples.

Face tracking is employed in many kinds of application in computer vision. In this thesis, we need face tracking for the gaze tracking system. At first, facial features such as eye, mouth, and nostrils are located in order to find gaze direction. We assume that there is only one user in front of the computer. We used the darkness of features, complete graph matching, and geometrical information for locating facial features. As an example of geometrical information, it may be pointed out that human faces are configured in the same manner. Even though the facial features depend upon the individual, race, sex, and age, human beings have a consistent distance between facial features By tracking the direction of a person's gaze, the band-width of communication between the user and the computer can be increased by using the information about what the person is looking at. And by employing the gaze information in the user interface, communication can be effected in a more convenient and user friendly manner than in a possible with the standard devices. Gaze-tracking system is used in a wide array of applications: Computer Interface, Virtual Reality and Games, Robot Control, Disabled Aid, Behavioral Psychology, Teaching and Presentation and so on. In this regard, considerable work has been done to develop gaze-tracking techniques and build human-computer interface using them in many different applications.

In this paper, we propose a facial features tracking and head pose estimation schemes in order to do construct a novel image-based human computer interface controlled by eye and head, which is a subtask of a multimodal and intelligent interface of a car navigation system.

This paper consists of following. A brief description of related work is contained in Section 2. Section 3 and Section 4 describe the proposed method of locating the facial features and of head pose estimation respectively. Experimental results are provided in Section 5. The paper concludes with Section 6.

2 Related Works

Due attention is being paid by the research community to face detection schemes, several kinds of approach such as template matching method, feature-based approach, color-based method, neural network method, and motion-based method to locate facial features have been proposed in this regard.

Template matching method that was introduced by Yuille D.S. uses deformable templates. This method is independent of size, slope, and illumination. But, at first, it requires knowledge of initial template of face.

And feature-based approach searches the image for a set of facial features and groups them into face candidates based on their geometrical relationship. Yow and Cipolla[4], Leung et.al. and Sumi and Ohta[6] employed this approach. And the color-based detection system[2] selects pixels that have similarity to skin color, and subsequently defines a subregion as a face if it contains a large of skin color pixels. [5] shows fast and effective face detection method using integral image.

A person's gaze direction is determined by two factors: the orientation of the head and the orientation of the eyes[7]. Techniques such as limbus tracking, pupil tracking, corneal and pupil reflection relationship use light (mainly infrared light) reflected by the eye (on the cornea or further in the eye). They concentrate on the orientation of the eye. For example, when the eye is panned horizontally or vertically, the relative positioning of the brightest spot by reflection, the glint and the center of the pupil can be used for calculating the direction of gaze [9]. In [8], they use neural network for gaze tracking. Based on captured grayscale eye images, the system effectively learns the gaze direction of a human user by modeling implicitly corresponding eye appearance-the relative positions of the pupil, cornea, and light reflection inside the eye socket. But, In [7], they have mainly employed head orientation for gaze finding. Then, the gaze estimation can be formulated as pose estimation. In this thesis, we concentrate on the head orientation. And we employed neural network for finding the head orientation.

3 Facial Features Locating

Facial features locating is needed for finding head pose. In this section, we describe the method of locating facial features. We assumed that one user is in front of computer. And we employed the dark information of features, geometric similarity matching and geometrical information for locating facial features. At first, the eyes is located using the dark information and geometrical information of the eyes. And the mouth is located using the information of eyes' position and size information. Finally, the nostrils are located using the information of eyes' and mouth's position and size information.

3.1 Locating the Eyes

3.1.1 Setting the threshold value and thresholding

It is important to find the proper threshold value in order to separate the eyes, nostrils and mouth from face. Among many methods to find the threshold value, we employed a heuristic P-Tile method [10]. After finding the weight center of histogram, the value is subtracted by constant value until the eyes is located for the first time. After finding threshold value, the image should be binarized by that value. An example of a binarized image obtained in this manner is shown in Figure 1.

