Computer Vision Sensing of Eye Movements:
A Communication Aid for the Severely Disabled

David Vernon,
Department of Computer Science,
University of Dublin,
Trinity College,
Dublin,
Ireland.

Abstract

A computer vision system which can track human eye movements is described. The system is based on low-cost personal computer technology and exploits normalised cross-correlation to effect the pattern recognition. Several other pattern recognition techniques were implemented and a brief comparison is provided. The system is designed for a person who suffers from severe cerebal palsy and possesses little voluntary muscular control. The vision system facilitates communication by monitoring the eye and by interpreting eye movements as answers to questions posed by a voice synthesis system. These questions facilitate simple conversations between the user and others; another set of questions allows the user to type by selectively identifying letters from groups of letters spoken by the voice synthesis system. Since the user for whom the system was designed exhibits severe involuntary random saccadic (spasmodic) eye motion, this paper addresses the particular problems encountered in making the system robust.

I. Introduction: the Disability and the Research Requirements

The research described in this paper was embarked upon with the intention of building a communication aid for a fourteen year old boy, Davoren Hanna,† who suffers from a severe form of cerebal palsy. The severity of his disability is such that he has little or no voluntary muscular control; the only muscles which he seemed to have an ability to control include are those associated with his eyes and his torso. In the latter case, he currently communicates with the aid of a helper by "falling" from a partially suspended sitting position toward an enlarged keyboard, gesturing with his hand at a particular key. The helper identifies the key and restores him to a vertical equilibrium from which he can again topple and gesture. Previous attempts to automate this process were unsuccessful and it was decided that the development of a communication aid associated with his eye movements offered the greatest possibility of success.

It was originally envisaged that the communication aid should be configured as simple switch: an appropriate eye movement (e.g. a blink or a gross shift in the gaze of the eye) would be interpreted as either a "yes" or "no" in response to a question posed by a voice synthesis system. By posing an appropriate (sequence) of questions and responding accordingly, Davoren can then either carry on an simple conversation or type text on a personal computer. The latter utility is facilitated by speaking groups of letters, a group containing the letter to be typed is selected by an eye movement, and each letter in the group is then spoken in turn. The final identification of the letter is accomplished by a second eye movement. However, as the research proceeded it became clear that Davoren exhibits severe involuntary random saccadic (spasmodic) eye motion and a significant amount of effort was expended in identifying a method of increasing the robustness of the system and increasing its tolerance to such noise.

† Since the system was designed for a specific individual, it seems appropriate to refer to him by name rather than in the third person, as would be more usual.

II. System Architecture

Two considerations influenced the configuration of the system. First, computer vision and image analysis are computationally demanding processes and it was not known at the outset whether or not simple techniques would suffice to fulfil the pattern recognition requirements of eye tracking. Thus, the computer system upon which the system was developed had to be powerful. Secondly, Davoren was, for the most part, confined to his home and was unable to visit regularly the Computer Vision Laboratory during the development of the software. This meant that the system delivered to him had to be compatible in every way with the development system in the laboratory and that it too was configured as a development system. As a result, the final system is probably over-specified and a successful implementation could be effected on a reduced system.

At present, the system (see diagram 1) comprises an IBM Model 30-286 hosting an INMOS Transputer card and an Imaging Technology PCVISION *plus* framestore. The Transputer system is a RISC (Reduced Instruction Set Computer) based processor with 2 Mbytes of DRAM (Dynamic Random Access Memory). The particular transputer used in this system is a T800 transputer with an on-chip floating point co-processor and is capable of 10 MIPS (Million Instructions per Second) and 2 MFLOPS (Million FLoating point OPerations per Second). The framestore provides an interface to a PULNIX miniature CCD camera, acquiring an image in 40 ms. The imaging resolution is 512x512 pixels with each pixel representing one of 256 levels of grey. The PULNIX camera is equipped with an 8mm focal length sub-miniature bayonet mount lens.

This would have completed the system, had it been possible to find an acceptable voice synthesis board for the IBM. Unfortunately, all of the available boards appear to be designed for blind users and are intended to vocally echo all screen output and

keyboard input rather than to be programmed to synthesise given text. Furthermore, the quality of the voice synthesis is, for the most part, not good. However, the Apple Macintosh supports a software voice synthesis package called *Macintalk*. This package comes complete with the tools to build a custom application; in this instance, a package was developed which monitors the serial port, reading incoming text, and converting it to spoken text. Thus the IBM was augmented with an Apple Macintosh which enunciates the messaged communicated to it by the IBM using a conventional serial line. It is obvious that the use of three CPUs (IBM, Apple, and Transputer) is wasteful and a target system could be significantly optimised.

