AUTOMATED READING ASSISTANCE SYSTEM USING POINT-OF-GAZE ESTIMATION

by

Jeffrey J. Kang

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science,
Graduate Departments of the Edward S. Rogers Sr. Department of Electrical and Computer Engineering
and the Institute of Biomaterials and Biomedical Engineering,
University of Toronto

Automated Reading Assistance System Using Point-of-Gaze Estimation

Jeffrey J. Kang

Master of Applied Science

Edward S. Rogers Sr. Department of Electrical and Computer Engineering

Institute of Biomaterials and Biomedical Engineering

University of Toronto

2006

Abstract

Head-mounted and remote point-of-gaze estimation methodologies were used in an automated reading assistant to vocalize unknown words in real-time. Points-of-gaze were mapped onto stationary and moving reading material using homographic mappings established from point correspondences between the reading material and images from a scene camera. Points-of-gaze on reading material were used to measure the processing time of each viewed word. Unknown words were detected based on processing time length. Viewed words could be uniquely identified when words on the reading material were separated by more than 15 mm. In a pilot study of four subjects, using remote point-of-gaze estimation and stationary reading material, the detector achieved a detection rate of 89% with a false alarm rate of 11%. In a pilot study with two subjects, using head-mounted point-of-gaze estimation to allow free head movement, the automated reading assistant provided vocalization for 94% of unknown words and 10% of known words.

Acknowledgements

I would like to thank my supervisor, Dr. Moshe Eizenman, for the immeasurable value of his guidance and assistance over the course of this work. Your advise and direction have helped me stay on course. I would also like to acknowledge the financial support provided by the Natural Sciences and Engineering Research Council.

I would like to express my gratitude to my fellow lab member Elias Guestrin for providing advise and assistance over the past two years. I would also like to acknowledge other lab members, past and present: Andrew, Bryon, Charudutt, Michael, Ron, Eric, Peggy, Jason, and Jerry. Each one of you has taught me something.

Finally, I would like to express thanks to my parents for their encouragement and support. I am also grateful to my sister Nancy and my girlfriend Emily. You are the two pillars I lean against.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

During reading, we visually examine words and convert letters to sounds via cognitive processing that activates word recognition. For skilled readers, this conversion is usually swift and effortless. However, for unskilled readers, such as children or students of a foreign language, the ability to perform these conversions is under-developed. We acquire this ability through repetitive exposure to word-to-sound mappings (Bettelheim and Zelan, 1982; Adams, 1990), usually mediated by a skilled reader (e.g. a teacher or a parent).

Word-to-sound mappings in the English language are governed by a set of grapheme (letter unit) to phoneme (sound unit) conversion rules. However, grapheme-to-phoneme conversions in English are characterized by many irregularities and inconsistencies. When a conversion cannot be successfully completed due to insufficient reader skill or irregular spelling, the reader must seek assistive intervention to learn the proper word pronunciation. Assistance of this nature facilitates the learning of new word-to-sound mappings.

This thesis presents the development of a system to detect when a reader encounters reading difficulty and provide automated reading assistance.

1.1 Monitoring Reading through Eye-tracking

The reading process has long been studied using eye-tracking technologies. During reading, we extract information embedded in text by fixating on a series of points. We then integrate information from these points to facilitate the visual perception that leads to cognition. The location of the point of fixation within the visual field, commonly referred to as the point-of-gaze, is a function of eye position. The process of moving the point-of-gaze through the visual field is facilitated by fast eye movements called saccades. These eye movements can be regarded as the overt manifestation of both conscious and subconscious shifts in visual attention. By observing eye movements, we may infer intent, study visual perception, and gain an understanding of the underlying cognitive processes. Eye-tracking technologies provide a means by which we may observe eye position, and hence estimate the point-of-gaze. In the context of reading, estimating the point-of-gaze can provide information regarding the processing time of each word.

1.2 Point-of-Gaze Estimation Applications

Eye-tracking, for the purpose of point-of-gaze estimation, has been used in a variety of applications, which can be divided into two broad categories: the study of visual attention in response to visual stimuli and the development of input devices to facilitate human-machine interaction. A study of selective visual attention was performed to quantify mood disorders (Yu and Eizenman, 2004; Eizenman et al., 2003). Results of the study showed that depressed subjects exhibited a bias in attention towards sad images. Another study investigated how children with childhood-onset

schizophrenia and attention-deficit/hyperactivity disorder viewed and perceived thematic imagery (Karatekin and Asarnow, 1999). The results suggested that schizophrenic children have impaired selective visual attention. Analysis of point-of-gaze in marketing research has been used to evaluate the effectiveness of advertisements in attracting visual attention (Loshe, 1997). Similarly, an evaluation of user interfaces has shown that poorly-designed interfaces result in inefficient visual search behaviour among users (Goldberg and Kotval, 1999). Eye-tracking has also been used extensively in the study of reading behaviour, as cognitive processing of text requires visual examination facilitated by sequential changes in point-of-gaze (Rayner, 1998).

As an input modality, point-of-gaze estimation is currently being used within Dr. Eizenman's research group to develop a visual aid that allows patients with age-related macular degeneration to control a magnifier on a computer screen using only eye movements. Point-of-gaze estimation has also been used to create input devices for physically impaired computer users (Hutchinson et al., 1989). In a more general case, point-of-gaze has been considered as one component of multimodal human-computer interfaces to relieve the input bottleneck created by dependence on keyboards and "mice" (Sharma et al., 1998).

1.3 Point-of-Gaze Estimation Methodology

When we observe a target of interest within the scene, light from the targeted object is focused by the cornea and the lens of the eye such that an image of the object falls upon the retina. The retina contains photoreceptor cells which convert light to neural signals. To make detailed observations, eye movements are made to direct the image of the target onto the highest acuity region of the retina known as the fovea, which

extends over 0.6 to 1 degree of visual angle (Young and Sheena, 1975). The visual axis is defined as the line that passes through the centre of the fovea and the nodal point of the eye (Figure 1.1). The point-of-gaze is then defined as the point of intersection of the visual axes of both eyes within the scene. When the scene is a planar object such as a computer screen or the page of a book, the point-of-gaze can be simplified as the intersection of the visual axis of one eye with the scene plane.

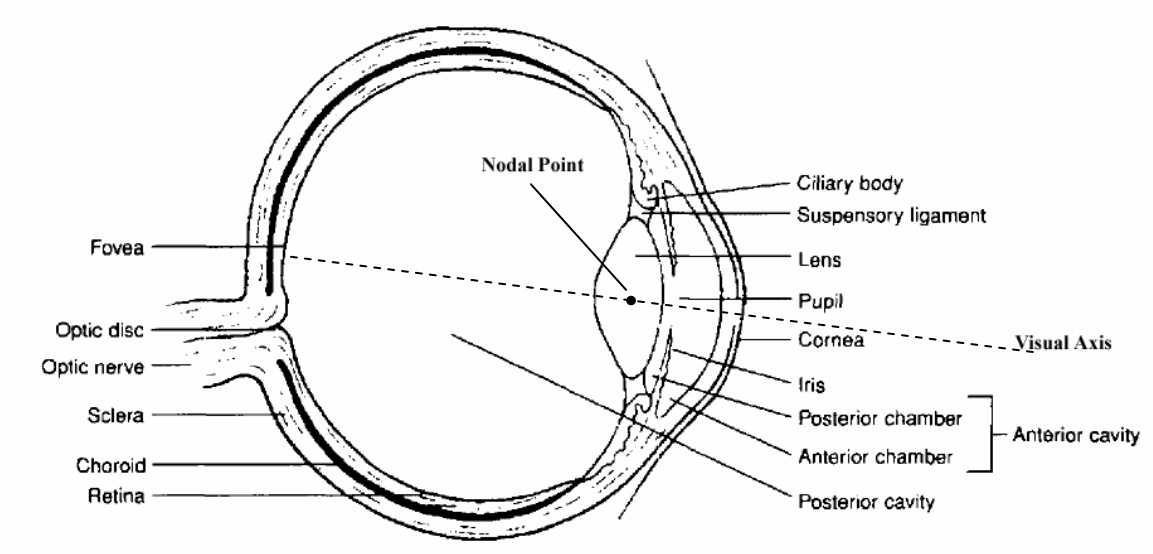

Figure 1.1: Diagram of the eye, adapted (Wilhelm et al., 2001).

Eye-tracking provides a means to determine eye position, from which the visual axis can be reconstructed. Modern eye-tracking systems are generally video-based, employing cameras to image the eye. Eye position is inferred from the measured image locations of eye landmarks such as the pupil centre, iris centre and iris-sclera boundary. A popular approach to estimating eye position, adopted by Dr. Eizenman's research group, uses measurements of the pupil centre and one or more "corneal reflections." A corneal reflection is generated by placing a discrete light source in front of the eye. Light

from this source reflects off of the outer corneal surface, which acts as a convex mirror, forming the corneal reflection as a virtual image of the source. This virtual image is also known as the first Purkinje image (Young and Sheena, 1975). Typically, near infra-red light sources are used for this purpose as they do not interfere with vision. The relative image locations of the corneal reflections and pupil centre are used to infer eye position.

Video-based eye-trackers using the pupil centre and corneal reflection approach can be classified into two categories: head-mounted and remote. In head-mounted systems, the light sources and the camera which images the eye are worn on the subject's head. Hence, the subject's head movement is not constrained. In remote systems, eye-tracking components are placed remotely and nothing is worn by the subject. However, head movement is constrained to the field of view of the remote-placed camera. Remote systems reduce subject fatigue and are more suitable for some investigations such as those involving young children.

1.4 Objectives

The objective of this thesis is to develop an automated reading assistance system that detects when a reader experiences reading difficulty, i.e. encounters an unknown word, and renders assistance by vocalizing the unknown word. This system requires two major components:

- The means to identify, in real-time, the word being read.
- The means to detect when an unknown word has been encountered by the reader.

A further requirement is for the system to operate within a natural reading setting, such that it may be used in a teaching capacity.

1.5 Structure of this Document

Chapter 2 describes the formulation and implementation of a system to determine the viewed word using a head-mounted eye-tracker. Chapter 3 proposes a method to allow the viewed word to be determined using a remote point-of-gaze estimation system. Chapter 4 develops detection criteria for automatically determining reader difficulty by analyzing the time spent viewing each word. Chapter 5 describes an experiment in which the viewed word is automatically vocalized when reader difficulty is detected. The main contributions of the thesis are summarized and directions for future work are proposed at the end of the chapter.

Chapter 2

Determining the Viewed Word

During reading, text processing is facilitated by visual examination of words falling on the high resolution area of the fovea. Therefore, to determine the word being processed, an estimate of point-of-gaze is required. To perform this estimation while allowing unconstrained head movement, it is necessary to use a head-mounted eye-tracker. This chapter describes a system that uses estimates of point-of-gaze made by a head-mounted eye-tracker to determine the viewed word.

Section 2.1 develops a method to map the point-of-gaze obtained from a head-mounted eye-tracker to a point on reading material with arbitrary pose. Section 2.2 presents the implementation of a system that performs the described mapping technique to determined the viewed word.

2.1 Mapping the Point-of-Gaze to the Reading Material

When using head-mounted eye-trackers, eye position is generally measured with respect to a coordinate system defined relative to the head (head coordinate system). Hence, point-of-gaze is estimated with respect to a moving frame of reference as the head moves. However, points of interest on an object, such as the position of words on

reading material, are defined with respect to an object coordinate system that does not move, or moves independently of the head. Therefore, to identify what the subject is looking at, the relationship between the head coordinate system and the object coordinate system must be known.

One solution to this problem is to measure head pose relative to the object coordinate system using a position sensor (Allison et al., 1996), from which a Euclidean transformation between the head coordinate system and the object coordinate system can be obtained. The main drawback of this approach is that the sensor must be calibrated with respect to a static object coordinate system. In other words, if the object coordinate system is defined with respect to reading material, the reading material may not move.

Another approach requires placing a camera on the head to image the scene containing the points of interest (Yu and Eizenman, 2004); in the discussion to follow, we refer to this camera as the scene camera. The position of the scene camera's imaging plane within the head coordinate system is constant, yielding a constant relationship between the head coordinate system and the scene camera's image coordinate system. This allows eye position to be easily transformed to a point-of-gaze within the scene camera's image coordinate system. Previous research has shown that using point correspondences between sets of images, points-of-gaze can be mapped to a reference image upon which points of interest are defined (Yu and Eizenman, 2004). In this chapter, we show that by using point correspondences between a scene image and the reading material, we can directly establish a mapping between the image coordinate system and the object coordinate system. Furthermore, the object coordinate system does not need to be static, and can be defined with respect to moving reading material. This

enables point-of-gaze estimates to be mapped onto reading material with arbitrary pose. The range of motion of the reading material is constrained to the field of view of the scene camera.

A method to obtain the mapping between the image coordinate system and the object coordinate system is described in the sections to follow by modeling the scene camera as a pinhole camera. We begin this description with a review of projective geometry and the pinhole camera model.

2.1.1 Projective Geometry

A camera projection is a mapping between the Euclidean 3-space ($\mathbb{R}^3$) of a scene and the Euclidean 2-space ($\mathbb{R}^2$) of an image. Projective geometry is used to allow the relationship between coordinates in $\mathbb{R}^3$ and $\mathbb{R}^2$ to be represented linearly in matrix form. To use projective geometry, coordinates in Euclidean space $\mathbb{R}^n$, referred to as inhomogeneous coordinates, must be expressed in projective space $\mathbb{P}^{n+1}$ as homogeneous coordinates. Homogeneous coordinates are only uniquely defined up to a scalar, hereafter denoted by a scale factor S; the coordinates $S(a, b, c)^T$ in $\mathbb{P}^3$ are considered to be the same homogeneous coordinate for any non-zero S. An inhomogeneous coordinate is converted to a homogenous coordinate by augmenting the vector with a final coordinate of 1. Hence, the $\mathbb{R}^2$ inhomogeneous coordinate $(x, y)^T$ can be expressed as the $\mathbb{P}^3$ homogeneous coordinate $S(x, y, 1)^T$.

Using projective geometry, we define **P** as a 3×4 homogeneous camera projection matrix, which relates an arbitrary 3D object point $\mathbf{M} = (X, Y, Z, 1)^T$ in the object

coordinate system to a 2D image point $\mathbf{m} = (x, y, 1)^T$ in the image coordinate system via the linear mapping described by (Faugeras, 2001):

$$S\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{P}_{3\times 4}\begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{2.1}$$

or simply as

$$S\mathbf{m} = \mathbf{PM} \tag{2.2}$$

where S is an arbitrary scalar. Hence, **P** is only defined up to a scale factor.

2.1.2 Pinhole Camera Model

Using the perspective projection model that characterizes a pinhole camera, the camera projection matrix **P** can be described as (Hartley and Zisserman, 2003):

$$\mathbf{P} = \mathbf{K}[\mathbf{R}\ \mathbf{T}]. \tag{2.3}$$

The matrix **P** encapsulates two transformations. The first is a Euclidean transformation, specified by a rotation, $\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$, and a translation $\mathbf{T} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}$, which transforms the 3D object point from the object coordinate system (axes denoted by *X*, *Y*, *Z;* origin denoted by **O**) to a camera-centred coordinate system (axes denoted by X_{cam}, Y_{cam}, Z_{cam}; origin denoted by $\mathbf{O}_{cam}$). **R** and **T** are often referred to as the extrinsic camera parameters. For brevity, the camera-centred coordinate system will be referred to as simply the camera coordinate system. $\mathbf{O}_{cam}$ corresponds to the camera's optical centre and is also commonly referred to as the camera's nodal point. The principal axis of the camera points straight down the Z_{cam} axis, while the imaging plane is parallel to the X_{cam}-Y_{cam}

plane. Figure 2.1 shows the relationship between the camera coordinate system and the object coordinate system.

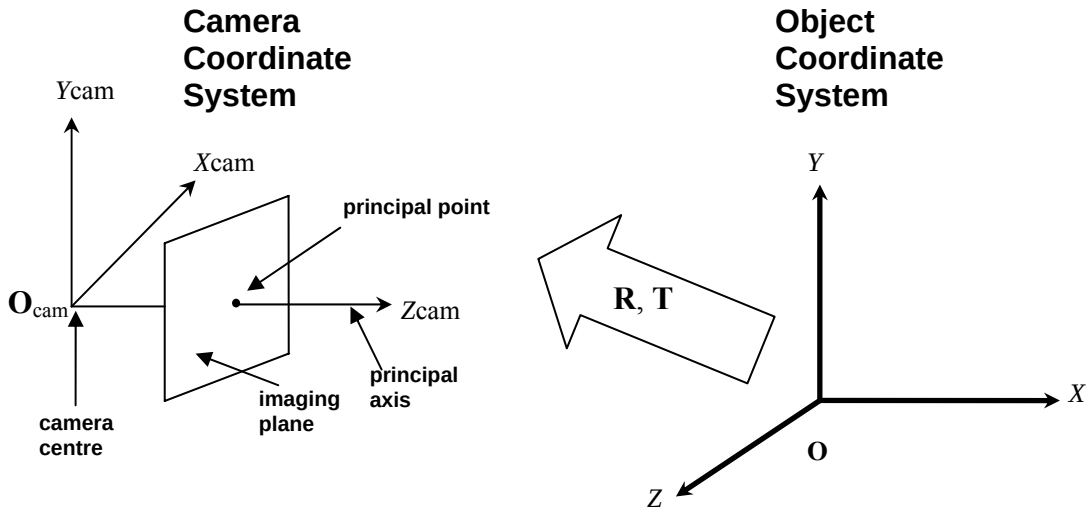

Figure 2.1: Relationship between camera coordinate system and object coordinate system.

The second transformation encapsulated in **P** is the perspective projection that transforms a 3D point described in the camera coordinate system to a 2D image point on the imaging plane. This perspective projection is fully described by **K**, called the intrinsic camera matrix. **K** is given by:

$$\mathbf{K} = \begin{bmatrix} \alpha_x & 0 & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix},$$

where (x_0, y_0) specifies the camera's principal point in the 2D image coordinate system of the imaging plane (in pixels); the principal point is the intersection of the camera's principal axis and the imaging plane; and α_x and α_y represent the focal length in pixels in the x and y directions, respectively. We obtain α_x and α_y using

$$\alpha_x = \frac{f}{p_x}, \text{ and} \qquad (2.4)$$

$$\alpha_y = \frac{f}{p_y}, \tag{2.5}$$

where f is the focal length of the camera in length units (typically in mm), while p_x and p_y specify size of each pixel on the imaging plane (in mm/pixel) in the x and y directions, respectively.

Although simple, the projective projection model of pinhole cameras accurately describes the geometry and optics of most modern CCD cameras (Faugeras, 2001). Using projective geometry, we can project an arbitrary 3D object point to a 2D image point using a linear mapping.