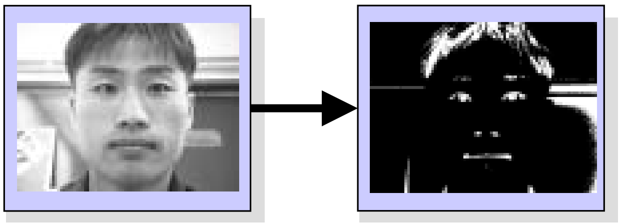


Figure 1. Binarized image

Edge detection may be employed to find the candidates for the eyes. But, it requires amount of computation intensive and it is difficult to find accurate edge pixels. So, thresholding method was employed instead to get the candidate of the eyes.

3.1.2 Finding the candidates of the eyes

After thresholding, we can assign unique a tag to each isolated block by labeling the binarized image. In finding the candidates of the eyes, eliminating the blocks that is not satisfied in condition of being the eyes is much efficient than the finding the proper block which is satisfied in condition of being the eye. So we need standard as follows:

Suppose that the two points [x1,y1],[x2,y2] are the top-left point and bottom-right point of a circumscribed rectangle respectively. Let l(x,y) be the tag of the pixel.

(i) $$\mathrm{Size}(i) = \sum_{x=x1}^{x2} \sum_{y=y1}^{y2} F(l(x, y))$$

$$(\text{if } l(x, y) = i \text{ then } F(i) = 1)$$

$$\mathrm{Min} \le \mathrm{Size}(i) \le \mathrm{Max}$$

(ii) Ratio = Max_Vertical / Max_Horizantal

$$\mathrm{Ratio} \le 1$$

If the block does not satisfy the conditions (i) and (ii) , then the block is eliminated from the candidate set. Condition (i) implies that the size of eye's block is between Max and Min value. Here, we define the Max and Min as 30 and 300 in size respectively by experimental results. By eliminating the blocks using the rough and simple size information, we could reduce the number of candidate blocks to a quarter. Condition (ii) means that the aspect ratio of the eye is less than 1.

3.1.3 Looking for similarity by geometric similarity matching

After eliminating the unsatisfactory blocks, a complete graph is composed with the candidate blocks and similarity for each pair is computed. Similarity is computed as follows:

$$1)\text{Normal\_size}(i, j) = \text{Size}(i)/\text{Size}(j)$$

$$2)\text{Normal\_Average}(i, j) = \frac{\text{Average\_gray}(i)}{\text{Average\_gray}(j)}$$

$$3)\text{Normal\_Aspect\_ratio}(i, j) = \frac{\text{A.R}(i)}{\text{A.R}(j)}$$

$$4)\text{Normal\_Angle}(i, j) = 1 - [y_{\text{distance}}/x_{\text{distance}}]$$

Normal_size(i,j) refers to similarity of two blocks in size while Normal_Average(i,j) and Normal_Aspect_ratio(i,j) refer to similarity of average gray value and aspect ratio between the blocks respectively. The small value is divided by larger value for normalization. Normal_Angle means the slope over x-axis. The pair of blocks that have the maximum sum of the above four factors are selected as the two eyes. Figure 2 shows the result of locating the two eyes.

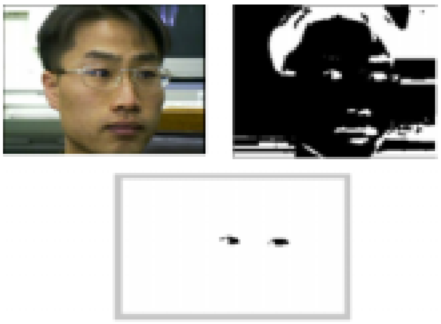


Figure 2. Locating the two eyes

3.2 Locating the Mouth and Lip-Corners

After locating eyes, We can define a rough region for the mouth by using the eye information. Figure 3 shows an example. We consider the largest blocks to be mouth in that region. After locating the mouth's block , we can find the lip-corners by scanning the first and last columns like Figure 4. If the face is tilt to the left like Figure 4, the first column is scanned from bottom to top and the last column is scanned from top to bottom. If the face is tilt to the right, it is scanned in opposition.