At present, the camera system is mounted centrally on the top of a light-weight leather helmet (see diagram 2); a small adjustable mirror is extended in front of the camera. This allows the assumption that the eye of the user is always in the centre of the camera's field of view, regardless of the position of his head. It is intended to relax this restriction in the future by mounting the camera remotely in front of the user and by incorporating additional software to identify the position of the head and eye in the field of view. At present, the system assumes that the eye lies somewhere in a small window in the centre of the screen (see diagram 3). The size of this window is variable but all of the work described in this paper assumes a window of either 64x64 pixels or 128x128 pixels.

III. Image Analysis

The central requirement of the image analysis phase is to monitor a sequence of images of a human eye and to identify some change in the presentation of the eye in order to activate a logical switch (indicating an affirmative or a negative response). Thus, we wish to identify some eye movement and to set a threshold on the degree of movement at which the movement will be deemed to be significant and voluntary. At this stage, it is the degree of movement which is important, rather than the directionality of the movement. As we will see, such restriction proved unrealistic.

In order to identify an appropriate pattern recognition paradigm to satisfy these objectives, five techniques were implemented. These include digital image subtraction [1], the Hough transform [2], projection analysis, and two techniques for template matching: one based on the Euclidean distance between two vectors and another based on normalised cross correlation [3]. We will review each technique in turn.

Digital Image Subtraction

The simplest approach to determining whether there has been a significant amount of eye movement in a sequence of images is to subtract two images I_1 and I_2, pixel by pixel, and compute the sum of absolute differences:

$$I_1 - I_2 = \sum_i \sum_j |I_1(i,j) - I_2(i,j)| \qquad (1)$$

If this value exceeds a certain threshold (which can be set by the user), then one can interpret this as a significant eye movement. A second, larger, threshold was used to discriminate between movements of the eye-ball and blinks. This technique worked

very well in laboratory conditions; however, as soon as it was tested in Davoren's home, it became clear that it was extremely sensitivite to the absolute level of the light intensity of the image and, in particular, to changes in the light intensity. Thus, for a static scene (no eye movement) a ray of sunshine or a shadow cast by a passer-by will produce significant differences in images and the sum of absolute differences in pixel values will exceed the "movement" threshold. This technique, then, was discarded at an early stage.

The Hough Transform

The Hough transform is a technique which is used to find curves of a given shape in an image. The classical Hough transform requires that the curve be specified in some parametric form and, hence, is most commonly used in the detection on lines, circles, and ellipses. The main advantages of the technique is that it is very tolerant of gaps in the object boundaries and it is relatively unaffected by noise. We will indicate here how it can be applied to the detection of circles and, in particular, to the detection of the centre of the circle delineating the boundary between the cornea and the iris.

The equation of a circle is given in parametric form by the equation:

$$(x - a)^2 + (y - b)^2 = r^2 \qquad (2)$$

where (a, b) defines the centre of the circle, (x, y) defines the points on the circle and and r is the radius of the circle.

If we have a point (x_i, y_i) on this circle, then

$$(x_i - a)^2 + (y_i - b)^2 = r^2 \qquad (3)$$

For a given circle, a, b, and r are constant. Suppose, however, that we do not know the precise coordinates of the circle but we do know the radius (i.e. a and b are unknown, but r the radius of the circle bounding the cornea is known) and we also know the coordinates of the point(s) on the circle, then we can consider a and b to be variable and x_i, y_i, and r to be constants. In this case, the equation

$$(x_i - a)^2 + (y_i - b)^2 = r^2 \qquad (4)$$

defines the values of a and b such that the circle of radius r passes through the point (x_i, y_i). If we plot these values of a and b, for a given point (x_i, y_i), on a graph, we see that we get a set of points in the (a-b) space, i.e. in a space where a and b are the variables. The transformation between the image plane (x and y coordinates) and the parameter space (a and b coordinates) is known as the Hough transform. Points in the image plane which all lie on a given circle will give rise to transform curves which all intersect in one point (or, rather, cluster at a point) since they share common a_i and b_i: they all belong to the circle given by

$$(x - a_i)^2 + (y - b_i)^2 = r^2 \qquad (5)$$

This, then, provides us with the means to detect circles. First of all we must sample the Hough transform space, i.e. we require a discrete representation of (a-b) space. Since a and b vary between 0 and the maximum resolution of the image or, in particular, between 0 and the maximum dimension of the window in the image in which we wish to search for the circle, this defines the sampling resolution of (a-b) space. For a window size of 64x64 pixels in the image space, the representation of (a-b) space is now simply a 2-D array of size 64x64, each element corresponding to a particular value of a and b. This is called an accumulator since we are going to use it to collect or accumulate evidence of curves given by particular boundary points (x, y) in

the image plane. For each boundary point (x_i, y_i) in the image we increment all accumulator cells such that the cell coordinates (a,b) satisfies the equation (4)