2.1.3 Homographic Mapping

The point-of-gaze estimation system provides the point-of-gaze as a 2D image point, **m**, from which we wish to find the corresponding 3D scene point **M**. From (2.2), we know that **M** can be calculated using

$$\mathbf{M} = S\mathbf{P}^{-1}\mathbf{m}. \tag{2.6}$$

Given a number of correspondences between 3D scene points $\mathbf{M}_i$ and 2D image points $\mathbf{m}_i$, $\mathbf{M}_i \leftrightarrow \mathbf{m}_i$, it is possible to calculate **P**. It can be shown that each point correspondence leads to two equations for the elements of **P** (Hartley and Zisserman, 2003). Since the matrix **P** has 12 elements, and 11 degrees of freedom (ignoring scale), 6 point correspondences are needed to solve for **P**.

P provides the mapping between an arbitrary 3D object point and its corresponding 2D image point. Now consider a special case when 3D object points are constrained to the 2D surface of our reading material (e.g. a page of a book or a reading card). Without loss of generality, we can assume that the 2D surface is on the plane Z=0

of the object coordinate system, i.e. $\mathbf{M} = (X, Y, 0, 1)^T$ (Zhang, 1999). From (2.1) and (2.3), we have

$$S\mathbf{m} = \mathbf{K}[\mathbf{R}\ \mathbf{T}]\begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_x & 0 & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} r_{11} & r_{12} & t_x \\ r_{21} & r_{22} & t_y \\ r_{31} & r_{32} & t_z \end{bmatrix}\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}. \tag{2.7}$$

We can omit the Z coordinate from **M** and use the notation $\mathbf{M}' = (X, Y, 1)^T$ to describe a point on this 2D surface. It then follows that a 2D object point **M'** and its corresponding image point **m** are related by a linear mapping

$$S\mathbf{m} = \mathbf{H}\mathbf{M}'. \tag{2.8}$$

H is a 3×3 matrix given by:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} = \lambda\begin{bmatrix} \alpha_x & 0 & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} r_{11} & r_{12} & t_x \\ r_{21} & r_{22} & t_y \\ r_{31} & r_{32} & t_z \end{bmatrix}. \tag{2.9}$$

Since the point **m** is only defined up to a scale factor S, **H** is also only defined up to a scale factor λ. This matrix describes a plane projective transformation and is commonly referred to as a homography. Given **m**, a 2D image point, we can find **M'**, the corresponding object point, using

$$\mathbf{M}' = S\mathbf{H}^{-1}\mathbf{m}. \tag{2.10}$$

There are many methods described in literature to solve for **H** (Tsai et al., 1982; Zhang, 1999; Criminisi et al., 1999). We employ a method called the Direct Linear Transformation algorithm (Hartley and Zisserman, 2003).

2.1.4 Direct Linear Transformation Algorithm

In this section, the Direct Linear Transformation (DLT) algorithm is used to solve for **H**. First, we establish N correspondences between scene points $\mathbf{M}_i' = (X_i', Y_i', 1)^T$ and image points $\mathbf{m}_i = (x_i, y_i, 1)^T$, where i = 1…N. From (2.10), the relationship between $\mathbf{m}_i$ and $\mathbf{M}_i'$ may be expressed as a vector cross product

$$S\mathbf{m}_i \times \mathbf{H}\,\mathbf{M}_i' = 0 \tag{2.11}$$

from which we can derive two linearly independent equations in the elements of **H**:

$$h_{11}x_i + h_{12}y_i + h_{13} - h_{31}x_iX_i' - h_{32}y_iX_i' - h_{33}X_i' = 0 \tag{2.12}$$

$$h_{21}x_i + h_{22}y_i + h_{23} - h_{31}x_iY_i' - h_{32}y_iY_i' - h_{33}Y_i' = 0. \tag{2.13}$$

If we have N point correspondences, then from (2.12) and (2.13) we can write the following equation to solve for the elements of **H**:

$$\mathbf{Ah} = \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1X_1' & -y_1X_1' & -X_1' \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -x_1Y_1' & -y_1Y_1' & -Y_1' \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2X_2' & -y_2X_2' & -X_2' \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -x_2Y_2' & -y_2Y_2' & -Y_2' \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N & y_N & 1 & 0 & 0 & 0 & -x_NX_N' & -y_NX_N' & -X_N' \\ 0 & 0 & 0 & x_N & y_N & 1 & -x_NY_N' & -y_NY_N' & -Y_N' \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \\ h_{33} \end{bmatrix} = \mathbf{0} \tag{2.14}$$

where **h** is a column vector containing the elements of **H**. We ignore the trivial solution of $\mathbf{h} = \mathbf{0}$. The non-zero solution for **h** is the null-space of **A**.

Since **h** has 9 elements and 8 degrees of freedom (ignoring scale), a unique and exact solution for **h** exists when **A** has rank 8. We obtain 2 equations in the elements of **h** from each point correspondence; thus, **A** has rank 8 when there are 4 point correspondences (N = 4) and no 3 $\mathbf{M}_i'$ are collinear. The presence of 3 collinear points

represents a degenerate configuration in which the 8 equations in the elements of **h** are not independent; hence **A** has rank < 8.

When there are more than 4 point correspondences, the system $\mathbf{Ah} = \mathbf{0}$ is over-determined. The degenerate configuration is avoided in the general case for N > 4 when no (N - 1) of the N $\mathbf{M}_i$'points are collinear. If the measurements for $\mathbf{M}_i$' and $\mathbf{m}_i$ are exact such that (2.6) is obeyed, then **A** still has rank 8. However, if the measurements are noisy, then **A** has rank > 8 and no exact non-zero solution for **h** exists. The DLT algorithm finds an approximate solution for **h** by minimizing the vector norm $\|\mathbf{Ah}\|$. An additional constraint $\|\mathbf{h}\| = 1$ is imposed to avoid the trivial solution $\mathbf{h} = \mathbf{0}$. It can be shown that the solution to this minimization problem is the eigenvector of $\mathbf{A}^T\mathbf{A}$ corresponding to the smallest eigenvalue. Equivalently, the solution is the unit singular vector corresponding to the smallest singular value of **A**, which can be found by performing the singular value decomposition of **A**. We rearrange the elements of **h** to reconstruct **H**. Using **H**, we can map the point-of-gaze provided by the eye-tracker as an image point to its corresponding point on the 2D surface of the reading material, described with respect to the object coordinate system.

2.2 System Description

In this section we describe a system that performs the mapping method described in Section 2.1 to determine the viewed word. We begin with an overview of the overall system architecture and its major components. This overview is followed by a description of a system implemented using a video-based head-mounted eye-tracker developed in Dr. Eizenman's research group.

2.2.1 System Overview

A block diagram of the system is presented in Figure 2.2, showing the decomposition of the system, and the flow of data between the components. The eye-tracker initiates the flow of data by capturing the image of the eye and the image of the scene containing the reading material. The scene image represents the visual field of the subject. The eye-tracker estimates a direction of gaze by analyzing the eye image, and calculates a point-of-gaze, $\mathbf{POG}_I = (x_{pog}, y_{pog}, 1)^T$ in the image coordinate system of the scene image, where x_{pog} and y_{pog} have units of pixels. To relate $\mathbf{POG}_I$ to a point in the object coordinate system of the reading material, $\mathbf{POG}_R = (X_{pog}, Y_{pog}, 1)^T$, where X_{pog} and Y_{pog} have units of length (e.g. mm), the position of the reading material within the scene image must be known. This position is established by analyzing the scene image to locate image points that correspond to known points on the reading material. From a set of point correspondences, a mapping between arbitrary image points and their corresponding reading material points is calculated. The point $\mathbf{POG}_I$ is then mapped to $\mathbf{POG}_R$ as shown in Figure 2.3.

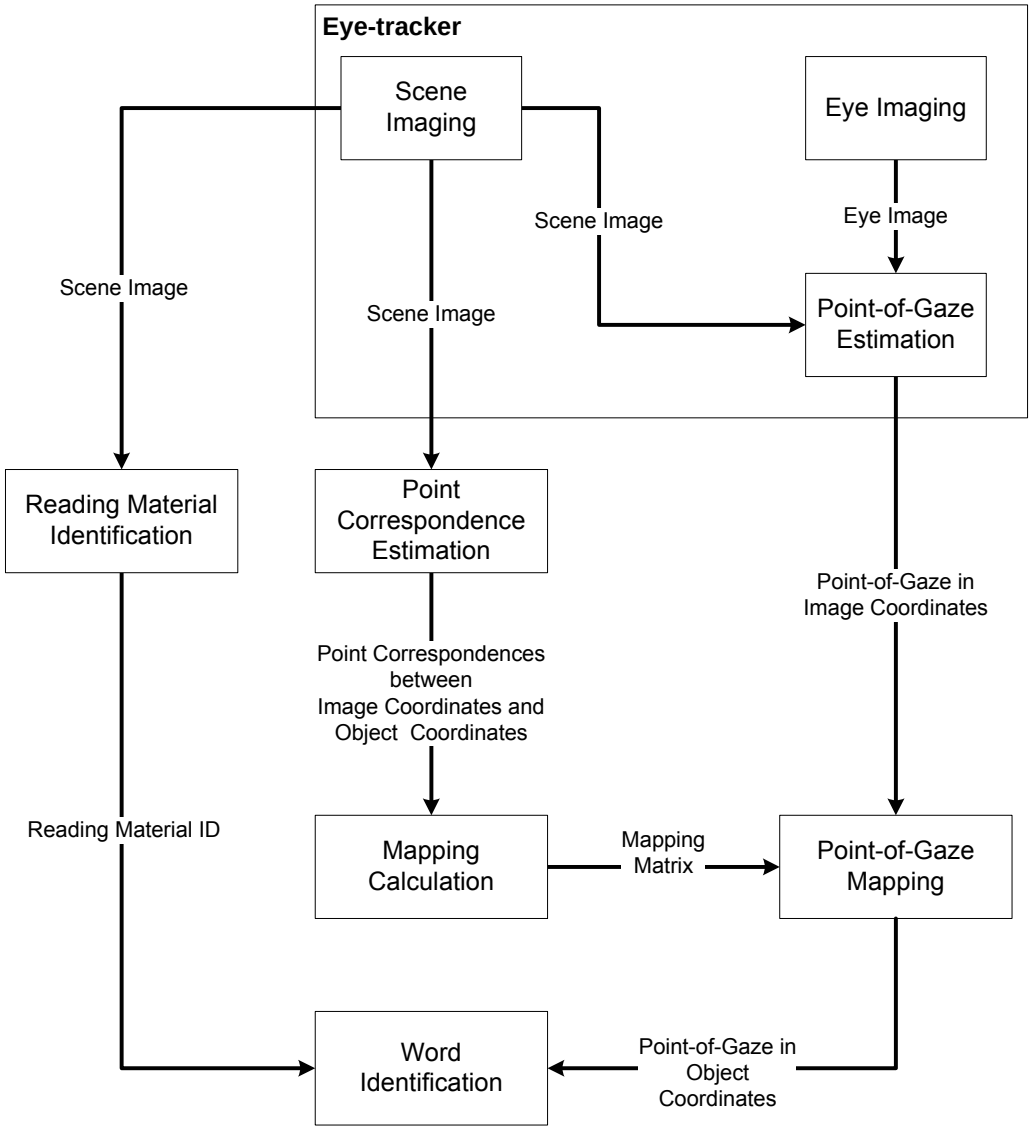

Figure 2.2: Block diagram of the system to determine the viewed word.

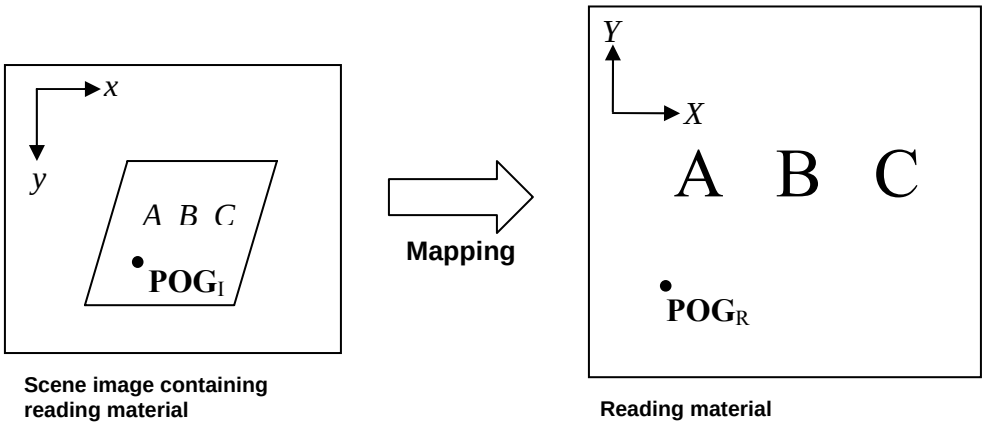

Figure 2.3: Mapping the point-of-gaze from the scene image to the reading material.

Reading materials are uniquely identified by affixed barcodes, which are later recovered from the scene image. After the reading material is identified, a lookup table containing the reading material's words is consulted. This table describes the position of each word with respect to the reading material. The viewed word is determined as a word that matches the position of the point-of-gaze, $\mathbf{POG}_R$.

2.2.2 Video-based Head-mounted Eye-tracker

The head-mounted video-based eye-tracker consists of an imaging unit mounted on a head-band carrying an eye-tracker assembly and a scene camera, and a PC-based processing unit. The eye-tracker assembly consists of three primary components: a small infra-red (IR) CCD video camera (eye camera), two IR light-emitting diodes (LEDs), and a hot mirror that allows visible light to pass through while reflecting IR light. The two IR LEDs are mounted above the right eye; light is reflected by the angled hot mirror onto the eye, providing illumination and generating the two corneal reflections. Images of the eye are reflected by the hot mirror up to the eye camera which is also positioned above

the eye. The scene camera is a CCD video camera that images the field of view of the subject. A photograph of the eye-tracker is shown in Figure 2.4.

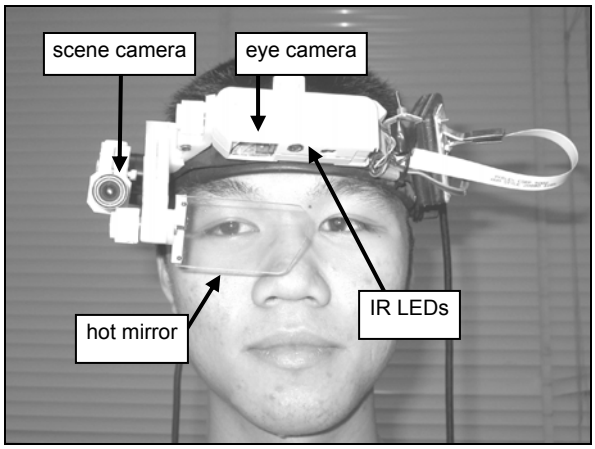

Figure 2.4: Photograph of the head-mounted video-based eye tracking system.

The processing unit is a PC workstation that performs image processing, data logging, point-of-gaze estimation, and other computational tasks. Images from the eye camera are captured at a rate of 50 Hz. Each image is processed to locate the corneal reflections and the pupil centre, as shown in Figure 2.5. The relative positions of these eye features are used to estimate the point-of-gaze on the image captured from the scene camera.

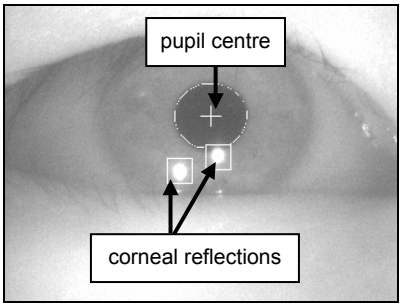

Figure 2.5: Sample image of the eye with located corneal reflections and pupil centre.

2.2.3 Generating Point Correspondences

The task of automatically locating corresponding points from images is a classic problem in machine vision and photogrammetry, and is frequently referred to as the correspondence problem (Marr and Poggio, 1976; Longuet-Higgins, 1981; Mouaddib, E. et al., 1997). A common approach to automating the measurement and identification of image points is to place coded targets within the scene (Ganci and Handley, 1998; Ahn et al., 2001). These targets are easily located in images using image processing techniques and allow image points to be precisely measured.

We adopt an approach using N circular coded targets, which are placed on the 2D surface of the reading material at known positions $\mathbf{M}_i'$, where $i = 1 \ldots N$, and $N \geq 4$. During reading, the scene camera images the reading material and the coded targets. From the scene camera images we estimate $\mathbf{m}_i$, the pixel positions of the targets. This establishes the N point correspondences from which **H** is calculated. Figure 2.6 shows a sample reading card displaying 4 targets.

Figure 2.6: Sample reading card with circular coded targets placed at the corners.

2.2.4 Evaluation of Mapping Accuracy

The mapping method presented above assumes that the scene camera is an ideal pinhole camera that projects scene points to image points via the linear relation described by (2.1). In reality, camera lenses introduce non-linear distortions to the projection. These distortions cause image points to deviate from the coordinates predicted under the pinhole camera model. Furthermore, image points cannot be measured exactly. Hence, the measured image point $\tilde{\mathbf{m}}_i = (\tilde{x}_i, \tilde{y}_i, 1)^T$ will deviate from the theoretical image point $\mathbf{m}_i = (x_i, y_i, 1)^T$. We estimate the homography $\mathbf{H}$ using a set of correspondences between $\tilde{\mathbf{m}}_i$ and $\mathbf{M}_i'$ by minimizing the algebraic distance $\|\mathbf{Ah}\|$ given in (2.14). In this section, we quantify the errors associated with mapping the point-of-gaze using a homography estimated from distorted, noisy image points.

To evaluate mapping performance in a typical experimental setup using reading cards like the one shown in Figure 2.6, we constructed a validation card containing 9 targets (see Figure 2.7) located at known reading material coordinates $\mathbf{M}_i'$ (i = 1…9). The card was held in a typical reading pose and imaged by the scene camera of the eye-tracker. Head movements and reading card movements were made to simulate the motions expected during a reading task. A sequence of 1000 scene images was captured and the image coordinates of the targets $\tilde{\mathbf{m}}_i$ (i = 1…9) were recovered from each image.

For each image, the 4 point correspondences at the corners of the validation card, $\mathbf{M}_i' \leftrightarrow \mathbf{m}_i$ (i = 1…4), were used to estimate the homography $\mathbf{H}$. This represents the configuration of 4 targets shown in the sample reading card. This homography was used to map the remaining 5 image points $\tilde{\mathbf{m}}_i$ (i = 5…9) to the corresponding reading material

points $\mathbf{H}^{-1}\tilde{\mathbf{m}}_i$. Using the sequence of 1000 scene images, a total of 5000 points were mapped in this manner. For each mapped point, a mapping error, defined as the Euclidean distance between $\mathbf{M}_i'$ and $\mathbf{H}^{-1}\tilde{\mathbf{m}}_i$ (i = 5…9), was calculated. Figure 2.8 shows a histogram of the mapping errors.