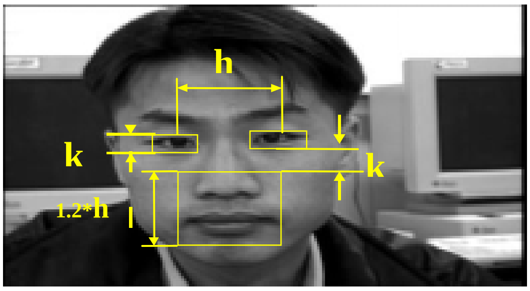


Figure 3. Defining the mouth region

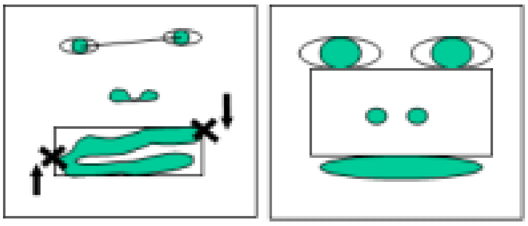


Figure4. Locating the lip-corners (left), Define the region for nostrils(right)

3.3 Locating Nostrils

We can define the region for nostrils using the two eyes and the mouth position information as in Figure 6. Like locating mouth, we used size information of blocks in defined region for locating nostrils. But we should examine whether the nostrils in the image are appeared in one block or not.

3.4 Verification

After locating the facial features such as the eyes, lip-corners and nostrils, we should check whether the facial features have been located correctly using the geometrical information. For example, we can prevent the eyebrow from being selected as the eyes using the information that there are eye blocks under the eyebrows.

4 Head Orientation based Gaze Tracking

A person's gaze is determined by two factors: the orientation of the head, and the orientation of the eyes. While the orientation of the head determines the overall direction of the gaze, the orientation of the eyes determines the exact gaze direction and is limited by the head orientation. In this paper, we concentrate on the orientation of the head. That is, we find a rough direction of gaze using only the orientation of head. For gaze tracking, Template Matching Method and Neural Network are employed.

4.1 Gaze Tracking using Template Matching Method

We employed template matching using the angles of pairs of facial features for head pose estimation. Each template consists of 6 angles like following Figure 5.

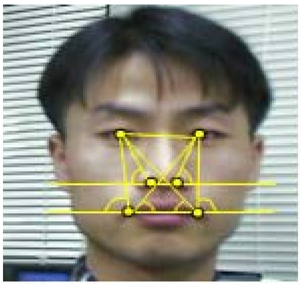


Figure 5. The angles of templates

We do not need consider the distance from user to camera because the angles are independent of the distance. We can create database of 11 templates from one person representing different poses. And each template indicates different gaze points. That is, we divided the computer screen into 11 blocks like Figure 6. We have made database of 55 templates from five persons.

Left_up		Up	Right_up	
Left_2	Left_1	Front	Right_1	Right_2
Left_down		Down	Right_down	

Figure 6. Screen Monitor Resolutions

We compare the angles of input image with those of each template for finding gaze point like following evaluation function:

$$E\_F(i) = \sqrt{\sum_{x=1}^{6} (T\_a(x) - I\_a(x))^2}$$

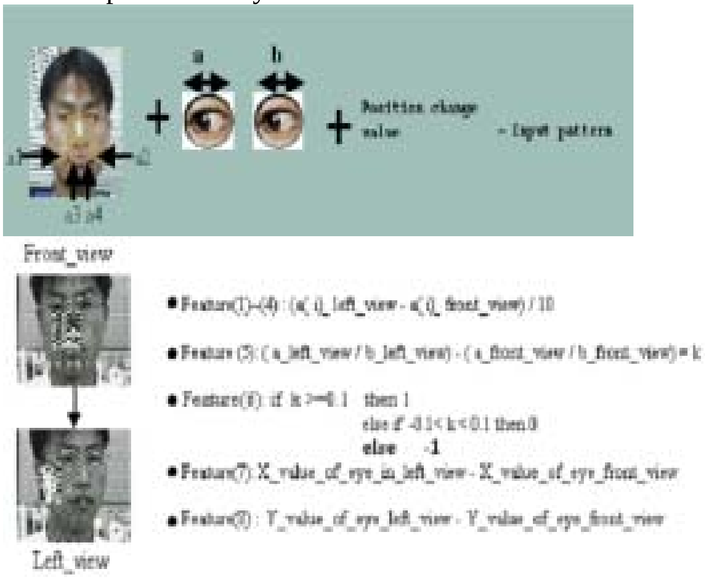
Where,
T_a(x) : xth angle of the template image
I_a(x) : xth angle of the input image
E_F(i) : evaluation value for the i th template

The template that has the minimum value among evaluation function values is selected as gaze point.

4.2 Gaze Tracking using Neural Network

We have employed the neural network for better results in finding the gaze block. In this section, we describe the learning phase of the neural network in the off-line system. In our neural network system, there are three layers: the input layer, the hidden layer, and the output layer. There are 8 nodes in the input layer, and 15 nodes in the hidden layer and 10 in the output layer. The Activation function is “tansig”.

Each node of the input layer receives the input pattern (Figure 7). For the features of the input pattern, we have used the angles between two facial features and the x axis , and the ratio of the width of left eye to that of right eye , and change value of position of eye’s center. Feature(1)~(4) consist of four angles (divided by 10 for normalization). Feature (5) means change value of the length ratio between length of left eye and length of right eye. Feature (6) is useful for distinguishing whether the head is moving to the left or the right. Features (7) and (8) refer to the change value of position of eyes’ center.



a
b
a1
a2
a3 a4
Position change value
= Input pattern
Front_view
Left_view
Feature(1)~(4) : (a(i)_left_view - a(i)_front_view) / 10
Feature (5) : (a_left_view / b_left_view) - (a_front_view / b_front_view) = k
Feature(6) : if k >=0.1 then 1
else if -0.1< k< 0.1 then 0
else -1
Feature(7): X_value_of_eye_in_left_view - X_value_of_eye_front_view
Feature(8) : Y_value_of_eye_left_view - Y_value_of_eye_front_view

Figure 7 The Composition of Input Pattern

In the output layer, there are 10 nodes for each gaze direction. The screen monitor is divided into 9 blocks: (3*3 output resolution). Each block of the screen is matched to each node of the output layer as in Figure 8. The tenth node in output layer means not moving situation. We used Mean Square Error for error function.

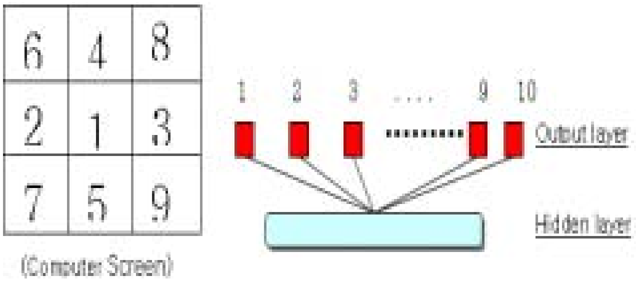


Figure 8 The Composition of Output Layer

5 Experimental Results

Experiments were conducted on a single-processor, 300MHz Pentium PC equipped with CCD camera and Coreco Ultra II frame grabber. Experimental results show that we can locate and track the eyes, the nostrils, and lip-corners in images with different resolutions and different illuminations in real-time(15+ Hz) as soon as the face appears in the field of the view of the camera. The accuracy is above 95% without any identifying mark on the user's face. We have also tested person wearing the glasses or not wearing the glasses. In the case of the subject' black glasses, unsatisfactory results are returned. And if the people have a mustache, then we can not locate the mouth exactly. Some experimental results are shown Figure 9 and you can see the result of gaze tracking using template matching in Figure 10. And the rectangles mean the gaze points.

