COMPUTER SIMULATION AND IMPLEMENTATION OF A VISUAL 3-D EYE GAZE TRACKER FOR AUTOSTREOSCOPIC DISPLAYS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

KUTALMIŞ GÖKALP İNCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

DECEMBER 2009

Approval of the thesis:

COMPUTER SIMULATION AND IMPLEMENTATION OF A VISUAL 3-D EYE GAZE TRACKER FOR AUTOSTREOSCOPIC DISPLAYS

submitted by **KUTALMIŞ GÖKALP İNCE** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İsmet Erkmen _____
Head of Department, **Electrical and Electronics Engineering**

Assoc. Prof. Dr. A. Aydın Alatan _____
Supervisor, **Electrical and Electronics Engineering Dept., METU**

Examining Committee Members

Prof. Dr. Aydan Erkmen _____
Electrical and Electronics Engineering Dept., METU

Assoc. Prof. Dr. A. Aydın Alatan _____
Electrical and Electronics Engineering Dept., METU

Prof. Dr. Gözde Bozdağı Akar _____
Electrical and Electronics Engineering Dept., METU

Prof. Dr. Uğur Halıcı _____
Electrical and Electronics Engineering Dept., METU

Assoc. Prof. Dr. Uğur Güdükbay _____
Computer Engineering Dept., Bilkent University

Date: 09.12.2009

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Kutalmış Gökalp İNCE

Signature :

ABSTRACT

COMPUTER SIMULATION AND IMPLEMENTATION OF A VISUAL 3-D EYE GAZE TRACKER FOR AUTOSTREOSCOPIC DISPLAYS

İnce, Kutalmış Gökalp

M.S., Department of Electrical and Electronics Engineering

Supervisor : Assoc. Prof. Dr. A. Aydın Alatan

December 2009, 102 Pages

In this thesis, a visual 3-D eye gaze tracker is designed and implemented to tested via computer simulations and on an experimental setup. Proposed tracker is designed to examine human perception on autostereoscopic displays when the viewer is 3m away from such displays. Two different methods are proposed for calibrating personal parameters and gaze estimation, namely line of gaze (LoG) and line of sight (LoS) solutions. 2-D and 3-D estimation performances of the proposed system are observed both using computer simulations and the experimental setup. In terms of 2-D and 3-D performance criteria, LoS solution generates slightly better results compared to that of LoG on experimental setup and their performances are found to be comparable in simulations. 2-D estimation inaccuracy of the system is obtained as smaller than 0.5° during simulations and approximately 1° for the experimental setup. 3-D estimation inaccuracy of the system along x- and y-axis is obtained as smaller than 2° during the simulations and the experiments. However, estimation accuracy along z-direction is significantly sensitive to pupil detection and head pose estimation errors. For typical error levels, 20cm inaccuracy along z-direction is observed during simulations, whereas this inaccuracy reaches 80cm in the experimental setup.

Keywords: 3-D eye gaze tracking, pupil detection, calibration.

ÖZ

OTO-STEREOSKOPİK EKRANLAR İÇİN GÖRSEL BİR 3-B BAKIŞ NOKTASI TAKİP SİSTEMİNİN BİLGİSAYAR SİMÜLASYONU VE UYGULAMASI

İnce, Kutalmış Gökalp

Yüksek Lisans, Elektrik Elektronik Mühendisliği

Tez Yöneticisi : Doç. Dr. A. Aydın Alatan

Aralık 2009, 102 Sayfa

Bu çalışmada görsel bir 3-B bakış noktası takip sistemi tasarlanmış, önerilen sistem, bilgisayar simülasyonları ve gerçek veriler ile test edilmek üzere uygulamıştır. Önerilen sistem, oto-stereoskopik ekranlar üzerinde izleyicinin ekrandan 3m uzakta olduğu durumda çalışmak üzere tasarlanmıştır. Bakış noktası kestirimi ve kişisel parametrelerin kalibrasyonu için bakış ekseni ve görüş ekseni çözümleri önerilmiştir. Önerilen sistemin 2-B ve 3-B kestirim başarımları bilgisayar simülasyonları ve deneyler ile gözlemlenmiştir. Hem 2-B hem de 3-B kestirimlerde, görüş ekseni çözümü, bakış ekseni çözümüne göre gerçek veriler ile biraz daha iyi sonuçlar üretirken, simülasyonlarda iki çözüm benzer sonuçlar üretmiştir. Sistemin 2-B çözümlerde kestirim hatasının simülasyonlarda 0.5°'den küçük, deneyler sırasında is 1° dolayında olduğu gözlemlenmiştir. Simülasyonlar ve deneyler sırasında, sistemin 3-B çözümlerde x- ve y- eksenlerindeki kestirim hatasının 2°'den küçük olduğu gözlemlenmiştir. Sistemin z- eksenindeki kestirim hassasiyeti, göz bebeği tespit hataları ve izleyicinin pozunun kestirimindeki hatalar ile birlikte ciddi biçimde azalmaktadır. Simülasyonlar sırasında beklenen hatalar kullanıldığında z- eksenindeki kestirim hatasının 20cm dolaylarında olduğu gözlemlenmiş ancak deneyler sırasında bu hatanın 80cm'ye kadar yükseldiği görülmüştür.

Anahtar Kelimeler : 3-B Bakış noktası takibi, göz bebeği tespiti, kalibrasyon.

ACKNOWLEDGEMENTS

I would like to express my appreciation to my supervisor, Assoc. Prof. Dr. A. Aydın Alatan for his wisdom, advice and guidance. I would like to thank to Prof. Dr. Umur Talaslı from whom I have learned basics of visual perception, which encourages me to start such a study. I would also like to thank to Yoldaş Ataseven for his brilliant ideas, my other colleagues in ASELSAN for their interest and brainstorming and my manager Hüseyin Yavuz for his support. Most of all, I would like to thank to my wife Miray Akyunus for her invaluable support.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

CS : coordinate system
CCS : camera coordinate system
DCS : display coordinate system
DPI : dual purkinje image
EOG : electro-oculography
EGT : eye gaze tracker
HCS : head coordinate system
HCI : human computer interaction
IROG : infra-red oculography
LoG : line of gaze, the line passing through pupil and eyeball center
LoS : line of sight, the line passing through pupil and fovea.
LS : least squares triangulation
MCS : mirror coordinate system
MP : midpoint triangulation
MOS : mean opinion score
PCR : pupil-corneal reflection
PT : polynomial triangulation
PTZ : pan-tilt-zoom
RCS : reflected coordinate system
SSCQE : single stimulus continuous quality evaluation

CHAPTER 1

INTRODUCTION

Eye gaze tracking is one of the important research efforts in machine vision as well as human machine interaction. Duchowski [3] summarizes various applications of eye gaze trackers (EGTs) and divides them into two categories based on their utilization purpose: *diagnostic* and *interactive*. While diagnostic gaze trackers are used to obtain objective and quantitative evidence of user's attention, interactive gaze trackers are used as an input device to visually-mediated environments. Relative to their contact with user, EGTs are categorized as intrusive (remote) or non-intrusive (in contact) [2].

Interactive EGTs are mostly visual and based on non-intrusive methods. Since they are designed for HCI, most of the studies about visual EGTs are focused on decreasing calibration requirements. On the other hand diagnostic EGTs require some special hardware (contact lenses etc.) and they have serious calibration requirements. Shih and Liu [16] classify gaze tracking methods as 2-D techniques, model based 3-D techniques and 3-D techniques. They also point out that any 2-D tracking system can be extended to a 3-D tracking system with the known 3-D position of eye [16]. Head movements are one of the most important problems in traditional EGTs, since such a movement dramatically decreases gaze estimation performance [9]. Traditional methods and most of the present EGTs assume that gaze point lies on a plane (screen), which is not a valid assumption for 3-D displays.

1.1 Motivation

Over the years a consensus has been reached that the introduction of 3-D TV can only be a lasting success, if the perceived image quality and the viewing comfort is at least comparable to conventional television [1]. About viewing comfort,

Sexton and Surman [25] point to a wide belief that 3-D TV must be autostereoscopic (unaided) and it should supply stereo to several viewers for a non-rigid viewing position. Considering perceived image quality, subjective assessments are thought to be a vital element for 3-D TV [18].

In order to evaluate studies about 3-D pictures in terms of perceived image quality, subjective assessment methods are described in [18]. Subjective assessments depends on mean opinion score (MOS), mean of scores given by subjects, which is related with the asked aspects of perceived image quality. These subjective assessments are conducted to measure effect of engineering efforts, such as transmission, compression, display, coding and 3-D reconstruction, on perceived image quality. However, apart from errors arise from imperfections in these steps, 3-D displays suffer from erroneous monocular cues and monocular-binocular conflicts to be explained in Appendix-A in detail. Erroneous monocular cues and monocular-binocular conflicts are related to the scene content, viewer's attention and viewer's movements. It is logical to expect these errors and conflicts affecting the perceived quality, as well as the scores of subjects on assessments. In other words, score of a subject will also be dependent to his/her attention, movements and scene content, in addition to other technical variables, such as 3-D coding or reconstruction. Although for a long period of viewing and a large number of subjects, getting MOS might reduce subjective variances, the result should still be dependent on the scene content and at least mean of erroneous monocular cues and monocular-binocular conflicts.

Taking the above discussion into account, if one can estimate the errors on monocular cues and monocular-binocular conflicts and relate them with the MOS, the significance of these conflicts could be observed on the perceived image quality, which might yield the maximum perceived quality that we could be achieved by the available content, for the case other variables are kept ideal and constant. Moreover, if the effects of error on monocular cues and monocular-binocular conflicts could be removed (or reduced) from MOS, one might obtain a score as a function of only desired variables, such as transmission, compression, coding, etc. In other words, one could obtain more reliable results with respect to subjective assessments.

1.2 Scope of the Thesis

Knowing 3-D gaze point, distance between viewer and display and tracking eye movements, one can obtain errors on monocular cues and monocular-binocular conflicts. To achieve this purpose, one will require a gaze tracker to work with autostereoscopic displays which estimates (or provides required information to estimate)

- 3-D gaze point,
- distance between viewer and display,
- position and orientation of eyeball.

Throughout this work, we will develop a visual 3-D eye gaze tracker which estimates gaze point in 3-D, distance between viewer and display and tracks eyeball position and orientation. We will mostly focus on obtaining 3-D gaze point accurately, while natural head movements are allowed. The viewer is positioned much far away from the display compared to present methods. Considering the complexity of the requirements and the problem, we will prefer an intrusive method. Such a selection will not create a serious drawback, the required gaze tracker is diagnostic. In Chapter 2, we will review related studies. Proposed EGT will be presented in Chapter 3. Finally, we will make conclusions about study in Chapter 4.

1.3 Notation

Throughout this text, matrices are shown with bold capital letters, whereas vectors with capital letters and scalars with non-capital letters.

a : scalar a

P : vector P

M : matrix M

Following is the list of notation for coordinate systems, special matrices and vectors related to coordinate systems.

$\{A\}$: coordinate system A

${}^{A}X$: arbitrary vector X viewed from $\{A\}$

${}^{A}\hat{x}_{B}$: unit vector on the x axis of $\{B\}$ viewed from $\{A\}$

T_{A}^{B} : transformation from $\{B\}$ to $\{A\}$

R_{A}^{B} : rotation from $\{B\}$ to $\{A\}$

t_{B}^{A} : translation from $\{B\}$ to $\{A\}$

CHAPTER 2

RELATED WORK

Before presenting the proposed 3-D eye gaze tracker, some related work from the literature is briefly examined in the subsequent sections.

2.1 Anatomy of Human Eye

In his eye gaze trackers review, Morimoto [2] summarizes human eye structure. Human eye can be approximated as a sphere with radius of 12.5mm. Visible parts of eye are sclera (white part), iris (colored part) and pupil (black part in the center of iris). Boundary between sclera and iris is called as limbus. Cornea covers the visible part of iris and has a spherical shape with radius of 7.5 mm. The line passing through center of eyeball, center of cornea and pupil is called optical axis or namely line of gaze (LoG). The most color sensitive part of the retina is called as fovea on which gaze object is projected. Fovea does not lie on optical axis, but it is slightly shifted. The line passing through fovea and pupil is called as visual axis or line of sight (LoS). Structure of human eye is presented in Figure 2.1.1.

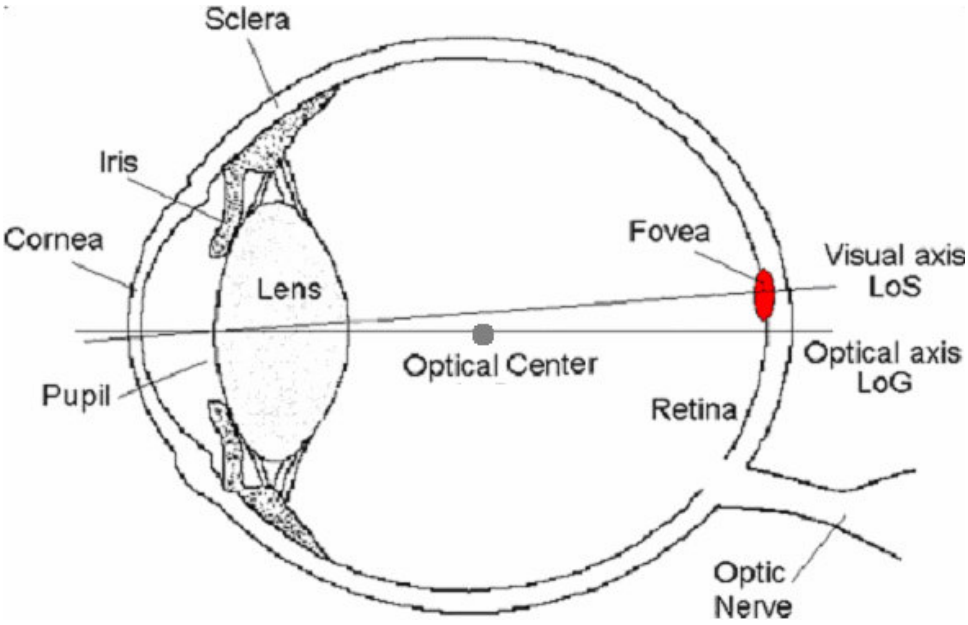


Figure 2.1.1 : Structure of human eye (adapted from [2])

2.2 Eye Tracking, Gaze Tracking and Gaze Estimation

2.2.1 Traditional Methods

Morimoto [2] presents a survey of both traditional and advanced methods for gaze tracking problem. In this survey, eight traditional methods are presented. Shih and Liu [16] classify gaze tracking methods as 2-D techniques, model based 3-D techniques and 3-D techniques. They also point out that any 2-D tracking system can be extended to a 3-D tracking system with a known 3-D position of eye [16]. The eight traditional methods presented by Morimoto [2] are all 2-D techniques. Head movements are one of the most important problems in traditional gaze trackers, since such a movement dramatically decreases gaze estimation performance [9]. Reported accuracies and comments about traditional methods are summarized in Table 2.2.1.

Contact Lens Method : In this method, user should wear a contact lens with a small coil in it and user's gaze is estimated by the voltage induced in coil by an external magnetic field.

Electro-Oculography (EOG) Method : By placing electrodes around the eye, eye movements of a user is measured by changes in its skin potential and related to the user's gaze.

IR Oculography (IROG) Method : Limbus of the user is illuminated via several IR LEDs and their reflection from limbus is sensed via photo-transistors. Orientation of the eye is determined from difference between nasal and temporal reflectance.

Purkinje Image Methods : Reflection of light from different layers of the eyeball is called Purkinje images. The first Purkinje image is the reflection from outer surface of cornea, whereas the second one is from inner surface of the cornea. The third image is reflection from the outer surface of lens and the fourth image is reflection from inner surface of the lens. The first Purkinje image is easier to detect and track, while detecting other Purkinje images require special hardware. Since the third and fourth images gives information about thickness of the lens, accommodation distance can be estimated by using the third and fourth Purkinje images. While eye is rotating, each Purkinje image moves through a different distance and hence, distance between Purkinje images is affected from orientation of eyeball. Dual Purkinje image (DPI) method estimate user's gaze from first and fourth Purkinje images.

Limbus/Pupil Tracking Methods : Center of iris/pupil is detected from an image and a mapping between iris/pupil position and visual targets is estimated. Since the position of iris/pupil changes with head movements, in this approach, user's head should be fixed. For allowing head movements, rather than center of iris/pupil, a vector defined from center of iris/pupil to a fixed reference point on viewers' head is used. Limbus tracking methods has a low vertical resolution due to occlusions of iris with eyelashes and eyelids. In pupil tracking methods, detection of pupil is challenging and IR illuminators are used to obtain bright or dark pupil effects and overcome this challenge. However both methods are sensitive to head movements.

Pupil - Corneal Reflection (PCR) Method : This method is a particular type of pupil tracking in which the reference point is defined as corneal reflection of an IR LED. Pupil is usually illuminated with an on-axis IR LED to get bright pupil effect. PCR

method is used to estimate gaze point on a screen. Mapping between pupil-corneal reflection vectors and the screen coordinates is obtained by a polynomial and a calibration process is performed in order to find coefficients of the polynomial (usually second order). PCR method is widely used [4, 5, 6, 7] due to its high accuracy relative to its simplicity. However, this method is also sensitive to head movements, i.e. accuracy of the mapping decays as user moves his/her head from its initial position.

Appearance Based Methods : These methods handle gaze estimation as a recognition problem. Rather than utilizing geometric features, such as pupil, iris or contours, a cropped image of the eye is used to estimate user's gaze. Intensity image of the eye is fed to a classifier which is trained with different orientations of eyeballs and the output of this classifier is the estimated gaze.

Table 2.2.1 : Traditional gaze trackers (adapted from [2])

Method	Accuracy	Properties
Contact lens	0.016°	Very intrusive, but fast and accurate
EOG	2°	Intrusive, but simple and low cost
IROG	0.033°	Head mounted, limbus tracking
DPI	0.016°	Not intrusive, but requires bite bar
Limbus tracking	1°	Visual, lower vertical accuracy
Pupil tracking	1°	Visual, hard to detect the pupil
PCR	1°	Visual, tolerate some head motion
Appearance-based	0.5-2°	Visual, requires training

2.2.2 EGT Approaches with Reduced Restrictions on Users

Further efforts in EGTs basically try to reduce the need of calibration per user session and the large restriction on head movements [2].

In order to increase estimation accuracy, size of eye on the image must be increased so that FOV must be decreased. However, in order to allow larger head movements, FOV must be increased. For the solution of this problem, employment of both wide and narrow view (pan-tilt-zoom) cameras is proposed [12, 13, 14, 23].

Zhu and Qiang [9] propose a method to compensate natural head movements in PCR method. They have reported an accuracy of 1.3° in horizontal and 1.7° in vertical for users 45cm away from camera with allowed head movements of 20x20x30cm (width-height-depth).

Liu and Talmi [23] use one wide view stereo camera and one narrow view PTZ camera to determine head pose and to detect pupil, respectively. They propose a method to compensate head movements for their PCR method. Only one eye is tracked and once gaze vector is estimated, gaze point is obtained by the intersection of the gaze vector with nearest object on the scene. Tracker is designed for interactive stereoscopic displays and provides gaze distance as well. For a user 60cm away from display, an accuracy of 0.7° is reported.

In the method proposed by Morimoto, Amir and Flickner [10], one camera and two IR LEDs are utilized. System allows free head motion and do not require per session calibration (per user calibration is required). Modeling cornea as a convex mirror and using reflection of IR LEDs from cornea, center of cornea is estimated in 3-D coordinates. Using this estimate and geometry of eye, pupil position is also computed in 3-D coordinates. Once center of cornea and pupil is obtained in 3-D, gaze direction can be obtained as the vector defined from center of cornea and to the pupil. In simulated data, an accuracy up to 3° is reported for users 30-80 away from camera.

Yoo and Chung [11] use one camera and five LEDs. One of the LEDs placed on optical axis to get bright pupil effect. Other LEDs placed on four corners of the

screen to obtain four glitters. From the glitters, a polygon is obtained and by using position of the pupil relative to this polygon, user's gaze is estimated on the screen. This method is calibration free and allows head movements. Gaze position is obtained in 2-D and an accuracy up to 2° is reported for users 30-40cm away from camera.

Beymer and Flickner [12] use one wide view and one narrow view (PTZ) stereo cameras for gaze estimation. Using wide view stereo camera, user's eye regions are extracted and narrow view camera is steered to get eye in FOV. From narrow view camera, first glitters and pupil are detected and located in 3-D. Then using geometric structure of cornea, center of cornea is estimated by using pupil and glitters. Gaze vector is estimated as the vector from center of cornea to pupil. A LoS/LoG correction term is added to gaze vector and line of sight is obtained. Finally user's gaze is estimated as the intersection of line of sight with screen. Transformation between camera coordinate system and screen coordinate system is estimated by placing a mirror between cameras and screen, which will be explained in Section 3.2.6 in detail. In the proposed system, per user calibration is required; user's gaze is estimated in 2-D and accuracy of 0.6° is reported for users 62cm away from the monitor.

Wang and Sung [13, 14] use a wide-view camera to determine head pose and locate eye regions. A narrow view pan-tilt camera is utilized to obtain the gaze vector. Using circular shape of iris and its elliptical projection on to image plane, they propose 3-D localization of iris center and its normal vector for a known iris radius. Since projection of the circle on to image plane results in two possible solutions, for the selection of the right circle, "two circle" algorithm is proposed and gaze estimate is obtained from intersection of right and left gaze vectors [13]. In another study, Wang and Sung [14] propose one circle algorithm to select right circle and gaze estimation from a single eye. One circle algorithm uses the distance between eyeball center and two eye corners to select right circle. In this study, gaze point is defined as the intersection of gaze vector with screen plane. It should be noted that in two circle algorithm for a user 150cm away from the camera, an accuracy of 2°, whereas in one circle algorithm for a user 60cm away from the camera, an accuracy of 0.6° is reported.

Matsumoto and Zelinsky [15] combines head pose estimation with eyeball orientation for the estimation of gaze direction. In their proposed method, facial features is determined and located by a stereo camera. Throughout a training sequence, a vector defining center of eyeball relative to the head coordinate system is obtained in addition to eyeball radius and iris radius. Then, gaze line is defined from center of eyeball to the center of the iris. However, they report poor accuracy of gaze lines and to reduce errors on gaze lines, rather than intersecting them, average of the gaze lines is used as a single gaze line. An accuracy of 3° is reported for a typical user 80cm away from the camera.

Shih and Liu [16] propose a gaze tracking system with a pair of stereo cameras and three IR LEDs. They estimate the optical axis of the eye (equivalently line of gaze) as a vector defined from center of cornea to pupil. Center of cornea is obtained in 3-D from first Purkinje images and spherical structure of the cornea. As a difference from the method in [10], Shih and Liu can locate pupil in 3-D without requiring a user dependent parameter, since they use stereo cameras. The only user dependent parameter of their method is the difference between the LoS and LoG. A clear procedure for estimating transformation between display and camera coordinate systems and positioning the LED's relative to camera is given. They report accuracy about 1° for a user 45cm away from display.

Ki and Kwon [24] use distance between pupils to estimate gaze distance and PCR method to estimate gaze point on the screen. They employ an ultrasonic sensor to compensate head movements. Their method is particularly designed for interactive stereoscopic display; however, only very small head movements are allowed (2cm for left/right and up/down and 5cm for forward/back). For a user 84cm away from display accuracy higher than 1° is reported.

Accuracy of the advanced gaze trackers with their working distance and 3-D gaze point detection capability is summarized in Table 2.2.2.

Table 2.2.2 : Advanced eye gaze trackers: accuracy, viewing distance and 3-D capability

Author	Accuracy	Distance(cm)	3-D Gaze Point
Zhu and Qiang [9]	< 2°	30-60	NO
Liu and Talmi [23]	0.7°	60	YES
Morimoto et al. [10]	3°	30-80	ADAPTABLE
Yoo and Chung [11]	2°	30-40	NO
Beymer and Flickner [12]	0.6°	62	ADAPTABLE
Wang and Sung [13] (two circle)	2°	150	YES
Wang and Sung [14] (one circle)	0.6°	60	ADAPTABLE
Matsumoto and Zelinsky [15]	3°	80	YES
Shih and Liu [16]	< 1°	45	ADAPTABLE
Ki and Kwon [24]	< 1°	84	YES

2.3 Iris and Pupil Detection

One approach for iris/pupil detection is using face detection and segmenting the eye region for locating iris/pupil [8]. As another method, dark or bright pupil images could be utilized [8]. When an IR light source is placed near the optical axis of camera (on axis), reflection of the light source from retina is visible and results in bright pupil effect. If light source is away from optical axis, then dark pupil effect is observed.