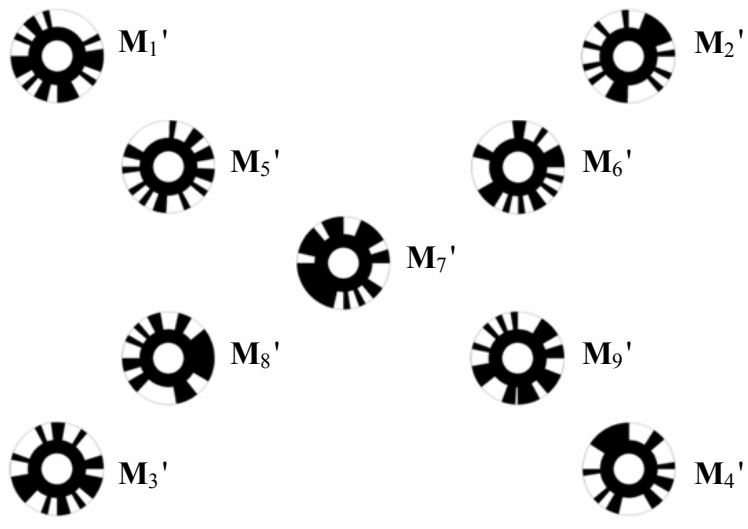

Figure 2.7: Validation card containing nine circular coded targets.

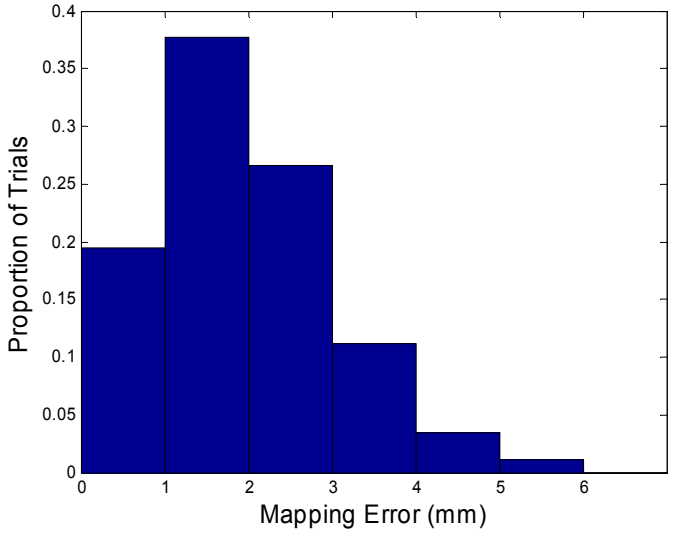

Figure 2.8: Histogram of mapping errors.

The RMS mapping error for the 5000 points was found to be 2.29 mm. More than 98% of the points yield a mapping error less than 5 mm. To put the magnitude of the mapping error into proper context, we can compare it to the magnitude of the point-of-gaze estimation error. Recall that the head-mounted eye-tracker provides gaze-angle estimates with an accuracy of approximate one degree of visual angle. When the reading material is placed 60 cm from the eye (a typical reading distance), the error in the gaze angle corresponds to approximately 10 mm on the reading material. The maximum error in the mapped point-of-gaze estimate is bounded by the sum of the estimation error and the mapping error. When designing or selecting reading material for use with the system, this upper bound is the minimum separation required between words to allow the viewed word to be uniquely identified. For example, if we assume the point-of-gaze estimation error to be a constant 10 mm and the mapping error to be less than 5 mm, we would expect word separations of 15 mm to be sufficient to allow proper identification of the viewed word.

2.2.5 Determining the Viewed Word

Using the described mapping procedure, we obtain a point-of-gaze in the object coordinate system attached to the reading material. To identify the viewed word based on this point-of-gaze, we need to identify the reading material and also know, in advance, the position of each word within the object coordinate system.

Figure 2.9 shows the sample reading card with an 8-bit barcode. In a typical experiment using reading cards, we do not expect the number of unique cards to exceed 50. A barcode of 6-bits is sufficiently long to identify up to 64 (2^6) unique reading cards.

A start bit and a stop bit are used to facilitate extraction of the barcode from the scene image.

The barcode is placed at a known position in the object coordinate system. Using the estimated homography, the image position of the centre of the start bit can be calculated. This pixel coordinate provides a starting point for the image processing algorithm used to recover the bit-values of the barcode. The width of each bit of the barcode on the reading material is set to 20 mm. Since the mapping error is expected to be less than 5 mm, this width ensures that the image processing algorithm begins within the boundaries of the start bit. Using this method, the barcode can be extracted from the scene image efficiently and reliably. The barcode is subsequently decoded to yield a "reading material ID."

Figure 2.9: Sample reading card with binary barcode.

Once the reading material is identified, the viewed word is identified via a lookup table containing the contents of each piece of reading material. The structure of this lookup table can be conceptualized as a three-dimensional grid (see Figure 2.10) with indices: row, column, and "reading material ID." Each cell of the grid contains a single character.

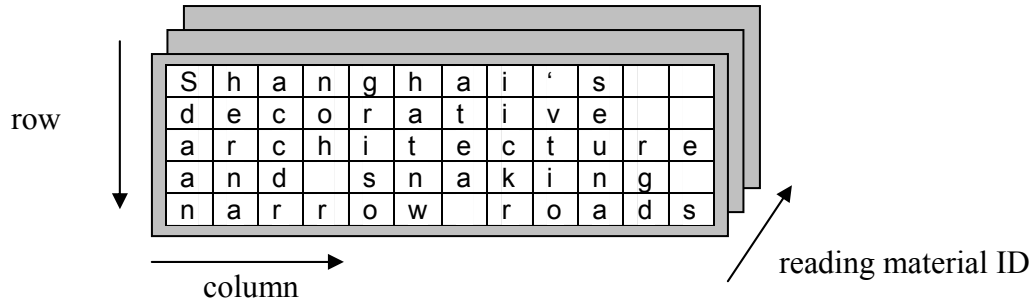

Figure 2.10: Structure of the reading material lookup table.

Using a priori knowledge of the character height and character width (in length units), and the point in the object coordinate system that corresponds to the top-left cell, the mapped point-of-gaze in the object coordinate system is converted to row and column indices. The “reading material ID” recovered from the barcode represents the third lookup table index. A search into the table for the cell specified by {row, column, reading material ID} returns the character corresponding to the point-of-gaze. The viewed word is reconstructed by scanning adjacent cells in the row.

2.3 Summary

This chapter described a system which determines the viewed word using a head-mounted eye-tracker that estimates point-of-gaze with respect to a image obtained from a scene camera. Using point correspondences established by using 4 coded targets, it was shown that point-of-gaze could be mapped from the image coordinate system of the scene camera to the object coordinate system of the reading material with an RMS error of 2.29 mm. This system was designed to tolerate unconstrained movement of the head, making it suitable for use within a natural reading setting.

In the next chapter, it is shown that if some constraints are placed on head movement, then a remote point-of-gaze estimation system can be used to identify the viewed word. Such a system requires no head-mounted components and is more suitable for some reading assistance applications.

Chapter 3

Extension to Remote Gaze Estimation

Chapter 2 described a system to determine the viewed word using a head-mounted eye-tracker. This system allows for unconstrained movement of the head, and can be used to facilitate automated per-word assistance within a natural reading setting. However, the use of the head-mounted eye-tracker is not suitable for applications requiring gaze to be monitored over long durations, as is the case for long reading tasks. Furthermore, it is not desirable to use head-mounted equipment in applications involving children. If head movement can be constrained to a degree that is still allows for reasonable movement required for typical reading tasks, then remote point-of-gaze estimation techniques can be applied to identify the viewed word.

A remote point-of-gaze estimation system with no head-mounted components has been developed in Dr. Eizenman's research group (Guestrin, 2006). This system performs point-of-gaze estimation with respect to a 2D scene object in a static pose. This chapter presents a theoretical framework for extending this system to estimate point-of-gaze with respect to a moving 2D scene object. Within the context of providing automated reading assistance, this development allows the viewed word to be determined

using a remote point-of-gaze estimation system while tolerating constrained head movement and reading material movement.

Section 3.1 describes the current remote point-of-gaze estimation system. Section 3.2 proposes a method to estimate the motion of a 2D object, such as a reading card, and hence how to estimate point-of-gaze with respect to a moving object. Section 3.3 evaluates, via simulation, the accuracy of the motion estimation and the corresponding effect on point-of-gaze estimation.

3.1 Remote Point-of-Gaze Estimation System

Like the head-mounted eye-tracker, the remote point-of-gaze estimation uses two infra-red (IR) light sources to illuminate the eye and produce two corneal reflections, and a video camera to capture images of the eye (eye camera). The system estimates point-of-gaze with respect to a fixed coordinate system using the image locations of the pupil centre and the corneal reflections. Since the position of the eye camera is fixed, head movement is constrained to the camera's field of view. Like the head-mounted eye-tracker, the accuracy of this system is approximately one degree of visual angle. Figure 3.1 shows an implementation of a system which estimates the point-of-gaze on a computer screen.

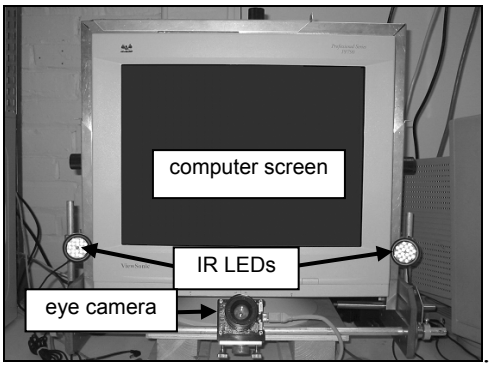

Figure 3.1: Remote point-of-gaze estimation system.

In general, the position and orientation of the visual axis of one eye must be known to estimate point-of-gaze with respect to a scene plane (e.g. a computer screen). Recall that the point-of-gaze is defined as the intersection of the visual axis with this plane, where visual axis is a line in space extending from the fovea through the eye's nodal point. In the case of the head-mounted eye-tracker, point-of-gaze is estimated with respect to images from a scene camera attached to the head. The position of the eye relative to the scene plane (the scene camera's imaging plane) is fixed regardless of head movements. Hence, point-of-gaze can be estimated using only the orientation of the visual axis with respect to the head.

In the case of the remote point-of-gaze estimation system, to accommodate head movement relative to the immotile scene plane, no assumptions are made about the position of the eye. Hence, both the orientation of the visual axis and the location of one point on the visual axis must be estimated with respect to a fixed coordinate system. The scene plane is assumed to coincide with the planar surface of an object such as a computer screen or a reading card; this object will henceforth be referred to as the 2D scene object.

A right-handed 3D Cartesian coordinate system (with axes *X*, *Y*, *Z* and origin **O** in Figure 3.2) is defined such at the 2D scene object is centred at the origin and occupies the plane Z=0. Using the terminology introduced in Chapter 2, this coordinate system will be referred to as the object coordinate system. Points of interest, such as the positions of words on the reading material, are defined relative to the object coordinate system.

The remote point-of-gaze estimation system uses a spherical corneal model. The nodal point of the eye is assumed to be coincident with the centre of corneal curvature. The system provides estimate of the point-of-gaze (**P**) and the centre of corneal curvature (**C**) in the object coordinate system. The visual axis is modeled as the line through **P** and **C**. We refer to the direction of the visual axis as the gaze vector **G** = (**P** – **C**).

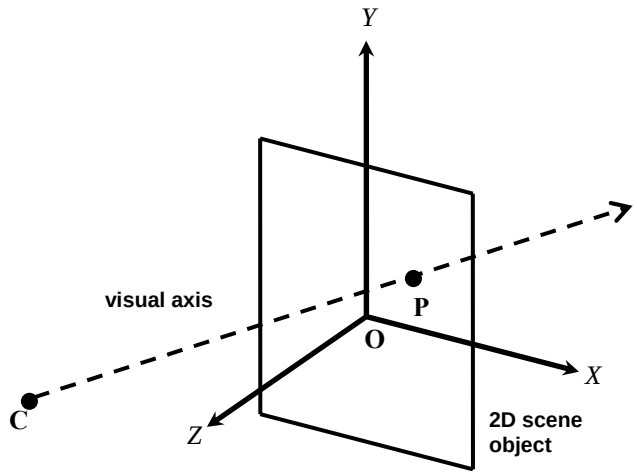

Figure 3.2: Intersection of visual axis with the 2D scene object.

If the 2D scene object moves from its assumed position, estimates for **P** will no longer be meaningful as the visual axis now intersects the scene object at a point **P'** (Figure 3.3). The points **C**, **P**, and **P'** are collinear, which is expressed in parametric form as

$$\mathbf{P}' = \mathbf{C} + \mu\,(\mathbf{P} - \mathbf{C}) = \mathbf{C} + \mu\,\mathbf{G} \qquad (3.1)$$

with parameter μ. In the next section, we present a method to estimate the motion of the 2D scene object. Using knowledge of this motion, we can calculate **P'** with respect to the true position of the scene object.

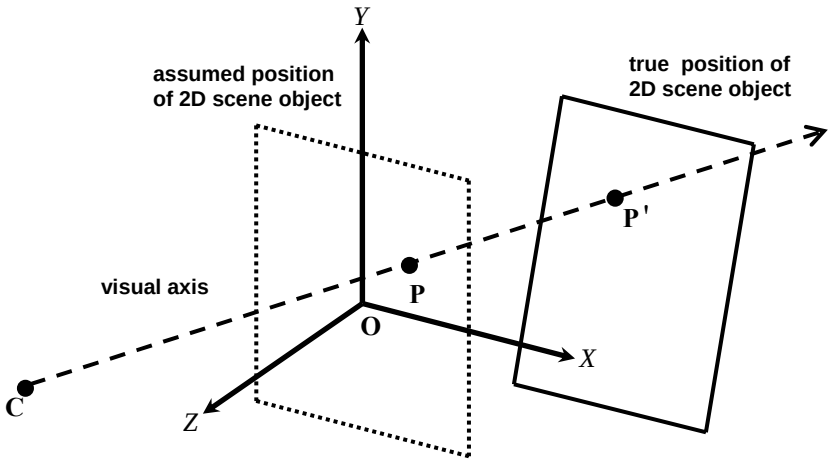

Figure 3.3: Intersection of visual axis with the 2D scene object after movement.

3.2 Estimating the Motion of a 2D Scene Object

In order to estimate the motion of a scene object, clearly a means to observe it is needed. We can use a scene camera to image the scene in which the object moves, as is done when using the head-mounted eye-tracker. However, since we desired not to encumber users with head-worn components, the scene camera should be placed remotely rather than on a head-band. In this section, we describe a method to estimate the motion of a 2D scene object using a static remotely-placed scene camera. Using this method, as long as movement of the scene object is constrained to the field of view of the scene camera, the position of the object can be determined.

3.2.1 Coordinate Systems

The remote point-of-gaze estimation system estimates gaze with respect to a fixed object coordinate system. Without loss of generality, we can assume that this fixed object coordinate system is attached to the 2D scene object in some initial pose at time t=0 (denoted by axes X_0, Y_0, Z_0 and origin $\mathbf{O}_0$ in Figure 3.4). As before, the scene object is centred at the origin and occupies the plane $Z_0=0$. At time t=1, the scene object has a new pose and a new coordinate system (denoted by axes X_1, Y_1, Z_1 and origin $\mathbf{O}_1$). The scene camera has its own coordinate system (denoted by axes X_{cam}, Y_{cam}, Z_{cam}) and has an origin $\mathbf{O}_{cam}$, where $\mathbf{O}_{cam}$ represents the camera's optical centre (or nodal point).

Recall that two coordinate systems are related by a Euclidean transformation described by a rotation matrix **R** and a translation vector **T**, where

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, \text{ and } \mathbf{T} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}.$$

When describing the relationship between a camera coordinate system and an object coordinate system, **R** and **T** are also referred to as the camera's extrinsic parameters.

Let us define the following notation for relating the three coordinate systems:

- $\mathbf{R}_{0\to C}$: Rotation from object coordinate system at t=0 to camera coordinate system,
- $\mathbf{R}_{1\to C}$: Rotation from object coordinate system at t=0 to camera coordinate system,
- $\mathbf{R}_{0\to 1}$: Rotation from object coordinate system at t=0 to object coordinate system at t=1,
- $\mathbf{T}_{0\to C}$: Translation from object coordinate system at t=0 to camera coordinate system,
- $\mathbf{T}_{1\to C}$: Translation from object coordinate system at t=1 to camera coordinate system, and
- $\mathbf{T}_{0\to 1}$: Translation from object coordinate system at t=0 to object coordinate system at t=1.

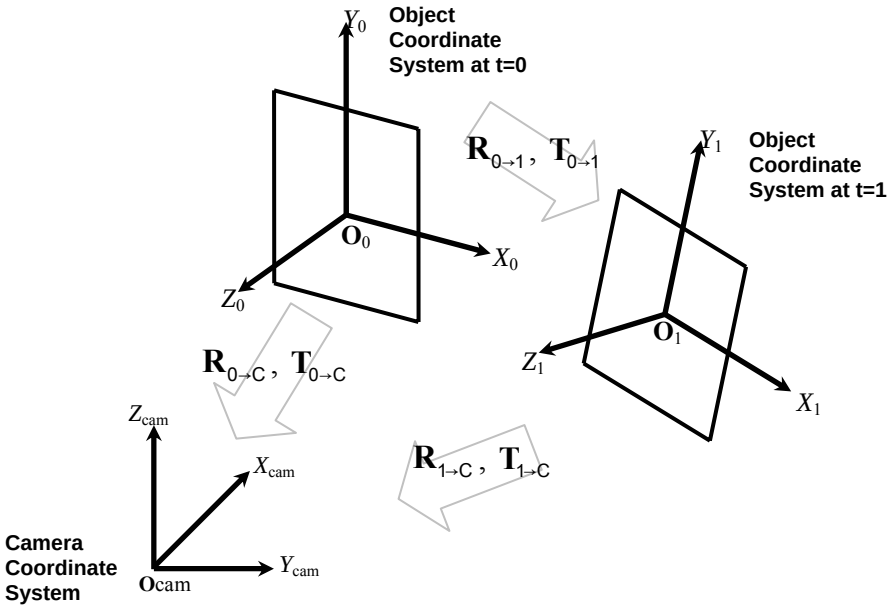

Figure 3.4: Camera coordinate system and the object coordinate systems at two time instances.

$\mathbf{R}_{0\to1}$ and $\mathbf{T}_{0\to1}$ specify the Euclidean transformation between two object coordinate systems. Since the two coordinate systems are both attached to the 2D scene object, $\mathbf{R}_{0\to1}$ and $\mathbf{T}_{0\to1}$ also specify the motion undergone by the object from t=0 to t=1.