When we have done this for all available (x_i, y_i) points we scan the accumulator searching for cells which have a high count since these will correspond to circle for which there are many points in the image plane. In fact, because there is likely to be some errors in the actual position of the x and y coordinates, given rise to errors in a and b, we search for clusters of points in the accumulator having high counts, rather than isolated points.

Before one can perform a Hough transform to identify the centre of the circle delineating the boundary between the cornea and the iris, one must process the grey-scale image, isolating all possible boundary points. This was accomplished, for the sake of comparison, using two edge detectors. In the first instance, the Marr-Hildreth edge detector [4,5] was exploited in which boundary points are isolated as zero-crossings in Laplacian of Gaussian filtered images. This yielded extremely good boundaries but required an unacceptably long period (10 seconds) to perform the computation. The second approach utilised the Sobel operator [6] which evaluates the gradient of the grey-scale image in a local (3x3) neighbourhood. This provided acceptable results in significantly less than one second.

The total period from image acquisition to identification of the centre of the circle, including the edge detection, Hough transform, and analysis of the accumulator, was approximately two seconds. Although this was a significant improvement on previous implementations, it was deemed to be unacceptable for a real-time eye-movement detector. It is conceivable that optimisation would yield further savings (e.g. using a hardware implementation of the edge detection) but this would have increased the total cost of the system. However, a second (unforseen) problem presented itself. Davoren has a tendency to drop his eye-lids while using the sensor. This resulted in

the reduction of the visible portion of the boundary between cornea and iris from an average of 70% of the circle to less than 20%, on occasions. While the Hough transform is indeed very robust and can tolerate large gaps in the circular boundary, such a reduction in data proved unsustainable and it was decided not to proceed with the development of the technique. Although it was shelved in favour of other approaches, the Hough transform remains one of the potentially most useful techniques and cannot be dismissed from consideration for other similar applications.

Projection Analysis

If one acquires an image of a human eye and projects the image onto its X and Y axes by summing the pixel grey levels along each row and column respectively,† one generates a signature of the variation in image intensity in each direction. An image of a human eye, comprising pupil, iris, cornea, skin, eye-brows and eye-lashes, will exhibit a lower intensity in the centre of the image, i.e. at a position corresponding to the pupil. This will be manifested as a minimum in the associated projection: the position of the two minima in the two orthogonal projections then identify the the centre of the eye. This is a particularly simple technique, from a computational point of view, and worked well in laboratory conditions. However, it occassionaly misclassified a miminum caused by an eye-brow as the centre of the eye and it was noted that, in less than ideal lighting conditions, such misclassification frequently occurred.

Template Matching

† In the image frame of reference, the X axis is vertically aligned so that projection on the X axis requires summation in a horizontal direction. Conversely, the Y axis is horizontally aligned so that projection on the Y axis requires summation in a vertical direction.

The problem of tracking the movements of an eye using digital imaging is essentially one of ascertaining whether a pre-defined sub-image (i.e. of an eye-ball) is contained within a test image and of ascertaining its whereabouts. This sub-image is called a template and should typically be an ideal representation of the eye. This can be generated by acquiring an image of a well-exposed eye, normally looking straight ahead. The template matching technique involves the translation of the template to every possible position in the image and evaluating a measure of the match between the template and the image at that position. If the similarity measure is large enough then the eye may be assumed to be present and its coordinates are those at which the maximum similarity occurred between template and test image. Although, several similarity measures are possible, two measures were evaluated to determine their suitability for tracking eye movements: one based on the Euclidean distance between image and template and one based on cross-correlation.

The standard Euclidean distance between two vectors (e.g. the template t(i,j) and the test image g(i,j)) is defined by:

$$E(m, n) = \sqrt{\sum_i \sum_j [g(i,j) - t(i-m, j-n)]^2)}$$

The summation is evaluated for all i, such that (i-m) is in the domain of definition of the template. The above definition amounts to translating the template t(i,j) to a position (m, n) along the test image and evaluating the similarity measure at that point. Thus, when searching for a template shape, the template is effectively moved along the test image and the above template match is evaluated at each position. The position (m, n) at which the smallest value of E(m, n) is obtained corresponds to the best match for the template.