Morimoto et. al. [8] proposes utilization of multiple light sources to improve the accuracy and robustness of pupil detection. In their proposed method, on-axis and off-axis IR light sources are synchronized with odd and even frames of the camera to obtain bright and dark pupil images, respectively. Using the difference of odd and even frames (bright and dark pupil images) and threshold, pupils could be detected. Once a candidate region for iris/pupil detection is segmented, ellipse/circle fitting algorithms or sometimes determining center of mass of the extracted region is used to locate the iris/pupil center precisely. Taylor and Robert [19] present a review of ellipse fitting methods. He et. al. [20] propose a method, namely “pulling and pushing”, to find center of iris/pupil from a rough estimate of iris/pupil center for partially occluded iris/pupil.

CHAPTER 3

PROPOSED APPROACH

3.1 Overview

We develop a gaze tracker to work with autostereoscopic displays which estimates (or provides required information to estimate)

- 3-D gaze point,
- distance between the viewer and the display,
- orientation of the eyeball.

A method satisfying the following requirements needed to be designed:

- It should work for a user up to 3m away from away from display (optimum viewing distance for stereoscopic displays [18])
- It should allow natural head movements (user should be able to trigger motion parallax mechanism, see Appendix-A),
- It must be able to find gaze point in 3-D (in order to quantify the convergence cue, see Appendix-A)
- It must be able to find distance between pupils and display (in order to quantify the accommodation cue, see Appendix-A),
- It must be able to find motion of pupil between consecutive frames (in order to quantify the expected parallax cue, see Appendix-A),
- It must be able to register scene content with measurements both in time and space.

As stated in the previous chapter, traditional methods and most of the previous efforts assume that gaze point lies on a plane (screen), which is not a valid assumption for 3-D displays. Hence, traditional methods are not applicable for the required system. The methods examined in Section 2.2.2 (except Zhu and Qiang [9] and Yoo and Chung [11]) could be possible candidates for the required

system. However, all of these approaches are related to HCI; hence, they are focused on decreasing calibration requirements and non-intrusive methods. Matsumoto and Zelinsky [15] points to the errors in gaze vectors and rather than using intersection of right and left gaze vectors, they propose using the average of them. Other methods require a detailed image of the eye in order to locate glitters or to observe detailed shape of iris. Moreover, their accuracy for a user at 3m is not reported. Considering their high resolution requirement, for a user 3m away from a display, a dramatic decrease in the gaze estimation accuracy should be expected. Most of the aforementioned methods find gaze point as intersection of gaze line by the screen plane; in other words, the resulting accuracy does not include triangulation errors. Hence, they are not applicable to the given problem in their current form; however, some of the ideas, such as using an unseen point (center of cornea or eyeball) as reference point [10, 12, 15, 16] or display-camera calibration by the help of a mirror [12, 16], could be borrowed.

We mostly focus on obtaining 3-D gaze point accurately, while natural head movements are allowed. Moreover, the user is positioned much far away from the display compared to present methods. Considering the complexity of the requirements and the problem, we will prefer an intrusive method. Such a selection does not create a serious drawback, since the required gaze tracker is diagnostic.

In this chapter, we propose a new gaze estimation method, which uses 3-D gaze vectors. We initiate our method by the definition of a gaze point [2] and an approximation:

- Line of sight (LoS) is defined by the line passing through fovea and pupil,
- Line of gaze (LoG) is defined by the line passing through center of eyeball and pupil.
- Gaze point is defined as intersection of left and right LoS.
- Gaze point can be approximated by intersection left and right LoG.

These measurements could be obtained by a gaze tracker, if the center of eyeball and pupil in 3-D for both eyes and some personal parameters are available. Therefore, we will first obtain 3-D position of the eyeball center (for both eyes); then by using geometric properties of the eye and projection of the pupil on the

image plane, we will find 3-D coordinates of the pupil and we will obtain LoG for both eyes. Next, as first solution, utilizing a triangulation algorithm, we will estimate gaze point in 3-D space. As a second solution, using the observed LoG and some personal parameters, we will find the current orientation of LoS, then as in the first solution, by the help of a triangulation algorithm, we will estimate gaze point in 3-D space. Finally, we will transform gaze point estimates, pupils and center of eyeball into the desired coordinate system in order to obtain estimates of accommodation, convergence and expected/received motion parallax, if these measurements will be used for analyzing human perception. The block diagram of this strategy is depicted in Figure 3.1.1.

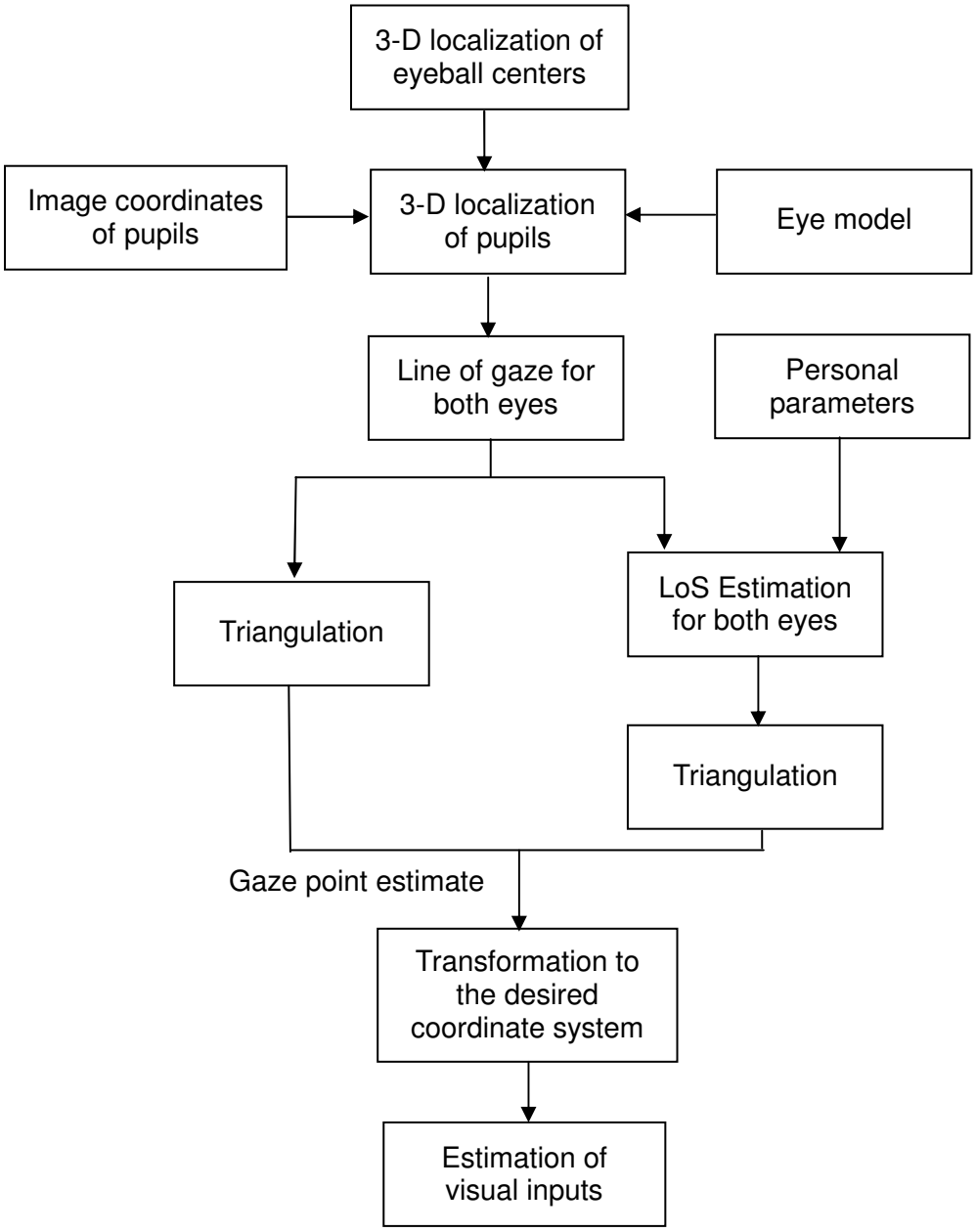


Figure 3.1.1 : Block diagram of the proposed method

3.2 Eye Model

When viewer is gazing onto a point X, X lies on the line of sight (LoS), defined by fovea (F) and pupil (P),

$$X = P + aS \tag{3.2.1}$$

where S is unit vector in the direction of LoS,

$$S = \frac{P - F}{|P - F|} \tag{3.2.2}$$

In addition, eyeball can be modeled as a sphere centered at *C* with radius *r* and *P* lies on this sphere

$$|P - C| = r \tag{3.2.3}$$

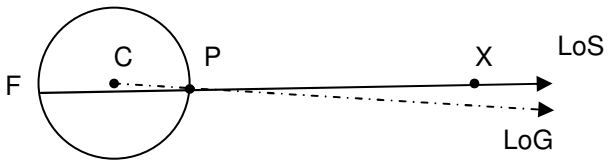


Figure 3.2.1 : Eyeball, pupil, line of sight, line of gaze and gaze point

As illustrated in Figure 3.2.1, gaze point does not line on line of gaze (LoG), defined by center of eyeball and pupil. Unit vector in the direction of LoG can be obtained by the following equation

$$G = \frac{P - C}{r} \tag{3.2.4}$$

The angle between LoS and LoG is around 2°. In some cases, assumption of gaze point lies on the LoG could be useful, since LoG can easily be obtained, once center of eyeball and pupil are located in 3-D.

It is known that movements of an eyeball are rotational only and rotation axis passes through the center of eyeball. Therefore, position of the eyeball center is

changed only by the head movements. However, eyeball orientation is affected both by head movements and eyeball movements. *Donder's Law* [29] states that for a fixed head orientation, in order to gaze on to a specific point, eyeball must have a unique orientation. In other words by observing gaze direction, orientation of eyeball can be uniquely determined; moreover, this orientation is independent of movement history [29]. *Listing's Law* [29] states that any observed eyeball orientation can be reached from a primary position with a single rotation about an axis which lies on the plane defined by LoG in primary position (Listing's Plane) [29]. Orientation of LoG and LoS at primary position will be named as *principal LoG* and *principal LoS*, respectively. Hestenes [30] expresses Listing's Law in quaternion rotations. Let a vector X be rotated about the rotation axis V with an angle of α, then

$$A' = QAQ^{-1} \tag{3.2.5}$$

where Q and A are quaternion representations of rotation and vector X, respectively ($Q = \cos(\alpha/2) + \sin(\alpha/2)n^T V$, $A = n^T X$ and $n^T = \begin{bmatrix} i & j & k \end{bmatrix}$).

Orientation of LoG at the ith instant can be expressed as,

$$n^T G_i = Q_i (n^T G_0) Q_i^{-1} \tag{3.2.6}$$

where G_0 is the orientation of the principal LoG. When one rearranges (3.2.6), the following relation is obtained.

$$G_i = \cos(\alpha) G_0 + \sin(\alpha)(G_0 \times V_i) \tag{3.2.7}$$

Once G_0 and G_i are known, one can determine the orientation of eyeball and form the quaternion rotation [32]

$$\alpha_i = \cos^{-1}(G_0 \cdot G_i) \tag{3.2.8}$$

$$V_i = \frac{(G_0 \times G_i)}{\left| G_0 \times G_i \right|} \tag{3.2.9}$$

Then, the orientation of primary LoS (S_0) or the orientation of LoS at the i[th] instant (S_i) could be obtained using each other,

$$n^T S_i = Q_i (n^T S_0) Q_i^{-1} \tag{3.2.10}$$

$$n^T S_0 = Q_i^{-1} (n^T S_i) Q_i \tag{3.2.11}$$

One can write the relation between S_0 and S_i with a rotation matrix also

$$S_i = R_{E,i} S_0 \tag{3.2.12}$$

$$S_0 = R_{E,i}^T S_i \tag{3.2.13}$$

3.3 Gaze Estimation

Let a coordinate system, namely *head coordinate system* (HCS) be connected rigidly to the viewer's head, and transformation from HCS to camera coordinate system (CCS) at the i[th] instant be known, as

$$T_{H,i}^C = \left[\begin{array}{c|c} R_{H,i}^C & t_{H,i}^C \\ \hline 0 & 1 \end{array}\right] \tag{3.3.1}$$

Let the center of eyeball in HCS be denoted by ${}^H C$. Since position of the eyeball center changes only with head movements, if the transformation from HCS to CCS can be obtained, we can obtain center of eyeball in CCS, ${}^C C$, by applying the following transformation:

$${}^C C_i = R_{H,i}^C {}^H C + t_{H,i}^C \tag{3.3.2}$$

Let the projection of pupil on to image plane and its normalized unit vector, U in CCS be both available. Then, one can obtain the line on which pupil lies as

$$^{C}P = aU \tag{3.3.3}$$

Line on which pupils are lying on (3.3.3) intersects eyeball sphere at two points and pupil lies on the point nearer to camera. In the following derivations, all equations are written in CCS; therefore, let ^{C}C be denoted by C for simplicity. If we consider (3.2.3) and replace P with (3.3.3),

$$(C - aU)\cdot(C - aU) = r^2 \tag{3.3.4}$$

$$\Rightarrow |U|^2 a^2 - 2(C\cdot U)a + |C|^2 - r^2 = 0$$

$$\Rightarrow a = \frac{C\cdot U \pm \sqrt{r^2 + (C\cdot U)^2 - |C|^2|U|^2}}{|U|^2}$$

since $|I|^2|J|^2 = (I\cdot J)^2 - |I\times J|^2$ and $|U| = 1$

$$\Rightarrow P = \left(C\cdot U \pm \sqrt{r^2 - |C\times U|^2}\right)U$$

the pupil is on the nearer intersection point to the camera,

$$\Rightarrow P = \left(C\cdot U - \sqrt{r^2 - |C\times U|^2}\right)U \tag{3.3.5}$$

When one replaces P with (3.3.5) on (3.2.4), normalized gaze vector can be obtained

$$G = \frac{\left(C\cdot U \pm \sqrt{r^2 - |C\times U|^2}\right)U - C}{r}$$

By using $I \times (J \times K) = I \cdot (J \cdot K) - J \cdot (I \cdot K)$ and $I \cdot (U \cdot U) = I$

$$\Rightarrow G = \frac{U \times (C \times U) - U\sqrt{r^2 - |C \times U|^2}}{r} \tag{3.3.6}$$

3.3.1 Line of Gaze Solution

If we assume that the gaze point lies on the intersection of left and right LoG, then using (3.3.6) and by the help of a triangulation method (see Appendix-D for a brief explanation of triangulation methods), one can find 3-D gaze point. If we assume gaze point lies on the screen plane, then a 2-D solution is required. 2-D solution can be obtained by the intersection of LoG with the screen plane. This solution is denoted as *2-D LoG solution*.

3.3.2 Line of Sight Solution

Once center of eyeball and LoG for both eyes are obtained, if the principal LoG and principal LoS are known, one can find alignment of current LoS by using (3.2.8) to (3.2.10) and (3.3.6). Then, by using pupil positions (3.3.5), one can form current LoS for both eyes. Since gaze point lies on the intersection of left and right LoS, by the help of a triangulation method, 3-D gaze point could be obtained. For the 2-D solution, similar to LoG case, intersection of LoS with the screen plane could be used. This solution is termed as *2-D LoS solution*.

3.3.3 Reducing Uncertainty in the Solutions

Triangulation methods should be employed for 3-D gaze estimation. However, Hartley and Zisserman [22] points to the triangulation uncertainty, as illustrated in Figure 3.3.1.

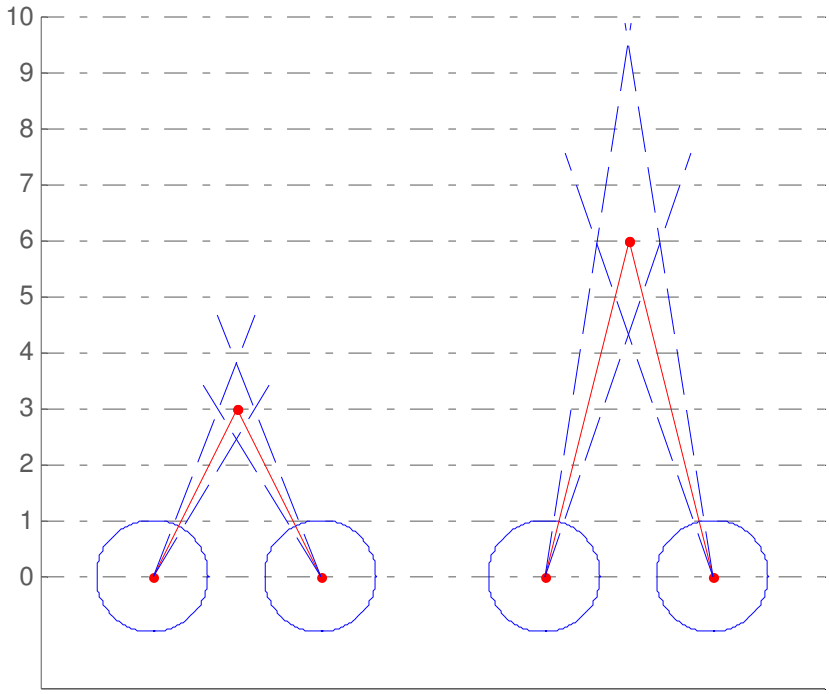


Figure 3.3.1 : Triangulation uncertainty (adapted from [22])

In Figure 3.3.1, triangulation uncertainty is illustrated while gaze distance is doubled. While the angle between two lines getting smaller, uncertainty of triangulation result in z- direction increases much faster with respect to the uncertainty in x- and y-directions. In proposed setup, the distance between two eyeball centers is approximately 65mm and distance between subject and gaze point is approximately 3m. Then, the angle between left and right LoS (or LoG) will be very small which should result in a high uncertainty in z- direction.

To reduce uncertainty, two different approaches might be employed. At each frame, left and right LoS (LoG) are obtained. However, subject's gaze point do not change at each frame. Average fixation duration of subjects is approximately 125ms which corresponds to 3 frames for a 25fps camera. Whenever fixation duration is longer than one frame, temporal data can be utilized to reduce uncertainty. During experiments, longer fixation durations are observed (300ms, 7-8 frames).

Temporal data may be used in two different ways.

- **Using multiple lines for triangulation:** Rather than using two lines (one frame) for triangulation, multiple lines can be used. When we have multiple lines, least squares triangulation is a good alternative for triangulation. This solution will be denoted as *multi-line least squares solution* (multi-line LS) (see Appendix-D). Multi-line LS is applicable for both LoS and LoG solutions.
- **Using maximum likelihood estimate:** Result of each frame could be obtained independently. Since we do not expect the gaze point change during a certain number of frames, one can average result of consecutive frames to obtain maximum likelihood estimate of gaze point. For triangulation of each frame, least squares triangulation could be employed. This solution will be denoted as *averaged least squares solution* (averaged LS). Similar to multi-line LS, this method is also applicable for LoG and LoS solutions.

Performance of these two alternatives will be compared during both simulations and experimental setup with different amount of temporal information. In summary, there are four different 3-D solutions, namely

- multi-line LS LoS solution,
- averaged LS LoS solution,
- multi-line LS LoG solution, and
- averaged LS LoG solution.

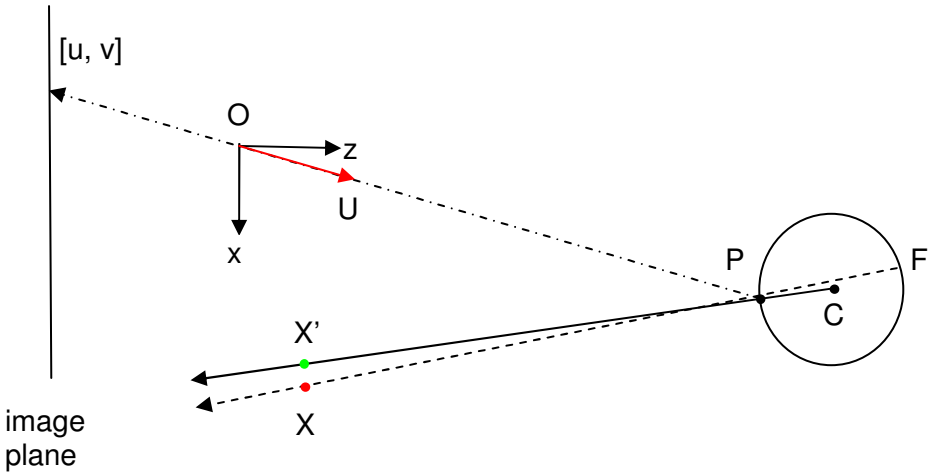
to be compared in simulations and experiments.

3.4 Calibration of Personal Parameters

In the previous section, two different methods for estimation gaze point are defined. However, both methods require some personal parameters. LoG solution requires center of eyeball and eyeball radius of the viewer. For LoS solution, principal LoG and principal LoS are also required. In this section, with a set of known gaze points, a procedure for calibrating personal parameters will be proposed.

3.4.1 Center of Eyeball and Eyeball Radius Estimation

For center of eyeball estimation, let's assume gaze point lies on the LoG. Let transformation from HCS to CCS, normalized unit vector of pupil and gaze point be available. Geometry of the scene is illustrated in Figure 3.4.1.



O : origin (the optical center of the camera)
[u, v] : the projection of the pupil on to the image plane
U : unit vector from origin to pupil
P : pupil
F : Fovea
C : the center of eyeball
X : gaze point
X' : the projection of gaze point on to LoG

Figure 3.4.1 : Geometry of the scene while viewer gazing on to a point X

The position of the pupil can be obtained by following equation

$$P_i = C_i + r\frac{X_i - C_i}{|X_i - C_i|} \tag{3.4.1}$$

Replacing P in (3.3.3) with this result, one could obtain

$$a_i U_i = C_i + r\frac{X_i - C_i}{|X_i - C_i|} \tag{3.4.2}$$

For $b_i = \frac{r}{|X_i - C_i|}$

$$(1-b_i)C_i + b_i X_i - a_i U_i = 0 \tag{3.4.3}$$

For $b_i' = b_i/(1-b_i)$ and $a_i' = a_i/(1-b_i)$

$$\Rightarrow R_{H,i}^{C}\,{}^{H}C + b_i' X_i - a_i' U_i = -t_{H,i}^{C} \tag{3.4.4}$$

As it can be observed in (3.4.4), there are three equations and it is required to solve two new parameters for each gaze point. Therefore, to obtain the eyeball center, at least 3 gaze points are required. For N gaze points, the eyeball center can be found with the pseudo-inverse solution of the following system of equations:

$$\begin{bmatrix} R_{H,1}^{C} & X_1 & -U_1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{H,N}^{C} & 0 & 0 & \cdots & 0 & X_N & -U_N \end{bmatrix} \begin{bmatrix} {}^{H}C \\ b_i' \\ a_i' \\ \vdots \\ b_i' \\ a_i' \end{bmatrix} = \begin{bmatrix} -t_{H,1}^{C} \\ \vdots \\ -t_{H,N}^{C} \end{bmatrix} \tag{3.4.5}$$

As the eyeball radius initial estimate, one can use either a theoretic value 12.5mm or using b_i', one can find eyeball radius for each gaze instant with the following relation:

$$r_i = \frac{b_i'}{(1+b_i')}|X_i - C_i| \tag{3.4.6}$$

Since we expect a constant eyeball radius, the averages of radius estimates could also be utilized

$$\Rightarrow r = \frac{1}{N}\sum r_i \tag{3.4.7}$$

The eyeball center and radius estimates are obtained by solving (3.4.5) and given in (3.4.7) can be directly used in the LoG solution. For the LoS solution, these values could only be initial estimates of another optimization procedure that will be defined in the next sections.