3.2.2 Motion from Extrinsic Camera Parameters

When the extrinsic parameters of the scene camera are unknown, a solution to the motion of a 2D object, described by $\mathbf{R}_{0\to1}$ and $\mathbf{T}_{0\to1}$, exists up to a two-fold ambiguity and a unknown scale factor in $\mathbf{T}_{0\to1}$ (Tsai and Huang, 1981; Faugeras and Lustman, 1988; Weng et al., 1991). These ambiguities can be eliminated if additional information regarding the object are known, but in general a partial recovery of the extrinsic camera parameters, such as the distance between the camera and the object, is required.

However, when the extrinsic parameters of the scene camera relative to the respective local coordinate systems of the 2D object at t=0 ($\mathbf{R}_{0\to C}, \mathbf{T}_{0\to C}$) and at t=1 ($\mathbf{R}_{1\to C}$ $\mathbf{T}_{1\to C}$) are known, a unique solution to $\mathbf{R}_{0\to 1}$ and $\mathbf{T}_{0\to 1}$ can be determined.

Let us define the following notation for describing points within the three coordinate systems:

- $\mathbf{M}_0$: 3D point described with respect to the object coordinate system at t=0,
- $\mathbf{M}_1$: 3D point described with respect to the object coordinate system at t=1, and
- $\mathbf{M}_C$: 3D point described with respect to the camera coordinate system.

Using the above notation, $\mathbf{M}_0$, $\mathbf{M}_1$, and $\mathbf{M}_C$ are related by:

$$\mathbf{M}_C = \mathbf{R}_{0\to C}\mathbf{M}_0 + \mathbf{T}_{0\to C}, \tag{3.2}$$

$$\mathbf{M}_C = \mathbf{R}_{1\to C}\mathbf{M}_1 + \mathbf{T}_{1\to C}, \text{ and} \tag{3.3}$$

$$\mathbf{M}_1 = \mathbf{R}_{0\to 1}\mathbf{M}_0 + \mathbf{T}_{0\to 1}. \tag{3.4}$$

From (3.2) and (3.3), we may obtain the relationship between $\mathbf{M}_0$ and $\mathbf{M}_1$ to be

$$\mathbf{M}_1 = \mathbf{R}_{1\to C}^{-1}\mathbf{R}_{0\to C}\mathbf{M}_0 + \mathbf{R}_{1\to C}^{-1}(\mathbf{T}_{0\to C} - \mathbf{T}_{1\to C}). \tag{3.5}$$

From (3.4) and (3.5), we have

$$\mathbf{R}_{0\to 1} = \mathbf{R}_{1\to C}^{-1}\mathbf{R}_{0\to C}, \text{ and} \tag{3.6}$$

$$\mathbf{T}_{0\to 1} = \mathbf{R}_{1\to C}^{-1}(\mathbf{T}_{0\to C} - \mathbf{T}_{1\to C}). \tag{3.7}$$

From (3.6) and (3.7), it is clear that $\mathbf{R}_{0\to C}, \mathbf{T}_{0\to C}$, $\mathbf{R}_{1\to C}$ and $\mathbf{T}_{1\to C}$ are sufficient to solve for the motion of the scene object in terms of $\mathbf{R}_{0\to 1}$ and $\mathbf{T}_{0\to 1}$.

3.2.3 Extrinsic Camera Parameters from the Homography

Chapter 2 introduced the planar projective transformation (homography) which describes a linear mapping between points on a 2D object and their corresponding image points. In this section we show a method to solve for the extrinsic camera parameters (**R** and **T**) from this homography based on a technique used for camera calibration (Zhang, 1999).

The equation (2.8), relating the homography matrix **H** with a camera's intrinsic and extrinsic parameters, is restated here for convenience:

$$\mathbf{H} = \lambda\mathbf{K}\begin{bmatrix} r_{11} & r_{12} & t_x \\ r_{21} & r_{22} & t_y \\ r_{31} & r_{32} & t_z \end{bmatrix}, \tag{3.8}$$

where λ is an arbitrary scalar and **K** is the intrinsic camera matrix. Let us denote **H** by its column vectors: $\mathbf{H} = [\mathbf{h}_1\ \mathbf{h}_2\ \mathbf{h}_3]$. Likewise, let us denote **R** by $\mathbf{R} = [\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3]$. Using this notation, (3.8) can be expressed as:

$$[\mathbf{h}_1\ \mathbf{h}_2\ \mathbf{h}_3] = \lambda\mathbf{K}[\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{T}]. \tag{3.9}$$

From (3.9), we have

$$\mathbf{r}_1 = \frac{1}{\lambda}\mathbf{K}^{-1}\mathbf{h}_1, \tag{3.10}$$

$$\mathbf{r}_2 = \frac{1}{\lambda}\mathbf{K}^{-1}\mathbf{h}_2, \text{ and} \tag{3.11}$$

$$\mathbf{T} = \frac{1}{\lambda}\mathbf{K}^{-1}\mathbf{h}_3. \tag{3.12}$$

To solve for the remaining rotation component $\mathbf{r_3}$ and to recover the scale λ, we must recognize some special properties of the matrix **R**.

R is a rotation matrix, and is hence orthogonal. A property of orthogonal matrices is that the column vectors are orthonormal; by definition:

1. The column vectors are mutually orthogonal, and

2. Each column vector has length (norm) equal to one.

Due to orthogonality of the column vectors of **R**,

$$\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2. \tag{3.13}$$

Furthermore, since the column vectors have length of one,

$$\|\mathbf{r}_1\| = \|\mathbf{r}_2\| = \|\mathbf{r}_3\| = 1. \tag{3.14}$$

From (3.10) and (3.11), we know

$$\|\mathbf{r}_1\| = \frac{1}{\lambda} \| \mathbf{K}^{-1}\mathbf{h}_1\|, \text{ and} \tag{3.15}$$

$$\|\mathbf{r}_2\| = \frac{1}{\lambda} \| \mathbf{K}^{-1}\mathbf{h}_2\|; \tag{3.16}$$

it then follows that

$$\lambda = \| \mathbf{K}^{-1}\mathbf{h}_1\| = \| \mathbf{K}^{-1}\mathbf{h}_2\|. \tag{3.17}$$

If the homography **H** and the intrinsic camera parameters, **K**, are known, then the extrinsic camera parameters **R** and **T** can be uniquely determined.

We may estimate **H** at each time instance using the method presented in Section 2.1. Coded targets can be placed on the 2D scene object at known positions $\mathbf{M}_i$, for i = 1…N, and N ≥ 4, defined relative to the object coordinate system. At each time instance, we measure the corresponding image coordinates of the coded targets as $\tilde{\mathbf{m}}_i$. From the point correspondences, we estimate a homography $\mathbf{H}_0$ at t=0, from which we obtain $\mathbf{R}_{0\to C}$ and $\mathbf{T}_{0\to C}$. Likewise, at t=1, at obtain $\mathbf{R}_{1\to C}$ and $\mathbf{R}_{1\to C}$ from a homography $\mathbf{H}_1$. Finally, using (3.6) and (3.7), we may solve for $\mathbf{R}_{0\to 1}$ and $\mathbf{T}_{0\to 1}$.

3.2.4 Finding the Point-of-Gaze After Motion

The point-of-gaze after movement is given by the point $\mathbf{P}'$, defined to be the point in space where the visual axis intersects the 2D scene object in its t=1 position. Since points of interest are defined with respect to the object coordinate system, it is clear that at t=1 we wish to determine $\mathbf{P}'$ with respect to the t=1 object coordinate system (with axes X_1, Y_1, Z_1 and origin $\mathbf{O}_1$). In other words, $\mathbf{P}'_1$ is needed rather than $\mathbf{P}'_0$. Equation (3.1) is restated for t=1 as

$$\mathbf{P}'_1 = \mathbf{C}_1 + \mu\, \mathbf{G}_1. \tag{3.18}$$

However, the current remote point-of-gaze estimation system estimates gaze with respect to the object coordinate system attached to the initial position of the 2D scene object, defined to be the t=0 object coordinate system (with axes X_0, Y_0, Z_0 and origin $\mathbf{O}_0$). Hence, only $\mathbf{C}_0$ and $\mathbf{G}_0$ are provided. Therefore, we must obtain $\mathbf{C}_1$ and $\mathbf{G}_1$ indirectly using knowledge about the motion of the scene object. By applying (3.4), we have $\mathbf{C}_1 = \mathbf{R}_{0\to1}\mathbf{C}_0 + \mathbf{T}_{0\to1}$ and $\mathbf{G}_1 = \mathbf{R}_{0\to1}\mathbf{G}_0 + \mathbf{T}_{0\to1}$. The remaining unknown is μ.

The value of μ is given by the intersection of the visual axis, expressed parametrically in (3.18), with the 2D scene object. Recall that at t=1 the 2D scene object is defined to occupy the Z_1=0 plane. This plane has a unit normal given by $\hat{\mathbf{n}} = [0, 0, 1]^T$, and is described by the equation

$$\hat{\mathbf{n}} \bullet (\mathbf{P}'_1 - \mathbf{Q}) = 0, \tag{3.19}$$

where $\mathbf{Q}$ is any other point on the plane. Since the plane includes the origin, for simplicity, we let $\mathbf{Q} = \mathbf{0}$, from which we obtain

$$\hat{\mathbf{n}} \bullet \mathbf{P}'_1 = 0. \tag{3.20}$$

Solving (3.18) and (3.20) for μ yields

$$\mu = \frac{\hat{\mathbf{n}} \bullet (-\mathbf{C}_1)}{\hat{\mathbf{n}} \bullet \mathbf{G}_1}, \tag{3.21}$$

which allows a unique solution for $\mathbf{P'}_1$ to be obtained.

3.3 Simulation of Point-of-Gaze Estimation Accuracy

The accuracy of the point-of-gaze estimate, $\mathbf{P'}_1$ on objects that have undergone an arbitrary motion depends on two factors:

- the accuracy of the estimate of gaze provided by the current remote point-of-gaze estimate system, in terms of $\mathbf{C}_0$ and $\mathbf{G}_0$, and
- the accuracy of the estimate of scene object motion in terms of $\mathbf{R}_{0\to1}$ and $\mathbf{T}_{0\to1}$.

Using the methods presented above, the estimation accuracy of $\mathbf{R}_{0\to1}$ and $\mathbf{T}_{0\to1}$ depends exclusively on how well the homographies $\mathbf{H}_0$ and $\mathbf{H}_1$ can be estimated and how accurately the intrinsic camera matrix $\mathbf{K}$ can be determined. In general, $\mathbf{K}$ can be obtained via the camera manufacturer's specifications or through camera calibration to a high degree of accuracy (Faugeras, 1993; Hartley and Zisserman, 2003; Zhang, 1999; Weng et al., 1991). Hence, we focus on errors in estimating $\mathbf{H}_0$ and $\mathbf{H}_1$, which result from deviations between the measured image locations of the coded targets, $\tilde{\mathbf{m}}_i$ (i = 1…N), and the image locations predicted using the pinhole camera model. These deviations are chiefly caused by measurement noise and lens distortions. The effect of these errors in $\tilde{\mathbf{m}}_i$ on the estimation accuracy for $\mathbf{R}_{0\to1}$ and $\mathbf{T}_{0\to1}$, and the corresponding effect on $\mathbf{P'}_1$, was studied through simulations.

3.3.1 Simulation Parameters

Motion of a moving 2D scene object was simulated under typical experimental conditions. Four targets were placed at the corners of the scene object, such that their coordinates in the object coordinate system were $[-150, 150, 0]^T$, $[150, 150, 0]^T$, $[-150, -150, 0]^T$ and $[150, -150, 0]^T$ (unit in mm). This scene object approximates a reading card with dimensions 300 mm by 300 mm.

Using the pinhole camera model, these targets were projected onto the imaging plane of a scene camera. The camera's intrinsic parameters were modeled after a Pixelink PL-A741 camera having a 1280x1024 CMOS sensor with of 6.7 μm square pixel and a lens with a focal length of 16 mm. Using equations (2.4) and (2.5), we obtain α_x and α_y with $p_x = p_y = 6.7$ μm/pixel, and f = 16 mm. The principal point was assumed to be located at the centre of the imaging plane. Hence, the following intrinsic parameters were used

- principal point (in pixels): x_0=640.5, y_0=512.5, and
- focal length (in pixels): $\alpha_x = 2400$, $\alpha_y = 2400$.

The resulting intrinsic camera matrix was

$$\mathbf{K} = \begin{bmatrix} 2400 & 0 & 640.5 \\ 0 & 2400 & 512.5 \\ 0 & 0 & 1 \end{bmatrix}.$$

The scene camera was placed 2 m from the initial position of the scene object (t=0), with the camera's principal axis (Z_{cam} axis) on the Z_0 axis of the object coordinate system at t=0, and the camera's X_{cam} and Y_{cam} axes aligned with the X_0 and Y_0 axes (recall Figure 4.3). Hence, the extrinsic parameters of the scene camera were: $\mathbf{T}_{0\to C} = [0, 0, 2000]^T$ (unit in mm) and

$$\mathbf{R}_{0\to C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

For each trial, the scene object is moved to a random pose within the scene at t=1. The translation is specified by $\mathbf{T}_{0\to1} = [t_x, t_y, t_z]^T$, where t_x, t_y, and t_z are uniformly distributed between -300 mm and 300 mm. To specify the rotation, we parameterize $\mathbf{R}_{0\to1}$ by a vector $\boldsymbol{\theta}_{0\to1}$ which has a direction parallel to the axis of rotation and a magnitude equal to the angle of rotation. This rotation vector is specified by $\boldsymbol{\theta}_{0\to1} = \theta\mathbf{W} = [\theta_x, \theta_y, \theta_z]^T$, where $\mathbf{W}$ is a direction vector generated as a random point uniformly distributed on the surface of a sphere with a centre at the origin and a radius of 1, and θ is an scalar angle uniformly distributed between -30° and 30°. $\mathbf{R}_{0\to1}$ and $\boldsymbol{\theta}_{0\to1}$ are related by the Rodrigues formula (Appendix A).

Gaze is modeled by an eye in fixed position, with a centre of corneal curvature $\mathbf{C}_0 = [0, 0, 600]^T$ (unit in mm). The gaze vector $\mathbf{G}_0$ is simulated such that the subject always fixates on the centre of the 2D scene object after the random movement, i.e. the point-of-gaze $\mathbf{P'}_1 = [0, 0, 0]^T$.

For each trial, the simulated image locations of the coded targets were used to estimate $\mathbf{H}_0$ and $\mathbf{H}_1$ at t=0 and t=1, respectively. The image locations of circular coded targets can be estimated to sub-pixel accuracy (Ahn et al., 2001; Clarke et al., 1994), with RMS estimation errors of approximately 0.1 pixels reported when using ellipse fitting techniques (et al., 1994). Errors in the image locations caused by non-linear lens distortions can be minimized using distortion correction techniques. Residual errors in image locations have been reported to be on the order of 0.01 pixels (Heikkila and Silven,

1997). To simulate errors in the image locations, Gaussian noise with zero mean and standard deviation between 0.1 and 1.0 pixels was added to the projected image coordinates.

$\mathbf{R}_{0\to1}$ and $\mathbf{T}_{0\to1}$ were calculated using the method presented in Section 3.2; the estimates obtained under noisy conditions are denoted by $\hat{\mathbf{R}}_{0\to1}$ and $\hat{\mathbf{T}}_{0\to1}$. Using $\hat{\mathbf{R}}_{0\to1}$ and $\hat{\mathbf{T}}_{0\to1}$, we obtained an estimate for the point-of-gaze, denoted by $\hat{\mathbf{P}}'_1$.

3.3.2 Simulation Results

The movement of the 2D scene object and the point-of-gaze after movement was estimated over 5000 simulated trials for each noise level. The error in the estimate of the translation is described by E_{tx}, E_{ty}, and E_{tz}, the RMS errors associated with each translational component. For example, E_{tx} is given by

$$E_{tx} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |\hat{t}_{x,i} - t_{x,i}|^2} \tag{3.22}$$

where N = 5000, and for the i-th trial, we have $\mathbf{T}_{0\to1,i} = [t_{x,i}\ t_{y,i},\ t_{z,i}]^T$ and $\hat{\mathbf{T}}_{0\to1,i} = [\hat{t}_{x,i}, \hat{t}_{y,i}, \hat{t}_{z,i}]^T$. To express the error in the estimate of the rotation, the rotation matrix $\hat{\mathbf{R}}_{0\to1}$ is converted to its equivalent rotation vector $\hat{\theta}_{0\to1}$. The error in the estimate of the rotation vector, therefore, is described $E_{\theta x}$, $E_{\theta y}$, and $E_{\theta z}$, the RMS errors associated with each rotational component. The RMS error in the estimate of the point-of-gaze is denoted by E_P. The RMS errors E_{ty}, E_{tz}, $E_{\theta x}$, $E_{\theta y}$, $E_{\theta z}$, and E_P are found using the method shown in (3.22) for E_{tx}.

Figure 3.5 shows the RMS error in the estimate of the translation (unit in mm) as a function of the standard deviation of the noise in the image coordinates of the coded

target centres. Likewise, Figure 3.6 shows the RMS error in the estimate of the rotation (unit in degrees). Note that noise with standard deviation of 0 corresponds to noise-free conditions.

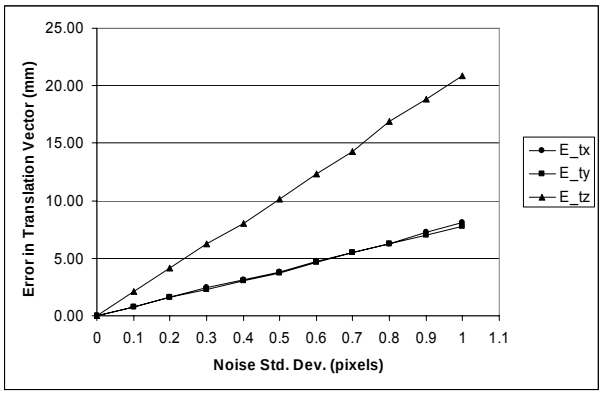

Figure 3.5: Error in the estimation of the translation vector.

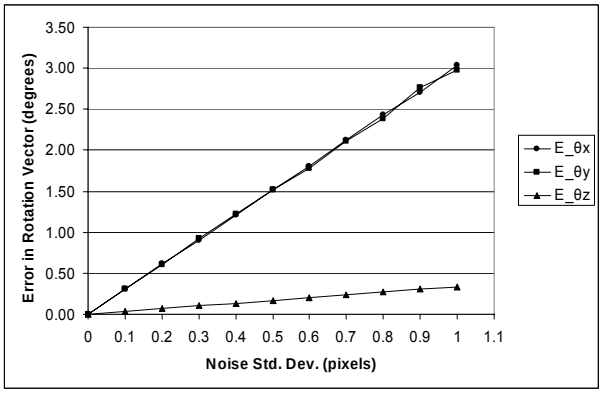

Figure 3.6: Error in the estimation of the rotation vector.