Since the evaluation of square roots and multiplication operations are computationally expensive, this template matching metric is often simplified by evaluating the sum of the absolute difference of g(i,j) and t(i-m, j-n) rather than the square of the difference. This metric is defined by:

$$S(m, n) = \sum_i \sum_j |g(i,j) - t(i-m, j-n)|$$

This technique was implemented but, as with the techniques based on image subtraction outlined above, it proved to be extremely sensitive to unforseen changes in ambient illumination.

In an attempt to overcome this sensitivity, we investigated template matching using normalised cross-correlation which is, in fact, related to the Euclidean distance metric. The square root in the Euclidean definition can be removed by squaring both sides of the equation and letting the similarity measure be $E^2(m, n)$. Hence:

$$E^2(m, n) = \sum_i \sum_j [g(i,j)^2 - 2\, g(i,j)\, t(i-m, j-n) + t(i-m, j-n)^2]$$

As before, the summation is evaluated for all i and j, such that (i-m, j-n) is in the domain of definition of the template. Note that the summation of the last term is constant since it is a function of the template only and is evaluated over the complete domain of the template. If it is assumed that the first term is also constant, or that the variation is small enough to be ignored, then $E^2(m, n)$ is small when the summation of the middle term is large. Thus, a new similarity measure might be R(m,n), given by:

$$R(m, n) = \sum_i \sum_j g(i,j)\, t(i-m, j-n)$$

again summing over the usual range of i and j. R(m,n) is the familiar cross-correlation function. The template t(i-m, j-n) and the section of g(i, j) in the vicinity of (m, n) are similar when the cross-correlation is large.

Since the assumption that the summation of g(i, j) is independent of m and n is not often valid, (i.e., the energy varies spatially across the image) an alternative to computing R is to compute the normalised cross-correlation N(m, n), given by:

$$N(m, n) = R(m, n) \Big/ \sqrt{\sum_i \sum_j g(i,j)^2}$$

summing over the usual range of i and j. Note that, by the Cauchy-Schwarz inequality,

$$N(m, n) \leq \sqrt{\sum_i \sum_j t(i-m, j-n)^2}$$

Hence, the normalised cross-correlation may be scaled so that it lies in the range 0 to 1 by dividing it by the above expression. Thus, the normalised cross-correlation may be redefined:

$$N(m, n) = R(m, n) \Big/ \left(\sqrt{\sum_i \sum_j g(i,j)^2} \sqrt{\sum_i \sum_j t(i-m, j-n)^2}\right)$$

This technique proved to be extremely robust: it is possible to change the ambien illumination significantly without causing any errors in the detection of the eye position. Futhermore, it is robust enough to deal with the variation of the shap iris as the gaze of the eye changes from fixation point to fixation point. This in shape arises because a circular iris will be projected onto the image plane ellipse when the eye is looking to the side. For a user with severe cerebal palsy, this is

particularly important since the eye often swivels to such an extent that the iris nearly disappears. Diagram 4 illustrates the ability of the system to track eye movements using correlation-based template matching.

IV. Discussion

A template image of the eye of 32x32 pixels in size proved adequate for recognition. In laboratory conditions, it was sufficient to match this template within a suitably placed 64x64 pixel window in the acquired image. The template itself is generated in a training phase by requiring a helper to position a cursor over the centre of the eye. The 64x64 pixel window is positioned in the centre of the 512x512 pixel image; a restriction facilitated by the placement of the mirror mounted in front of the camera and reflecting the image of the eye. Normalised cross-correlation in such a configuration would require in excess of two million multiplications for each image acquired. Since the correlation is effected in software, this is quite unacceptable. To alleviate this computational load, only a sample of the template values are used in the evaluation of the measure of normalised cross-correlation. Several sampling frequencies were investigated and it emerged that a sampling frequency of four pixels in both the horizontal and vertical directions is the most suitable for this application. Thus, only every fourth pixel is used in estimating the normalised cross-correlation. Furthermore, the template itself is translated by an increment of four pixels between successive estimates of the correlation measure. Again, this applies to both the horizontal and vertical directions. Thus, the normalised cross-correlation measure which was adopted may be expressed:

$$N(m, n) = R(m, n) \Big/ \left(\sqrt{\sum_i \sum_j g(i,j)^2} \sqrt{\sum_i \sum_j t(i-m, j-n)^2} \right)$$

where m, n, i, and j vary incrementally in steps of four pixels.

