# Gaze Detection via Self-Organizing Gray-Scale Units

**Jun Kawai**

Submitted to the Computer Science Department in
partial fulfillment of the requirements for the degree of
Bachelor of Arts
with departmental honors at
Boston College
May 1999

Author........................................................................................................
Jun Kawai

Accepted by................................................................................................
Margrit Betke
Assistant Professor
Thesis Supervisor

**Abstract**

*We present a gaze estimation system that detects an eye in a face image and estimates the gaze direction by computing the position of the pupil with respect to the center of the eye. Our system is based on unsupervised learning. It creates a map of self-organized gray-scale image units that collectively describe the eye outline. Our approach is information-conserving.*

# 1 Introduction

In the near future, standard desktop computers will be equipped with cameras that will augment traditional human-computer interfaces such as keyboard and mouse. Cameras pointed at the computer user can capture the user's gaze direction, facial expression, lip movement, head orientation, etc. Being able to analyze and understand such image sequences automatically, reliably, and in real time has been and continues to be the topic of exciting research that is aimed at developing a new human-computer interface.

Our research focuses on the problem of gaze estimation, which has previously been approached by applying neural networks [3, 9], morphable models [8], and other techniques [5]. We developed an unsupervised learning method that is based on Kohonen's self-organizing maps [6, 2]. Self-organizing feature maps have previously been used in computer vision, for example, in image compresssion [1], medical image processing [10], and face recognition [7] applications. Instead of feature maps, we follow the "information-conserving" approach discussed in Ref. [4] and use gray-scale subimages as the building blocks of our self-organized recognition system. They are the "units" that learn to arrange themselves around the eye of a trial image in order to estimate the eye center and pupil position.

# 2 System Overview

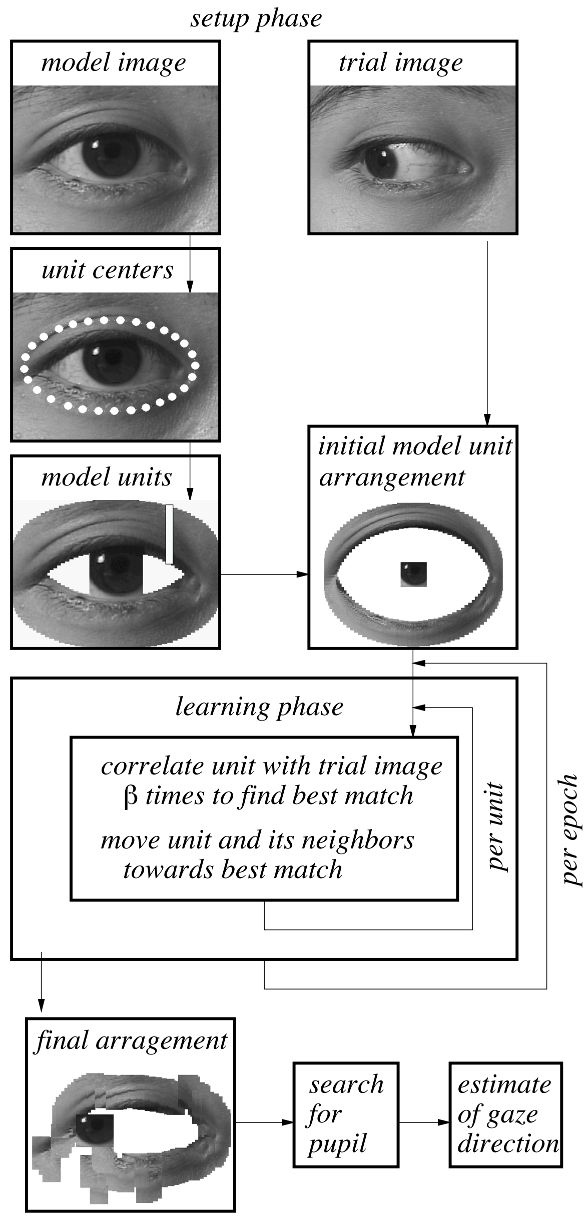


Figure 1: Gaze Recognition System.