3.4.2 Estimation of Principal LoG

As an initial estimate of principal LoG, z-axis of HCS can be used.

$$\hat{G}_0 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T \tag{3.4.8}$$

A better estimate can be obtained by following procedure. Once eyeball center and radius (or their initial estimates) are obtained, for each gaze instant, current LoG and LoS can be found. (3.2.11) can be rearranged as

$$S_0 = \cos(\alpha_i) S_i - \sin(\alpha_i)(S_i \times V_i) + (1 - \cos(\alpha_i))(S_i \cdot V_i) V_i \tag{3.4.9}$$

In this relation, one could replace α_i and V_i with (3.2.8) and (3.2.9) in (3.3.9) to obtain

$$S_0 = S_i (G_0 \cdot G_i) - S_i \times (G_0 \times G_i) + \frac{(G_0 \times G_i)(S_i \cdot (G_0 \times G_i))}{1 + G_0 \cdot G_i} \tag{3.4.10}$$

(3.4.10) can be expressed in following form

$$\begin{bmatrix} G_0^T & 1 \end{bmatrix} M \begin{bmatrix} G_0 \\ 1 \end{bmatrix} + \begin{bmatrix} G_0^T & 1 \end{bmatrix} N \begin{bmatrix} S_0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \end{bmatrix} \tag{3.4.11}$$

where M and N are 4x4 matrices determined by G_i and S_i.

The solution for (3.3.11) could be obtained either by a non-linear optimization or assigning new variables to quadratic terms and applying pseudo-inverse solution to obtain initial estimate for principal LoG.

3.4.3 Calibrating Personal Parameters for LoS Solution

Using the solutions of (3.4.5), (3.4.7) and either (3.4.8) or (3.4.11), one can obtain the initial estimates for the eyeball center, the eyeball radius and the principal LoG. Using these initial estimates, the pupils in 3-D and the orientation of the eyeball are determined. Calibration of personal parameters can be expressed as an optimization problem with the following error and cost functions,

$$e_i = X_i - X_i' \tag{3.4.12}$$

$$J = \sum |e_i|^2 \tag{3.4.13}$$

where X_i is the i[th] known gaze point and X_i' is the projection of gaze point on the estimated current LoS. The projection of gaze point on to estimated current LoS is given by following equation

$$X_i' = \hat{P}_i + ((\hat{P}_i - X_i) \cdot \hat{S}_i)\hat{S}_i, \tag{3.4.14}$$

which minimizes $|X_i - X_i'|$.

Once the eyeball and the pupil are located, one can get S_i with (3.2.1). Let X_i' and X_i be replaced with (3.2.1) in (3.4.12),

$$\begin{aligned} e_i &= \hat{P}_i + aS_i - \hat{P}_i - b\hat{S}_i \\ e_i &= aS_i - b\hat{S}_i \end{aligned} \tag{3.4.15}$$

$$\Rightarrow |e_i|^2 = a^2 + b^2 - 2ab(S_i \cdot \hat{S}_i) \tag{3.4.16}$$

In (3.4.16), since $a > 0$, $b > 0$ and $S_i \cdot \hat{S}_i > 0$, $\min(|e_i|^2)$ can be obtained with $a = b$, then

$$e_i = a(S_i - \hat{S}_i) \tag{3.4.17}$$

When we multiply both sides of (3.3.17) with $R_{E,i}^T$, $|e_i|^2$ remains same,

$$e_i' = a(R_{E,i}^T S_i - \hat{S}_0)$$

S_0 which minimizes cost function can be obtained by solving following equation system

$$\begin{bmatrix} I \\ \vdots \\ I \end{bmatrix} \hat{S}_0 = \begin{bmatrix} R_{E,1}^T S_1 \\ \vdots \\ R_{E,N}^T S_N \end{bmatrix} \tag{3.4.18}$$

where I is 3x3 identity matrix. Solution to (3.4.18) is simply averaging $R_{E,i}^T S_i$ for *N* gaze points. Then, for the given eyeball center, radius and principal LoG, principal LoS which minimizes cost function are given below

$$\hat{S}_0 = \frac{1}{N} \sum R_{E,i}^T S_i \tag{3.4.19}$$

It is required to find ten personal parameters for LoS solution (C, r, G_0, S_0). However, once eyeball center, eyeball radius and principal LoG are set, the optimum principal LoS can be obtained by (3.4.19). Therefore, we have a six parameter non-linear optimization problem (since principal LoG has unit magnitude, we need to solve two parameters for it) whose error and cost functions are defined in (3.4.12) and (3.4.13). To solve this problem, Levenberg-Marquardt minimization algorithm could be employed.

3.5 Head Pose Estimation

In the previous sections, we have investigated the methodology for estimating gaze point and calibration of personal parameters, when HCS and transformation from HCS to CCS are known. The next problem is to define HCS and find transformation form HCS to CCS. We might define such a coordinate system relative to a calibration object, which is rigidly connected to viewer's head. It is also possible to utilize some fixed points on the face as well. For simplicity and

robustness, a calibration object is preferred and this object is selected as a checkerboard (it may be glasses as well). We define x-axis from lower left corner to lower right corner, whereas y-axis form lower left corner to upper left corner of checkerboard as shown in Figure 3.5.1.

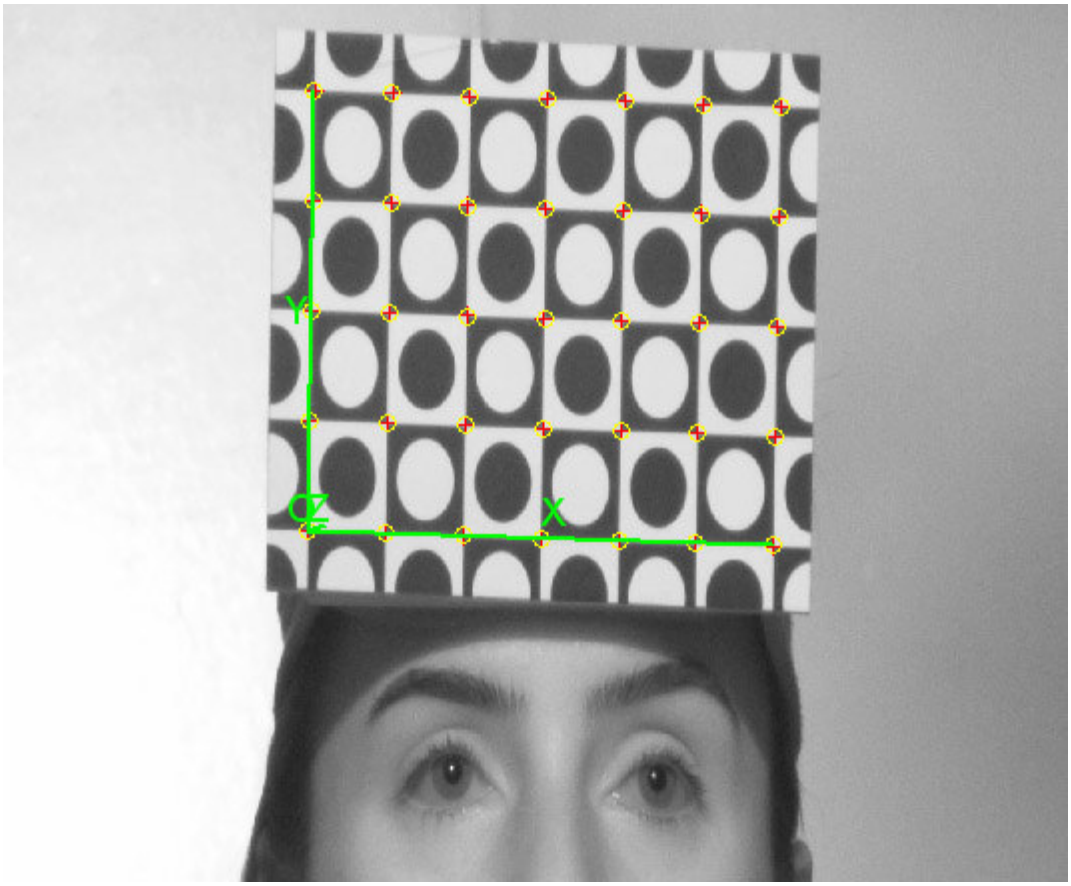


Figure 3.5.1 : Calibration object mounted on the viewer's head and the head coordinate system

Once we define a coordinate system relative to checkerboard as well as the spacing between grids on checkerboard, by using the methods that are mentioned in Appendix-D, one can determine the transformation between HCS and CCS at any instant with given internal camera parameters. This transformation is required to transform eyeball center, principal LoG and principal LoS into the CCS.

3.6 Pupil Detection

Knowing center of eyeball and projecting it onto the image plane, one can bound the region in which the pupil is searched. Since one can obtain a candidate region for pupil detection and real-time processing is not required, an ellipse detection algorithm [34] in the candidate area could be executed to extract pupil centers. It should be noted that for pupil detection, illumination of the environment is critical. Since we need to see checkerboard as well as pupils, we use 24 IR LEDs placed nearly on optical axis of camera to obtain bright pupil effect. Another option for environment illumination may be to use an off-axis IR light source and illuminating whole environment. A sample (dark pupil) image with projections of eyeball centers and candidate pupil detection regions is given in Figure 3.6.1.

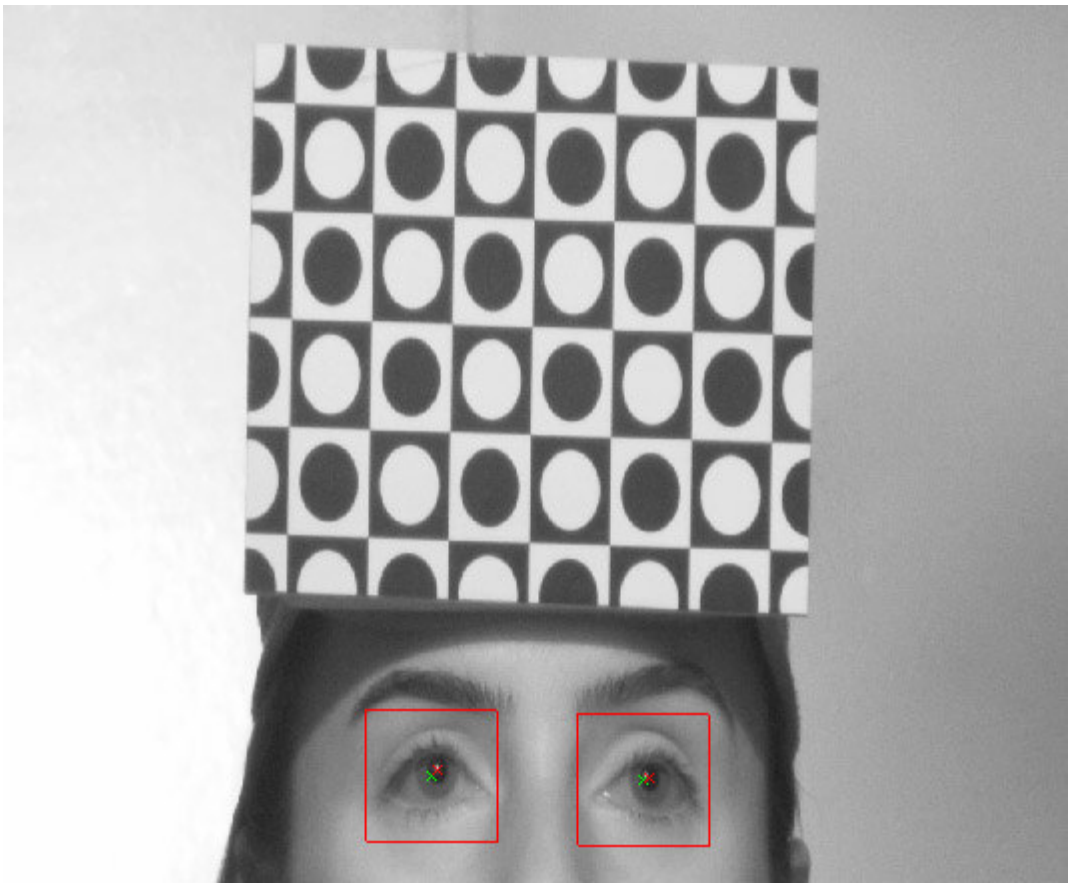


Figure 3.6.1 : The candidate pupil regions (red rectangles), the projection of the eyeball centers (green crosses) and pupil centers (red crosses).

A sample pupil detection result is presented in Figure 3.6.2.

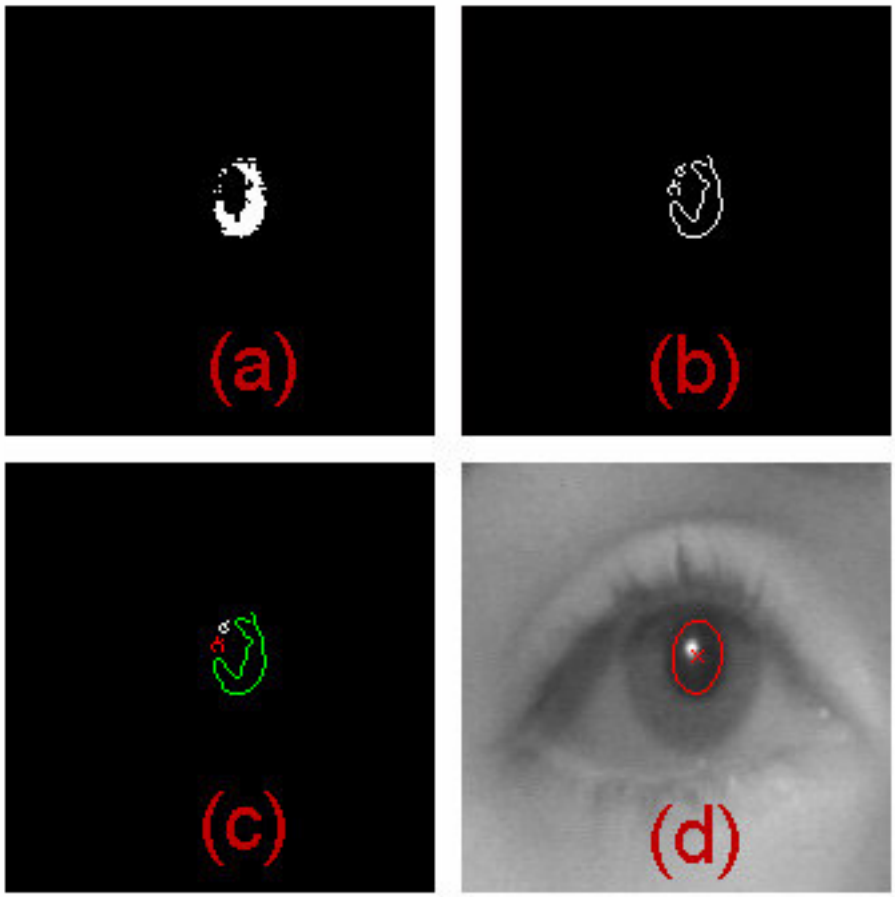


Figure 3.6.2 : A sample pupil detection result, **(a)** binary thresholding of candidate pupil region **(b)** edge detection results **(c)** connected component labeling **(d)** extracted pupil center

3.7 Display Camera Pose Estimation

In the proposed method, reference coordinate system is CCS; in other words, it estimates 3-D gaze point in the camera coordinate system. In order to relate the outputs of this method to the scene content, one requires determining the transformation between the reference coordinate system and coordinate system relative to which scene content is defined. If we want to relate our gaze estimation results to content on a display, then a method to estimate transformation between CCS and display coordinate system (DCS) is required.

In Appendix-C, some methods for finding transformation between CCS and a calibration object are presented. If the display is in the field of view of camera, we should simply place/project a calibration object onto this display and use known calibration methods to determine display-camera transformation. However, in our case, the display is not placed in the field of view. In order to solve this problem, Beymer and Flickner [12] proposed placing a reflective mirror between camera and display. Shih and Liu [16] also use a similar approach for positioning LEDs relative to camera. The method proposed by Beymer and Flickner [12] is employed by placing a mirror between the camera and the display. Hence, we will be able to present a calibration object on the display, while its reflection in field of view is also kept.

A point X and its reflection on a planar reflective surface X' has related with following equation [12]:

$$X' = X - 2\big((X - J)\cdot N\big)N \tag{3.7.1}$$

where N is normal to the reflective surface and J is an arbitrary point on the surface. As this equation implies, the reflection of a point on a mirror is dependent to the position and orientation of the mirror. Another calibration object must be placed on to mirror in order to estimate its position and orientation. At this point, we define four coordinate systems that are camera coordinate system (CCS), display coordinate system (DCS), mirror coordinate system (MCS) and reflected coordinate system (RCS).

For this transformation estimation, 2D calibration objects are utilized in both mirror and display. Display coordinate system is defined by its origin on the display, whereas x- and y-axis are horizontal and vertical borders of the display and z-axis is pointing out of the display. On the other hand, MCS is defined by its origin on its reflective surface; x- and y-axis are lying on this surface and z-axis pointing out of this reflective surface. Reflected coordinate system is defined as its origin on reflection of origin of DCS, $\hat{x}_R$ and $\hat{y}_R$ as reflection of $\hat{x}_D$ and $\hat{y}_D$ respectively and $\hat{z}_R = \hat{x}_R \times \hat{y}_R$. A sample image utilized for display pose estimation is

presented in Figure 3.7.1 on which mirror and reflected coordinate systems are both marked.

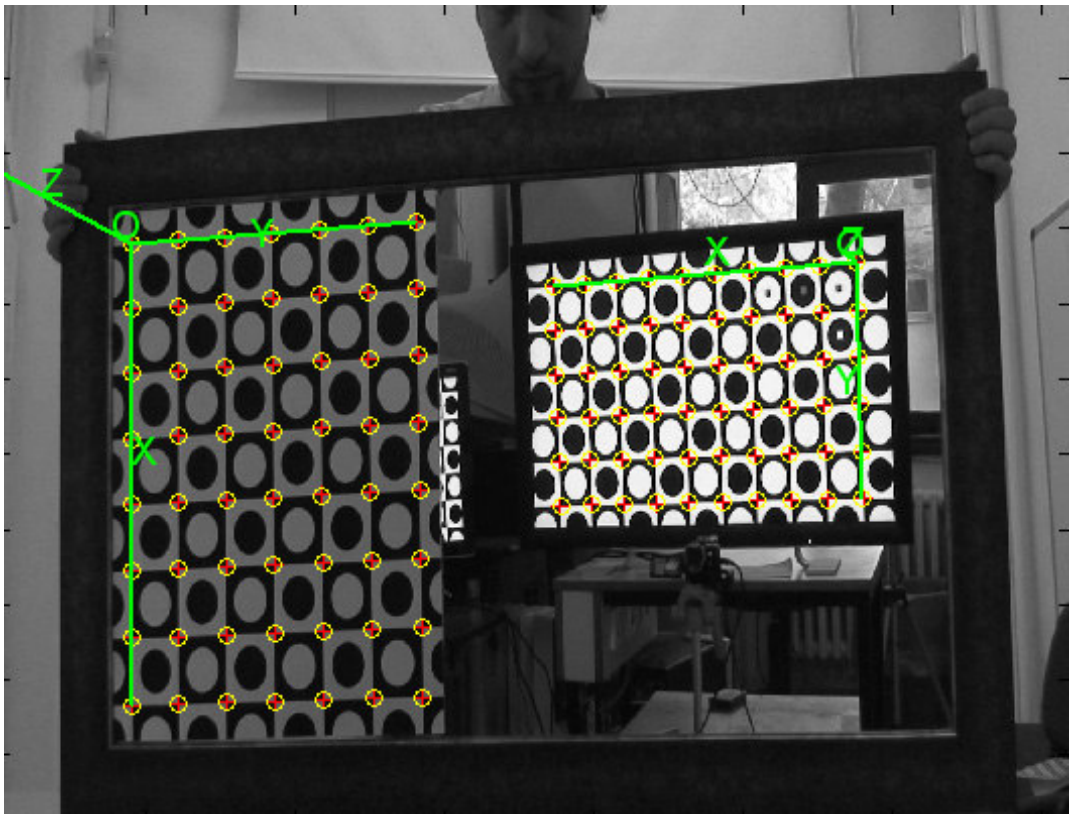


Figure 3.7.1 : An image used for display-camera pose estimation. Mirror and reflected coordinate systems are marked.

In these coordinate systems, a point in DCS, ${}^{D}X$, and its reflection in MCS, ${}^{M}X'$ are related with the following equation [12]:

$$ {}^{M}X' = diag(1 \quad 1 \quad -1 \quad 1) T_{D}^{M} \begin{bmatrix} {}^{D}X \\ 1 \end{bmatrix} \tag{3.7.2} $$

Transformation from DCS to MCS can be defined as

$$ T_{D}^{M} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.7.3} $$

Then using (3.7.2) and (3.7.3) and definition of RCS, one can find ${}^{M}O_R$, ${}^{M}\hat{x}_R$, ${}^{M}\hat{y}_R$ and ${}^{M}\hat{z}_R$ in terms of T_D^M.

$$ {}^{M}O_R = {}^{M}O'_D \quad \Rightarrow {}^{M}O_R = \begin{bmatrix} t_1 & t_2 & -t_3 \end{bmatrix}^T $$

$$ {}^{M}\hat{x}_R = {}^{M}\hat{x}'_D \quad \Rightarrow {}^{M}\hat{x}_R = \begin{bmatrix} r_{11} & r_{21} & -r_{31} \end{bmatrix}^T $$

$$ {}^{M}\hat{y}_R = {}^{M}\hat{y}'_D \quad \Rightarrow {}^{M}\hat{y}_R = \begin{bmatrix} r_{12} & r_{22} & -r_{32} \end{bmatrix}^T $$

$$ {}^{M}\hat{z}_R = {}^{M}\hat{x}_R \times {}^{M}\hat{y}_R \quad \Rightarrow {}^{M}\hat{z}_R = \begin{bmatrix} -r_{13} & -r_{23} & r_{33} \end{bmatrix}^T $$

Since transformation matrix is defined as $T_R^M = \left[\begin{array}{ccc|c} {}^{M}\hat{x}_R & {}^{M}\hat{y}_R & {}^{M}\hat{z}_R & {}^{M}O_R \\ \hline 0 & 0 & 0 & 1 \end{array}\right]$,

$$ T_R^M = \begin{bmatrix} r_{11} & r_{12} & -r_{13} & t_1 \\ r_{21} & r_{22} & -r_{23} & t_2 \\ -r_{31} & -r_{32} & r_{33} & -t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.7.4} $$

At this point, the transformation from RCS to CCS, T_R^C, and the transformation from MCS to CCS, T_M^C, are required and one can find them with methods defined in Appendix-C. Then, by using T_R^C and T_M^C, T_R^M can be found

$$ T_R^M = \left(T_M^C\right)^{-1} T_R^C \tag{3.7.5} $$

Using (3.7.4) and (3.7.5), one can find T_D^M, and finally, using T_D^M and T_M^C, T_D^C can be obtained as

$$ T_D^C = T_M^C T_D^M \tag{3.7.6} $$

Although, it seems that only one image is enough to estimate T_D^C, since pose estimation is minimization of reprojection errors rather than an absolute solution, a series of images should be utilized with different mirror orientations, to find T_D^C.

3.8 Computer Simulations

In order to observe the performance of the proposed system in a controlled environment, a simulator is implemented with the following capabilities:

- generating subjects,
- simulating the head movements of a subject,
- simulating the eyeball movements of a subject,
- projecting the HCS and pupils on to image plane,
- simulating the errors in pupil detection and head pose estimation,
- simulating the errors in internal camera calibration, and
- simulating the errors in display camera pose estimation.

Before computer simulations, a set of initial experiments are conducted and some useful data is recorded to be used by the simulator.

Personal parameters: Using the personal calibration procedure defined in Section 3.4.3, eyeball centers, eyeball radius, principal LoG and principal LoS are obtained for different subjects with the data recorded in the initial experiments. Then, in order to generate a random subject, a random transformation is applied to eyeball centers, a random scaling is applied to the eyeball radius and distance between eyeball centers and a random rotation is applied to the principal LoG and LoS. Applied translation, scaling and rotation parameters are selected through a normal distribution with parameters given in Table 3.8.1.

Table 3.8.1 : Parameters for random subject generation (normal distribution, mean ± std)

Distance between eyeball centers (mm)	65 ± 2
Eyeball radius (mm)	12.5 ± 0.3
Rotation of eyeball centers (ZYX Euler angles in degrees)	[0 0 0] ± [4 4 4]
Translation of eyeball centers (mm)	[0 0 0] ± [10 10 10]
Rotation of principal LoG and LoS (ZYX Euler angles in degrees)	[0 0 0] ± [4 4 4]
Angle between LoG and LoS (θ, degrees)	2 ± 0.3

Main steps of the random subject generation is listed below.