The results show the RMS error in the estimates of the motion parameters increases linearly with the standard deviation of the noise. With respect to the translation vector, the error is greatest in estimates of translation along Z_0 axis. At each noise level, E_{tz} is higher than E_{tx} and E_{ty}, which are approximately equal. With respect to the rotation vector, the error is smallest in estimates of rotation about the Z_0 axis. At each noise level, $E_{\theta z}$ is lower than $E_{\theta x}$ and $E_{\theta y}$, which are approximately equal. The errors with respect to

the Z_0 axis are different because this is also the principal axis of the scene camera. It is harder to estimate translations along the principal axis, as such translations result in the smaller changes in pixel coordinates than translations of the same magnitude along the X_0 or Y_0 axes. Conversely, it is easier to estimate rotations along the principal axis because such rotations result in greater changes in pixel coordinates than rotations of the same magnitude about the X_0 or Y_0 axes.

Figure 3.7 shows the RMS error in the point-of-gaze estimation. Note that in the simulation, we assume that $\mathbf{C}_0$ and $\mathbf{G}_0$ are error-free. Thus, the error in the point-of-gaze result solely from errors made in estimating $\mathbf{R}_{0\to1}$ and $\mathbf{T}_{0\to1}$.

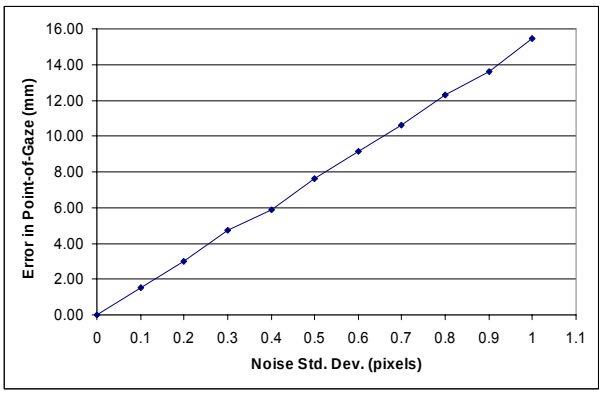

Figure 3.7: Error in the estimation of the point-of-gaze.

The results show that the error in the point-of-gaze increases linearly with the standard deviation of the noise. The results were obtained from simulations in which a subject fixates on $\mathbf{P'}_1 = [0, 0, 0]^T$ at t=1 for all trials. The simulations were repeated for other fixation points, and similar results were obtained.

Based on the aforementioned studies, we may reasonably expect the RMS error in the image coordinates of the circular coded targets to be approximately 0.1 pixels. For this noise level, the simulated RMS error in the point-of-gaze was 1.51 mm.

3.4 Summary

This chapter presented a methodology to estimate the rotation parameters of a 2D scene object. This development allows point-of-gaze to be estimated on moving reading material. Unlike the system described in Chapter 2, this system requires no head-mounted components, making it suitable for use in long reading tasks, and in applications involving children.

Within the context of reading assistance, the determination of the point-of-gaze $\mathbf{P'}_1$ allows the viewed word to be determined using the methodology presented in Section 2.2. The RMS error in $\mathbf{P'}_1$ resulting from the motion estimation varied linearly with the standard deviation of the noise in the measured image position of the circular coded targets. For noise having standard deviation of approximate 0.1 pixels, the simulated error in $\mathbf{P'}_1$ was comparable to the RMS mapping error of 2.29 mm reported in Chapter 2 using the head-mounted system. Thus, the proposed remote system has approximately the same resolution as the head-mounted system.

Chapter 4

Quantitative Criteria for Detecting Reading Difficulty

In the previous two chapters, methods have been presented to determine the viewed word. By monitoring the viewed word over time, eye movements during reading can be characterized on a per-word basis. In this chapter, quantitative criteria based on eye movements are developed to detect reading difficulty on a per-word basis.

Section 4.1 provides an overview of previous eye movement research in reading tasks. Section 4.2 presents the factors that affect reading performance along with a model of the reading process. Section 4.3 describes an experiment conducted to determine quantitative differences in eye movements when a reader encounters difficulty. Sections 4.4 through 4.7 propose and evaluate a method for detecting reading difficulty on a per-word basis. Within the context of a reading assistance application, such detection would enable the development of a need-activated response mechanism.

4.1 Basic Characteristics of Eye Movements in Reading

During reading, we continually make eye movements called saccades to bring new text under foveal inspection. Saccadic velocity is a function of saccade distance, and can reach velocities as high as 500° of visual angle per second (Rayner, 1998). A

typical saccade during reading covers a distance of 2° and takes around 30 ms (Abrams et al., 1989). In between saccades, our eyes remain relatively still during fixations of about 200-300 ms.

Other eye movements such as pursuit, vergence, and vestibular eye movements bear mentioning. Pursuit eye movements occur when the eyes follow a moving target; vergence eye movements occur when the eyes move to fixate on objects at different distances; finally, vestibular eye movements occur when the eyes move to compensate for head motion. However, these types of eye movements stabilize images on the retina during reading, while saccadic eye movements are greater in magnitude and are responsible for moving the point-of-gaze. For the reading tasks described in the remainder of this chapter, we will only consider saccadic eye movements.

Saccades bring new regions of text into foveal view for detailed analysis. Not all words are fixated during reading as foveal processing of every word is not necessary. The probability of a word being fixated increases with its length; words 2-3 letters in length are only fixated 25% of the time (Rayner and McConkie, 1976). Longer words are sometimes fixated more than once.

Although reading English text is performed left to right, 10-15% of saccades are made right to left, or upwards to previously read lines (Rayner, 1998). These saccades in the direction opposite of text flow are called regressions. Some regressions are made because the reader made an overshoot in a saccade, necessitating a short corrective movement to the left. In-word regressive saccades may indicate difficulty processing the fixated word. Longer saccades back to previous lines may indicate comprehension difficulty.

4.2 Processing Time During Reading

Reading performance is affected by a number of factors, the most important of which are the reader's skill level, and, conversely, the difficulty of the text. For difficult text, the reader's fixation durations increase, saccade lengths decrease, and frequency of regressions increases (Rayner, 1998). Within a body of text, there are also variations in processing times for individual words. To understand the factors that cause these variations, one must rely on a model of the reading process.

4.2.1 Dual-Route Reading Model

The most popular theory of visual word recognition is the dual-route theory. According to this theory, reading involves two separate processing routes, one to read irregular words such as "yacht" and one to read pronounceable non-words such as "yat." For regular words, these two processes compete to produce word recognition. A model based on the dual-route theory is presented in Figure 4.1 (Coltheart et al., 2001).

The first stage of reading is associated with orthographic analysis, whereby information regarding a word's length, shape, and letter units is extracted from the text. Typographical factors such as print quality, font face, line spacing, letter spacing, and line length may affect this analysis (Kolers et al., 1981; Morrison and Inhoff, 1981). However, under normal reading conditions, the typographical effect should be small. After orthographic analysis, the model diverges, splitting off into the lexical route and the non-lexical (sub-lexical) route. In the lexical route, the word's letter units are processed visually in parallel and a match is found within the reader's orthographic lexicon, containing all words the reader knows. In effect, the word is recognized as a

single unit, rather than through an examination of its individual letters. A mapping from the orthographic lexicon to the phonological lexicon, i.e. from a word image to a pronunciation, is then performed. When there is a one-to-many relationship between word image and pronunciation, such as for the word "read" which has two pronunciations depending on its tense, the semantic system disambiguates on the basis of meaning and context.

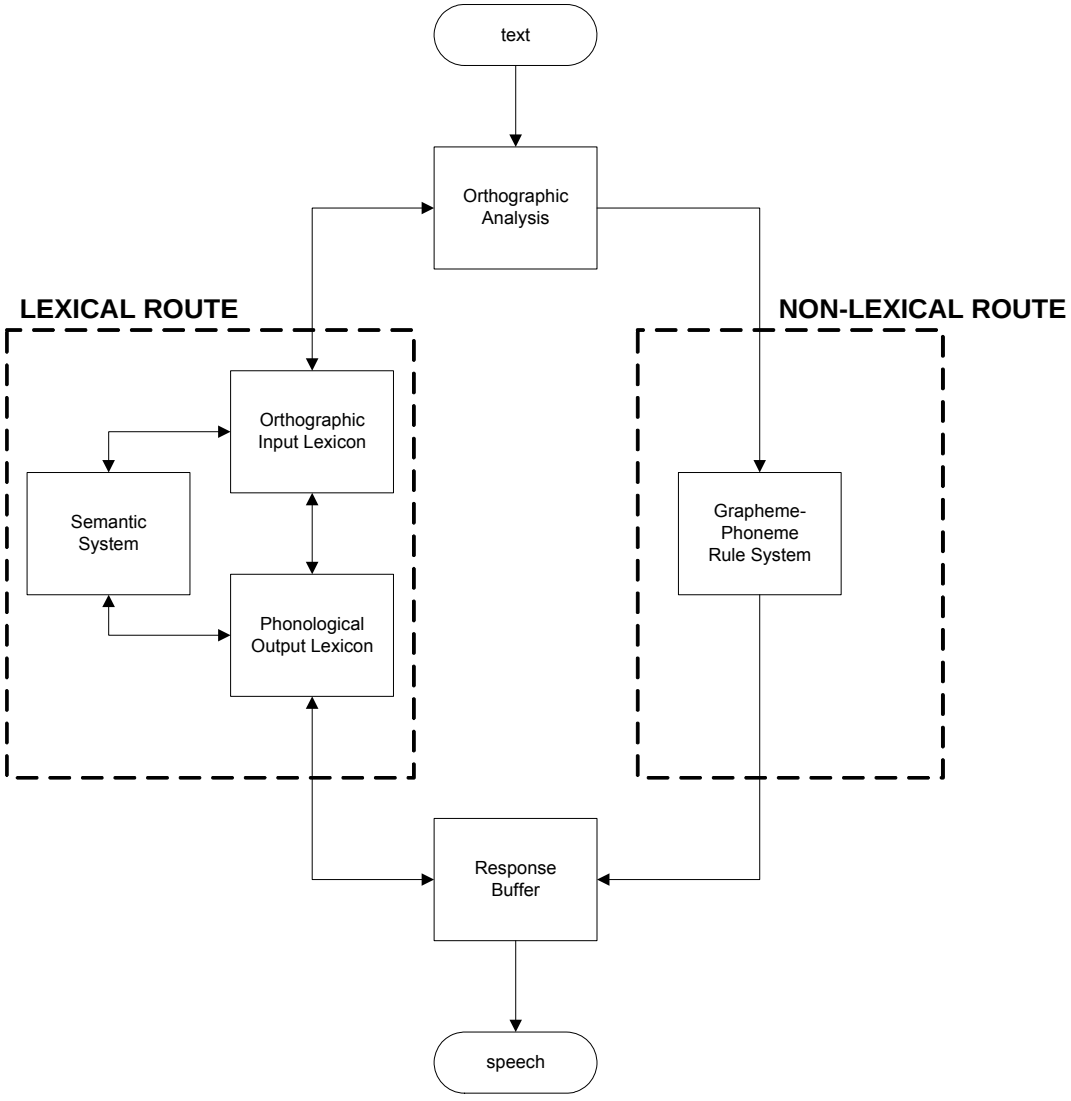

Figure 4.1: Model of dual-route cascaded model of visual word recognition and reading aloud (adapted from Coltheart et al., 2001).

In the non-lexical route, a letter string is converted into a phoneme string via a grapheme-to-phoneme conversion. A phoneme is the smallest unit of sound by which words can be distinguished; hence, changing one phoneme in a word yields a different word. A grapheme is the letter or set of letters that represent a particular phoneme. This conversion is performed one grapheme at a time, and requires a sequential examination of the letters in the word. The conversion is performed based on a set of mapping rules. For example, the grapheme c in the word "cat" is mapped to the phoneme /k/, while the same grapheme in "cent" is mapped to the phoneme /s/.

The bidirectional arrows in the lexical-route are significant and represent the positive feedback that is provided to prior stages. For example, production of an output phoneme in the response buffer induces an excitation effect in phoneme searching within the phonological output lexicon. This positive feedback may propagate all the way back to the orthographic analysis stage. Similar feedback effects have been proposed for the non-lexical route, but have not yet been successfully modeled.

The lexical route is faster since each word is recognized as a whole, rather than through an examination of its constituent letters. The speed of the lexical route depends on how often the reader had previously encountered the word. Commonly encountered words are recognized more quickly. Hence, the frequency of the word affects how long a reader spends looking at the word; this has been described as the "frequency effect" (Inhoff, 1984; Inhoff and Rayner, 1986; Just and Carpenter, 1980). When the word is encountered at "low-frequency", the non-lexical route may be faster. When the word is unknown to the reader, and thus not within the reader's lexicon, the lexical route fails.

Not all English words conform to standard grapheme-to-phoneme mapping rules, and the non-lexical route fails to produce the proper phoneme string in these cases.

After a phoneme has been activated, it is placed in the response buffer. This response buffer has a finite size and imposes a limit on the reading rate. The rate at which output from the response buffer can be consumed depends on whether the reading is performed silently or aloud. When silently reading text suitable for their reading ability, skilled readers are able to read at a rate in excess of 300 words per minute (wpm). When reading aloud, reading rate is limited by the speed of the vocalization. The eyes tend to move ahead of the voice, and the reader must periodically wait for the voice to "catch up" (Rayner, 1998). Hence, reading speed depends on whether the reading is performed silently or aloud.

Based on the model described above, reading performance is a function of how quickly the lexical and non-lexical routes can activate phoneme generation. The performance of the lexical route depends on the size of the reader's lexicon and how quickly that lexicon can be searched, while the performance of the non-lexical route depends on the reader's ability to apply acquired grapheme-to-phoneme conversion rules. For "low-frequency" words, readers will tend to fall back upon the non-lexical route, which increases processing time. This processing time is further increased if the word is long, or if the reader is unable to efficiently apply known grapheme-to-phoneme rules due to word irregularity. By measuring the processing time of a word, we can attempt to detect when the lexical route fails, and hence determine when a reader encounters an unfamiliar word.

4.2.2 Measuring Processing Time

Eye movements provide an efficient and objective measure of reading performance on a per-word basis. The two most frequently used measures are "gaze duration" and "first fixation duration" (Rayner, 1998). Gaze duration is the sum of all fixations made upon a word prior to a saccade to another word. First fixation duration is the duration of the first fixation upon a word, regardless of how many subsequent fixations are made upon that word. Results of past studies show that first fixation duration and gaze duration yield similar results.

When using these measures to quantify processing time, it is important to note that fixation times do not directly correspond to the underlying cognitive processes. Parafoveal processing (Fisher and Shebilske, 1985), whereby clues regarding word shape and the length of the next word to be processed are acquired using peripheral vision, tend to result in slight underestimation of the processing time. Saccadic latency (Abrams and Jonides, 1998; Rayner et al., 1983), associated with the time required to program and activate the next saccade, tend to result in slight overestimation of the processing time. Nevertheless, fixation times provide a reliable and convenient means to indirectly quantify cognitive processes.

4.3 Reading Task Experiment

4.3.1 Objective

An experiment was conducted to observe eye movements during a reading task. The objective of the experiment was to determine quantitative differences in eye fixations when a reader encounters difficulty reading a word. Within the context of a

reading assistance application, detection of such differences on a per-word basis would enable reading assistance to be rendered when needed.

Readers are expected to experience difficulty when encountering a word that they do not recognize (low-frequency word), and hence must rely on the non-lexical reading route. We hypothesized that this difficulty would result in a longer gaze duration upon the word being read, and would also be characterized by a higher rate of in-word regressions. From these quantitative differences, a set of criteria can be established to automatically determine reader difficulty on a per-word basis.

4.3.2 Apparatus

The reading material used in this experiment consisted of short passages of text, between twenty and fifty words in length. The passages were selected from English literature such that each passage contained at least one low-frequency word, varying from 5 letters to 12 letters in length. Word frequency was measured using the British National Corpus (BNC) metric (Leech, 2001), which counts the number of word occurrences within a database of 100 million words sampled from present-day English. For example, "the"—the most common word in the English language—has a BNC Frequency of 6187267. This can be expressed as an occurrence rate of 6.187267% within the corpus. A low-frequency word was defined as one having a BNC Frequency less than 100. A sample passage follows: "Perhaps most amazingly, votaries of diversity insist on absolute conformity." Table 4.1 shows the corresponding BNC Frequencies of each word from the sample passage. For this text, "votaries" is the low-frequency word expected to be unknown to the reader.

Table 4.1: BNC Frequency of Words in a Sample Passage

Word	BNC Frequency
Perhaps	35039
Most	54966
Amazingly	377
Votaries	4
Of	2941444
Diversity	1387
Insist	881
On	647344
Absolute	3480
Conformity	473

The experiment was conducted using a standard desktop PC operating the remote point-of-gaze estimation system described in Chapter 3. Point-of-gaze was estimated with respect to a static scene (a computer monitor); hence, the proposed modifications for performing an experiment using moving reading material were not needed. The remote system was used instead of the head-mounted system to obtain point-of-gaze estimates free of errors caused by movement of the head or the reading material. To this end, a chin-rest was used to stabilize the head. Point-of-gaze was estimated and recorded at 30 Hz.

Subjects were positioned at a distance of 65 cm from a 19" computer monitor; text was presented on the monitor with a character height of 1.5 cm, subtending a visual angle of 1.3°. The character height was chosen to accommodate the point-of-gaze estimation system, which is accurate to within 1° of visual angle under normal operating conditions. The selected text size can be read comfortably by individuals with 20/20 vision. When visual acuity is not a factor, text size does not affect reading speed, as the

number of letter spaces traversed by saccades is largely invariant to text size (Morrison and Rayner, 1981).

4.3.3 Methodology

Four subjects participated in the experiment. Subjects all possessed university-level reading ability. All four subjects had normal visual acuity (20/20) or achieved equivalent visual acuity using corrective lenses.

Forty passages were presented to each subject. Subjects read twenty passages silently and twenty passages aloud in alternating fashion to reduce biasing effects of fatigue. To alleviate fatigue, subjects were given brief rest periods between each set of ten passages. Subjects were asked to disregard text comprehension during the reading task.

After each passage was read, subjects were asked to name the words they did not recognize. These unknown words were foreign to the orthographic lexicon of the subject. Based on the dual-route model of reading, subjects processed these words using the non-lexical reading route. We assume known words were processed using the faster lexical reading route. Using this information, words fixated upon during reading can be labelled as either "known" or "unknown." Therefore, words from the forty passages can be divided into four sample populations: silent known, silent unknown, aloud known, and aloud unknown. By analyzing these four sample populations separately, one can determine distinct fixations behaviours that are associated with each population.

4.3.4 Differences in Performance between Silent and Aloud Reading

As shown in Table 4.2, aloud reading was slower than silent reading for all subjects. Analysis of reading performance with respect to fixation time and fixation rate (fixations per word) confirmed some results from previous studies (Rayner, 1984; 1998). For aloud reading, fixation rate was higher, while fixation duration was approximately the same or higher. The high variability in fixation time and fixation rate between subjects is expected; these two factors combine to produce variations in reading speed. It is interesting to note that Subject V.S. had the longest fixation times, but achieved the fastest reading speed by also having the lowest fixation rate, suggesting the use of more efficient saccadic programming.