These modifications reduce the computational complexity by a factor of 256; a substantial saving with minimal effect on the robustness. It is important to note, however, that it imposes a limitation on the accuracy with which one can identify the

position (m_{max}, n_{max}) of the eye; errors of up to ± 2 pixels in the x and y directions are possible. This error is significant when one considers that we are operating within a total window of 64x64 pixels. It is eliminated by refining the estimate of the position of maximum correlation, evaluating the normalised cross correlation measure for all positions (m, n) of the template in a 2x2 neighbourhood around this nominal maximum: that is for m and n such that

$$m_{max} - 2 \leq m \leq m_{max} + 2 \text{ and } n_{max} - 2 \leq n \leq n_{max} + 2.$$

Taking into consideration image acquisition on the PCVISION framestore, transfer of the image to the transputer system, estimation of the normalised cross correlation for all template positions, refinement of the position of the maximum correlation, and analysis of the eye movement, it is possible to achieve an update of eye position four times every second with this approach.

The use of a 64x64 pixel viewing window, though adequate in laboratory conditions, caused several problems when being operated in Davoren's home: slight displacements of the helmet due to the involuntary movement of his head and a greater degree of eye movement than that of the author forced the use of a 128x128 pixel window. The current system now reports the eye position once every second.

Had Davoren been able to shift his gaze from position to position voluntarily and with consistency, the system as described would have been complete. Indeed, it was recognised at the outset that an improvement in this respect was required and it was anticipated that practice with the eye movement sensor would lead to improved performance. Unfortunately, this was not the case: Davoren continued to exhibit random saccadic eye motion and consequently each spasmodic movement was then misinterpreted by the system resulting in very erratic behaviour. The solution is quite straightforward and is facilitated by the ability of the system to *track* eye movements rather than merely to *detect* movements. Davoren is, in fact, able to move his eyes to

the left (signifying "yes") and to the right (signifying "no") but such movements take a considerable period of time (between five and ten seconds). Furthermore, the movement is not smooth and several spasmodic movements are usually exhibited. Quite often the voluntary movement to the left is preceded by a (preparatory) movement to the right. Thus, in order to successfully distinguish between "yes" and "no", it is necessary to consider the entire trajectory of the eye over an extended period, typically ten seconds, rather than using single eye movements. Davoren is now beginning to learn how to use and control the system and, despite the severity of his disability, it appears that the system offers much promise.

V. Conclusions

It has been established that it is feasible to exploit non-intrusive image analysis to interpret human eye movements and to use them, in conjunction with a voice synthesis system, to allow the severely disabled to communicate. It emerged that, in the case of a user suffering from severe cerebal palsy, it is not sufficient to base the interpretation upon a single voluntary eye movement and it is mandatory to analyse the trajectory of the eye over an extended period of time. This places a severe limitation on the efficiency of the system (in terms of its effective rate of communication) but, for those with no means of independent communication, this still represents a significant improvement. Such a restriction would not apply, for example, to disabled persons who possess normal gaze control, such as those suffering from quadiplegia.

VI. Acknowledgments

The author would primarily like to acknowledge the courage and determination of Davoren Hanna, for whom the system was designed. Many people contributed to the success of the project; those in the the Computer Vision Laboratory in Trinity College Dublin: Mairead Flanagan, Paul Healy, James Mahon, and David Stokes; those in companies who provided equipment: Pauline Knight and Peter Lawless in IBM Ltd., Brenda Kirwin in Apple Computer Ltd., Pat O'Leary in MicroMarketing Ltd, Fred Kennedy in Captec Ltd.. However, one person alone instigated the project and provided, through the Department of Industry and Commerce in Ireland, the financial backing: Agnes Aylward, *sine qua non.*

VII. References

1. A. Rosenfeld and A. Kak, *Digital Picture Processing*, Academic Press, New York, 1982.

2. P. V. C. Hough, "Methods and Measures for Recognising Complex Patterns", *U.S. Patent 3069654*, 1962.

3. R. O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley and Sons, New York 1973.

4. D. Marr, "Early Processing of Visual Information", *Philosophical Transactions of the Royal Society of London*, Vol B275, pp. 483-524, 1976.

5. D. Marr and E. Hildreth, "Theory of Edge Detection", *Proceedings of the Royal Society of London*, Vol B207, pp. 187-217, 1980.