1. Scale the distance between eyeball centers by translating one of the eyeball centers.
2. Scale each eyeball radius independently.
3. Rotate and translate eyeball centers together in order to simulate different orientations of calibration object on viewers' head.
4. Rotate and translate principal LoG in order to simulate different orientations of calibration object on viewers' head (use a different random transformation than Step-3, so simulate the subjective variances).
5. Pick a random LoG/LoS angle (θ) for each eye.
6. Scale the parallel and perpendicular components of LoS to LoG with cos(θ) and sin(θ).

Head and eyeball movements: During the initial experiments, the head pose of different subjects, while gazing on to different points on screen, is recorded. Then, by using the mean and standard deviation of the recorded head pose parameters, the range in which the head can move is defined. While subjects are gazing on to different points, the angle between the principal LoG and current LoG is recorded. This value is used to define the range in which subjects' eye might rotate. Recorded parameters are presented in Table 3.8.2.

Table 3.8.2 : Head pose and eyeball orientation parameters during and between gaze instants (normal distribution, mean ± std)

Translation between HCS and CCS (mm)	[-57 67 1373] ± [9 4 13]
Rotation between HCS and CCS (ZYX Euler angles in degrees)	[1.4 -7.7 -167] ± [0.7 2.3 2.0]
Change in translation during gaze (mm)	[0 0 0] ± 10^{-2} [12 6 17]
Change in orientation during gaze (ZYX Euler angles in degrees)	[0 0 0] ± 10^{-2} [22 34 37]
Angle between principal LoG and current LoG (α) (degrees)	7 ± 3.1

Internal camera parameters: Internal camera parameters as well as their uncertainties are also recorded during the initial experiments. Internal parameters and their uncertainties are presented in Table 3.8.3. Uncertainties of internal parameters are used to simulate errors in camera calibration. Errors are modeled with zero mean Gaussian random variables and given uncertainties are used as standard deviations of the random variables.

Table 3.8.3 : Estimated internal camera parameters with uncertainties (mean ± std)

Focal length	[4363.4 5820.9] ± [2.24 2.90]
Principal point	[729 546] ± [5.58 4.84]
Radial and tangential distortion	[.023 -.717 -.0011 .0026] ± [.011 .414 .0002 .0004]

Display-Camera Transformation: Display-camera transformation parameters and their uncertainties are also recorded during initial experiments. The recorded transformation values and their uncertainties are presented in Table 3.8.4. Uncertainties of transformation parameters are used to simulate errors in display-camera pose estimation. Errors are modeled with zero mean Gaussian random variables and given uncertainties are used as standard deviations of the random variables.

Table 3.8.4 : Estimated display camera transformation parameters with uncertainties (mean ± std)

Translation between DCS and CCS (mm)	[443.4 806.7 1341.5] ± [31.5 15.1 15.5]
Rotation between HCS and CCS (ZYX Euler angles in degrees)	[179.4 -1.4 -162.4] ± [0.17 0.70 2.25]

The flowchart of the simulator is presented in Figure 3.8.1. All uncertainties are modeled with zero mean Gaussian random variables. Random values are generated based on the values presented in Tables 3.8.2, 3.8.3 and 3.8.4. Variables a_1, and a_2 are used to set pupil detection and corner detection uncertainty respectively. Variables a_3 and a_4 are used to adjust the uncertainties in internal camera parameters and display camera transformation, by scaling standard deviations of random variables.

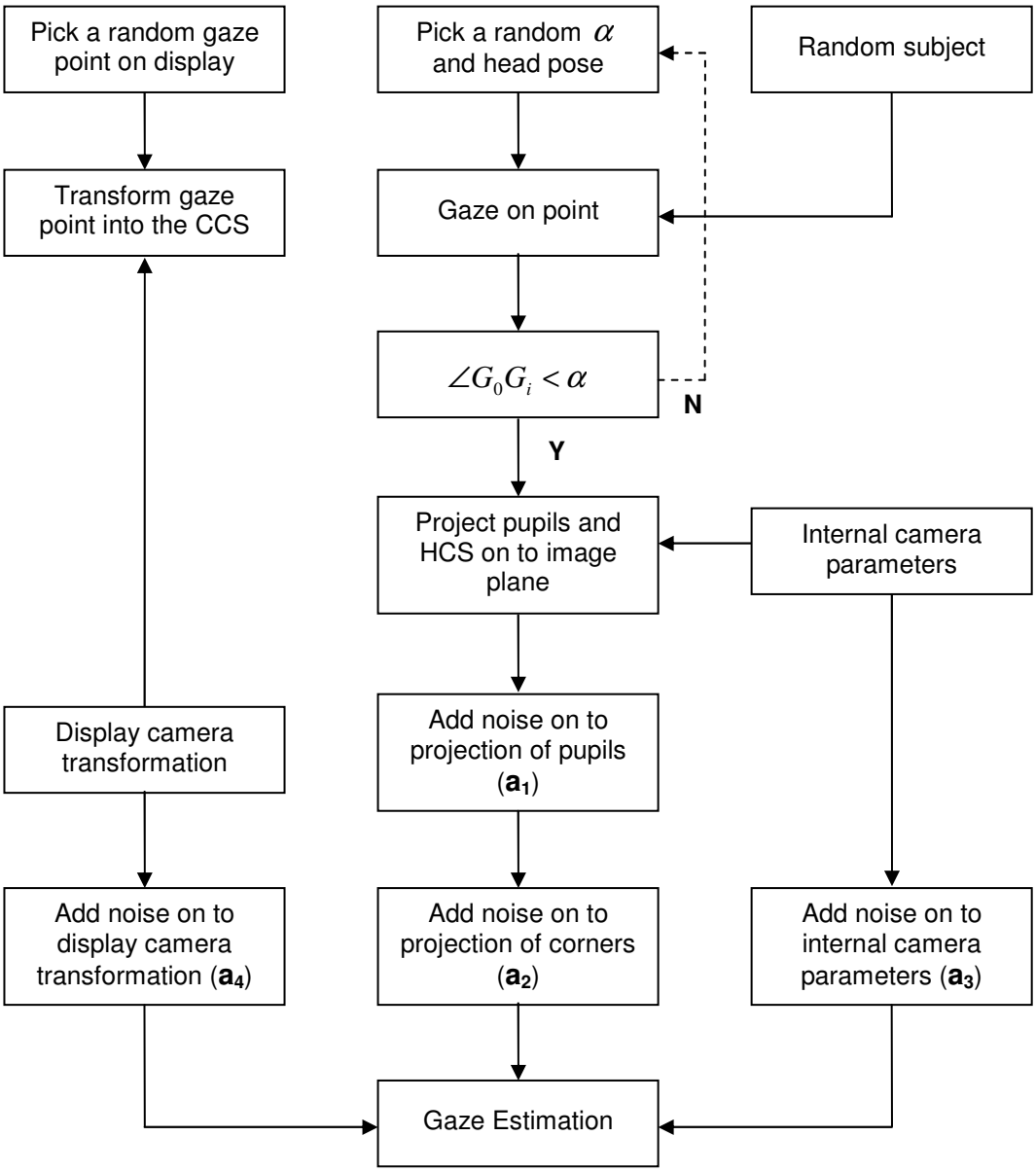


Figure 3.8.1 : The flowchart of the simulator

Main steps of the simulation is listed below.

1. Generate a random subject.
2. Pick a random point on display to gaze on.
3. Projected this point in to the CCS with noise free display camera transformation (use expected values given in Table 3.8.4).
4. In order to subject gaze on to this point, select a random head pose and principal LoG current LoG angle using Table 3.8.2.
5. Compute the orientation of the eyeball while gazing onto this point.
6. If the angle between principal LoG and current LoG is smaller than the angle selected in Step-4 continue, otherwise return to the Step-4.
7. Project corners on calibration object and pupils onto the image plane with the transformation matrix selected in Step-4 and noise free internal camera parameters (use expected values given in Table 3.8.3).
8. Add zero mean Gaussian noise onto the projection of pupils and corners with standard deviations a_1 and a_2.
9. Add zero mean Gaussian noise to the internal camera parameters. Standard deviation should be obtained by scaling uncertainties presented in Table 3.8.3 with a_3.
10. Add zero mean Gaussian noise to the display camera transformation parameters. Standard deviation should be obtained by scaling uncertainties presented in Table 3.8.4 with a_4.

During simulations, first step and last two steps presented above should be performed once for each experiment. Other steps should be repeated for each gaze point.

3.9 Implementation Details

3.9.1 Setup

As capturing device, a HD camera sensitive to low length IR is utilized, (brand & model : SONY HDR-HC3). An autostereoscopic display is placed 3 meters away from the viewer and HD camera is located at mid-point between display and viewer as shown in Figures 3.9.1 and 3.9.2.

Figure 3.9.1 : Display-camera orientation

Figure 3.9.2 : Display-viewer-camera orientation

At display-camera calibration step, we use camera in day mode with zoom out. During gaze estimation, we use camera in night mode and zoom in. During gaze estimation step, the environment is also illuminated by 24 on-axis low length IR LEDs in order to detect pupils with more ease.

The camera is connected to a PC through IEEE1394 (firewire) connection, and it is triggered by a software that records mouse events. Since the system is expected to work offline, it is necessary to register recorded video with mouse events and displayed content.

Before we start gaze tracking, the viewer is obliged to wear the calibration object and gaze on the 40 different known points on the screen. At each gaze towards the known points, the viewers are asked to click the mouse to record gaze instants. The viewer is allowed to move his head $\pm[20 \quad 10 \quad 20]cm$ during the gaze estimation stage; such a limitation is required to keep the viewer and the calibration object in the field of view.

Main steps of a typical experimental data collection process are listed below.

1. Adjust the orientation of the camera, such that the subject will be in the FOV in zoom in mode.
2. Zoom out the camera and turn into the day mode. Disable auto-focus option, if enabled.
3. Calibrate internal parameters of the camera in zoom out (wide view) mode.
4. Perform display-camera pose estimation.
5. Place a checkerboard in FOV and record pose estimate. During next step keep this checkerboard fixed.
6. Zoom in camera and switch to the night mode and adjust focus such that a range of 80cm centered at expected position of the subject is sharpened. Such an operation is required, since different poses of a calibration object will be required for internal calibration of the parameters.
7. Record the checkerboard placed in previous steps, then remove this checkerboard and perform calibration of internal parameters.
8. Find pose estimate of the checkerboard recorded in previous step and find the transformation between two views of the camera, *switching transformation*. Then using this transformation update display camera pose estimation.
9. Display known points on display and ask subject to gaze on these points. Simultaneously record the mouse events.
10. Using the data obtained in previous step, estimate personal parameters.
11. Start gaze estimation.

During data collection, first eight steps should be performed once for each setup. Last three steps should be performed for each subject.

3.9.2 Calibration and Pose Estimation

For internal camera calibration and pose estimation for display-camera calibration and head pose estimation, Camera Calibration Toolbox [31] is used. Details about toolbox are given in Appendix-C.

3.9.3 Capturing HD Video

In order to record HD video, an open source software DVGRAB [17] working on Linux is used. Linux and DVGRAB is preferred, since Linux allows to start DVGRAB by a script easily and run in background.

3.10 Simulation Results

As shown in Figure 3.8.1, there are four different error sources is available within the proposed system:

- errors in internal camera parameters,
- errors in display-camera pose estimation,
- errors in head pose estimation, and
- pupil detection errors.

Throughout the simulations, random setups and subjects are generated to observe effect of each noise source on performance of the system. For each configuration 2-D and 3-D performances of the system are evaluated.

Error bars and mean absolute errors for different solutions are presented in the following figures with different error sources and levels. Errors are defined in DCS. For multi-line LS LoS and LoG solutions, seven consecutive frames (14 lines) are used. For averaged LS LoS solution, averaged result of seven consecutive frames is used. Performance of each solution is expressed with a different color:

- blue for averaged LS LoS solution,
- red for multi-line LS LoS solution,
- cyan for averaged LS LoG solution and
- green for 2-D LoS solution.

For different pupil detection accuracies, performance of the system on x-, y- and z-axis (relative to DCS) are presented in Figures 3.10.1, 3.10.2 and 3.10.3.

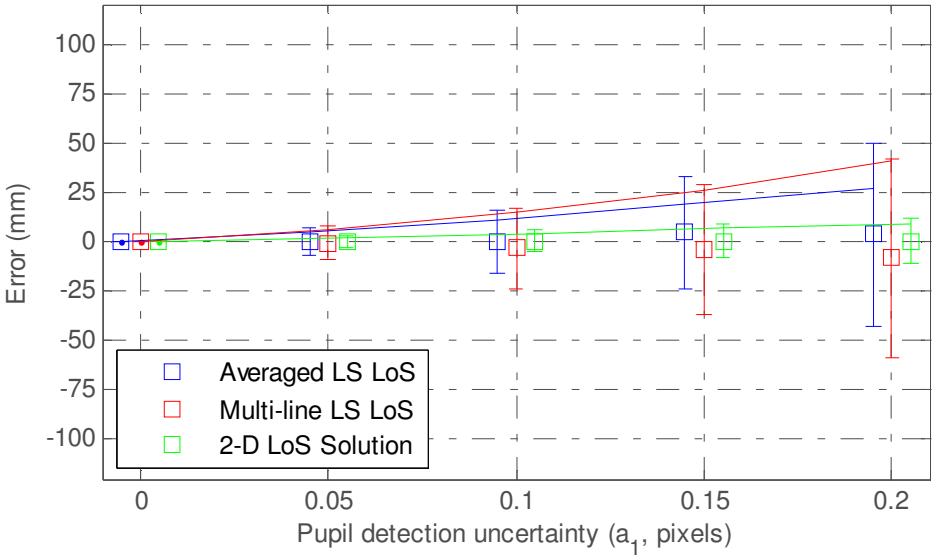


Figure 3.10.1 : Pupil detection uncertainty (a_1) vs. x-axis performance (a_2, a_3, a_4 = 0)

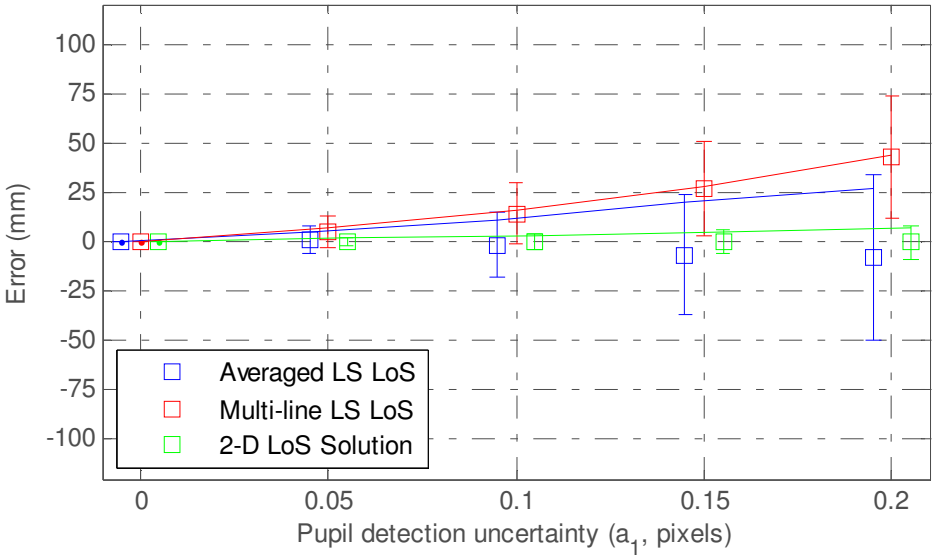


Figure 3.10.2 : Pupil detection uncertainty (a_1) vs. y-axis performance (a_2, a_3, a_4 = 0)

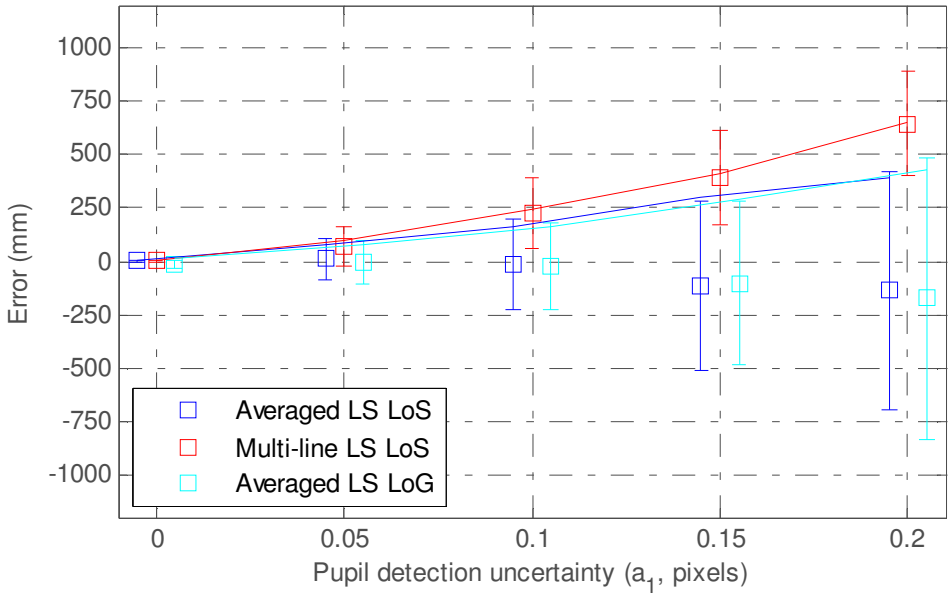


Figure 3.10.3 : Pupil detection uncertainty (a_1) vs. z-axis performance (a_2, a_3, a_4 = 0)

2-D LoS solution provides significantly better results on x- and y-axis for all levels of pupil detection uncertainty. When performance of multi-line LS LoS solution is considered, as pupil detection uncertainty increases, there is an increasing bias for this solution on y- and z-axis. During these simulations camera was standing at midpoint between viewer and display, 40cm below the viewer's head. Therefore in order to keep subject in FOV, camera is tilted up. Due to camera orientation, pupils are always below the principal point on y-axis. A zero mean pupil detection noise results in a non-zero mean angular error on y-z plane of the CCS, that pupils are estimated at a lower position on y axis on the average. When camera orientation is changed (translated upwards on y-axis and rotated around x- axis counter clockwise) so pupils are kept around the principal point, performance of the system on z-axis is presented in Figure 3.10.4. This orientation is named as *forward-looking* camera position.

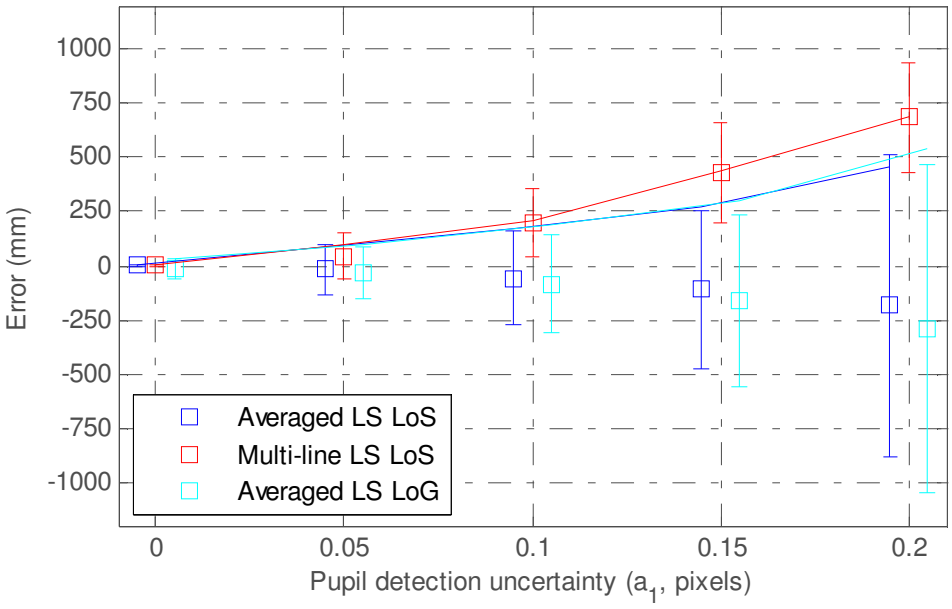


Figure 3.10.4 : Pupil detection uncertainty (a_1) vs. z-axis performance (a_2, a_3, a_4 = 0) (forward looking camera position)

When compared to results obtained with tilted camera position (Figure 3.10.3), changing camera location does not yield a significant difference when only pupil detection error is present (tilted position is even better). For tilted and forward looking camera positions, effect of utilized amount of temporal data is presented in Figures 3.10.5 and 3.10.6. Results for tilted camera position is slightly better than forward looking camera position. Moreover, as the utilized number of frames are increased, bias on the multi-line LS LoS solution increases for both camera positions.

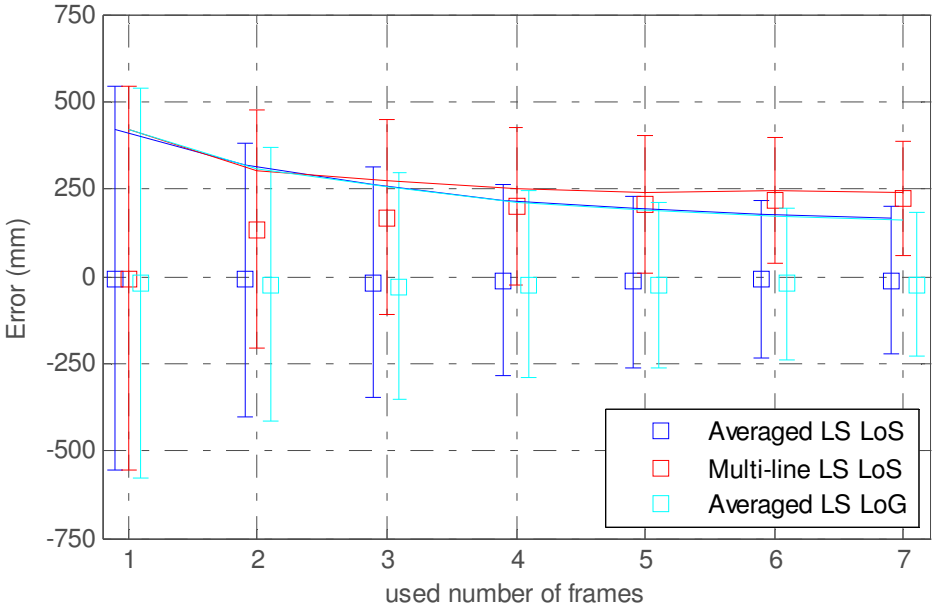


Figure 3.10.5 : Used number of frames vs. z-axis performance with pupil detection uncertainties (a_1 = 0.1, a_2, a_3, a_4 = 0)

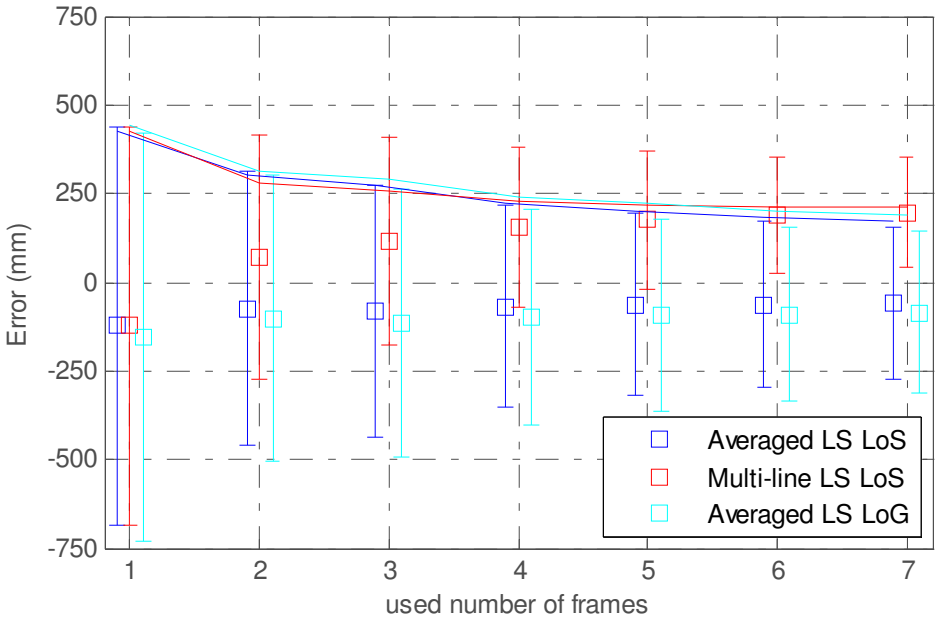


Figure 3.10.6 : Used number of frames vs. z-axis performance with pupil detection uncertainties (a_1 = 0.1, a_2, a_3, a_4 = 0) (forward looking camera position)

For different corner detection uncertainties, performance of the system is presented in Figures 3.10.7, 3.10.8 and 3.10.9.