Figure 1 gives an overview of the gaze recognition system. Given a model and trial image of an eye as inputs, the system computes an estimate of the user's gaze direction in the trial image.

The system has two phases – an initial setup phase and a learning phase. In the setup phase, the system uses the model image to create and arrange gray-scale subimages, or units, in an elliptic pattern. The units are then correlated with the trial image at locations that are determined in the learning phase. The learning phase consists of a number of epochs. In each epoch, the units move towards the trial eye. Each unit and its neighborhood learn their best positions and organize themselves in a final arrangement. The center of the final arrangement is an estimate of the position of the eye center in the trial image. The best-correlating pupil position in the trial image is then determined. The location of the pupil center with respect to the eye center is used as an estimate of the gaze direction and is the system output.

# 3 Setup Phase

In the setup phase, learning units are created from a model image of the eye and arranged around the eye of a trial image.

### 3.1 Initial Arrangement of Units in Model Image

For each subject, a model image $m(x, y)$ of one of the subject's eyes is created in the setup phase. We ask the subject to look straight into the camera, so that the eye is imaged with the pupil in its the center. We then determine the coordinates $(p_{xm}, p_{ym})$ of the pupil center $\mathbf{p}$, and the parameters $a_m$ and $b_m$ that describe an ellipse

$$\mathbf{r} = (a_m \cos\theta + p_{xm})\hat{\mathbf{x}} + (b_m \sin\theta + p_{ym})\hat{\mathbf{y}} \quad (1)$$

fitted around eye, where $\hat{\mathbf{x}} = (1, 0)$, $\hat{\mathbf{y}} = (0, 1)$ are unit vectors, $\mathbf{r}$ is the vector between pupil center and any point on the ellipse and $\theta$ is the angle between $\mathbf{r}$ and the $x$-axis.

The image regions at the outline of the ellipse are then organized into $n$ units. The unit centers are placed along the outline of the ellipse at equally spaced intervals. The image regions that surround these centers are used as the model units. The image region containing the pupil is used as a gray-scale pupil model.

Figure 1 shows a model image, an image, where every other unit center is shown as a white dot, and an image that contains the pupil model and arrangement of model units, which overlap each other. One of the units is shown in white.

### 3.2 Initial Arrangement of Units in Trial Image

The model units are rearranged to form a larger ellipse, so that an overlay of this new arragement onto the trial image would surround the eye in the trial image, as shown in Fig. 1. The trial image dimensions are used to determine how much the model units are spread out in the initial arrangement. In particular, the ellipse parameters $a$ and $b$ of the new arragement are chosen to be 40% of the width and height of the trial image, respectively.

# 4 Learning Phase

Since the system does not know the position or size of the eye in the trial image, we expect the initial arragement of model units to poorly describe the eye in the trial image. A learning phase therefore follows, in which the units organize themselves and move into positions that better describe the eye in the trial image.

### 4.1 Self-Organization of Units

Within $\xi$ epochs, the model units organize themselves into a final arrangement that describes the eye in the trial image. In each epoch, each of the $n$ model units is chosen as the center of a neighborhood of $\mu$ units that collectively learn better descriptions of the trial image.

The neighborhood centers are selected sequentially in clockwise order starting with unit 0, which is the rightmost unit in the arrangement. Unit $i$ has a neighborhood of units

$(i - \mu/2) \bmod n, \ldots, (i + \mu/2) \bmod n$. The results of the learning process of unit $i$ and its neighboring units are immediately encorporated into the unit arrangement, so that any unit $j$ that is processed after unit $i$ in the same epoch, i.e., $0 \leq i < j \leq n-1$, and its neighborhood make use of the newly learned unit arrangement. Similarly, at the beginning of an epoch, unit 0 and its neighborhood use the unit arrangement obtained in the previous epoch from the learning process of unit $n-1$ and its neighborhood.