The results show that eye fixation behaviour differs between silent and aloud reading. Since the reading assistance system should operate for both silent and aloud reading, we will examine these two reading modes separately in the sections to follow.

Table 4.2: Reading Speed and Fixation Differences between Silent and Aloud Reading

	Reading Speed (wpm)		Fixation Time (ms)				Fixation Rate (fpw)	
			Silent		Aloud			
Subject	Silent	Aloud	Mean	S.D.	Mean	S.D.	Silent	Aloud
E.G.	165	142	228	122	221	122	1.35	1.57
M.E.	135	115	222	131	238	146	1.63	1.80
P.L.	186	134	254	151	266	164	1.08	1.43
V.S.	219	134	277	117	347	184	0.87	1.14

4.3.5 Differences in Gaze Durations between Known and Unknown Words

To analyze reading performance on a per-word basis, the gaze duration for each word was calculated. Some words were viewed more than once via regressions; for unknown words, only the first viewing was considered under the assumption that word are no longer unfamiliar after they have been processed once. The mean gaze duration of unknown words was compared to the mean gaze duration of known words. Since processing time is affected by word length (Trueswell, Tanenhaus and Garnsey, 1994), we examine words of different length separately. Figures 4.2 and 4.3 show mean gaze duration as a function of word length for each subject during silent and aloud reading, respectively. The error bars in Figures 4.2 and 4.3 represent one sample standard deviation. Consecutive estimates of the means and standard deviations for unknown words show high variability because the sample size of each set of unknown words is small (less than five words). Nevertheless, the results show that mean gaze durations for known words are shorter than mean gaze durations for unknown words, reflecting faster processing of known words.

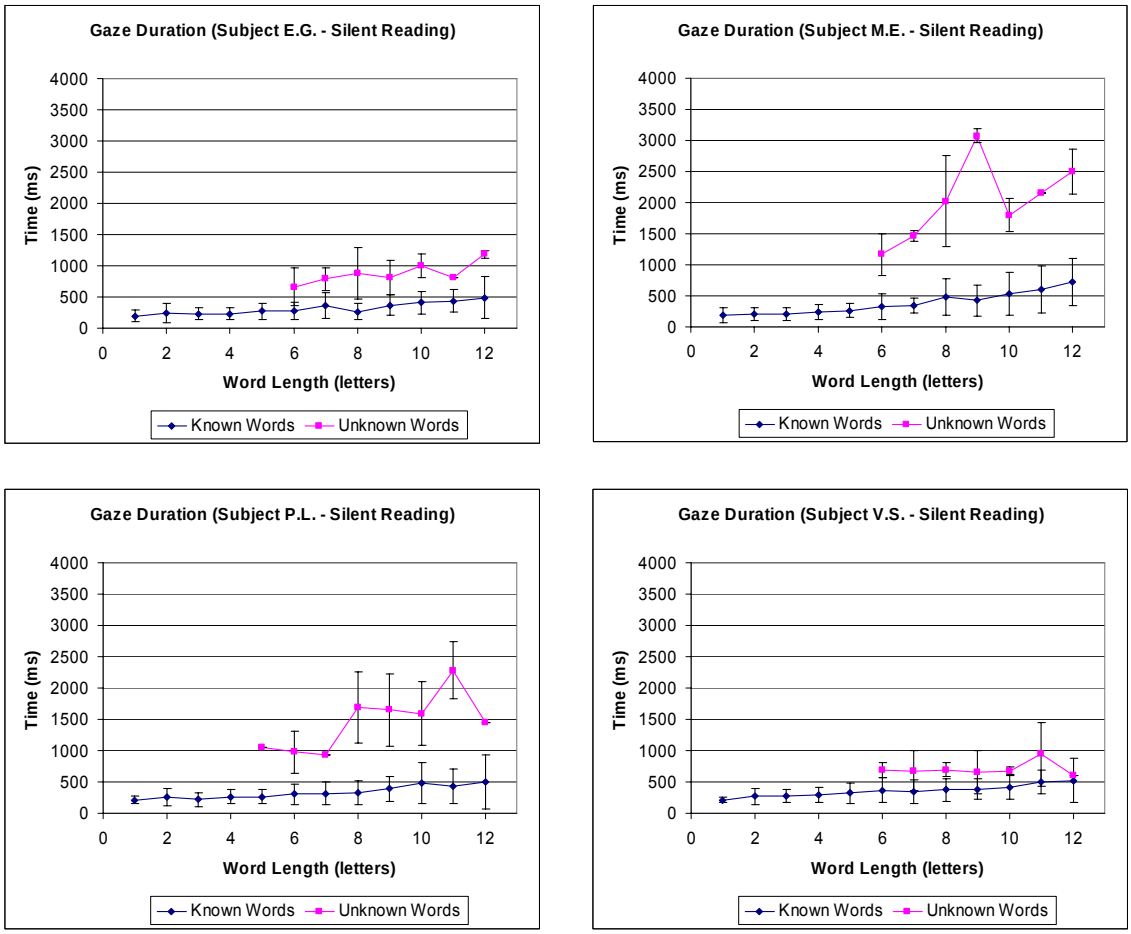

Figure 4.2: Gaze duration as a function of word length for known and unknown words during silent reading.

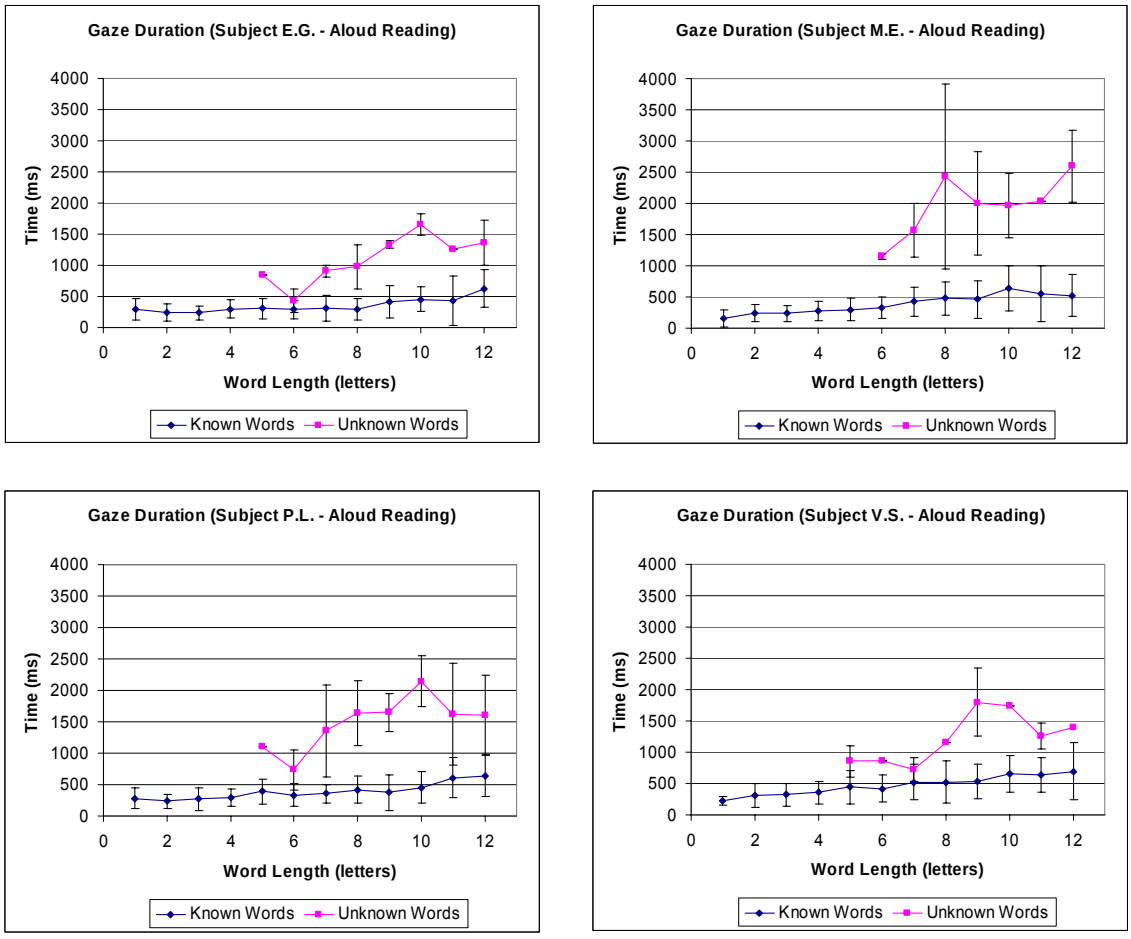

Figure 4.3: Gaze duration as a function of word length for known and unknown words during aloud reading.

4.3.6 Differences in In-Word Regressions between Known and Unknown Words

The recorded fixations were analyzed to determine regression behaviour during the reading tasks. A regression is defined as a saccade made by the reader to the left, or upward to a previous line of text. An in-word regression is a saccade made to an earlier portion of the same word. Prior studies suggested that in-word regressions may indicate that the reader is experiencing difficulty processing the fixated word (Kennedy and Murray, 1987). The rates of in-word regressions are summarized in Tables 4.3 and 4.4.

Table 4.3: In-Word Regression Rate for Silent Reading

Subject	Known Words			Unknown Words		
	Words	Regressions	Regression Rate	Words	Regressions	Regression Rate
E.G.	608	23	0.04	26	10	0.38
M.E.	712	41	0.06	24	10	0.42
P.L.	669	15	0.02	24	6	0.25
V.S.	676	8	0.01	21	5	0.24
Total	2665	87	0.03	95	31	0.33

Table 4.4: In-Word Regression Rate for Aloud Reading

Subject	Known Words			Unknown Words		
	Words	Regressions	Regression Rate	Words	Regressions	Regression Rate
E.G.	689	28	0.04	17	7	0.41
M.E.	623	41	0.07	24	15	0.63
P.L.	509	19	0.04	25	5	0.20
V.S.	686	17	0.02	20	5	0.25
Total	2507	105	0.04	86	32	0.37

The results show that the in-word regression rate was ten times higher for unknown words, for both silent and aloud reading. This confirms expectations, since we expect unknown words to require re-examination more often, and this is facilitated by in-word regressions. However, not all unknown words required an in-word regression to process. This is clearly the case for short words, since they can be fully re-examined without using in-word regressions. The in-word regression rate also varies greatly between subjects. Subject M.E. showed the highest in-word regression rate at 0.63 during aloud reading. In contrast, subject P.L. exhibited an in-word regression rate of only 0.20 during aloud reading.

4.3.7 Detecting Reading Difficulty on a Per-Word Basis

The results from the above experiments show that readers exhibit differences in fixation behaviour when reading unknown words. These differences are measurable, and

have been quantified as reading with longer gaze durations and a higher rate of in-word regressions. However, even though the rate of in-word regressions is much higher, most unknown words do not require an in-word regression to process. Hence, the solitary use of in-word regressions does not provide a reliable means to detect reading difficulty. We consider the use of a combined detection criterion using both in-word regressions and gaze durations. We evaluated the merits of such an approach by examining the measured gaze durations for words processed with and without using an in-word regression. These gaze durations are reported in Tables 4.5 and 4.6 for silent and aloud reading, respectively.

Table 4.5: Gaze Durations for Words Processed With and Without Using an In-word Regression for Silent Reading

	Gaze Duration			
	No In-word Regression (ms)		In-Word Regression (ms)	
Subject	Mean	S.D.	Mean	S.D.
E.G.	284	187	540	280
M.E.	302	231	1099	811
P.L.	322	271	1159	901
V.S.	332	177	664	249
Total	308	222	839	657

Table 4.6: Gaze Durations for Words Processed With and Without Using an In-word Regression for Aloud Reading

	Gaze Duration			
	No In-word Regression		In-Word Regression	
Subject	Mean	S.D.	Mean	S.D.
E.G.	295	193	591	381
M.E.	328	270	1126	934
P.L.	398	344	733	559
V.S.	448	283	1041	437
Total	362	278	818	655

The results in Tables 4.5 and 4.6 show that the mean gaze duration is much higher when an in-word regression is used to process words. An in-word regression

naturally increases gaze duration by increasing the number of fixations. We conclude that a combined detection criteria using in-word regressions and gaze duration is unlikely to provide any additional information over the use of gaze durations alone since the two metrics are highly correlated.

If gaze duration is used as the sole detection criterion, a threshold can be set to define the maximum gaze duration expected during the processing of a known word. Reading difficulty is detected when any gaze duration exceeds this threshold. In the next section, we will discuss a method for determining this threshold.

4.4 Proposed Detection System

As reading speed is a function of reader skill and text difficulty, the appropriate gaze duration threshold must vary with the subject and the text. Furthermore, the threshold should depend on the length of the word being examined. We can apply classical detection theory to set an appropriate threshold for detecting unknown words.

For each word length, gaze duration may be modeled by two Gaussian processes: $N(\mu_k, \sigma_k^2)$ for known words and $N(\mu_u, \sigma_u^2)$ for unknown words, where $\mu_u > \mu_k$. The observed gaze duration, r, is modeled as a random variable that can be classified using a binary hypothesis test. The two hypotheses are:

$$H_0 : r \sim N(\mu_k, \sigma_k^2), \tag{4.1}$$

$$H_1 : r \sim N(\mu_u, \sigma_u^2). \tag{4.2}$$

A standard Bayes test is inappropriate for this problem since we do not know the a priori probabilities of an unknown word and a known word being encountered. Furthermore, the costs associated with detection, rejection, false alarm, and miss are

application and subject dependent. Hence, it is appropriate to use the Neyman-Pearson Criterion to design a test that maximizes the probability of detecting an unknown word while constraining the probability of a false alarm. In the context of a reading assistance application, a false alarm triggers assistance when it is not needed. This type of interruption slows the reading process.

The probability of a false alarm is a conditional probability given by:

$$P_F = P\{r \geq T|H_0\}, \tag{4.3}$$

where T is the threshold above which an unknown word is detected; T varies with word length. Since H_0 states that r is a Gaussian random variable with mean μ_k and variance σ_k, the conditional density of r can be expressed as

$$P\{r = R|H_0\} = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(\frac{-(R - \mu_k)^2}{2\sigma_k^2}\right). \tag{4.4}$$

Hence, from (4.3) and (4.4), it follows that the probability of a false alarm is

$$P_F = \int_T^{\infty} \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(\frac{-(R - \mu_k)^2}{2\sigma_k^2}\right)dR. \tag{4.5}$$

P_F is the area of the shaded region shown in Figure 4.4, and is a continuous function of T. For some specified probability of false alarm, T can be expressed as:

$$T = \sigma_k\sqrt{2}\,\mathrm{erf}^{-1}(1 - 2P_F) + \mu_k. \tag{4.6}$$

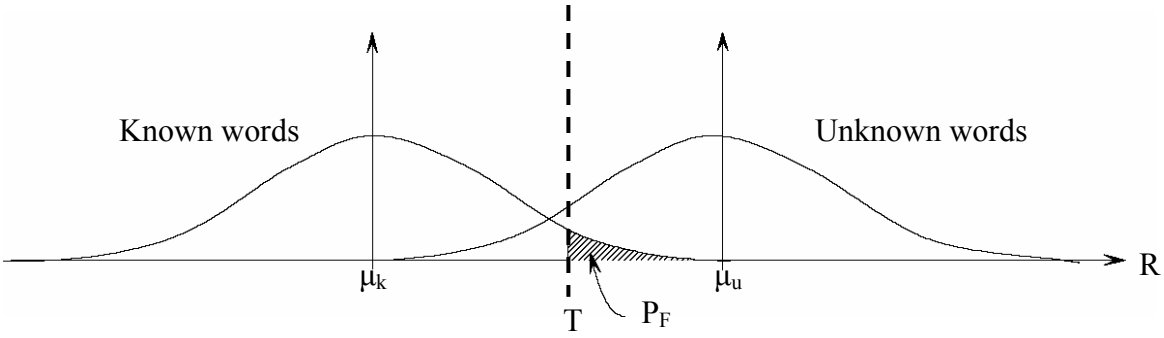

Figure 4.4: Probability of a false alarm, P_F, in the binary hypothesis test.

The calculation of T does not depend on μ_u or σ_u. This is advantageous because μ_u and σ_u are more difficult to estimate than μ_k and σ_k due to the smaller sample sizes of unknown words. We can consider this approach as using μ_k and σ_k to characterize a subject's normal reading behaviour, and detecting deviations from this behaviour.

The choice of P_F is application dependent; the effect of this selection on detection performance is discussed in the section to follow. Figure 4.5 shows two examples of threshold curves for P_F equal to 0.01 (left) and 0.1 (right). Each point on the threshold curve is calculated independently for each word length using (4.6).

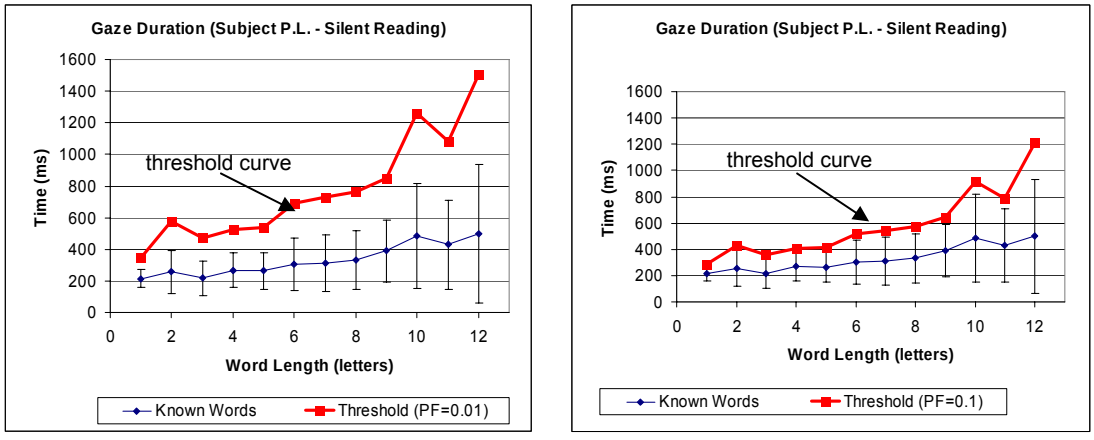

Figure 4.5: Example thresholds for a P_F values of 0.01 (left) and 0.1 (right).

4.5 Training Set Detection Performance

In this section we estimate the statistics of the distribution of gaze durations for known words (μ_k and σ_k^2) for each subject, and calculate the detection thresholds by equation (4.6). We consider this set of gaze durations measured in the experiment described in Section 4.3 to be the "training set" used by the detection system to learn subject-specific reading behaviour. By applying the detection thresholds on this set of gaze durations, we characterize the detector's training set detection performance.