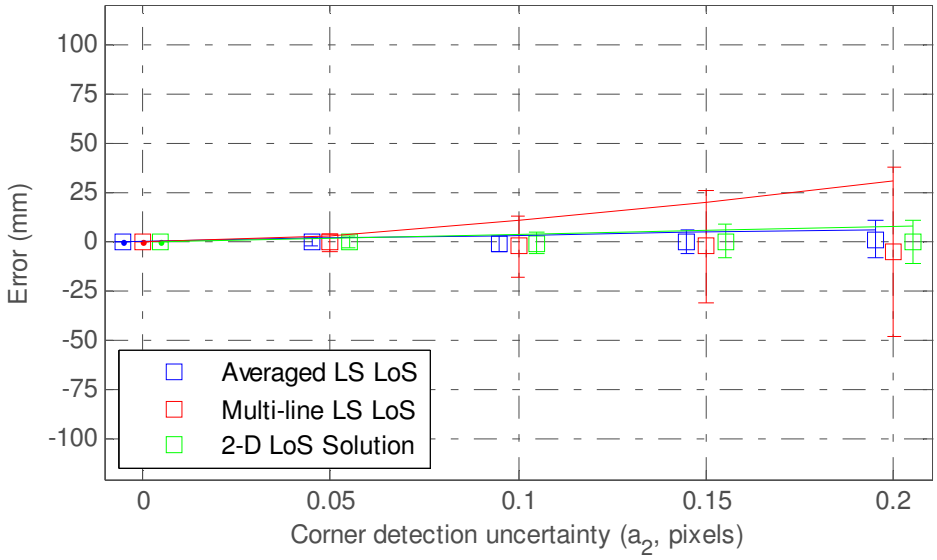


Figure 3.10.7 : Corner detection uncertainty (a_2) vs. x-axis performance (a_1, a_3, a_4 = 0)

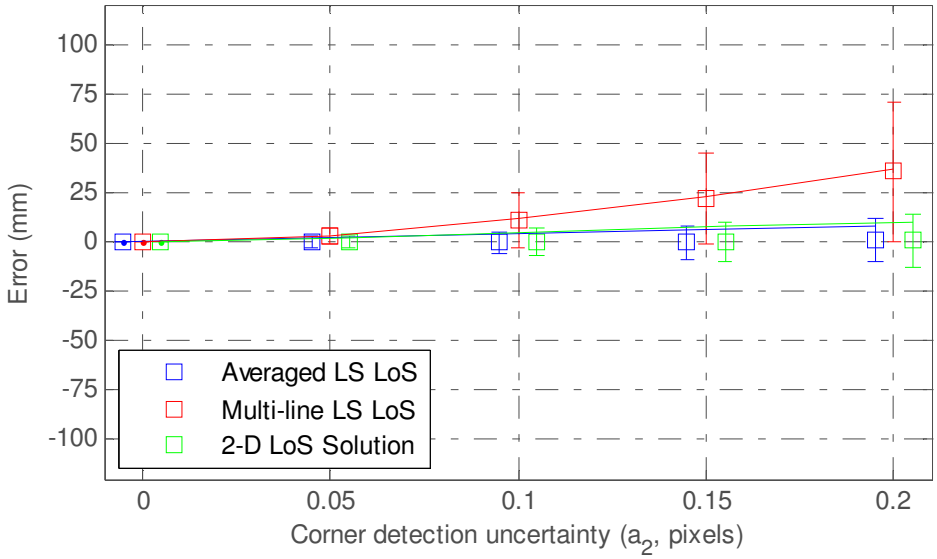


Figure 3.10.8 : Corner detection uncertainty (a_2) vs. y-axis performance (a_1, a_3, a_4 = 0)

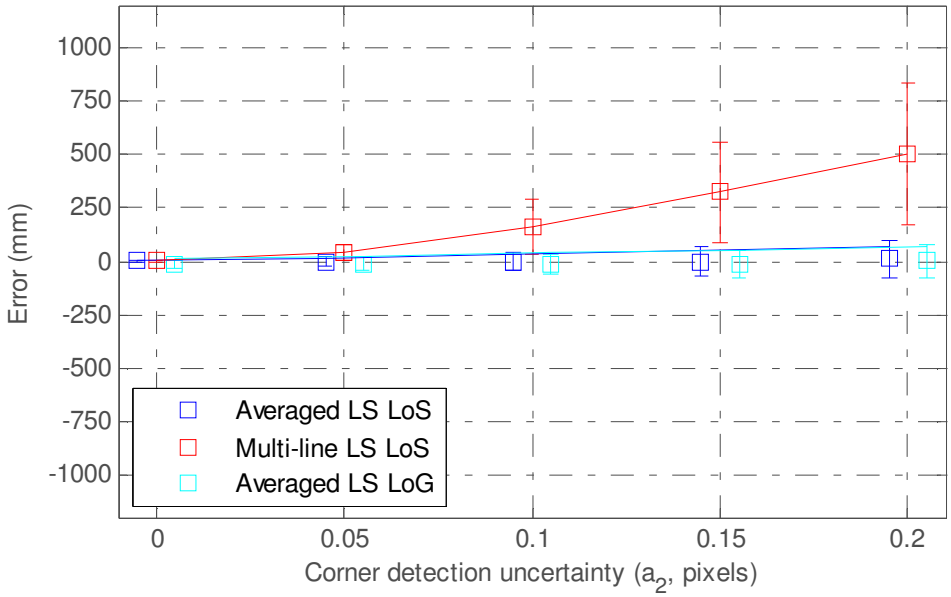


Figure 3.10.9 : Corner detection uncertainty (a_2) vs. z-axis performance (a_1, a_3, a_4 = 0)

When corner detection errors are considered, averaged LS solutions generate better results than multi-line LS LoS solution. A bias on multi-line LS LoS solution is observed for corner detection errors as in pupil detection errors. Performance of the system on z-axis at forward looking camera position for different corner detection uncertainties is presented in Figure 3.9.10.

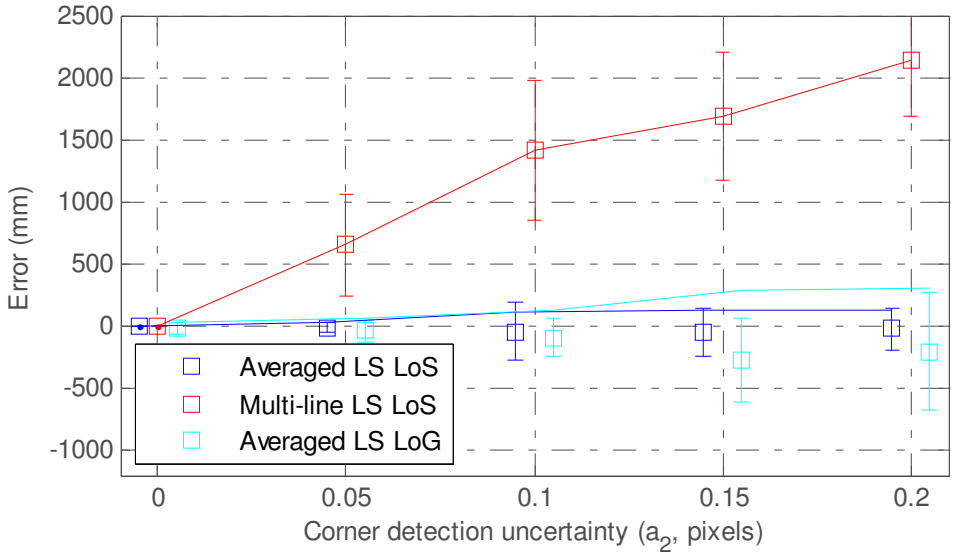


Figure 3.10.10 : Corner detection uncertainty (a_2) vs. z-axis performance (a_1, a_3, a_4 = 0) (forward looking camera position)

At tilted camera position, most of the corners on checkerboard on viewers' head are above the principal point. However, at forward looking camera position, since orientation of the camera is adjusted to keep pupils around principal point, corners on checkerboard are projected to further points to the principal point. In such a setup, zero mean corner detection errors result in a non-zero angular error, which makes the head pose estimate to be significantly biased. From the Figures 3.10.4 and 3.10.10, one can conclude that camera position should be adjusted to keep checkerboard around the principal point, rather than pupils.

For different internal calibration uncertainties, the performance of the system on z-axis is presented in Figure 3.10.11. Errors on x- and y- axis are approximately 10% of the errors on z-axis.

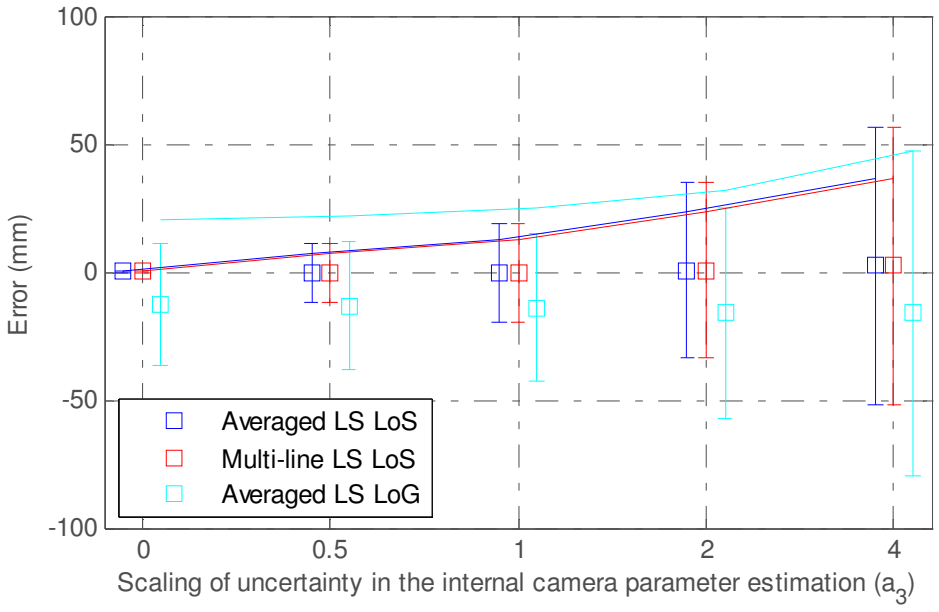


Figure 3.10.11 : Internal calibration uncertainty (a_3) vs. z-axis performance (a_1, a_2, a_4 = 0)

For different display-camera pose estimation uncertainties, the performance of the system on z-axis is presented in Figure 3.10.12. Errors on x- and y- axis are approximately 10% of the errors on z-axis.

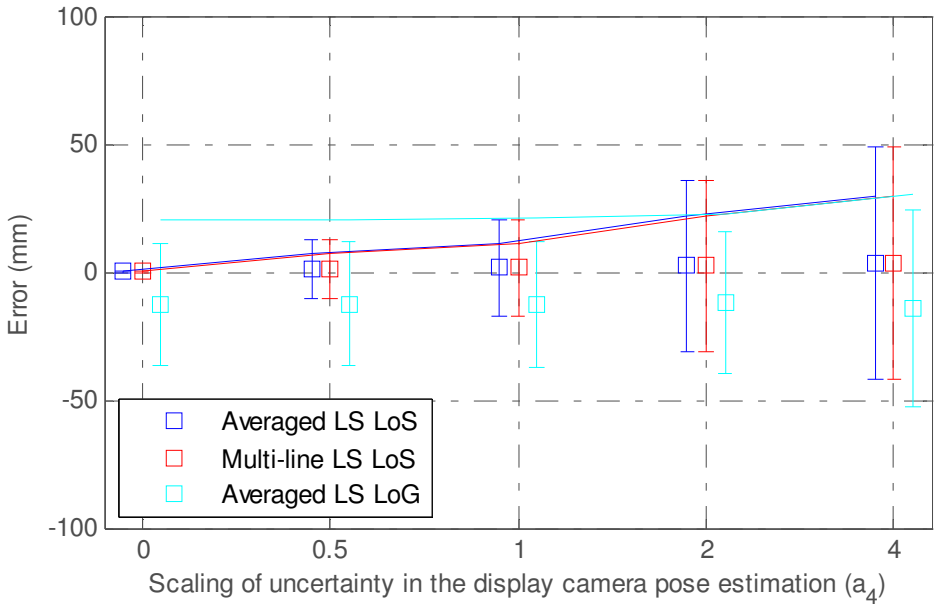


Figure 3.10.12 : Display-camera pose estimation uncertainty (a_4) vs. z-axis performance (a_1, a_2, a_3 = 0)

As seen in Figures 3.10.1 and 3.10.2, performance of the 2-D LoS solution is better than any 3-D solution when only pupil detection errors are present. When corner detection errors are considered, performance of the averaged LS LoS solution is comparable to the one of 2-D LoS solution. Effects of the camera calibration error and the display-camera pose estimation error are not significant for the 2-D LoS solution.

When 3-D performance is considered, the pupil and corner detection accuracies become quite critical. Compared to the pupil and corner detection accuracies, effect of camera calibration and display-camera pose estimation errors are not significant, in the given range of uncertainty. For pupil detection errors, multi-line least squares LoS solution, reduces standard deviation of error better than averaged least squares LoS solution; however, it introduces a non-zero mean error. For corner detection errors, the averaged least squares LoS solution

generates significantly better results than the multi-line least squares LoS solution.

Hartley and Zisserman [22] argue that intersection of two lines can be detected with an accuracy of 0.1 pixels. Moreover, when the pose of the checkerboard on subject's head is estimated and grids of the board are projected on to image plane standard deviation of the reprojection error is observed to be between 0.1-0.2 pixels. Considering the effect of estimation errors in internal camera parameters (which also causes reprojection errors), one can expect 0.15 pixel corner detection error in real setup. For internal camera parameters and display-camera pose estimation, there is no reason to assume a different error than the provided uncertainties. Therefore, we will assume unit value for error scaling factors a_3 and a_4 and 0.15 for a_2 as regular uncertainties. For these regular values performance of the system with different pupil detection accuracies is presented in Figures 3.10.13, 3.10.14 and 3.10.15.

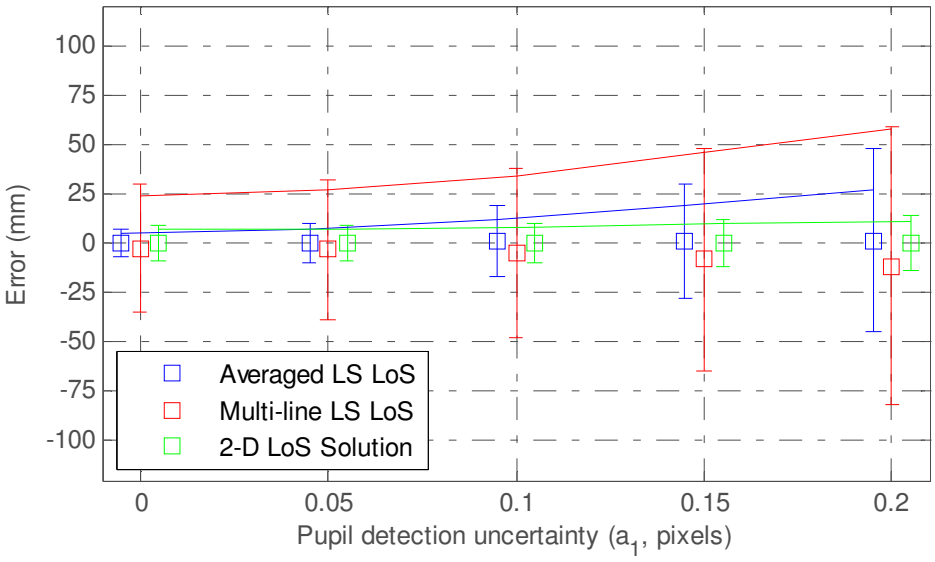


Figure 3.10.13 : Pupil detection uncertainty (a_1) vs. x-axis performance with regular uncertainties (a_2 = 0.15, a_3, a_4 = 1)

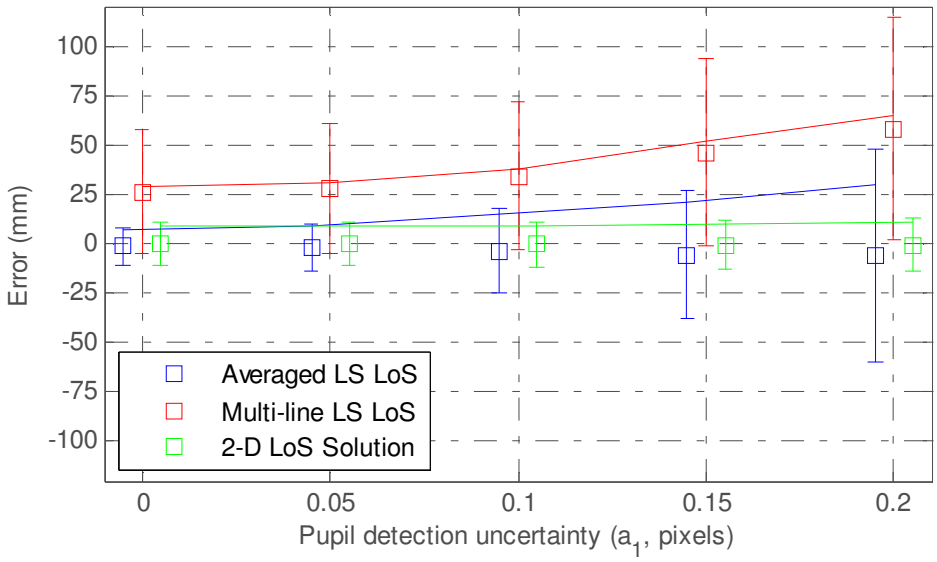


Figure 3.10.14 : Pupil detection uncertainty (a_1) vs. y-axis performance with regular uncertainties (a_2 = 0.15, a_3, a_4 = 1)

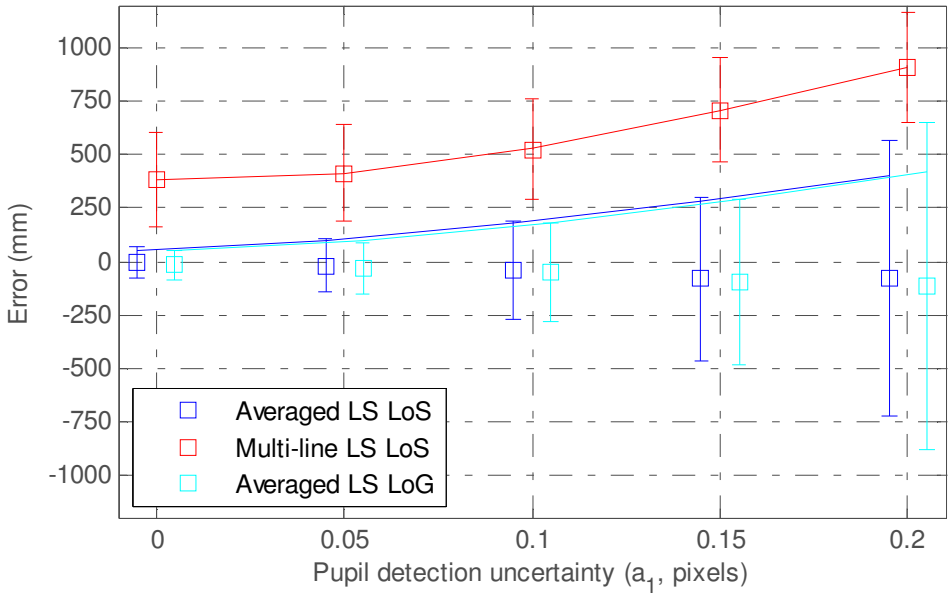


Figure 3.10.15 : Pupil detection uncertainty (a_1) vs. z-axis performance with regular uncertainties (a_2 = 0.15, a_3, a_4 = 1)

For tilted camera position and 0.1 pixels pupil detection accuracy, mean absolute error on x- and y-axis obtained by 2-D LoS solution is smaller than 25mm which corresponds to an angular inaccuracy smaller than 0.5°. When performance of the 3-D solutions on x- and y-axis are considered, an angular inaccuracy smaller than 1° is observed. Averaged LS LoS and LoG generates similar results and inaccuracy on z-axis is smaller than 250mm for 0.1 pixel pupil detection uncertainty. To reduce this inaccuracy up to 100mm, pupil detection uncertainty should be around 0.05 pixels.

Under 0.1 pixels pupil detection uncertainty, 3-D performances of multi-line least squares LoS, averaged least squares LoS and averaged least squares LoG solutions with different number of used frames are presented in Figures 3.10.16, 3.10.17 and 3.10.18, respectively.

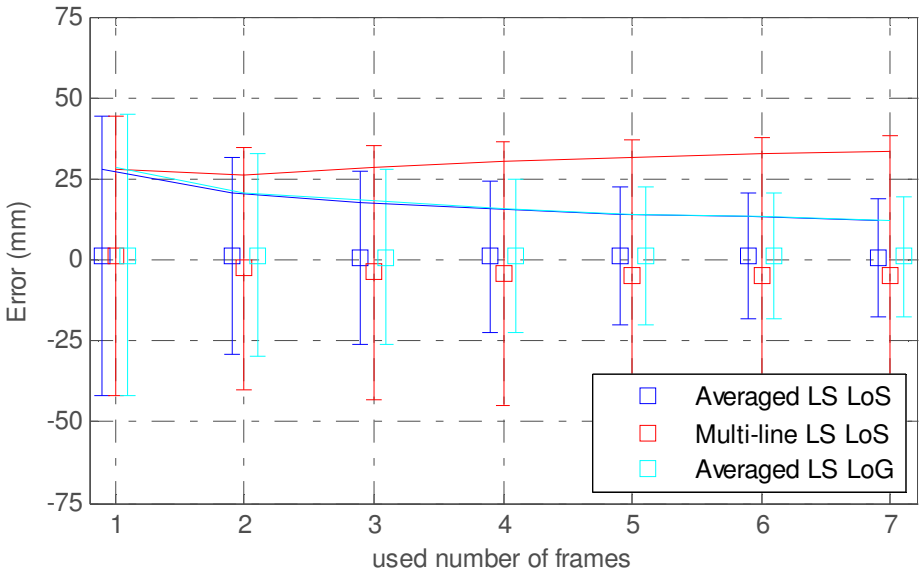


Figure 3.10.16 : Used number of frames vs. x-axis performance with regular uncertainties (a_1 = 0.1, a_2 = 0.15, a_3, a_4 = 1)

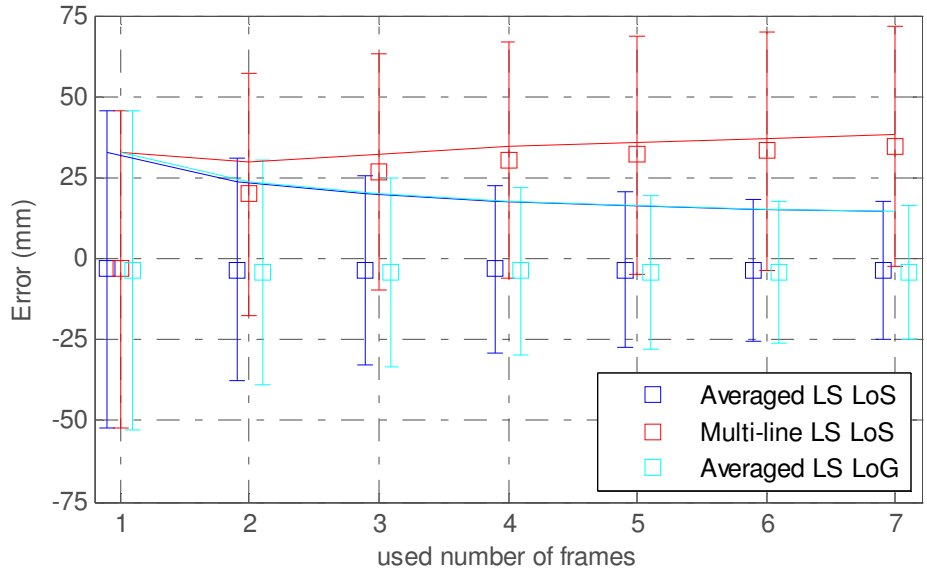


Figure 3.10.17 : Used number of frames vs. y-axis performance with regular uncertainties ($a_1 = 0.1$, $a_2 = 0.15$, a_3, $a_4 = 1$)

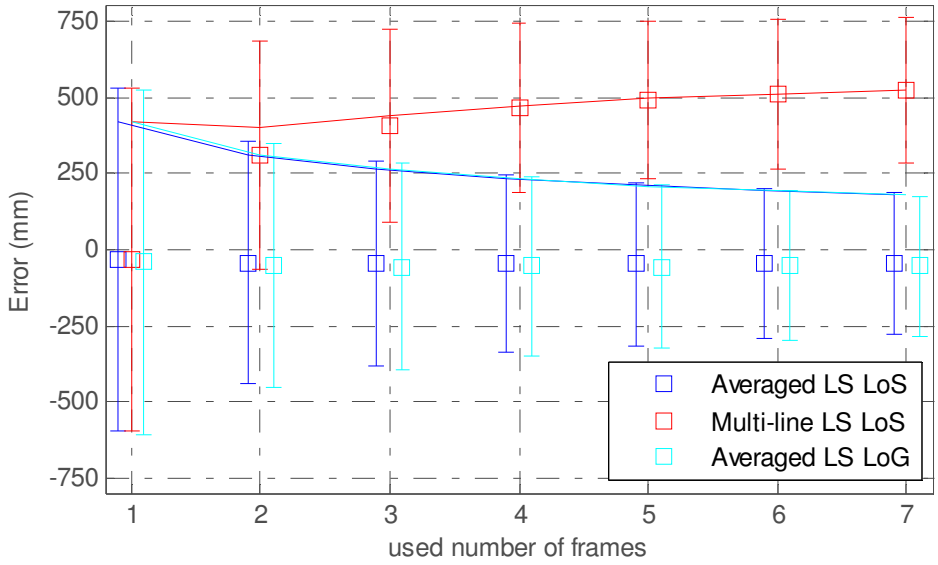


Figure 3.10.18 : Used number of frames vs. z-axis performance with regular uncertainties ($a_1 = 0.1$, $a_2 = 0.15$, a_3, $a_4 = 1$)

Under 0.1 pixels pupil detection accuracy, z-axis performance of the system for forward looking camera position is presented in Figure 3.10.19.