In each epoch, the learning process of a center unit $i$ and its neighborhood consists of several steps. First, a line through the center $\mathbf{c}_i$ of unit $i$ and the image center and $\beta$ test points on this line are determined that are equally spaced with distance $\Delta d$ from each other. Out of the $\beta$ test points, 80% are chosen to lie between the center $\mathbf{c}_i$ and the image center. The remaining 20% are taken on the line starting at center $\mathbf{c}_i$ and going outwards, and spaced at the same intervals $\Delta d$.

At each test point $\mathbf{p}$, unit $i$ is then correlated with the underlying subimage $t$ of the trial image, such that the center $\mathbf{c}_i$ is matched with test point $\mathbf{p}$, and the subimage $t$ has the same size as unit $i$. The normalized correlation coefficient

$$r(i,t) = \frac{A \sum i(x,y)t(x,y) - \sum i(x,y) \sum t(x,y)}{\sigma_i \sigma_t} \tag{2}$$

is used, where $A$ is the number of pixels in unit $i$, $\sigma_i = \sqrt{A \sum i(x,y)^2 - (\sum i(x,y))^2}$, and $\sigma_t = \sqrt{A \sum t(x,y)^2 - (\sum t(x,y))^2}$. The test point $\mathbf{p}_{\text{best}}$ with the highest correlation coefficient among the $\beta$ coefficients is determined and its distances $d(\mathbf{p}_{\text{best}}, \mathbf{c}_k)$ to the centers $\mathbf{c}_k$ of all units $k$ that are in the neighborhood of unit $i$ are computed.

The position $\mathbf{c}_k$ of a trial unit is shifted towards $\mathbf{p}_{\text{best}}$ by a fraction $f(G,\eta)$ of distance $d(\mathbf{p}_{\text{best}}, \mathbf{c}_k)$, i.e.,

$$\mathbf{c}_k^{(\text{new})} = \mathbf{c}_k + f(G,\eta)(\mathbf{p}_{\text{best}} - \mathbf{c}_k), \tag{3}$$

where

$$f(G,\eta) = \alpha \, \exp(-\frac{\eta^2}{G^2}) \tag{4}$$

is the *neighborhood kernel*, which is a function of the kernel-width parameter $G$ that is updated in each epoch by $G^{(\text{new})} = (1-\gamma)G$, where $\gamma$ is the decay factor, and $\eta$, the difference of unit indices $i$ and $k$. The *learning-rate* $\alpha$ is also updated in each epoch by $\alpha^{(\text{new})} = (1-\gamma)\alpha$. Note that the center unit $i$ always moves towards best matching test point $\mathbf{p}_{\text{best}}$ by a fraction $f(G,\eta) = 1/\sqrt{2}$ of the distance $d(\mathbf{p}_{\text{best}}, \mathbf{c}_i)$, while the units in $i$'s neighborhood move by smaller fractions.

Figure 2 shows the fraction $f(G,\eta)$ as a function of $\eta$ for a given kernel width $G$. It illustrates that the fraction $f(G,\eta)$ is large for a small $\eta$, i.e, if index $k$ of a neighbor unit is close to $i$, and small for a large $\eta$, i.e, for neigbors further away. Closer neighbors are stronger influenced by unit $i$'s move than distant neighbors. The size of parameter $G$ determines how fast fraction $f(G,\eta)$ falls off to zero when $\eta$ increases.

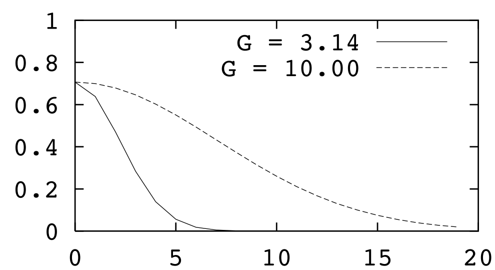


Figure 2: Function $f(G,\eta)$, as defined in Eq. 4, shown as a function of the index difference $\eta$ of neighoring units, and for a neighborhood size of $\mu = 20$ and kernel-width parameters $G = 3.14$ and $G = 10$.