We have assumed for convenience that the process of reading known words is Gaussian, while literature suggests that gaze durations are more accurately modeled as Gamma processes (Rayner, 1998). By comparing the measured false alarm rate to P_F, the specified probability of false alarm, we can evaluate the error introduced by using a Gaussian model. Figure 4.6 shows the measured false alarm rate for each subject under both silent and aloud reading as a function of P_F. The theoretical false alarm rate should have a slope of 1, and pass through the origin.

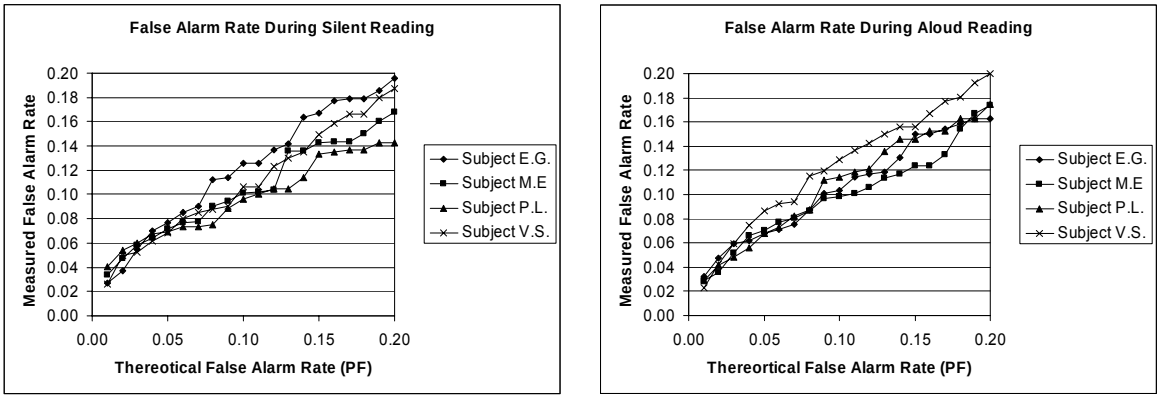

Figure 4.6: False alarm rate as a function of P_F for silent reading (left) and aloud reading (right).

The measured results follow theoretical expectations. The slope of each curve is approximately 1. Although the results vary between subjects, for P_F between 0.05 and 0.10, the false alarm rates closely match P_F. For example, for P_F equal to 0.10, the false alarm rate was between 0.095 and 0.125 for all subjects. We conclude that the Gaussian model of gaze durations for normal words does not introduce significant error in the false alarm rates.

We next evaluate the detection rate of unknown words. Figure 4.7 shows the detection rate for unknown words for each subject as a function of P_F, under both silent and aloud reading conditions.

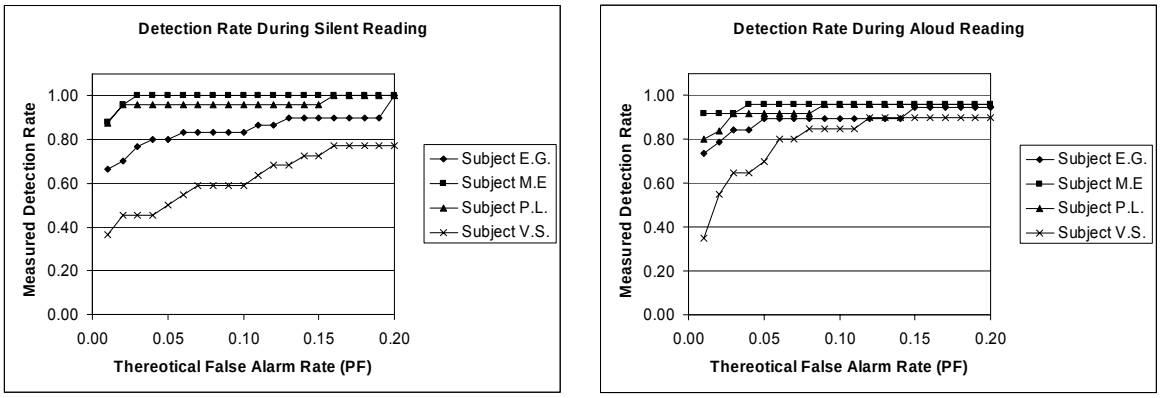

Figure 4.7: Detection rate as a function of P_F for silent reading (left) and aloud reading (right).

As expected, the results in Figure 4.7 show that as P_F increases, the detection rate increases. In other words, if we permit a higher false alarm rate, we can achieve a higher detection rate. However, there is considerable variation, from subject to subject and between aloud and silent reading, in the manner by which the detection rate increases. The detection rate is inversely correlated with the distance between the mean gaze duration of known words and the mean gaze duration of unknown words, as shown in

Figure 4.3. For example, the smallest distance between the two means was found for Subject V.S.; accordingly, the detection rate was lowest for this subject for all values of P_F. This result is expected since smaller distances between the means of the two distributions (for known and unknown words) result in more distribution overlap and a higher Bayes classification error. For this reason, detection performance is better for aloud reading than for silent reading. Under aloud reading conditions, for a P_F value of 0.10, the measured detection rate was greater than 0.80 for all subjects.

4.6 Approximating the Detection Threshold

Recall that the detection threshold curve consists of a set of threshold (T) values. Each threshold value specifies the gaze duration, which, if exceeded, causes words of a specific length to be classified as an unknown word. Each value is calculated independently using the observed mean (μ_k) and variance (σ_k^2) of the gaze durations recorded for words of that length. However, if each threshold value is a function of word length, then the values are clearly not independent. This suggests that the detection threshold may be approximated by some function of word length. This would allow each set of threshold values to be described in a more succinct manner. Furthermore, thresholds could then be estimated for words of arbitrary length. Figure 4.8 shows the detection thresholds for each subject under aloud and silent reading conditions, calculated for P_F equal to 0.10.

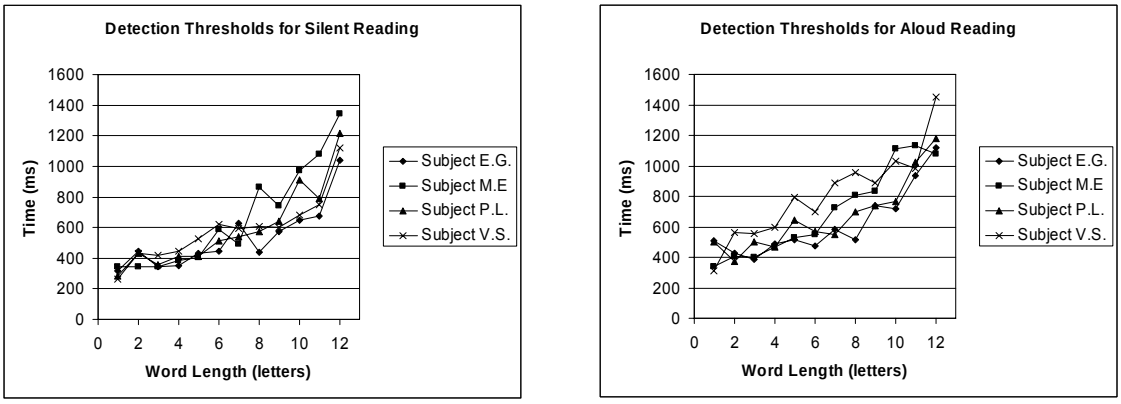

Figure 4.8: Detection thresholds for silent reading (left) and aloud reading (right).

The shapes of the detection threshold curves suggest that a first-order or second-order approximation is appropriate. For the sake of simplicity, we try a first-order approximation, and then evaluate its validity. Consider the following approximation of the threshold curve:

$$\tilde{T}(L) = B \times L + A, \qquad (4.7)$$

where $\tilde{T}$ is the approximated threshold and L is the word length. The parameters A and B are the coefficients of a least-squares regression line fitted to the threshold values. Figure 4.9 shows the first-order approximations of the detection threshold shown in Figure 4.8.

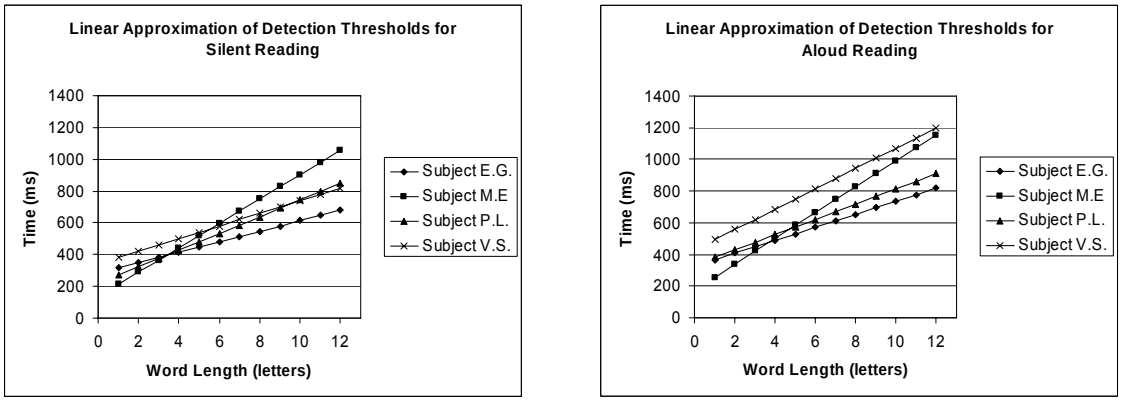

Figure 4.9: First-order approximations of the detection thresholds for silent reading (left) and aloud reading (right).

The validity of the first-order approximation can be evaluated by examining detection performance obtained using the approximated thresholds. The performance characteristics of the approximated thresholds are shown in Figure 4.10, which shows the false alarm rate, and Figure 4.11, which shows the detection rate.

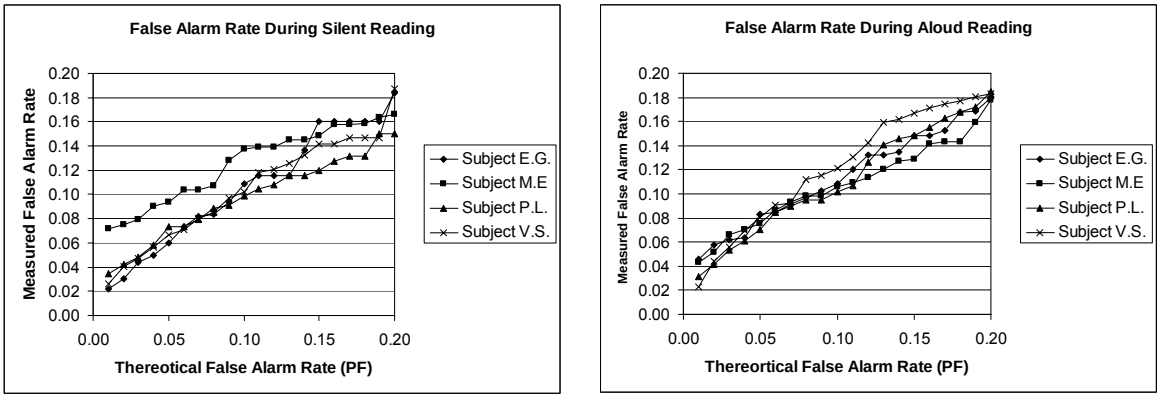

Figure 4.10: False alarm rate as a function of P_F when using the approximated detection threshold for silent reading (left) and aloud reading (right).

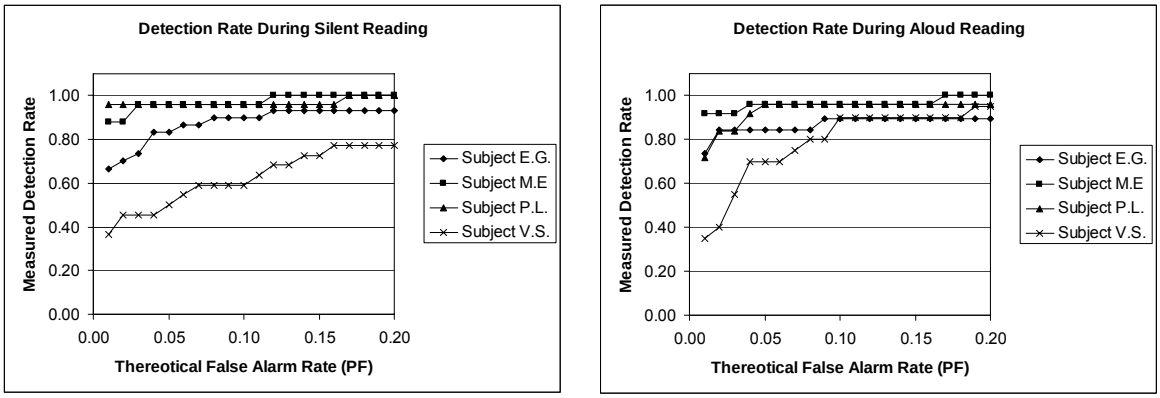

Figure 4.11: Detection rate as a function of P_F when using the approximated detection threshold for silent reading (left) and aloud reading (right).

The results shown in Figures 4.11 and 4.12, obtained using the approximated thresholds, can be compared directly to the results shown in Figures 4.6 and 4.7, obtained using the exact thresholds. The magnitudes of the changes in the false alarm rates do not exceed 0.04 under any conditions. The magnitudes of the changes in the detection rates are generally within 0.05. The results show that the use of the first-order approximation for the detection threshold does not substantially affect detection performance. This approximation allows each detection threshold to be described using only two parameters, while allowing the threshold to be estimated for words of arbitrary length.

4.7 Test Set Detection Performance

In this section, we apply the detection thresholds obtained from the training set on a new test set. Gaze durations were collected in the manner described in Section 4.3 for an additional twenty passages read aloud and twenty passages read silently. Detection performance for the new test set of gaze durations are compared to the training set results

of Section 4.6. Detection rate and false alarm rate for both the training set and the test set are summarized in Tables 4.7 and 4.8 for silent and aloud reading, respectively.

Table 4.7: Detection Performance in Simulated Reading Assistance Application for Silent Reading

Subject	Detection Rate		False Alarm Rate	
	Training Set	Test Set	Training Set	Test Set
E.G.	0.90	0.78	0.11	0.10
M.E.	0.96	1.00	0.14	0.16
P.L.	0.96	0.93	0.10	0.08
V.S.	0.64	0.74	0.10	0.12
Total	0.87	0.86	0.11	0.12

Table 4.8: Detection Performance in Simulated Reading Assistance Application for Aloud Reading

Subject	Detection Rate		False Alarm Rate	
	Training Set	Test Set	Training Set	Test Set
E.G.	0.89	0.86	0.11	0.09
M.E.	0.96	1.00	0.11	0.12
P.L.	0.96	0.94	0.10	0.10
V.S.	0.90	0.86	0.12	0.11
Total	0.93	0.92	0.11	0.10

The detection rate and the false alarm rate show high correlation between the training sets and the test sets. In general, the false alarm rate closely adheres to the specified P_F value of 0.10 for both sets. The detection rate exhibits more variation from subject to subject; however, this variation is to be expected since the detection threshold only controls the false alarm rate and not the detection rate. The results suggest that the detection threshold obtained from a small training set of gaze durations can be successfully applied to gaze durations measured when reading new text.

4.8 Detection Performance Based on Number of Words Read

In the previous sections we have discussed detection rates and false alarms rates in terms of the percentage of gaze durations which exceed the detection threshold. For the sake of brevity in the following discussion we will refer to each set of fixations for which each gaze duration is measured as a "gaze." During the reading tasks, the correspondence between words read and gazes was not one-to-one. In this section, we normalize detection rates and false alarm rates based on the number of words read.

We examine the relationship between gazes and words separately for known and unknown words. In the case of known words, many short low-content words were processed parafoveally (no recorded gazes), while other words were read more than once (multiple recorded gazes). By comparison, all unknown words were fixated upon. Furthermore, when an unknown word was read more than once, we only considered the first-pass gaze, based on the assumption that the word was no longer unfamiliar to the subject upon subsequence re-processing. Table 4.9 reports the gaze rate (gazes per word) for known words calculated from the gaze durations measured in the above experiment. Note that the gaze rate for unknown words, based on the discussion above, is always 1.

Table 4.9: Gaze Rate for Known Words During Silent and Aloud Reading

	Silent Reading			Aloud Reading		
Subject	Gazes	Words	Gaze Rate (gpw)	Gazes	Words	Gaze Rate (gpw)
E.G.	580	634	0.91	652	668	0.98
M.E.	479	509	0.94	606	544	1.11
P.L.	433	644	0.67	470	646	0.73
V.S.	411	654	0.63	384	513	0.75

In general, the gaze rate is lower for silent reading than for aloud reading. Furthermore, the gaze rate varies significantly between subjects. The detection rates and false alarm rates presented in Tables 4.7 and 4.8 were normalized based on the number of words read. The normalized word-based rates are presented Tables in 4.10 and 4.11 alongside the old gaze-based rates, for silent and aloud reading, respectively.

Table 4.10: Detection Performance Based on Number of Words and Number of Gazes for Silent Reading

	Detection Rate		False Alarm Rate	
Subject	Gaze-Based	Word-Based	Gaze-Based	Word-Based
E.G.	0.78	0.78	0.10	0.09
M.E.	1.00	1.00	0.16	0.15
P.L.	0.93	0.93	0.08	0.06
V.S.	0.74	0.74	0.12	0.08
Total	0.86	0.86	0.12	0.09

Table 4.11: Detection Performance based on Number of Words and Number of Gazes for Aloud Reading

	Detection Rate		False Alarm Rate	
Subject	Gaze-Based	Word-Based	Gaze-Based	Word-Based
E.G.	0.86	0.86	0.09	0.09
M.E.	1.00	1.00	0.12	0.13
P.L.	0.94	0.94	0.10	0.07
V.S.	0.86	0.86	0.11	0.09
Total	0.92	0.92	0.10	0.09

Since the gaze rate for unknown words is always 1, the detection rate does not change after normalization. However, the gaze rates for known word tends to be less than 1 (there are fewer gazes than words). Hence, the normalized word-based false alarm rate tends to be less than the gaze-based false alarm rate. Detection performance normalized by the number of words read provides a more accurate comparison of results between different subjects, and between aloud and silent reading.

4.9 Summary

In this chapter, the measurement of eye movements was introduced as an indirect method of estimating per-word processing time. It was hypothesized that by measuring gaze duration on the viewed word we can detect when a reader encounters an unknown word. In the context of the dual-route reading model, this difficulty represents a failure in the lexical reading route. It was shown that the mean gaze duration used to process an unknown word is greater than the mean gaze duration used to process a known word of the same length. Based on these experimental results, a method for detecting when a reader encounters an unknown word was designed based on the principle of constraining the false alarm probability. The detection system was validated for both aloud and silent reading tasks. Using detection thresholds calculated from a small training set for a false alarm probability of 0.10, a detection rate of 0.89 and a false alarm rate of 0.11 were obtained on a test set. The next chapter describes the use of the detection system within a reading assistance application that facilitates reading within a natural setting.