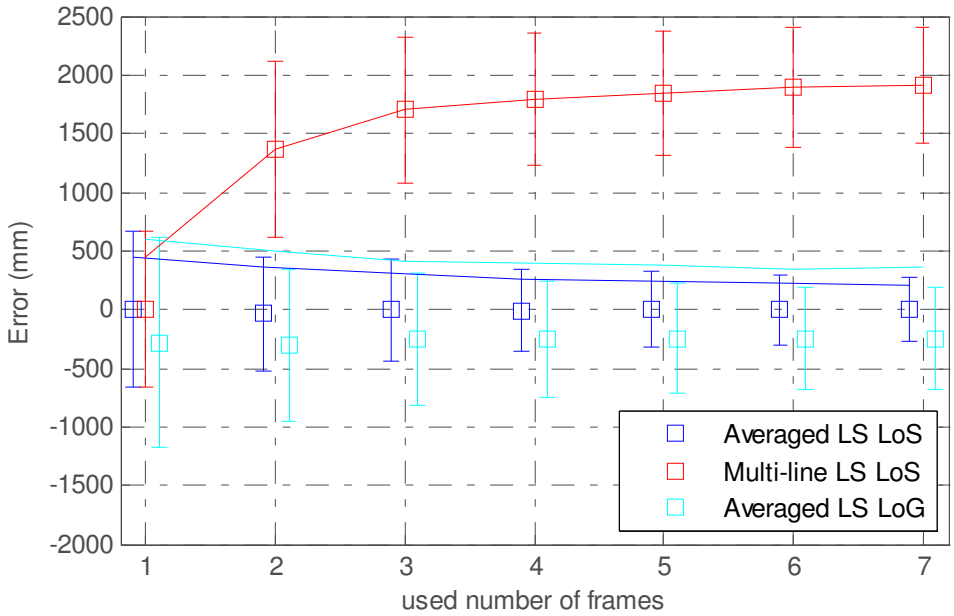


Figure 3.10.19 : Used number of frames vs. z-axis performance with regular uncertainties (a_1 = 0.1, a_2 = 0.15, a_3, a_4 = 1) (forward looking camera position)

When Figures 3.10.16, 3.10.17 and 3.10.18 are considered, as long as the utilized number of frames are increased, inaccuracies are reduced. When Figures 3.10.18 and 3.10.19 are compared, results obtained with tilted camera position is significantly better. Such a result is expected and it is consistent with previous observations (Figures 3.10.9 and 3.10.10).

3.11 Experimental Results

In this section, accuracies of the display-camera pose estimation and gaze estimation are presented. Three experiments conducted to measure performance of proposed system.

In first experiment, transformation between a 22” PC monitor and camera (camera is above the monitor and distance between viewer and display is equal to distance between viewer and camera) is estimated for 18 different mirror orientations. In the same experiment, LoG solution with different triangulation methods is inspected for a single subject. Eyeball centers of a subject are

estimated while subject was 1.5m away from camera and display. For personal calibration set, subject is asked to gaze on optical center of the camera from different locations. In this manner, center of eyeball estimation accuracy independent of display-camera pose estimation is observed. As the last part of first experiment, the subject is asked to gaze six different points on screen three times and gaze point is estimated.

In the second experiment, the transformation between a 3-D TV (brand & model: Philips-3D6W02 42") and camera (camera is below the display) is estimated for 19 different mirror orientations. In this experiment, only the LoG solution is tested for three subjects. Personal parameters of a subject are estimated while subject was 2.7m away from camera and display. For personal calibration set, subject is asked to gaze on optical center of the camera from different locations. As the last part of second experiment, the subject is asked to gaze nine different points on screen three times and gaze point is estimated.

In the third experiment, the transformation between a 3-D TV and camera (camera is below the display and at the midpoint of viewer and display, see Figures 3.9.1 and 3.9.2) is estimated for 25 different mirror orientations. In this experiment, both LoG and LoS solutions are tested for two subjects. Personal parameters of subjects are estimated, while subject was 1.4m away from camera and 2.8m away from the display. For personal calibration set, 40 points on display is utilized. Finally, the subject is asked to gaze on 60 different points on display to evaluate gaze estimation accuracy.

Since LoG solution is employed in the first and the second experiments, among the all personal parameters, only center of eyeballs are estimated. For radius estimate, a constant value of 12.5mm is used. Estimation results are presented in Tables 3.11.1 and 3.11.2. In Table 3.11.1 estimated eyeball center coordinates relative to head coordinate system, mean of minimum distance between used lines and estimated center of eyeball and standard deviation of distances are presented. In the first column, experiment id / subject id are given. In "used" column number of taken and used frames; in "L/R dist." column, distance between two eyeball centers is given. Results in Table 3.11.1 are given in millimeters. In Table 3.11.2 reprojection errors of estimated eyeball center is given in pixels.

Table 3.11.1 : Center of eyeball estimation results (mm)

Exp. 1	**x**	**Y**	**z**	**mean**	**std**	**used**	**L/R dist.**
Left Eye	99.863	-48.761	-41.358	0.2475	0.111	9 of 15	66.537
Right Eye	33.846	-48.49	-49.656	0.242	0.1186	11 of 15	
Exp. 2 / S1	**x**	**Y**	**z**	**mean**	**std**	**used**	**L/R dist.**
Left Eye	103.111	-43.509	-45.6	0.109	0.0402	10 of 17	66.8362
Right Eye	36.667	-42.096	-52.696	0.1369	0.0628	11 of 17	
Exp. 2 / S2	**x**	**Y**	**z**	**mean**	**std**	**used**	**L/R dist.**
Left Eye	104.899	-33.232	-40.891	0.2335	0.1251	8 of 17	62.7085
Right Eye	42.206	-34.626	-40.725	0.3059	0.1127	8 of 17	
Exp. 2 / S3	**x**	**Y**	**z**	**mean**	**std**	**used**	**L/R dist.**
Left Eye	107.447	-50.335	-43.072	0.2378	0.1434	7 of 19	61.822
Right Eye	45.628	-50.93	-42.862	0.1921	0.1352	7 of 19	

Table 3.11.2 : Reprojection errors for center of eyeball estimation (pixels)

Exp. 1	**Mean(x)**	**std(x)**	**mean(y)**	**std(y)**
Left Eye	-0.0178	0.4431	0.0245	0.9135
Right Eye	-0.025	0.4637	0.0335	0.9153
Exp. 2 / S1	**Mean(x)**	**std(x)**	**mean(y)**	**std(y)**
Left Eye	0.0032	0.3458	-0.0012	0.5022
Right Eye	-0.0023	0.5049	0.0101	0.6137
Exp. 2 / S2	**mean(x)**	**std(x)**	**mean(y)**	**std(y)**
Left Eye	-0.0265	0.9047	0.0505	1.0008
Right Eye	0.0202	1.1383	0.0707	1.1941
Exp. 2 / S3	**mean(x)**	**std(x)**	**mean(y)**	**std(y)**
Left Eye	-0.0155	0.5691	0.0262	1.5002
Right Eye	-0.0253	0.6294	0.0082	1.1184

As it can be observed from these tables, the eyeball centers are estimated very accurately during the first experiment and the first subject in second experiment. It should be noted that, since calibration object placed on subjects' head has a slightly different orientation in each experiment, coordinates of eyeball centers in the head coordinate system might be different. What makes the results

meaningful is smaller distance between gaze lines and eyeball center and smaller reprojection errors. Another clue is the distance between eyeball centers, which is approximately 65mm, but might vary among the subjects. The subject in the first experiment and the first subject in the second experiment is the same person. Since the results belong to the same subject, the distance between eyeball centers is expected to be constant. The reprojection errors given in the Table 3.11.2 can be assumed to be caused mainly by pupil detection errors. The detection process is manual and, as seen in the table, reprojection errors are mostly below the detection sensitivity. The differences in L/R distance and distance between lines and center of eyeball is also thought to be mainly caused by pupil detection errors but the effect of the calibration errors and pose estimation errors should also be considered.

The results of the first experiment and the first subject in the second experiment seem to be promising to continue with the next step. However, for the second and the third subjects in the second experiment, even though a significant number of the lines are eliminated due to errors, reprojection errors are still relatively high. These results do not seem acceptable like the previous ones. In addition to the semi-automatic detection of pupils, manual detection is also performed for the second and third subjects during Experiment 2. However, neither manual pupil detection nor averaging manual and semi-automatic pupil detection yields acceptable results. One possible explanation of this problem is unfamiliarity of the subjects with system. In other words, in eyeball center estimation sequence, they might be gazing on to a slightly different point rather than the optical center of the camera. This may be also caused by LoG/LoS difference, which is not considered in eyeball center estimation. Moreover, there were problems with internal camera parameters during these trials, which might be another possible reason for the errors. To solve these problems, rather than the optical center of the camera, known points on the display are utilized for calibration of personal parameters in the following experiment. As a second improvement, a complete LoS solution is implemented to use in the following experiment. The last improvement is to locate the camera on the mid-point between the display and the viewer, so FOV of the camera becomes wider, which eases the calibration of the internal parameters.

The results of display-camera pose estimation are presented in Table 3.11.3, Figures 3.11.1. and 3.11.2. The mean and standard deviations of the estimated transformation parameters in experiments are given in Table 3.11.3. The rotation angles are given in ZYX Euler angles and in degrees. The translation parameters are given in millimeters. The result of the first experiment is obtained from 18 images with different mirror orientations but same display-camera pose. The result of the second experiment is obtained from 19 images with different mirror orientations. The result of the third experiment is obtained from 29 images with different mirror orientations.

Table 3.11.3 : Display camera pose estimation results

Exp. 1	**mean(rot)**	**std(rot)**	**mean(tr)**	**std(tr)**
x	-178.79	0.5769	204.76	22.39
y	-0.45	0.8481	96.6	16.8
z	178.37	0.1928	180.5	9.56
Exp. 2	**mean(rot)**	**std(rot)**	**mean(tr)**	**std(tr)**
x	179.63	0.22	371.89	34.71
y	-3.21	1.47	-564.35	19.6
z	-169.56	1.58	-243.25	13.67
Exp. 3	**mean(rot)**	**std(rot)**	**mean(tr)**	**std(tr)**
x	179.56	0.17	443.4	31.53
y	-1.46	0.7	-806.7	15.14
z	-162.43	2.25	-1341.5	15.54

Display camera pose estimation results are consistent with manual measurements. However, since the manual measurements are not sensitive, it might be better to inspect the variation of the pose estimation during different mirror orientations. Since the position of the display relative to the camera is fixed, it is expected to obtain same transformation parameters for different mirror orientations.

When the reference points on the display are transformed into the CCS by using the transformation matrices obtained with different mirror orientations during the first and second experiments, the resulting points on x-y plane are shown in Figures 3.11.1 and 3.11.2.

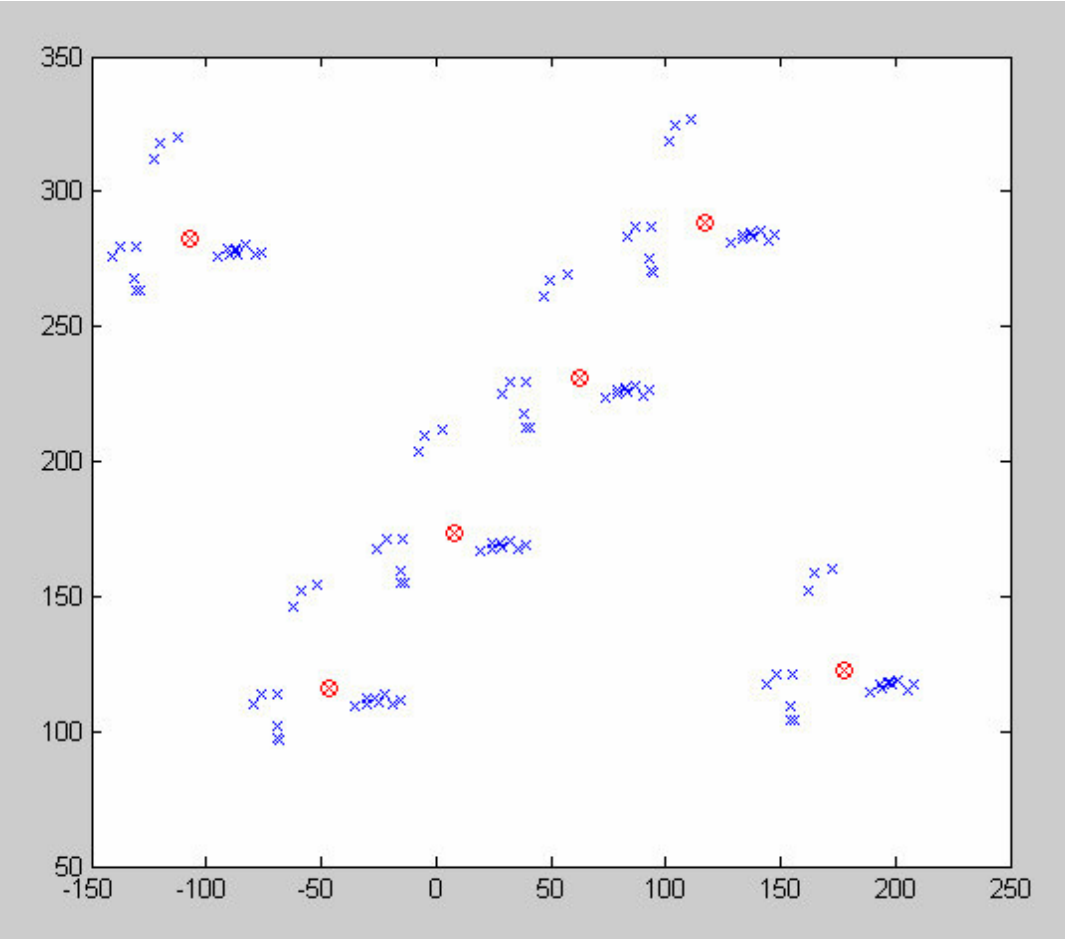


Figure 3.11.1 : Transformation of reference points in to the CCS for Experiment 1. Results obtained from different mirror orientations (blue crosses), average of transformations (red circles), transformation of points with averaged transformation parameters (red crosses)

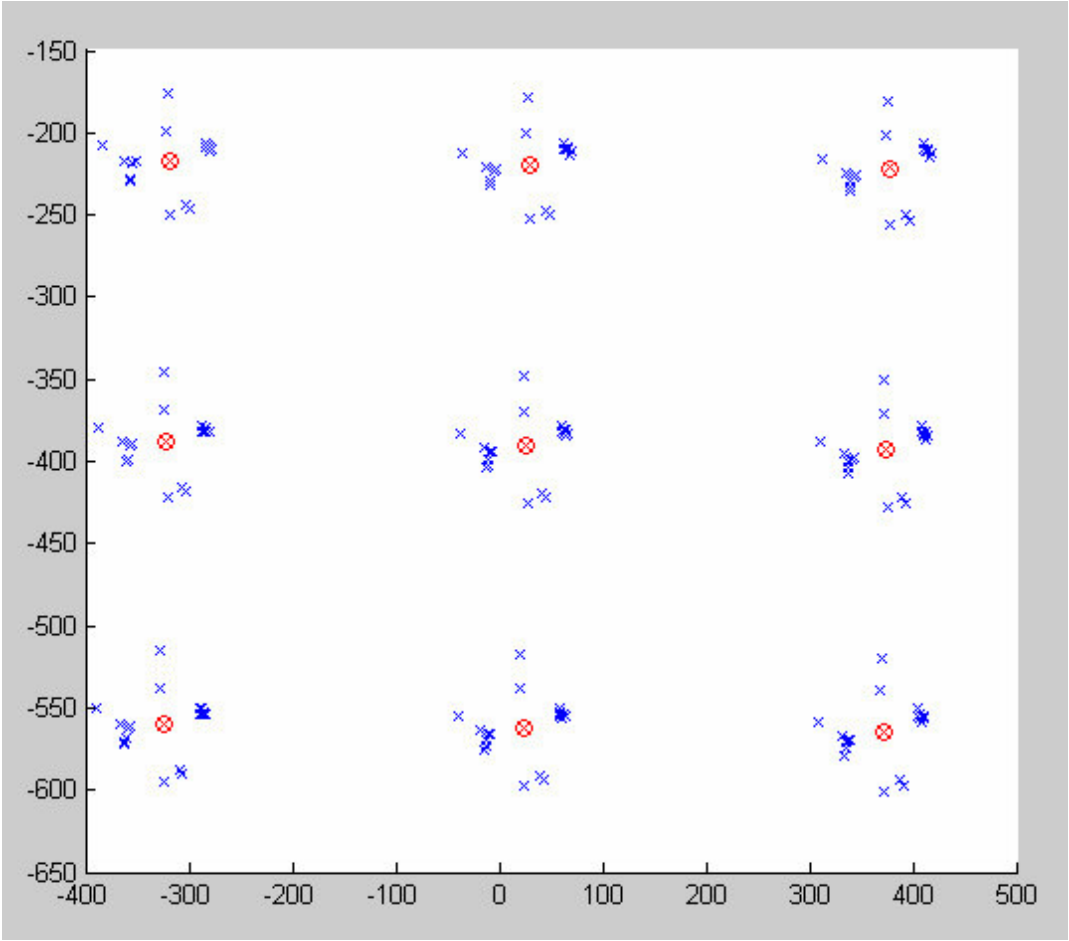


Figure 3.11.2 : Transformation of reference points in to the CCS for Experiment 2. Results obtained from different mirror orientations (blue crosses), average of transformations (red circles), transformation of points with averaged transformation parameters (red crosses)

During simulations, the display camera pose estimation parameters obtained on experimental setup are used with uncertainties observed on experimental setup. For these observed uncertainties, errors in display camera pose estimation step does not have a significant effect on performance of the system, when compared to the other error sources. Obtaining similar (insignificant) uncertainties for display camera pose estimation on three different setups, it can be concluded that performance of the display camera pose estimation algorithm is well enough to be employed in such a system.

Gaze estimation accuracy in Experiment 1 is given as mean absolute error (mm) in Table 3.11.4 and as degrees in Table 3.11.5. Gaze estimation is performed in five different ways: midpoint triangulation (MP), least-squares triangulation (LS), polynomial triangulation (PT), intersection of right gaze vector with screen plane (RG) and intersection left gaze vector with screen plane (LG).

Table 3.11.4 : Gaze estimation accuracy in Experiment 1 as mean absolute error (mm)

mean absolute error	MP	LS	PT	RG	LG
x	23.3	23.8	46.1	16.1	17.7
y	32.2	32.2	30.6	23.9	28.9
z	261.1	261.7	499.4	-	-

Table 3.11.5 : Angular gaze estimation accuracy in Experiment 1 (degrees)

angular error	MP	LS	PT	RG	LG
x	0.89	0.91	1.76	0.62	0.68
y	1.23	1.23	1.16	0.91	1.1

The performance of gaze estimation in x- and y-axis are quite acceptable, however z estimations are not appropriate. These errors are thought to be caused by errors in pupil detection.

To achieve a perfect gaze estimate with LoG solution, gaze line should be defined as the line drawn from eyeball center to gaze point. Then pupil should be on the intersection of this line with eyeball sphere. When this pupil position projected on to image plane, assuming error free estimations in other steps, this reprojection should be the ground-truth for pupil detection. The difference between detected and reprojected pupils are given in Table 3.11.6.

Table 3.11.6 : Reprojection errors of pupils for Experiment 1

	mean(P_R)	std(P_R)	mean(P_L)	std(P_L)
x	-0.1582	0.5714	-0.3553	0.5928
y	-0.8659	0.7884	-1.1443	0.6147

Since the pupil detection process is manual, a difference between the detected and reprojected pupils is expected. Moreover, the differences are quite small and above the accuracy of manual detection. Considering manual detection problem, reprojection of the pupils are accepted as the correct coordinates and gaze estimation simulations performed by adding noise to the correct coordinates. For three different head orientations correct coordinates of pupils are determined. Then, one of the true pupil coordinates is selected for 2000 times and simulated pupil detection result is obtained by adding noise. For three different noise levels, the mean absolute error (mm) and its standard deviation are presented in Table 3.11.7. The errors are given in millimeters whereas standard deviation of added noise is given in pixels.

Table 3.11.7 : Required pupil detection accuracy for Experiment 1

mean absolute error	0.1px	0.2px	0.3px	std	0.1px	0.2px	0.3px
x	10.9	22.2	33.8	x	8.4	19	34.3
y	2.5	5.2	7.9	y	1.9	4.4	7.5
z	78.5	159	241	z	60.4	137	244

The values observed at Table 3.11.7 is valid when distance between subject and gaze point is 1.5m. For 100mm accuracy in z- direction, required pupil detection accuracy is 0.1pixels. When distance between subject and gaze point increased to 3m, during simulations we find 0.05 pixel pupil detection accuracy is required. While distance between subject and gaze point decreases, one can expect the

required pupil detection accuracy to decrease. From this point of view, results in Table 3.11.7 and Figure 3.10.15 can be regarded as consistent.

The 3-D performance of averaged LS LoS, multi-line LS LoS and averaged LS LoG as well as 2-D performance of LoS solution at Experiment-3 for two subjects are presented in Figures 3.11.3 to 3.11.8 for the different number of utilized frames.