The parameters $n, \xi, \mu, \beta, \alpha, G, \gamma$ of the learning process are carefully chosen such that after $\xi$ epochs the units have converged into a final arrangement that describes the eye in the trial image well. Our measure of success is the quality of the eye center estimate that we can obtain from this final arrangement.

### 4.2 Eye Center Estimation

To estimate the eye center in the trial image, the units are paired by indices, so that unit $i$ is paired with unit $(i + n/2) \bmod n)$, for $0 \leq i \leq n/2$. The pairs will lie approximately opposite to each other in the final unit arrangement. So averaging the midpoints between the centers of all unit pairs results in an estimate of the eye center in the trial image. Note that the number $n$ of units is large, so inconsistencies in the learned unit arrangement due to a small number of unit pairs do not have a notably adversary effect on the estimate.

### 4.3 Pupil Estimation

The pupil model obtained from the model image is used to find the pupil position in the trial image. The pupil is compared to various regions of the trial image that are surrounded by the final unit arrangement using the normalized correlation coefficient, as defined in Eq. 2).

Since the pupil model is taken when the subject looks straight into the camera, the pupil appears smaller in a trial image that captures the subject looking to the left or right. The model pupil may therefore not correlate highly with the trial pupil. The model pupil is therefore transformed into templates of various sizes that are then correlated with the trial image [4]. The template choice depends on the distance of the test position to the eye center in the trial image. For example, if the subject looks all the way to its left, we found that the best matching pupil template is a transformation of the pupil model that is subsampled in its width to 2/3 of the original width of the pupil model.

## 5 Experimental Results

We tested our system on a 450 MHz Pentium II PC running Linux. Our database contains a total of 5200 eye images. It includes 13 Asian and Caucasian, male and female subjects. Each person was asked to look straight into the camera so that a model image could be taken. Then the subjects were asked to change their gaze direction. Three different lighting directions were tested. The eyes and pupils are imaged at various sizes in the trial images. Table 1 shows the values for the learning parameters chosen in our experiments.

Table 1: Initial Parameter Values

| Parameter | Initial Value |
|---|---|
| number of units $n$ | 100 |
| unit width/height | 0.5 |
| number of epochs $\xi$ | 10 |
| number of tests $\beta$ | 60 |
| neighborhood size $\mu$ | 10 |
| kernel width $G$ | 10 |
| learning rate $\alpha$ | $1/\sqrt{2}$ |
| decay $\gamma$ | 0.25 |

Figure 3 shows the model images of our 13 subjects, one trial image per subject, and the corresponding learned unit arrangements. The trial images displayed are chosen to illustrate the variety of images in our database. It includes images of left and right eyes, blinking eyes, and eyes that are looking into various directions.

We also tested our system model and trial images of different people. Our preliminary results indicate that the gaze direction can be detected, as long as the eyes of the people involved look similar. So we may not need to create a model image of each user, but can