Chapter 5

Natural Setting Reading Assistance

The primary objective of this investigation was to develop the means to provide automated reading assistance in a natural setting. Chapters 2 and 3 described methods to determine the viewed word using head-mounted and remote point-of-gaze estimation techniques, respectively. Chapter 4 presented a means to detect reading difficult on a per-word basis. This detector and the viewed-word identification techniques developed in Chapter 2 were combined to implement a system to provide automated reading assistance while tolerating head and reading material movement.

Section 5.1 describes an experiment to verify the principle of operation of the system, and evaluates system performance. Section 5.2 summarizes the main contributions of this investigation and makes recommendations for future work.

5.1 Reading Assistance Experiment

5.1.1 Objective

The proposed automated reading assistance system was implemented by combining: (1) the system to monitor the viewed word using a head-mounted eye-tracker, and hence measure per-word processing time, and (2) the processing time thresholds for

detecting when the reader encounters an unknown word. The objective of the experiment described in this section was to verify the principle of a novel automated reading assistance system.

5.1.2 Apparatus

The reading material used in the experiment was composed of short passages of text, totalling 615 words. The passages were selected from English literature in the manner described in Section 4.3. The reading material was printed on standard letter-sized (8"×11.5") paper, spanning 64 bound single-sided sheets. Character height was 15 mm, allowing a maximum of 70 characters (5 rows of 14 characters) to be presented on each page. In addition to the text, each page contained four coded targets to facilitate tracking of reading material position and a barcode to encode the page number.

The head-mounted system described in Chapter 2 was used to perform point-of-gaze estimation, and identify, in real-time, the word viewed by the subject. A scene camera with a focal length of 4 mm and a field of view of 92.1° by 69.1° (horizontal by vertical) was used to image the scene in which the reading material was presented. At a distance of 50 cm, the reading material occupied approximately 20% of the field of view. This provides a region in which the reading material may move with respect to the head-mounted scene camera without leaving the camera's field of view. The size of reading material with respect to the field of view can be seen in Figure 5.1, in which a sample image captured by the scene camera is shown.

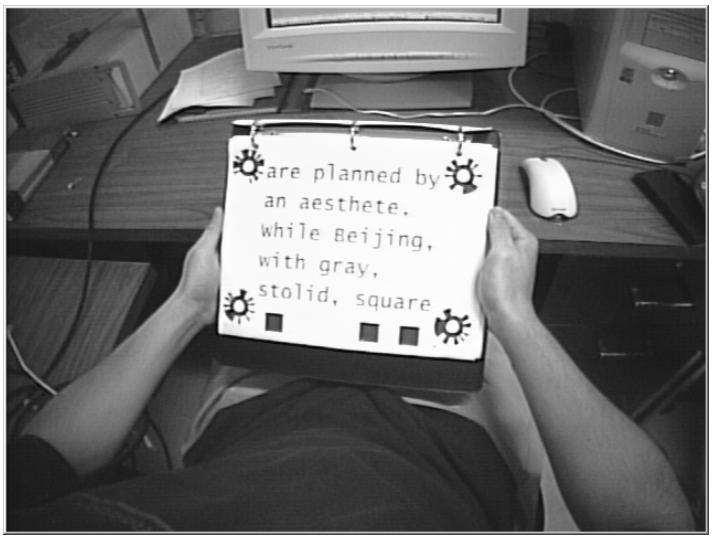

Figure 5.1: Image of the reading material used in the reading experiment captured by the scene camera.

When the position of the scene camera is fixed, the reading material may move within a region with dimensions of approximately 60 cm by 40 cm by 10 cm (horizontal by vertical by depth). The constraint in depth is imposed by the image processing used to track the four coded targets. When the position of the reading material is fixed, the range of tolerable head movement is approximately $\pm$ 20° by $\pm$ 15° (horizontal by vertical). Using this configuration, the system is expected to accommodate the head movements and reading material movements typically encountered during a reading task.

Customized monitoring software was used to calculate gaze durations and trigger assistance when gaze durations exceeded the specified threshold. Assistance was rendered in the form of computer vocalization of the viewed word. Computer vocalization was generated using the Lernout and Hauspie ® TruVoice text-to-speech engine, a state-of-the-art vocalization system for personal computers.

5.1.3 Methodology

Two subjects participated in this experiment; both had previously participated in the experiments described in Sections 4.3, and thus the appropriate detection thresholds were known.

While performing the reading task, subjects sat in a natural reading pose and held the reading material at a comfortable distance (approximately 50 cm). Head position was not restrained, allowing for natural head motions during reading tasks. These head motions typically consist of small changes in tilt and pan angle, and were expected to be within the tolerances described in Section 5.1.2. The reading material was held by the subject, and thus moved with the subject. This configuration simulates a natural reading setting. Figure 5.2 shows a subject in a typical reading pose while the system is in operation.

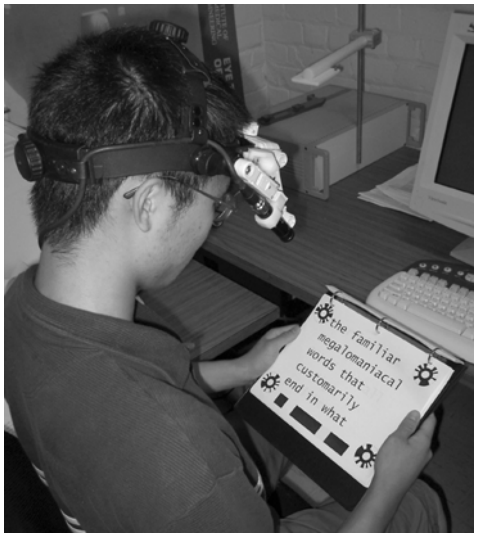

Figure 5.2: Photograph of a subject in a typical reading pose.

Subjects read the reading material aloud while per-word gaze duration was measured. The subject-specific thresholds for detecting unknown words were set based

on the linear threshold approximations described in Section 4.6. The threshold levels for each word length were selected for a false alarm probability of 0.10. When gaze duration for any word exceeded the length-appropriate threshold, the reading assistance system performed computer vocalization of the word. The system operates under the assumption that vocalization assistance is needed for all unknown words. At the end of each page, subjects were asked to identify any unknown words.

5.1.4 Results and Discussion

Detection performance achieved during operation of the automated reading assistance system is presented in Table 5.1. The detection rate corresponds to the percentage of unknown words for which the system provided needed vocalization assistance; while the false alarm rate corresponds to the percentage of known words for which the system provided unneeded vocalization.

Table 5.1: Natural Reading Setting Detection Performance

Subject	Detection Rate	False Alarm Rate
M.E.	0.94	0.10
P.L.	0.95	0.09

These results can be compared to those obtained in the evaluation of the detection system performed in Section 4.7, whereby gaze durations were measured using a remote point-of-gaze estimation system for fixed head and reading material positions. For the same two subjects on the aloud reading test set, the aggregate detection rate was 0.96 and the aggregate false alarm rate was 0.10. Detection performance during operation of the reading assistance system closely conformed to these earlier results. This suggests that

the proposed point-of-gaze mapping method accommodated the range of head and reading material movement without significantly reducing system performance.

5.2 Conclusions

5.2.1 Contributions of the Thesis

In developing an automated reading assistance system, some contributions have been made which are more generally applicable. The main contributions of this thesis are as follows:

- Development of a method to map point-of-gaze estimates obtained from a head-mounted eye-tracker to an object coordinate system attached to a moving 2D scene object. Using a set of four point correspondences, the mapping method was found to have an RMS error of 2.29 mm.
- Development of a method to estimate the motion of a 2D scene object within a scene when the intrinsic parameters of a statically-placed scene camera are known. The theoretical framework for extending the operation of an existing remote point-of-gaze estimation system from static 2D scene objects to moving 2D scene objects was developed. This enables the creation of a system to identify the viewed word on moving reading material using no head-worm components. Performance of the system was characterized through simulation. If four point correspondences between the scene image and the 2D scene object can be established with image noise having standard deviation of 0.1 pixels, then the proposed motion estimation method introduces a 1.51 mm RMS error to point-of-gaze estimates.

- Development of a method to detect when a reader encounters an unknown word. A method was developed to learn subject-specific reading behaviour using a small training set. This method establishes per-word thresholds for gaze durations such that an unknown word is detected when the reader's gaze duration exceeds the threshold. The detection sensitivity is specified by the desired probability of false alarm. For a specified probability of false alarm of 0.10, the system achieved a 0.89 detection rate and a 0.11 false alarm rate.
- Demonstrated the principle of operation for an automated reading assistance system that provides need-triggered computer-generated vocalization of unknown words to readers. The system operates in a natural reading setting, tolerating unconstrained head movement.

5.2.2 Future Work

The theoretical basis for an automated reading assistance system based on a remote gaze estimation system has been presented, but the implementation remains to be completed. This work requires interfacing a scene camera to an existing remote point-of-gaze estimation system (Guestrin, 2006) and implementing the 2D object motion estimation algorithm developed in Chapter 3. A reading assistance system with no head-worn components offers more comfort to the reader, and would be more suitable for long reading tasks. Remote point-of-gaze estimation is also preferable to head-mounted systems in applications involving children, for whom the reading assistance system may offer benefit in a teaching capacity.

Although the principle of operation of a reading assistance system has been demonstrated, further investigation is required to validate the efficacy of the system as a

teaching tool. The effectiveness of the system in teaching grapheme-to-phoneme mappings requires study, especially in direct comparison to current teaching techniques in which assistive intervention is provided by a human monitoring the reader's performance (e.g. a teacher or a parent).

Other forms of assistive intervention may be evaluated. In some applications, providing vocalization of an unknown word may be insufficient or inappropriate. Other possible forms of automated assistance include providing the definition of the word and translation of the word to a different language. The principle of the system to be conserved is the automatic detection of the reader's need followed by context-appropriate assistance.

Bibliography

Abrams, R.A. and Jonides, J. (1988). "Programming saccadic eye movements." *Journal of Experimental Psychology: Human Perception and Performance*, 14, 428-443.

Abrams, R.A., Meyer, D.E., and Kornblum, S. (1989). "Speed and accuracy of saccadic eye movements: Characteristics of impulse variability in the oculomotor system." *Journal of Experimental Psychology: Human Perception and Performance*, 14, 428-443.

Adams, M.J. (1990). Beginning to Read: *Thinking and Learning about Print*. Cambridge: MIT Press.

Ahn, S.J., Rauh W., and Kim S.I. (2001). "Circular coded target for automation of optical 3d-measurement and camera calibration." *International Journal of Pattern Recognition and Artificial Intelligence*, 15, 905-919.

Allison, R., Eizenman, M., and Cheung B. (1996). "Combined head and eye tracking system for dynamic testing of the vestibular system." *IEEE Transactions on Biomedical Engineering*, 43, 1073-1082.

Bettelheim, B. and Zelan, K. (1982). *On Learning to Read*. New York: Alfred A. Knopf.

Clarke, T.A., Cooper, M.A., and Fryer J.G. (1994). "Estimator for the random error in subpixel target location and its use in the bundle adjustment." In *Optical 3D Measurement Techniques II*, *International Society for Optical Engineering*, Proc.Vol. 2252, 161-168.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review*, 108, 204-258.

Criminisi, A., Reid, I., and Zisserman, A. (1999). "A plane measuring device." *Image and Vision Computing*, 17, 625-634.

Eizenman, M., Yu, L.H., Grupp, L., Eizenman, E., Ellenbogen, M., Gemar, M., and Levitan, R.D. (2003). "A naturalistic visual scanning approach to assess selective attention in major depressive order," *Psychiatry Research*, 188, 117-128.

Faugeras, O. and Lustman, F. (1988). "Motion and structure from motion in a piecewise planar environment," International Journal of Pattern Recognition and Artificial Intelligence, 2, 485-508.

Faugeras, O. (1993). *Three-dimensional Computer Vision*, MIT Press.

Fisher, D.F. and Shebilske, W.L. (1985). "There is more that meets the eye than the eyemind assumption." In R. Groner, G.W. McConkie, and C. Menz (Eds.), *Eye movements and human information processing*. Amsterdam: North Holland.

Ganci, G. and Handley, H.B. (1998). "Automation in videogrammetry." *International Archives of Photogrammetry and Remote Sensing*, 32, 53-58.

Goldberg, J. and Kotval, X. (1999). "Computer interface evaluation using eye movements: methods and constructs," *International Journal of Industrial Engineers*, 25, 631-645.

Guestrin, E. D. and Eizenman, M. (2006) "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Transactions on Biomedical Engineering*, in press.

Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*, 2nd Ed. Cambridge University Press.

Heikkila, J. and Silven, O. (1997). "A four-step camera calibration procedure with implicit image corrections." In *Conference on Computer Vision and Pattern Recognition 1997*, Proc, 1106-1113.

Hutchinson, T.E., White, K.P., Martin, W.N., Reichert, K.C., and Frey, L.A. (1989). "Human-computer interaction using eye-gaze input," *IEEE Transactions on Systems, Man, and Cybernetics*, 19, 1527-1534.

Inhoff, A.W. (1984). "Two stages of word processing during eye fixations in the reading of prose." *Journal of Verbal Learning and Verbal Behavior*, 23, 612-624.

Inhoff, A.W. and Rayer, K. (1986). "Parafoveal word processing during eye fixations in reading: Effects of word frequency." *Perception and Psychophysics*, 40, 431-439.

Just, M.A. and Carpenter, P.A. (1980). "A theory of reading: From eye fixations to comprehension." *Psychological Review*, 87, 329-354.

Karatekin, C. and Asarnow, R.F. (1999). "Exploratory eye movements to pictures in childhood-onset schizophrenia and attention-deficit/hyperactivity disorder," *Journal of Abnormal Child Pyschology*, 27, 35-49.

Kennedy, A. and Murray, W.S. (1987). "The components of reading time: Eye movement patterns of good and poor readers." In J.K. O'Regan and A. Lévy-Schoen (Eds.), *Eye movements: From physiology to cognition*. Amsterdam: North Holland.

Kolers, P.A., Duchnicky, R.L., and Ferguson, D.C. (1981). "Eye movement measurement of readability of CRT displays." *Human Factors*, 23, 517-527.

Leech, G., Rayson, P., and Wilson, A. (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman.

Longuet-Higgins, H.C. (1981). "A computer algorithm for reconstructing a scene from two projections." *Nature*, 293, 133-135.

Loshe, G. (1997). "Consumer eye movement patterns of Yellow Pages advertising," *Journal of Advertising*, 26, 61-73.

Marr, D. and Poggio, T. (1976). "Cooperative computation of stereo disparity." *Science*, 194, 283-287.

Morrison, R.E. and Inhoff, A.W. (1981). "Visual factors and eye movements in reading." *Visible Language*, 15, 129-146.

Morrison, R.E. and Rayner, K (1981), "Saccade size in reading depends upon character spaces and not visual angle," *Perception & Psychophysics*, 30, 395-396.

Mouaddib, E., Batlle, J., and Salvi, J. (1997). "Recent progress in structured light in order to solve the correspondence problem in stereovision." In *IEEE International Conference on Robotics and Automation 1997*, Proc., 20-25.

Rayner, K. and McConkie, G.W. (1976). "What guides a reader's eye movements." *Vision Research*, 16, 829-837.

Rayner, K. Slowiaczek, M.L., Clifton, C. and Bertera, J.H. (1983). "Latency of sequential eye movements: Implications for reading." *Journal of Experimental Psychology: Human Perception and Performance*, 9, 912-922.

Rayner, K. (1984). "Visual selection in reading, picture perception, and visual search: A tutorial review. In H. Bouma & D. Bouwhuis (Eds.), *Attention and performance (Vol. 10)*. Hillsdale, NJ: Erlbaum.

Rayner, K. (1998). "Eye movements in reading and information processing: 20 years of research." *Psychological Bulletin*, 124, 372-422.

Sharma, R., Pavlovic, V.I., and Huang, T.S. (1998). "Towards multimodal human-computer interface," *Proceedings of the IEEE*, 86, 853-869.

Shortis, M.R., Clarke, T.A., and Short, T. (1994). "Comparison of some techniques for the subpixel location of discrete target images." In proceedings of *Videometrics III, International Society for Optical Engineering*, Vol. 2350, 239-250.

Trueswell, J.C., Tanenhaus, M.K., and Garnsey, S.M. (1994). "Semantic influences on parsing: use of thematic role information in syntactic ambiguity resolution." *Journal of Memory and Language*, 33, 285-318.

Tsai, R.Y. and Huang, T.S. (1981). "Estimating three-dimensional motion parameters of a rigid planar patch." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29, 1147-1152.

Tsai, R.Y., Huang, T.S., and Zhu, W. (1982). "Estimating three-dimensional motion parameters of a rigid planar patch, II: Singular value decomposition." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30, 525-534.

Weng, J., Cohen, P., and Herniou M. (1991). "Camera calibration with distortion models and accuracy evaluation." IEEE Transactions on Pattern Analysis and Machine Intelligence, 14, 965-980.

Wilhelm, P.B., Rhees, R.W., and Van De Graaff, K.M. (2001). *Human Anatomy and Physiology*. New York: McGraw-Hill.

Young, L.R. and Sheena, D. (1975). "Methods and designs: survey of eye movement recording methods," *Behavior Research Methods and Instrumentation*, 7, 397-429.

Yu, L.H. and Eizenman, M. (2004). "A new methodology for determining point-of-gaze in head-mounted eye tracking systems." *IEEE Transactions on Biomedical Engineering*, 51, 1765-1773.

Zhang, Z. (1999). "Flexible camera calibration by viewing a plane from unknown orientations." In *ICCV 1999*, 666-673.

Appendix A

Rodrigues' Rotation Formula

One method of representing a rotation is the angle-axis representation, whereby a rotation by a angle θ about a fixed axis specified by a unit vector $\mathbf{W} = [w_x\ w_y\ w_z]^T$ is fully described by a vector $\boldsymbol{\theta} = \theta\mathbf{W}$. The equivalent rotation matrix $\mathbf{R}$ of this rotation is given by Rodrigues' rotation formula (Hartley and Zisserman, 2003):

$$\mathbf{R}=\begin{bmatrix} \cos\theta+w_x^2(1-\cos\theta) & w_xw_y(1-\cos\theta)-w_z\sin\theta & w_y\sin\theta+w_xw_z(1-\cos\theta) \\ w_z\sin\theta+w_xw_y(1-\cos\theta) & \cos\theta+w_y^2(1-\cos\theta) & -w_x\sin\theta+w_yw_z(1-\cos\theta) \\ -w_y\sin\theta+w_xw_z(1-\cos\theta) & w_x\sin\theta+w_yw_z(1-\cos\theta) & \cos\theta+w_z^2(1-\cos\theta) \end{bmatrix}. \qquad (A.1)$$