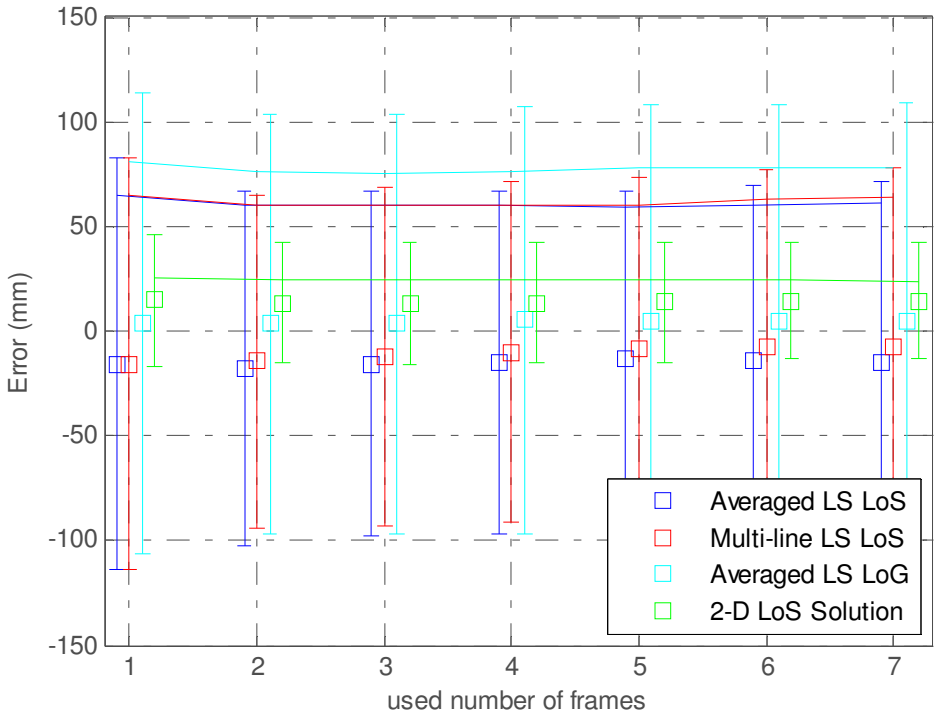


Figure 3.11.3 : Experiment-3 / Subject-1 x-axis performance

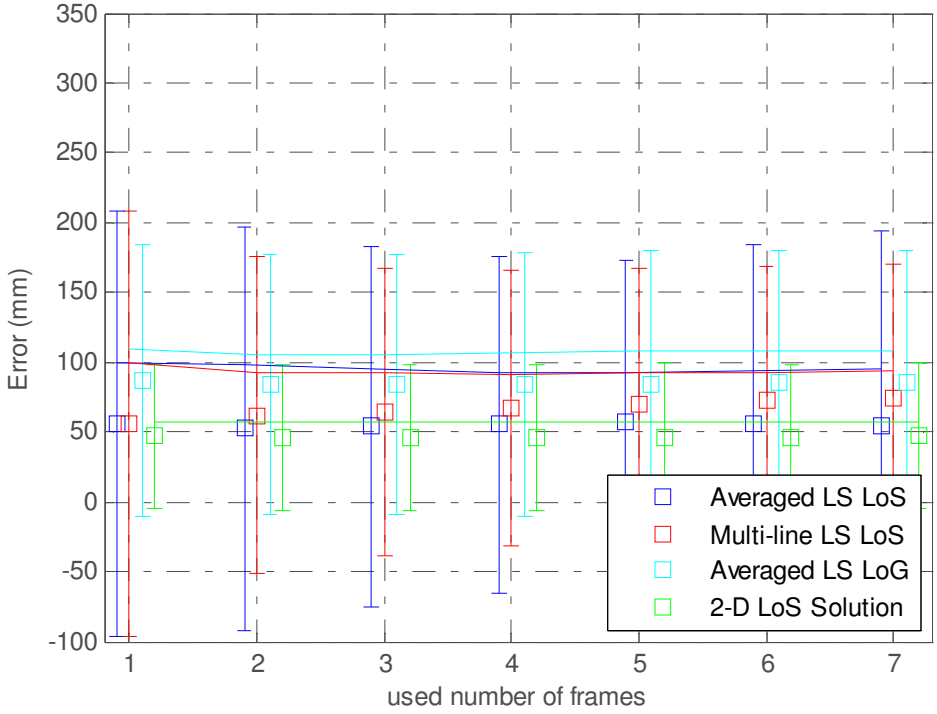


Figure 3.11.4 : Experiment-3 / Subject-1 y-axis performance

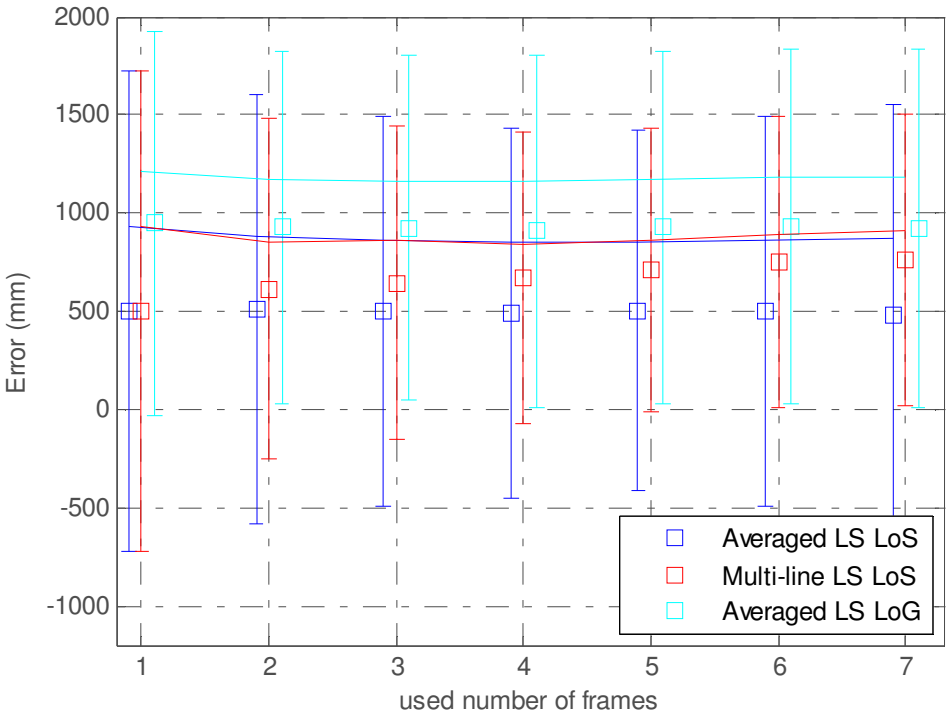


Figure 3.11.5 : Experiment-3 / Subject-1 z-axis performance

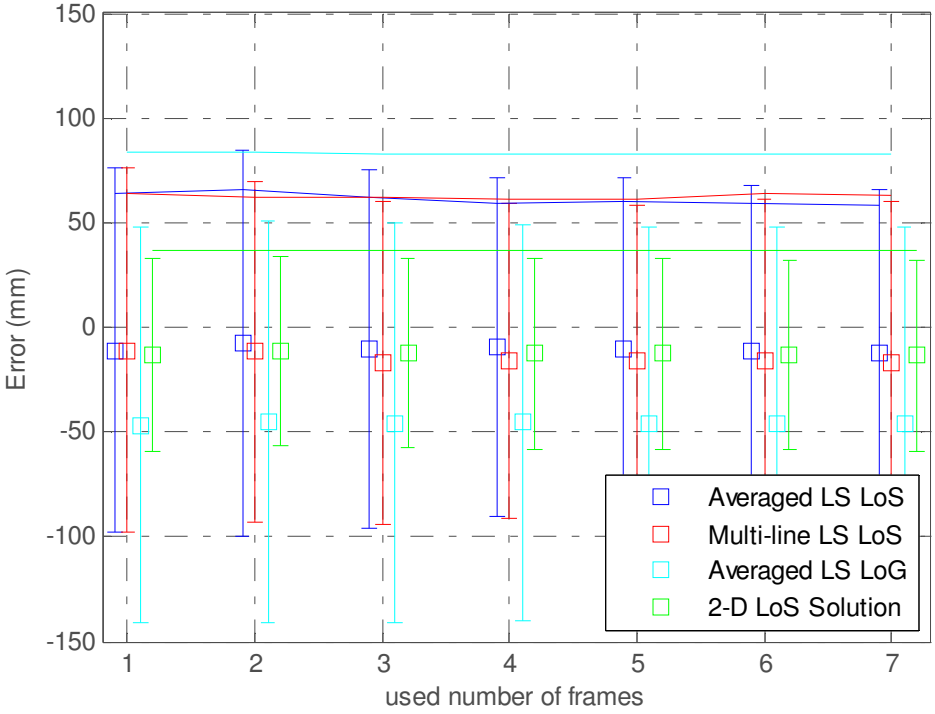


Figure 3.11.6 : Experiment-3 / Subject-2 x-axis performance

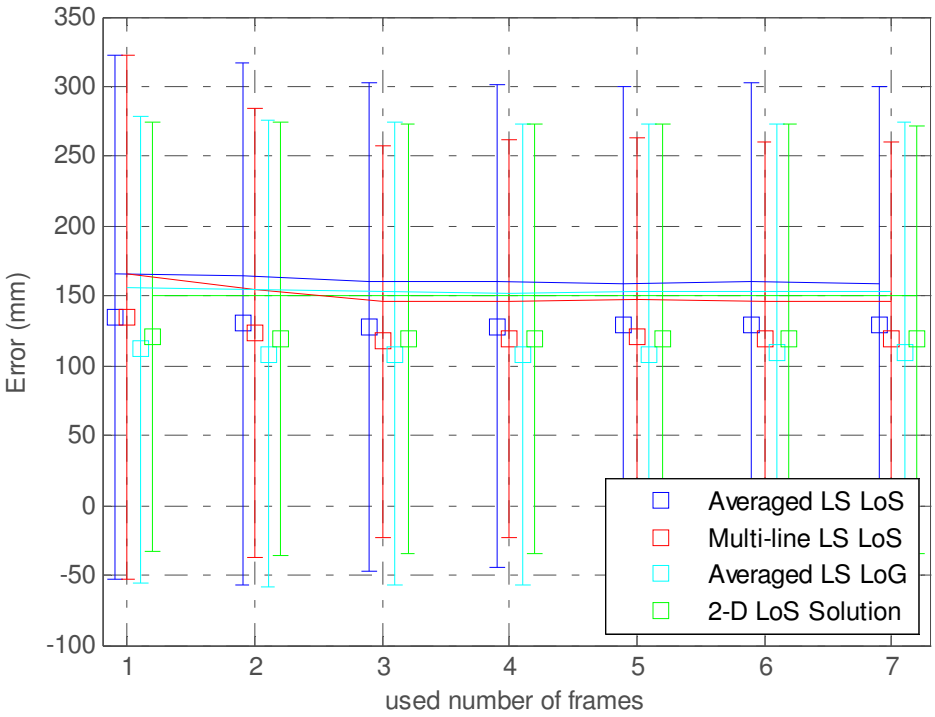


Figure 3.11.7 : Experiment-3 / Subject-2 y-axis performance

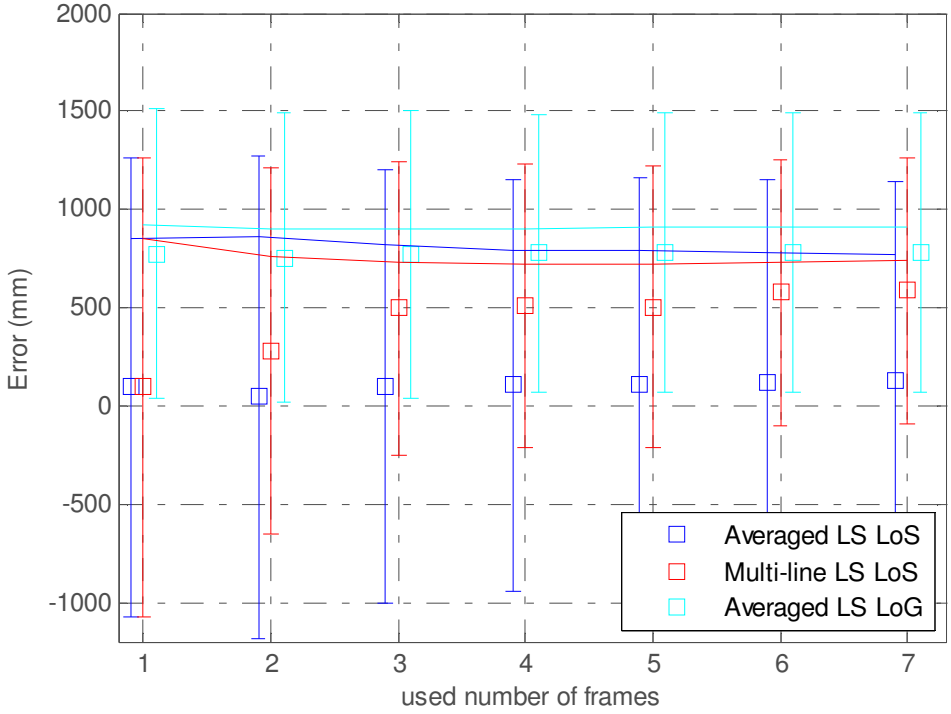


Figure 3.11.8 : Experiment-3 / Subject-2 z-axis performance

As seen in the Figures 3.11.5 and 3.11.8, the best accuracy on z-axis is obtained at 5 frames which implies fixation durations of the subjects are slightly shorter than the fixation duration used in simulations. Accuracies in x- and y-directions do not change significantly as the used number of frames is changed. Different from the simulation results, for averaged LS LoS solution a non-zero mean error is observed for first subject. Similar to simulation results, multi-line LS LoS solution decreases standard deviation of the error more effectively than averaged LS LoS, however it introduces an additional mean of 200mm for first and 500mm for second subject. Mean absolute error on x-axis obtained by 2-D LoS solution is smaller than 50mm, which corresponds to an angular inaccuracy smaller than 1° for a viewer 3m away from the display. On y-axis an average of 75mm inaccuracy is observed for 2-D LoS solution, which corresponds to an angular accuracy of 1.5°. When performance of 3-D solutions on x- and y axis are considered, mean absolute error is around 100mm, which corresponds to an angular inaccuracy smaller than 2°. Inaccuracy of the estimates on z-axis is round 800mm.

There is a significant difference between experimental results and simulation results. For regular errors, a zero mean error is observed for averaged LS LoS during simulations; however, in experimental results, an error with mean value of 50cm is observed for first subject. As another distinction, for regular errors, standard deviation of the error is observed to be reduced up to 20 cm for multi-line LS LoS solution; however, in experimental results, this value is greater than 70cm. These differences might be due to the mismatch between assumptions during simulations and experiments. First of all, during simulations, for pupil detection error zero mean Gaussian distribution is assumed; however, due to illumination conditions and perspective of the camera, during experiments, this pupil detection error may not have zero mean value. Moreover, the standard deviation of the pupil detection error may be even higher than 0.2 pixels.

CHAPTER 4

CONCLUSION

4.1 Summary of Thesis

Throughout the study, a new visual 3-D eye gaze tracker for autostereoscopic displays is proposed. Before presenting the proposed approach, related studies are reviewed in Chapter 2. In Chapter 3, LoG and LoS solutions for gaze estimation are derived. To observe the performance of the system, a simulator is developed which simulates whole tracking environment. The proposed system is implemented and tested both with computer simulations and on an experimental setup. The 2-D and 3-D performances of the LoG and LoS solutions are observed and presented.

4.2 Discussions

In terms of 2-D and 3-D performance criteria, LoS solution generates slightly better results compared to that of LoG. 2-D estimation accuracy of the system is found to be smaller than 1° in simulations and around 1° on experimental setup, which is compatible with present EGTs. 3-D estimation inaccuracy of the system on x- and y-axis is found to be smaller than 2° in simulations and experiments. Estimation accuracy in z- direction is significantly decreasing with increasing pupil detection and head pose estimation errors. For regular noise levels and 3m subject-gaze point distance, 20cm inaccuracy is observed in simulations, however in experiments inaccuracy increased up to 80cm. When distance between the viewer and display decreases (i.e. at 1.5m), the system is able to produce acceptable (20cm inaccuracy in z-direction) results in 3-D estimation too. This difference is thought to be caused by the triangulation uncertainty. The performance of the display-camera pose estimation method is observed to be well enough to be employed in such a system.

The simulations demonstrate that for current camera configuration, if corner detection accuracy of 0.1 pixels and pupil detection accuracy of 0.05 pixels can be achieved, then proposed system is expected to estimate 3-D gaze point with an accuracy 10cm in z-direction for a user 3m away from the display. Comparing experimental results with simulation results, current pupil detection accuracy is estimated to be around 0.2 pixels and current corner detection accuracy is around 0.15 pixels.

When camera orientation is considered, keeping calibration object around the principal point of the camera seems to be the best choice.

The data provided by proposed system can be easily converted to estimates of some perceptual inputs. Distance between pupils and 2-D solution of gaze points can be directly used as accommodation estimate. Distance between 3-D gaze point and pupil can be used as convergence estimate. If 3-D scene that viewer is looking at is defined, since at any instant orientation of the eyeball is known, motion parallax that viewer receives or expects to receive can be estimated. However with present configuration, tracker is poor in estimating gaze depth or namely convergence.

Hardware requirement of proposed tracker is quite acceptable. It is limited with IR LEDs and a night view HD camera without any need of synchronization or a controller. Since it is not designed for HCI, system is not suitable for such applications, due to its offline processing and calibration requirements. However, for a vision laboratory, if a proper calibration is provided and required pupil detection accuracy is achieved, as an accurate and low cost device with easy to obtain and use hardware, it could be accepted as a convenient solution.

Main idea in proposed method is defining a coordinate frame rigidly connected to viewer's head, estimating center of eyeball, principal LoG and principal LoS relative to this coordinate system by gazing on known points on display. Then pupils can be positioned in 3-D using head pose, center of eyeball and eyeball radius. Once eyeball center and pupil are known in 3-D, with a known principal LoG and LoS, present LoS can be constructed. In this aspect, center of eyeball, principal LoG, principal LoS and eyeball radius must be obtained per user. In the

current configuration, the proposed method is intrusive; however, by replacing the head pose estimation part by a non-intrusive head pose estimation method, an accurate and non-intrusive EGT could also be obtained. Rather than an HD camera, one wide and one narrow view (PTZ) cameras might be utilzed in order to allow larger head movements and to obtain a more detailed image of eye. For diagnostic proposes, rather than a checkerboard, glasses may be used as a more comfortable calibration object.

4.3 Future Work

In order to use proposed system to examine human perception on autostereoscopic displays, 3-D estimation performance of the tracker should be improved (especially on z- direction). To improve 3-D performance, pupil detection accuracy should be enhanced. For this purpose, performance of different pupil detection algorithms should be tested with present setup. If all the pupil detection algorithms are unable to reach the desired accuracy, then another option could be to restricting head movements of the subject and zooming in the camera or reducing distance between subject and camera with current camera configuration. Another way to increase pupil detection accuracy may be to use a head mounted narrow view camera or one wide view and one narrow view PTZ camera, by getting a more detailed image of the pupil.

REFERENCES

[1] W.A. IJsselsteijn, P.J.H. Seuntiëns and L.M.J. Meesters, *State-of-the-art in human factors and quality issues of stereoscopic broadcast television*, ATTEST-IST-2001-34396, Agust 2002.

[2] Carlos H. Morimoto, Marcio R.M. Mimica, *Eye gaze tracking techniques for interactive applications*, Computer Vision and Image Understanding, Vol. 98, 2005, pp. 4-24.

[3] Andrew T. Duchowski, *A Breadth-First Survey of Eye Tracking Applications*, Behavior Research Methods, Instruments and Computers, Vol. 34, No.4, 2002, pp. 445-470.

[4] LC Technologies, Inc., http://www.eyegaze.com/, last visited on September 2009.

[5] SensoMotoric Measurements GmbH, http://www.smivision.com/, last visited on September 2009.

[6] Applied Science Laboratories, http://asleyetracking.com/site/, last visited on September 2009.

[7] Tobii Eye Tracking, http://www.tobii.com/, last visited on September 2009.

[8] C.H. Morimoto, D. Koonsb, A. Amir, M. Flickner, *Pupil detection and tracking using multiple light sources*, Image and Vision Computing, Vol. 18, 2000, pp. 331-335.

[9] Zhiwei Zhu, Qiang Ji, *Eye Gaze Tracking under Natural Head Movements*, Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, 2005, pp. 918-923.

[10] C.H. Morimoto, D. Koonsb, A. Amir, *Detecting eye position and gaze from a single camera and 2 light sources*, Proceedings of the 16th International Conference on Pattern Recognition, Vol. 4, 2002, pp. 314-317.

[11] Dong Hyun Yoo, Myung Jin Chung, *Vision based eye gaze estimation system using robust pupil detection and corneal reflections*, International Journal of Human-friendly Welfare Robotic Systems, Vol. 3, No. 4, 2002, pp. 2-8.

[12] David Beymer, Myron Flickner, *Eye Gaze Tracking Using an Active Stereo Head*, Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03), Vol. 2, 2003, pp.451-458.

[13] Jian-Gang Wang, Eric Sung, *Gaze determination via images of irises*, Image and Vision Computing , Vol. 19, 2001, pp. 891-911.

[14] Jian-Gang Wang and Eric Sung, *Study on Eye Gaze Estimation*, IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, Vol. 32, No. 3, 2002, pp. 332-350.

[15] Yoshio Matsumoto, Alexander Zelinsky, *An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement*, Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 499-504.

[16] Sheng-Wen Shih, Jin Liu, *A Novel Approach to 3-D Gaze Tracking Using Stereo Cameras*, IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics, Vol. 34, No. 1, 2004, pp.234-245.

[17] Linux Digital Video, http://www.kinodv.org/, last visited on September 2009.

[18] ITU, *Subjective Assessment of Stereoscopic Television Pictures*, ITU-R BT.1438, 2000.

[19] R. M. Taylor, P. J. Probert, *Range Finding and Feature Extraction by Segmentation of Images for Mobile Robot Navigation*, Proceedings of the IEEE International Conference on Robotics and Automation, Vol. 1, 1996, pp. 95-100.

[20] Zhaofeng He, Tieniu Tan, Zhenan Sun, Xianchao Qiu, *Toward accurate and fast iris segmentation for iris biometrics*, IEEE Transactions on Pattern Analysis ans Machine Intelligence, Vol. 31, No. 9, 2009, pp. 1670-1684.

[21] Richard I. Hartley, Peter Sturm, *Triangulation*, Computer Vision and Image Understanding, Vol.68, No.2, 1997, pp. 146-157.

[22] Richard Hartley, Andrew Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2nd Ed., 2003.

[23] Kay Talmi, Jin Liu, *Eye and gaze tracking for visually controlled interactive stereoscopic displays*, Signal Processing: Image Communication, Vol. 14, 1999, pp. 799-810.

[24] Jeongseok Ki, Yong-Moo Kwon, *3D Gaze Estimation and Interaction*, Proceedings of the 3DTV Conference'08, 2008, pp. 373-376.

[25] I Sexton, P Surman, *Stereoscopic and autostereoscopic display systems*, IEEE Signal Processing Magazine, May 1999, pp. 85-89.

[26] Umur Talaslı, *A Flaw In The Explanation Of The Stereokinetic Cone Illusion and Two Guiding Hypotheses For New Research*, Perceptual and Motor Skills, Vol. 76, 1993, pp. 823-829.

[27] Umur Talaslı, *On The Possibility of Intermittence During Fixations: A Conceptual Assesment*, Perceptual and Motor Skills, Vol. 77, 1993, pp. 323-329.

[28] Irvin Rock, *An Introduction to Perception*, Macmillan, 1st Ed., 1975.

[29] Amir A. Handzel, Tamar Flash, *Geometry of Eye Rotations and Listings Law*, Advances in Neural Information Processing Systems, Vol. 8, 1995, pp. 117-123.

[30] David Hestenes, *Invariant Body Kinematics:I. Saccadic and Compensatory Eye Movements*, Neural Networks, Vol. 7, No. 1, 1994, pp. 65-77.

[31] Camera Calibration Toolbox for MATLAB, http://www.vision.caltech.edu/bouguetj/calib_doc/, last visited on December 2009.

[32] Ken Shoemake, *Animating Rotation with Quaternion Curves*, ACM SIGGRAPH Computer Graphics, Vol. 19, No. 3, 1985, pp. 245-254.

[33] Zhengyou Zhang, *Flexible Camera Calibration By Viewing a Plane From Unknown Orientations*, Proceedings of International Conference on Computer Vision (ICCV'99), Vol. 1, 1999, pp. 666.

[34] Description of fit_ellipse, http://www.chronux.org/downloads/chronux/chronux/documentation/chronux_2_00/fly_track/FTrack/functions/fit_ellipse.html, last visited on December 2009.

APPENDIX-A

DEPTH PERCEPTION IN HUMAN VISUAL SYSTEM

Rock [28] presents a comprehensive introduction to the visual perception. The perception of size, motion and third dimension, as well as basics of visual perception is all explained in detail. The material presented in this appendix is mainly adapted from Rock [28].

A.1 Top-Down and Bottom-Up Processes

Perception is a combination of top-down and bottom-up processes. Top-down processes are conceptually driven processes and use the information named *world knowledge*. World knowledge is a non-subjective knowledge like "windows are rectangular", "if I let the pen, it drops", "snow is white" etc. Bottom-up processes are driven by the data coming from the receptors.