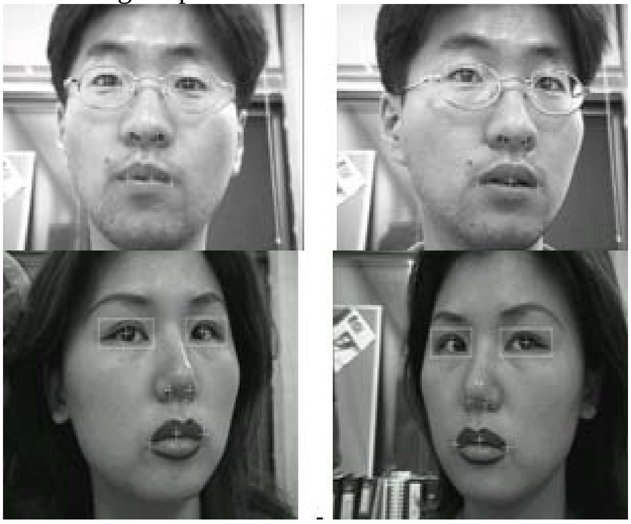


Figure 9. Results of facial feature locating

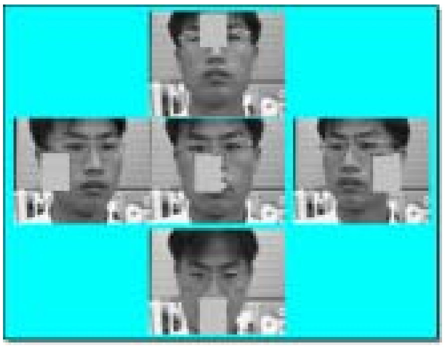


Figure 10. Results of gaze tracking using template matching

And we show the results with 3 by 3 output resolution using neural network. That is, the screen monitor is divided into 3 by 3 blocks. In the gaze tracking system, we have acquired input patterns from 5 persons. We can get 40 patterns from each person. And we should also consider the static situation. Thus, the total number of input patterns is 40 * 5 + 1 = 201. And in the off-line system, the neural network learned these input patterns. Figure 11-(up) shows the result of decreasing of error value