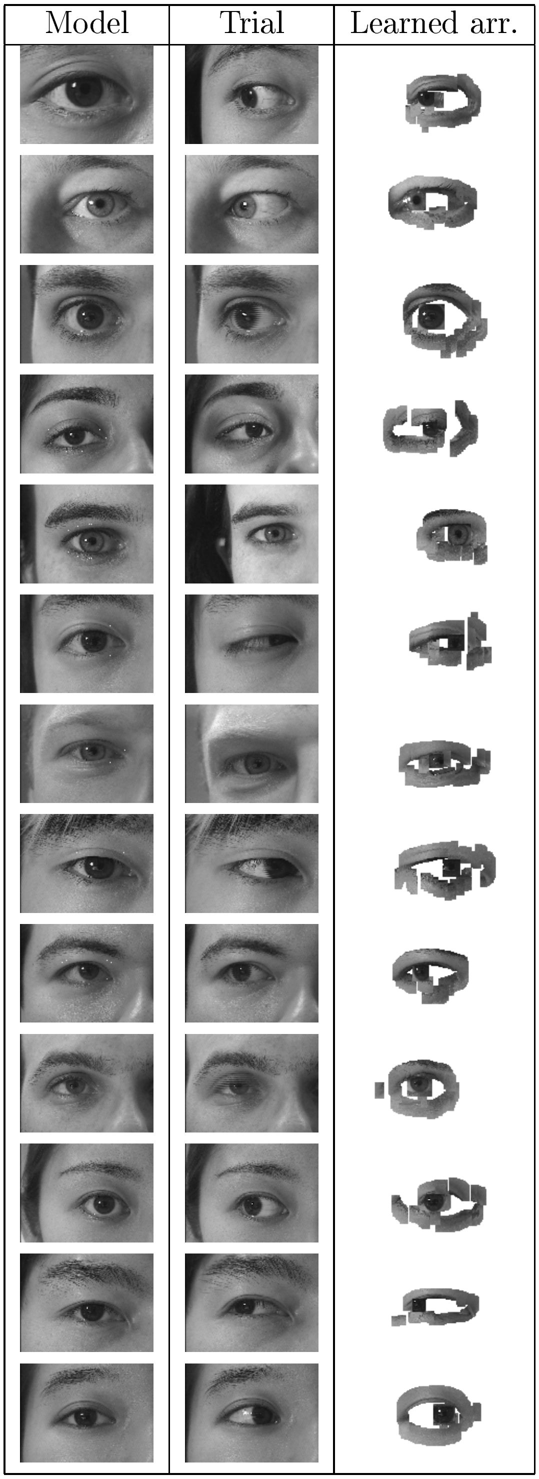


Figure 3: Learning results: Examples of model and trial images and their corresponding learned arrangement of the gray-scale units.

offer a set of images from which the user can chose a similar looking eye.

We found that the location of the pupil is easier to recognize for eyes with brown irises than for eyes with blue, grey or green irises, because the normalized correlation coefficient is invariant to the uniform variations in shading that may appear of brown irises, but not to nonuniform scale changes that occur in lighter irises.

Our system can be used to track the center of an eye and the position of a pupil with respect to this center in videos. Figure 4 shows the results of tracking the pupil's distance from the eye center over time. The sequence includes a few frames during which the subject blinked while moving his eye.

## 6 Summary and Conclusions

We have developed an information-conserving, unsupervised learning approach to the problem of estimating the gaze direction of a person. Our method has been successfully applied to estimate the gaze directions of male and female, Asian and Caucasian subjects. The center of a subject's eye is determined by averaging the eye center estimates derived from gray-scale model units that learn to arrange themselves around the eye. Even if some units do not match the test image, the large number of units make the eye center estimate reliable.

## References

[1] C. Amerijckx, M. Verleysen, P. Thissen, J.-D. Legat, "Image Compression by Self-Organizing Kohonen Map," *IEEE Trans. on Neural Networks,* Vol. 9, No. 3, 503–507, 1998.

[2] B. Angéniol, G. de la Crox Vaubois and J. Le Texier, "Self-Organizing Feature Maps and the Travelling Salesman Problem," *Neural Networks*, Vol. 1, pp. 289-293, 1988.