6. R. Nevatia, *Machine Perception*, Prentice-Hall, 1984.

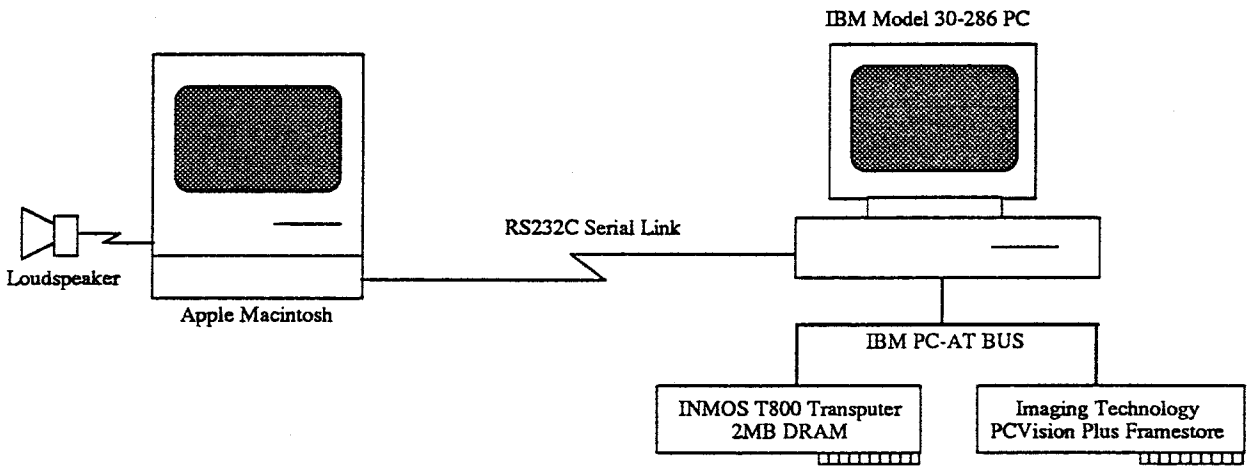

Diagram 1: System Architecture.