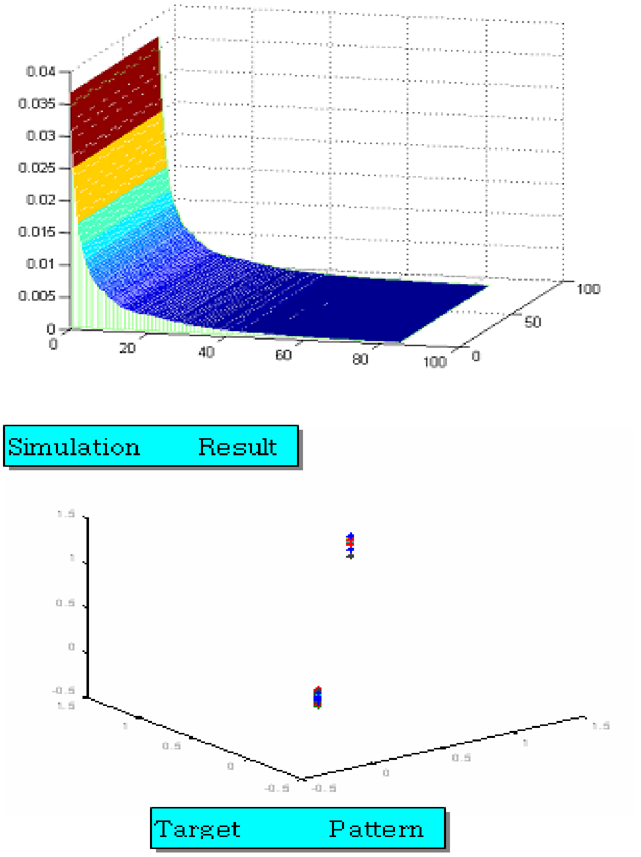
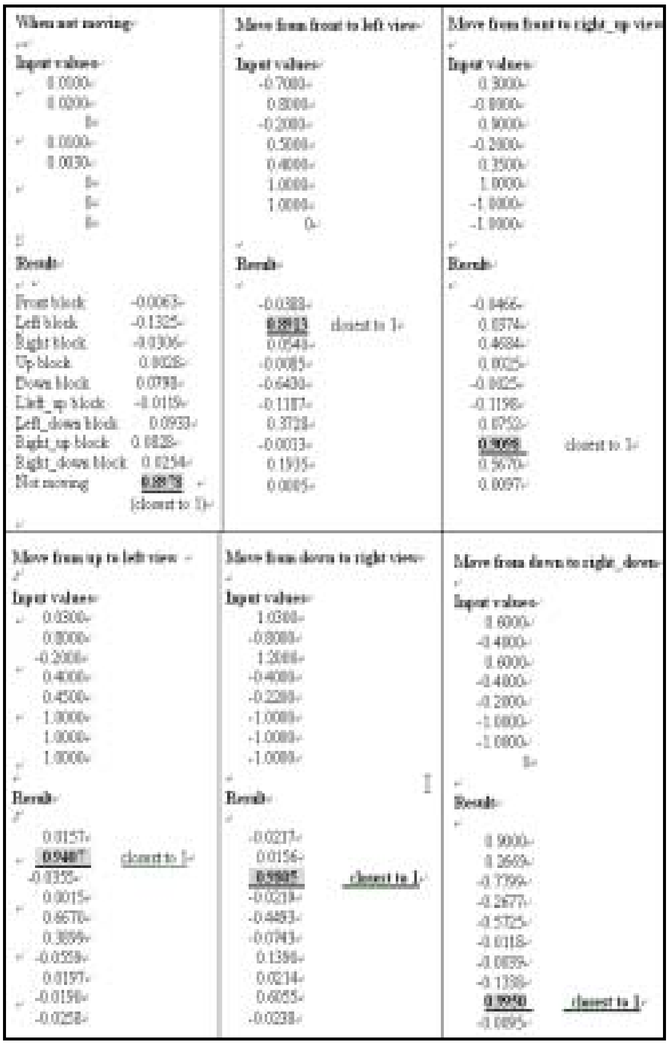


Figure 11. Results of decreasing error value and neural network learning result

Figure 11(down) shows the result obtained with input patterns after the learning process. The result should be close to zero or one. If all the results are close to zero or one, then we can conclude that the neural network has learned the patterns well. Figure 12 shows the results of neural network with new input images. Among the 10 nodes in the output layer, the nodes having the value closest to one is selected as the gaze direction.



When not moving

Input values
0.0000
0.0200
0
0.0000
0.0030
0
0
0

Result
Front block -0.0063
Left block -0.1325
Right block -0.0306
Up block 0.0028
Down block 0.0798
Left_up block -0.0119
Left_down block 0.0933
Right_up block 0.0828
Right_down block 0.0254
Not moving 0.8878
(closest to 1)

Move from front to left view

Input values
-0.7000
0.8000
-0.2000
0.5000
0.4000
1.0000
1.0000
0

Result
-0.0383
0.8813 closest to 1
0.0548
-0.0085
-0.6430
-0.1187
0.3728
-0.0013
0.1935
0.0085