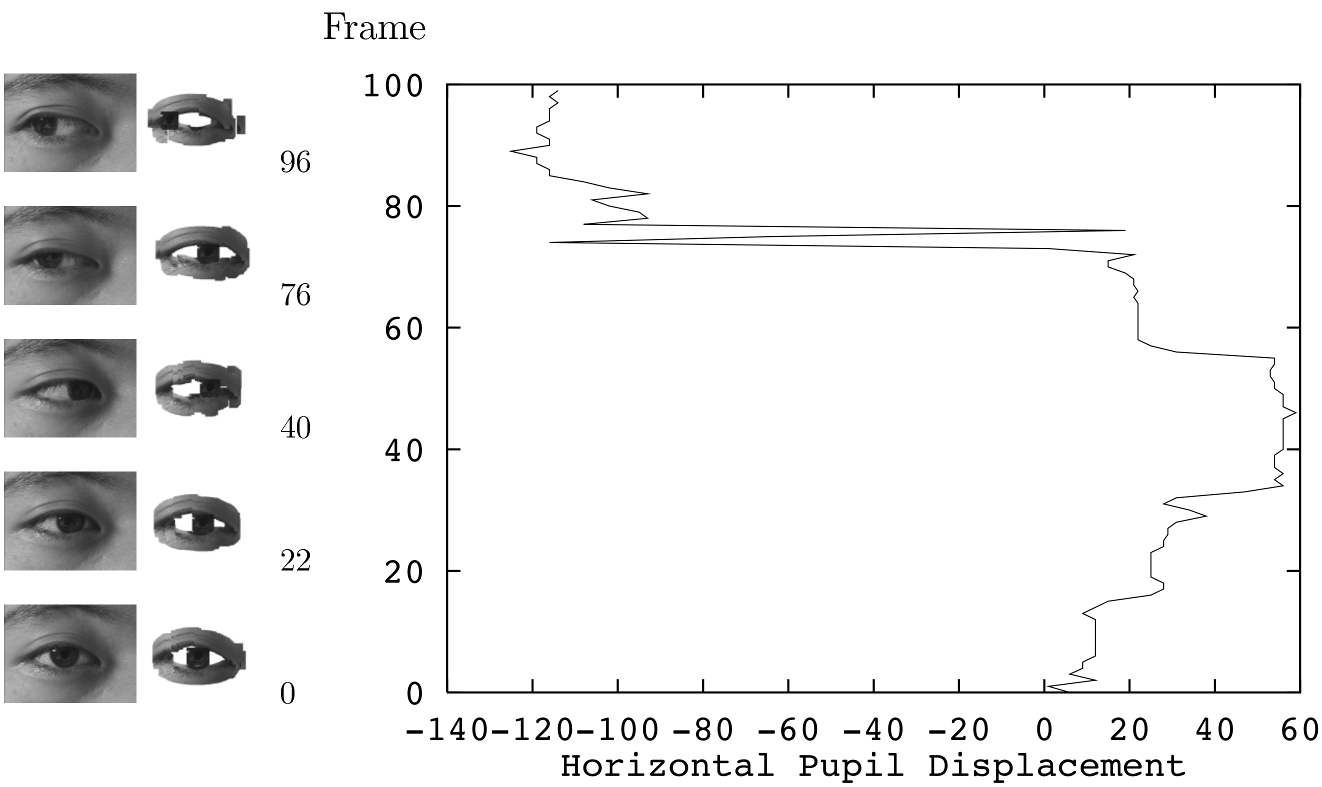


Figure 4: Tracking results: For each frame in the video sequence, the pupil's distance from the eye center is shown. The sequence includes a blink that starts at frame 73. At frame 74, the eye is almost closed and at frame, 80 the eye is completely open again.

[3] S. Baluja, D. Pomerleau, "Non-Intrusive Gaze Tracking Using Aritficial Neural Networks," *Advances in Neural Information Processing Systmes (NIPS) 6*, 1994.

[4] M. Betke, N. C. Makris, "Information-Conserving Object Recognition," *IEEE Int. Conference on Computer Vision,* 145–152, 1998.

[5] A. Gee, R. Cipolla, "Determining the Gaze of Faces in Images," *Image and Vision Computing,* Vol. 12, No. 18, pp. 639–647, 1994.

[6] T. Kohonen, "Self-Organizing Maps," Springer-Verlag, 1995.

[7] S. Lawrence, C. Giles, A. Tsoi, A. Back, "Face Recognition: A Convolutional Neural-Network Approach," *IEEE Trans. on Neural Networks,* Vol. 8, No. 1, pp. 98–111, 1997.

[8] T. Rikert, M. Jones, "Gaze Estimation using Morphable Models," *Int. Conference on Automatic Face- and Gesture-Recognition,* 1998.

[9] B. Schiele, A. Waibel, "Gaze Tracking Based on Face-Color," *Int. Workshop on Automatic Face- and Gesture-Recognition,* 1995.

[10] E. Tsao, W.-C. Lin, C.T. Chen, "Constraint Satisfaction Neural Networks for Image Recognition," *Pattern Recognition,* Vol. 26, No. 4, pp. 553–567, 1993.