Diagram 3: Position of the Mirror and Eye within the Field of View

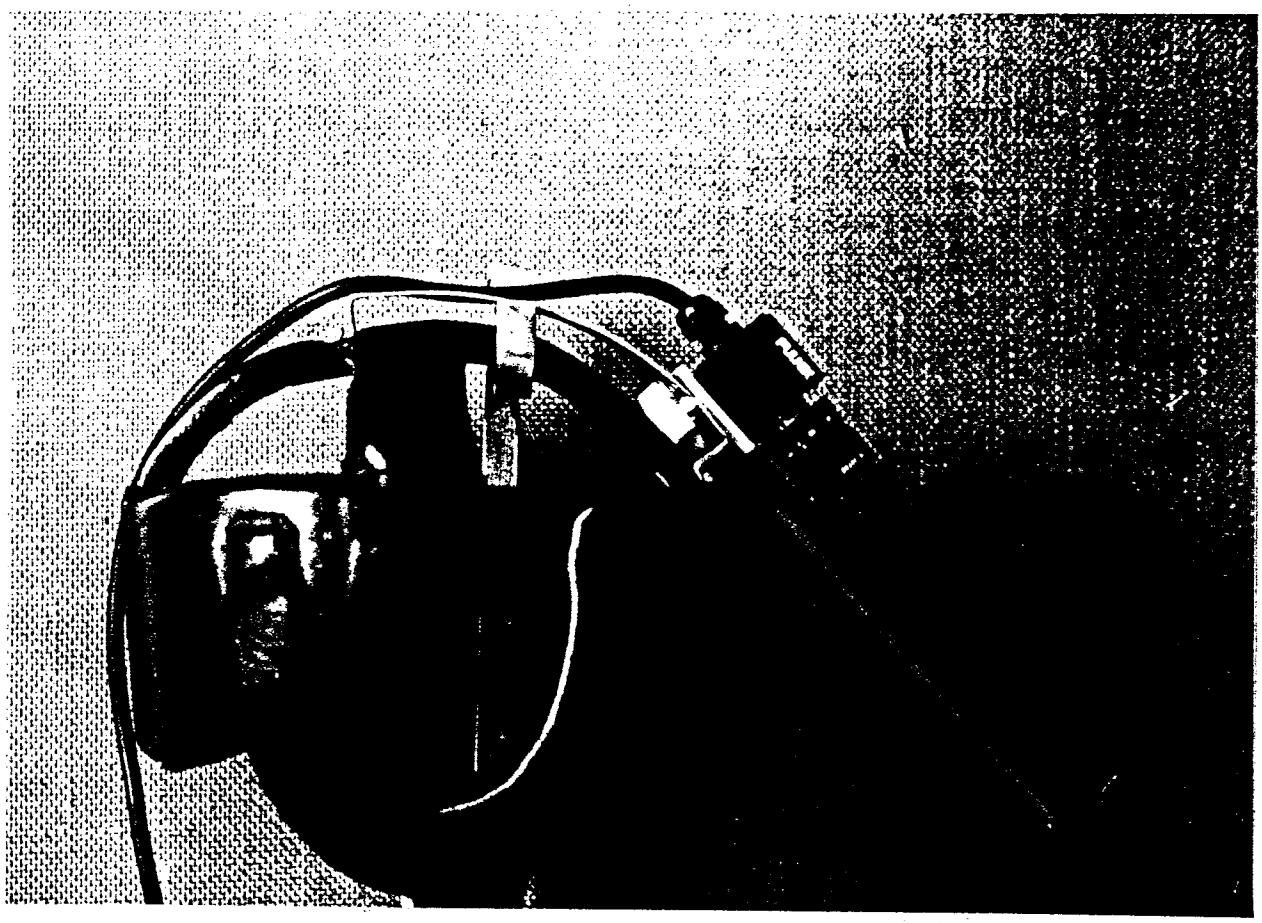

Diagram 2: Camera Headset.