Move from front to right_up view

Input values
0.3000
-0.8000
0.9000
-0.2000
0.3500
1.0000
-1.0000
-1.0000

Result
-0.0466
0.0374
0.4684
0.0025
-0.0025
-0.1198
0.0752
0.9098 closest to 1
0.5670
0.0097

Move from up to left view

Input values
0.0300
0.8000
-0.2000
0.4000
0.4500
1.0000
1.0000
1.0000

Result
0.0157
0.9407 closest to 1
-0.0355
0.0015
0.6670
0.3899
-0.0559
0.0197
-0.0198
-0.0258

Move from down to right view

Input values
1.0300
-0.8000
1.2000
-0.4000
-0.2200
-1.0000
-1.0000
-1.0000

Result
-0.0217
0.0156
0.9805 closest to 1
-0.0219
-0.4483
-0.0743
0.1396
0.0214
0.6055
-0.0238

Move from down to right_down

Input values
0.6000
-0.4000
0.6000
-0.4000
-0.2000
-1.0000
-1.0000
0

Result
0.9000
0.2669
-0.7799
-0.2677
-0.5725
-0.0118
-0.0059
-0.1338
0.9950 closest to 1
-0.0895

Figure 12. Results of Neural network on new input images

6 Conclusion

Real-time facial feature tracking and a Head Orientation based-Gaze tracking system using neural network has been proposed for Human-Computer Interface. We have located facial features using the darkness information of features along with geometrical information. By binarizing the image, we have separated the facial features from the background. Among the facial features, the eyes are easy to find and less deformable than other facial features. We have, therefore, located the eyes first. The mouth and the nostrils are located thereafter using the information about the eye's location. But, more intelligent gray-level thresholding methods and verification techniques are desirable.

We concentrated on the head orientation for finding the gaze direction. We have employed head orientation for finding a rough gaze direction. We used neural networks for matching the gaze direction and head orientation. The output resolution was 3*3. It is not sufficient for gaze tracking. We should enhance the output resolution in order to obtain more exact gaze direction. In general, the gaze tracking system consists of eye orientation based gaze tracking and head orientation based gaze tracking. We can get more accurate and flexible gaze tracking system by combing the two approaches.

More robust initialization and calibration for different users, different computers and different environments are essential for these techniques to be effectively employed routinely in computer interfaces.

References:

[1] Ming-Hsuan Yang, David. J. K. Narendra A., "Detecting Faces in Images: A Survey", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol.24, No.1, 2002.

[2] S. Mckenna, S.Gong, Y. Rajs, " Modelling Facial Colour and Identity with Gaussian Mixtures," *Pattern Recognition*, vol. 31, no 12, pp. 1883-1892, 1998.

[3] Kyung-Nam Kim "Contributions to Vision-Based Eye-Gaze Tracking," Thesis for Master's Degree, 1999.

[4] K. C. Yow, R. Cipolla, "Feature-based Human Face Detection", Image and Vision Computing, vol. 15, no. 9, pp. 713-735, 1997.

[5] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," Computer Vision and Pattern Recognition 2001", 2001

[6] K. K. Sung and T. Poggio, "Example-based learning for view-based human face detection," Technical Report A.I. Memo 1521, CBLC Paper 112, MIT, 1994.

[7] Rainer Stiefelhagen, "Gaze Tracing for Multimodal Human-Computer Interaction," Diplomarbeit, Universitat Karlsruhe, Sep. 1996

[8] Li-Qun Xu, Dave M. and P. Sheppard. "A Novel Approach to Read-time Non-intrusive Gaze Finding," BMVC98, 1998

[9] A. J. Glenstrup and T. E. Nielsen, "Eye Controlled Media: Present and Future State," Thesis, University of Copenhagen, DIKU, June, 1995

[10] X. Xie, R. Sudhakar and H. Zhuang, "Estimation of eye features from facial images," in Proc. 4th Annu. Conf. Recent Advances Robot., Boca Raton, FL, pp. 73-80, 1991