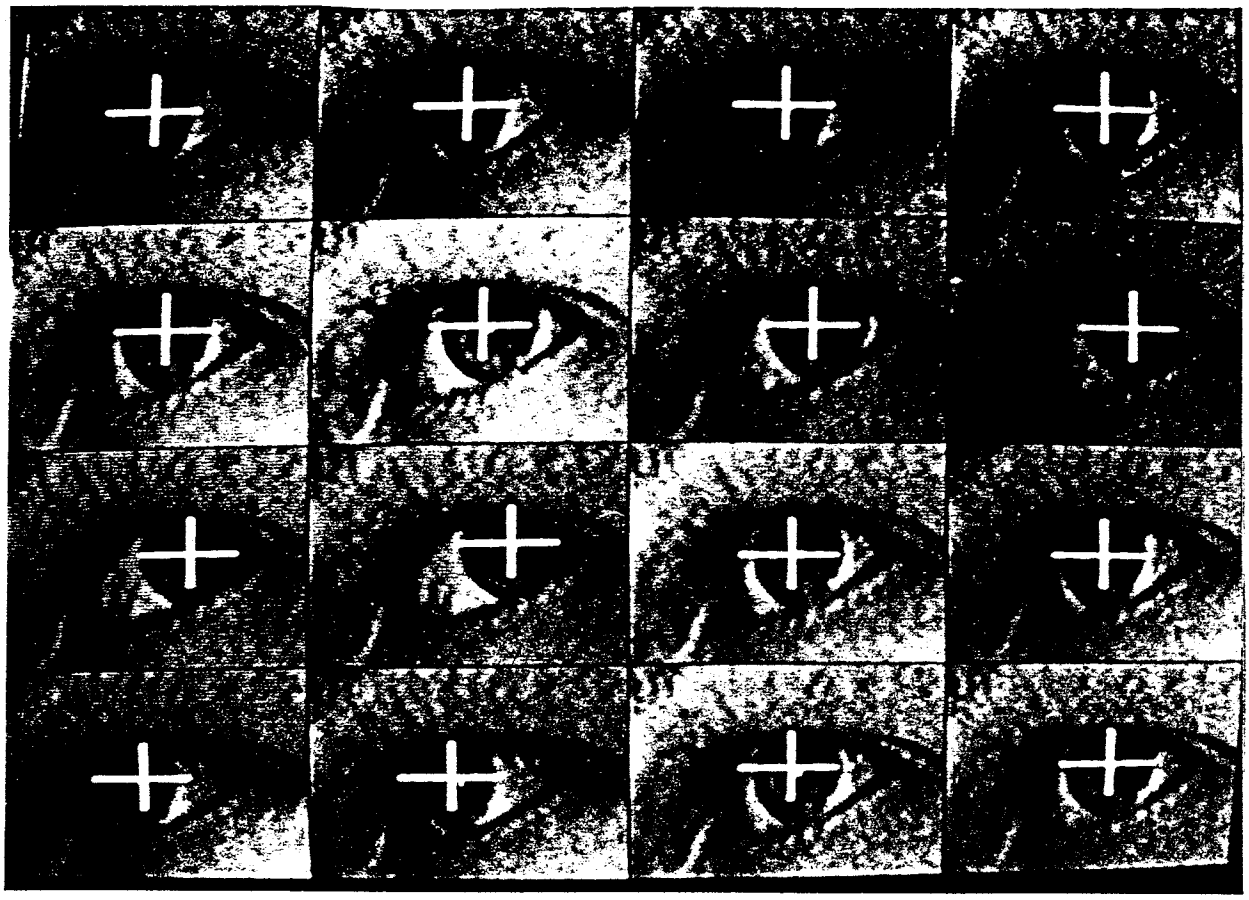

Diagram 4: Eye Tracking Using Normalised Cross-Correlation.

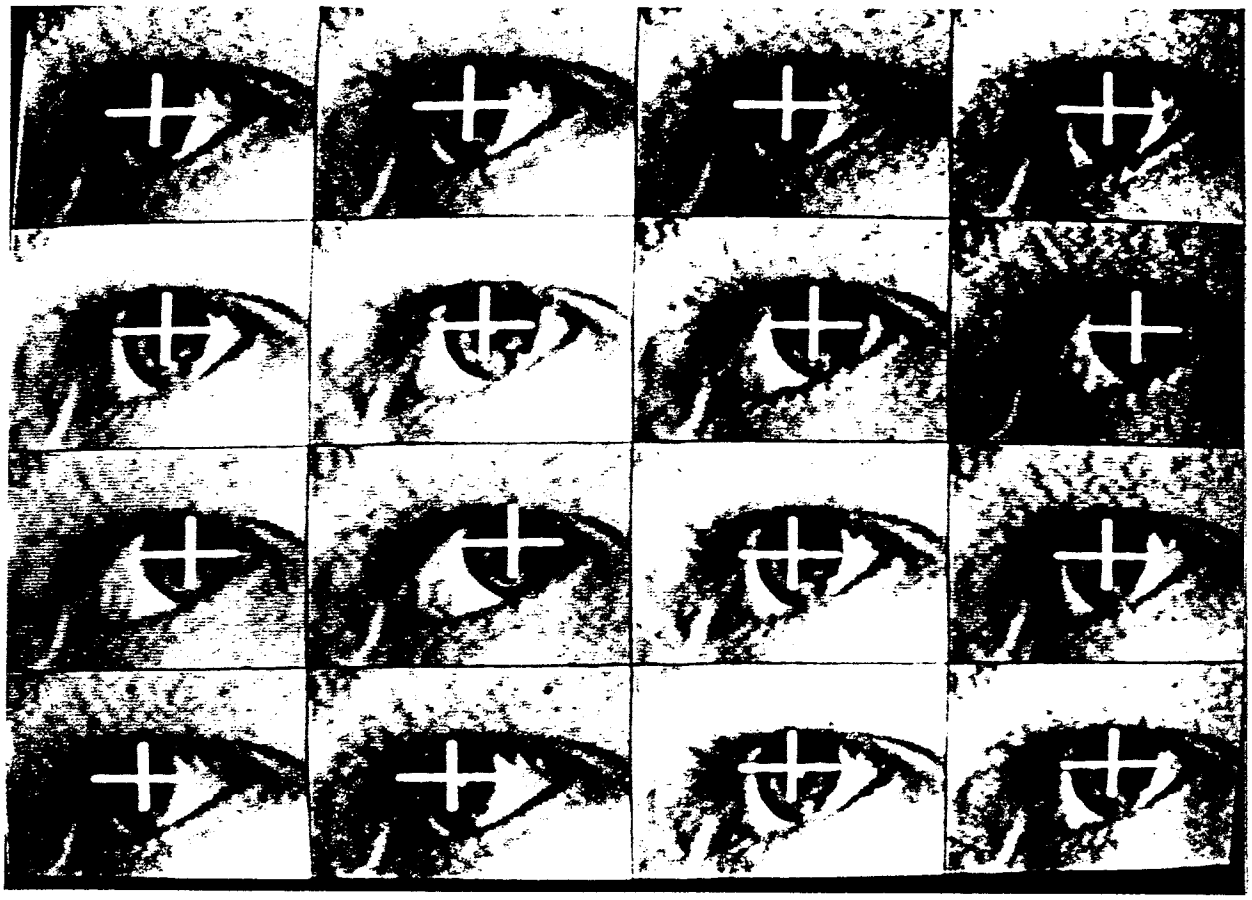

Diagram 4: Eye Tracking Using Normalised Cross-Correlation.