A.2 Retina, Retinotropic Projection, Corresponding Retinal Points and Horopter

Human visual system is discrete in space domain. Retina is composed of cone and rod cells that are sensitive to the light and color, respectively. However, there are spaces between these cells; therefore, data recorded by the retina is not continuous. Actually it is a mosaic, namely *retinal mosaic*. Every scene point projected on to the retina is transmitted to the brain without a loss, which is named as *retinotropic projection*. If a person picks a point in one eye, there will be a corresponding point in the other eye; which is denoted as *corresponding retinal points;* i.e. corresponding point of left eye's nasal 5 degree is right eye's temporal Corresponding retinal points are connected to the same location in the brain. When intersections of each corresponding retinal point pairs are connected, they form an arc named *horopter*. Corresponding retinal points and horopter are shown in Figure A.1.

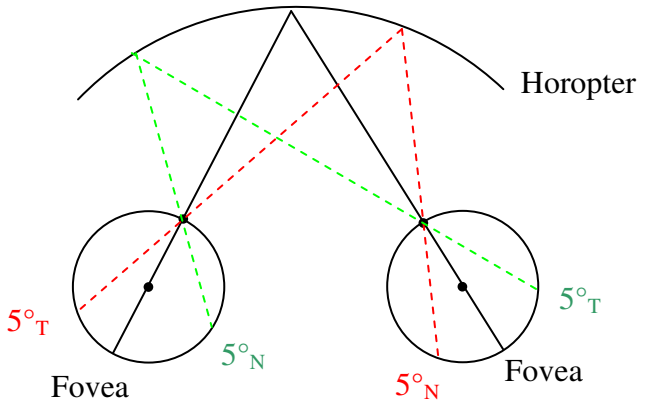


Figure A.1 : Corresponding retinal points and horopter

A.3 Latency

Signals send from different retinal points have different arrival times to the brain [26]. Latencies of retinal points draw a saw-tooth wave pattern (Figure A.2). Since two eyes latency difference is in opposite direction, when one looks at the world with both eyes, the latencies are averaged out and finally a flat latency is obtained.

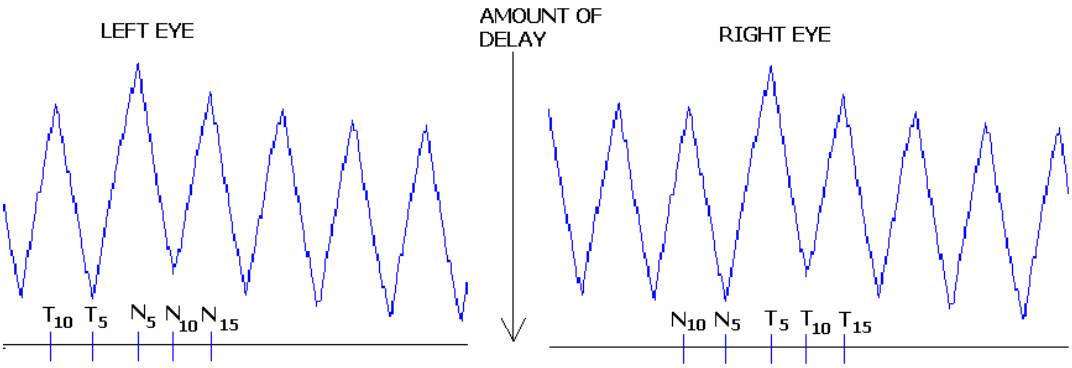


Figure A.2 : Latencies on left and right eyes

A.4 Intermittence, Flicker Fusion and Stoboscopic Motion Mechanism

Human eye has a high frequency movement, called *High Frequency Eye Tremor*, approximately at a frequency about 80 Hz and retinal data is submitted to the

cortex only during fixation [27]. Average span of movement is 1.5 times of cone cell diameter. Since human visual system is discrete in time domain, two mechanisms are employed to experience continuity and reconstruct motion. Flicker fusion mechanism fuses up the images of successive snapshots and in this way, it generates the continuity of perception, but not necessarily the motion. In other words, flicker fusion mechanism works towards the registration of the stimuli that is continuous in nature. This fusion begins from 60Hz in HVS. Movement of objects is not directly perceived by HVS; instead stroboscopic integration of successive snapshots generates the perceived motion. It is necessary to destroy the motion to make form perceivable. This mechanism explains how objects in stationary frames are perceived as moving when the frames are presented with rapid succession. In this mechanism, motion perception begins from 5Hz.

A.5 Position Constancy

Although basically movement perception is explained by the changes in the position of the objects across retina, movement of the objects in retinal image does not always results in perceived movement of the objects. The movement of the body, head or eyes of the observer causes the displacement of all the stationary objects in the retinal image; however, any object perceived as it is moved. This phenomenon is called *position constancy*.

A.6 Depth Perception and 3-D Cues in Real 3-D Environment

The following 3-D cues in HVS are utilized for depth perception: disparity, convergence, motion parallax, accommodation and pictorial cues.

A.6.1 Disparity

Disparity is a binocular cue, caused by the horizontal separation of left and right eyes. Fixation point is projected on to fovea for both eyes, points at the same distance with fixation point have zero disparity, further points have uncrossed and nearer points have crossed disparity (Figure A.3). The small region around the horopter is called as Panum's Fusional Area. Only the disparity pairs in *Panum's*

Fusional Area are matched and processed. Disparities that are greater than 2 degrees are ignored (the limit for Panum's Fusional Area).

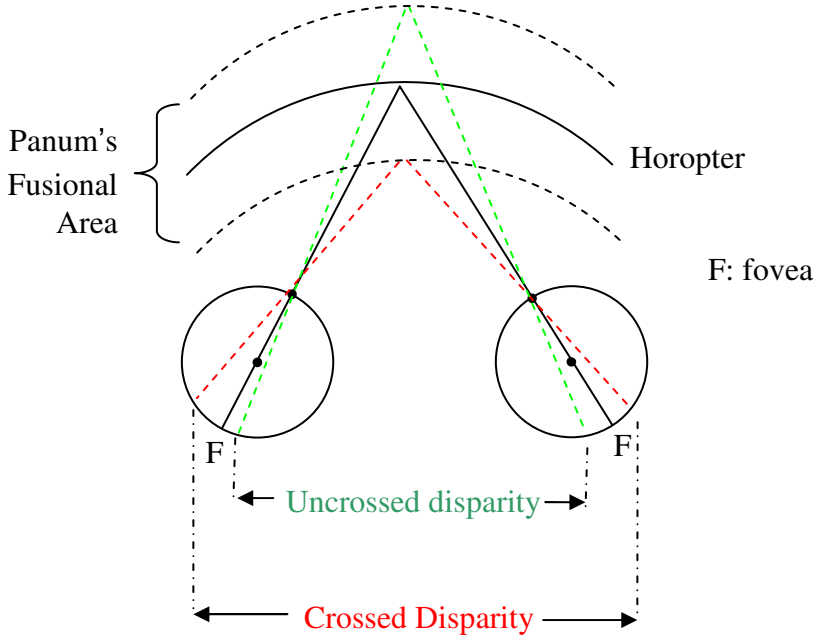


Figure A.3 : Crossed and uncrossed disparities

Number of disparity pairs that can be processed is limited and its limit is around 200.000 pairs.

Using the orientations of eyes and the disparity information, depth relative to the fixation point is calculated. Therefore, in order to find the exact distance of an object, the distance of the fixation object must be available.

A.6.2 Convergence

Convergence is a binocular 3-D cue, obtained by data coming from both eyes. For each different alignment of right and left eyes, in order to fixate to a point (project a point on to fovea), there is a unique angle between left and right eye. From this angle and alignment of the eyes, distance of the fixation point is calculated by the brain (Figure A.4). However, convergence is not a precise data due to insufficient feedback mechanism. Signals that are sent from central nerve system to the effectors are called as *efferent signals* and feedback signals from effectors to

central nerve system are denoted as *afferent signals*. In convergence, afferent signals are neglected and system makes calculations relative to efferent signals.

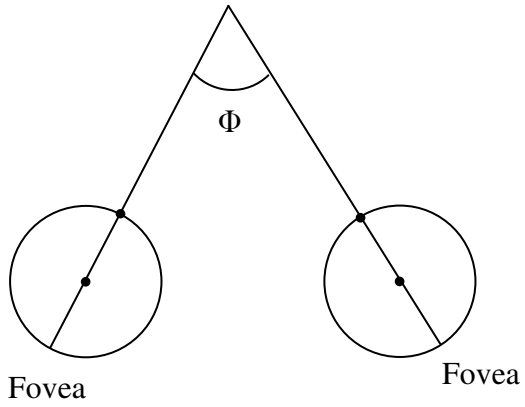


Figure A.4 : Convergence

A.6.3 Motion Parallax

Motion Parallax is a monocular cue, caused by any relative motion between the scene and the eye. During this relative motion, the fixation object remains on fovea, nearer objects move in opposite direction, with increasing speed with decreasing distance, whereas further objects move in same direction with increasing speed with increasing distance (Figure A.5). The object at infinite distance, such as moon or sun will move at a speed equal to the speed of eye. From the direction and the speed of the retinal projection of an object, its depth relative to the fixation point is calculated. In order to find the distance of an object, the distance of the fixation object must be available.

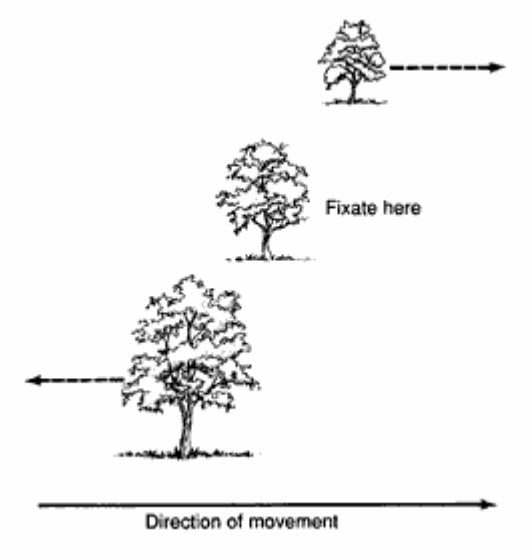


Figure A.5 : Motion parallax

A.6.4 Accommodation

Accommodation is a monocular cue, obtained from a single eye. For each different distance, in order to fixate a point at that distance, lens must have a unique thickness. From this thickness information, the distance of the fixation point is calculated. Similar to convergence, in accommodation, afferent signals are also neglected.

A.6.5 Pictorial Cues

Pictorial cues are defining the further-nearer relations between objects. Shadows, occlusions, perspective etc. are typical pictorial cues. The contrast of the scene, illumination etc. might affect the pictorial cues. Pictorial cues are evaluated by using the world knowledge through top-down processes.

Effect of the pictorial cues on depth perception may be easily seen on Ponzo illusion. In this illusion, two identical objects are placed on converging lines and subjects are asked to select the one with greater size. Since converging lines gives an impression of increasing depth, one object is perceived further. Since two objects' retinal size are equal and real size is obtained by using distance between object and viewer and retinal size according to Emert's Law, one of the objects is perceived larger. A sample drawing for this illusion is given in Figure A. 6.

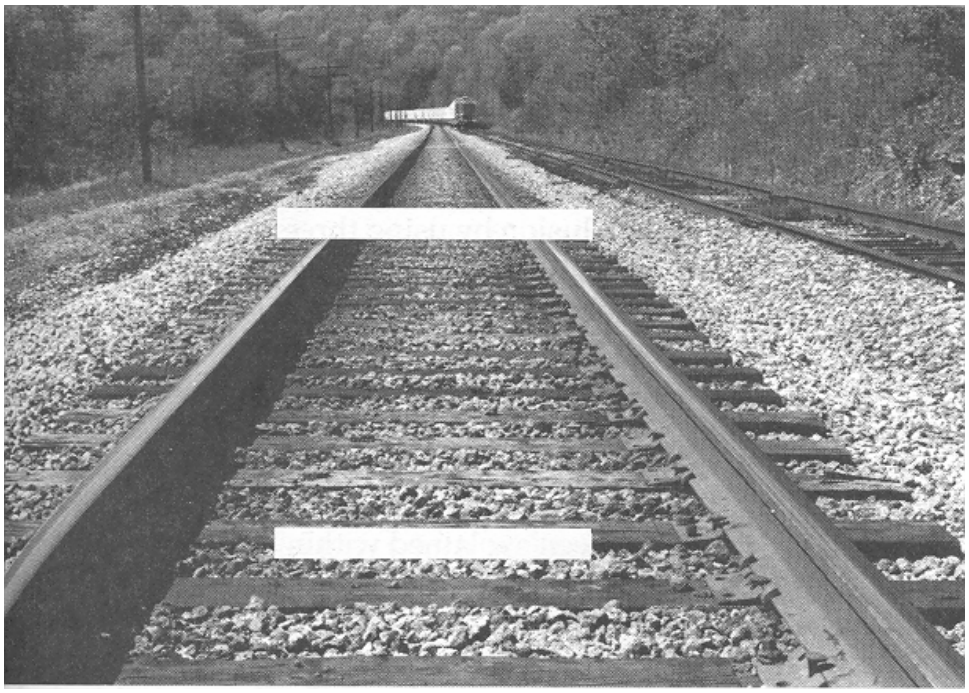


Figure A. 6 : Ponzo illusion

A.7 3-D Displays and Depth Perception

A.7.1 Binocular Cues

In 3-D displays, in order to provide disparity and convergence information, for objects desired to be on screen, position of the object in the left and right images should be equal, whereas for the objects to be perceived inside the screen, the object on left image should be shifted to left to generate uncrossed disparity and finally, for the objects outside of the screen, the one on left image should be

shifted to the right to generate crossed disparity. If the following assumptions are correct then the disparity and convergence inputs provided by a 3-D display should be same with the ones in a real 3-D environment:

- left and right images are perfectly separated,
- disparity pairs lies on Panum's Fusional Area,
- amount of disparity pairs are kept below the processing limit,
- perfect depth extraction and coding are available in the system, and
- there are no compression or transmission errors.

A.7.2 Monocular Cues

Since left and right image lies on the plane of the 3-D display, in order to focus on the images, lens is set to sharpen the objects on the screen plane. Therefore, accommodation is expected to give distance between viewer and screen plane. In a real 3-D environment, focusing on an object, nearer or further objects will be gradually blurred based on their distance due to limited depth focus [_23]. However in a stereoscopic display, all objects lie on the screen with maximum resolution, which possibly makes viewer understand the planarity. Similarly, when viewer moves, the resulting motion parallax is related with the distance between user and screen plane rather than distance between objects and viewer. Actually, the motion parallax that is provided in 3-D displays, when the viewer moves, will be same with those in 2-D displays, which yields viewer to understand all objects in the scene lying on the same plane. This specific situation for 3-D displays will be denoted as *lack of motion parallax* [1].

When a camera moves, objects at different distances will move by various speeds, producing a parallax effect. However, in this case, if a viewer is stationary (or is not moving exactly the same manner with camera), evaluation of this parallax might be ambiguous due to the position constancy, since the parallax that viewer receives is not caused by his/her movements. This situation will be termed as *unexpected motion parallax*. Lack of motion parallax and unexpected motion parallax is simply erroneous monocular cues that are referred in Chapter 1. In summary, 3-D displays does not yield any monocular cues to HVS, the perception will be same as the ones in 2-D displays.

A.7.3 Conflict between Monocular and Binocular Cues

In real 3-D world, data coming from convergence and accommodation are expected to be similar (not exactly same due to rejection of afferent signals). In 3-D displays, since data coming from accommodation and convergence is different, an ambiguity arises, namely *accommodation-convergence conflict* [1]. This conflict itself might have some effects on the perceived depth; moreover, since disparity and motion parallax generates relative depth information, requiring distance between viewer and object on fovea, it might also have some effects on evaluation of disparity and motion parallax. Some researches point a possible feedback mechanism between accommodation and convergence [1]. Accommodation-convergence conflict is assumed to break this feedback mechanism and causes eye strain [1].

Motion parallax and disparity are also expected to generate similar results from a real 3-D scene. However, for a 3-D display, when the camera is stationary, motion parallax provides a planar scene, whereas disparity results by a 3-D scene, which is another conflict named *disparity-parallax conflict*. The only way to prevent this conflict is to move the camera which causes unexpected parallax, i.e. an ambiguity. The subjective tests show that viewers perceive depth better for the scenes with motion with respect to the stationary scenes; in other words, they prefer unexpected motion parallax to the lack of motion parallax [1].

A.7.4 Other Variables Affecting Depth Perception

Monocular and binocular cues that a viewer is expected to receive while gazing on to a 3-D display are already defined, while the other variables are kept ideal. However, the display and the examined content cannot be ideal. Quantization of depth is the first error source that one can face with. Errors during depth estimation, coding, transmission and compression should affect depth perception. In order to generate disparity, 3-D displays project two different images to the left and right eyes; however, every display has its own optical limitations for filtering the left and right eye's images. It is expected to observe a leakage which will cause two images to be mixed up.

For pictorial cues, SNR, illumination and contrast of the scene may be thought to have some effect on depth perception; however, considering the evaluation of pictorial cues, even these will not be significant, as long as the form information on retinal mosaic is not distorted significantly. The research results support this assumption, image size, MPEG-2 coding and low-pass filtering are reported to have no significant effect on depth perception [1].

APPENDIX-B

A SAMPLE USAGE OF PROPOSED SYSTEM: DEPTH PERCEPTION EVALUATION IN AUTOSTEREOSCOPIC DISPLAYS

In this appendix, a sample usage of proposed system to evaluate depth perception in autostereoscopic displays is defined. Depth perception should be evaluated both in subjective and objective simulations. The aim is to be able to quantify depth perception using estimation of perceptual inputs in order to be able to separate effects of scene content and processing on depth perception.

B.1 Subjective Evaluation

For the subjective evaluation of the scene, subjects are asked to rate perceived depth on the scene. Subjective evaluation methods of a scene in 3-D displays are defined in ITU-R BT.1438 [18]. For long video sequences, Single Stimulus Continuous Quality Evaluation Method (SSCQE) is considered to be a convenient method [1]. Subjects rate perceived depth for given content by a mouse scroll, as increasing or decreasing impression of depth.

B.2 Objective Evaluation

Results of subjective evaluation of a scene is a function of some parameters, such as scene content, subject's attention and errors in transmission, depth estimation, coding and display. Subjective assessments are conducted in order to estimate performance of engineering efforts, such as depth estimation, coding etc., while scene content is controlled. However, in commercial 3-D TV, variables similar to scene content and subjects' attention could not be controlled; therefore, in order to make results of the subjective assessments applicable, content and subject independent scores are required.

In subjective evaluation strategy, for reducing subjective variances, mean opinion score (MOS) among many subjects is obtained. Similarly, repeating subjective assessments with different contents is used to reduce effect of the content on MOS. However, obtaining MOS (or repeating assessments with different content) might reduce the subjective variances (effect of the content); on the other hand, effect of the subjects' attention (content) should be still present in the scores, at least in the mean. As explained in Appendix-A, erroneous monocular cues and monocular-binocular conflicts are dependent on the scene content as well as the subjects' attention and they might have a significant effect on perceived depth. If this is the case, there may be a ceiling effect on MOS, caused by both content and subjects' attention. Effect of engineering efforts might be measured as distortion rate in perceived quality. However, in order to obtain this rate, the best score that can be obtained as a result of the available content and current subject should be estimated. In this section, an objective evaluation of the scene defined as the estimation of effects of the erroneous monocular cues and monocular-binocular conflicts on perceived quality is introduced. Proposed depth quality evaluation block diagram is given in Figure B.1.

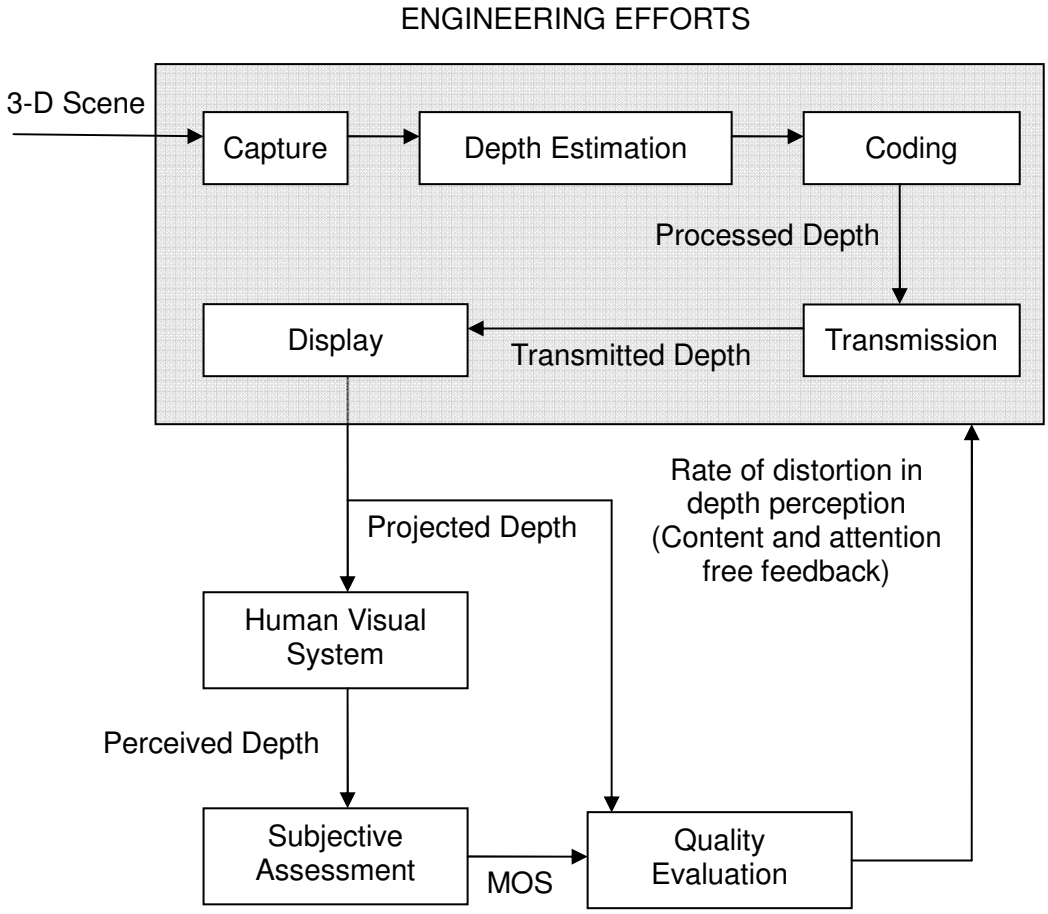


Figure B.1 : Depth Quality Evaluation Block Diagram

Accommodation-convergence conflict is one of the factors affecting the perceived depth. While all other parameters kept constant, for different accommodation-convergence conflict levels, by obtaining subjective evaluation of the scene, rate of perception distortion can be related to the conflict level.

Motion parallax information provided by a 3-D display is not different the one from provided by a 2-D display, as explained in Appendix-A. Motion parallax is received, or expected to be received, when either camera or viewer moves. If neither camera not viewer is moving, then system neither receives nor expects to receive motion parallax; therefore, disparity-parallax conflict does not arise. Therefore, disparity-parallax conflict is defined only when camera or viewer is

moving and it is defined as the difference between depth maps obtained via disparity and motion parallax. Since motion parallax can be provided only by camera movements, relative movement between camera and viewer should cause unexpected motion parallax or lack of motion parallax. If it is assumed that subject is able to process parallax provided by camera movements (can handle the unexpected parallax), then disparity-parallax conflict becomes a function of scene content only and assumed to be independent of viewer. Lack of motion parallax is defined when viewer is moving and camera is stationary. Unexpected parallax is defined when camera is moving. Both will be defined as the difference between pupil motion vector and camera motion vector. Mentioned conflicts and errors cues and their measurements are given in Table B.1

Table B.1 : Conflicts and errors to be measured

Conflict/Error	**Pre-requisite**	**Measurement**
Accommodation-convergence conflict	none	convergence - accommodation
Lack of motion parallax	zero camera motion	pupil motion
Unexpected parallax	non-zero camera motion	pupil motion - camera motion
Disparity-parallax conflict	non-zero camera or pupil motion	disparity depth map - motion depth map

For a translational camera motion with an appropriate speed, disparity-parallax conflict should converge to zero, whereas unexpected motion parallax is expected to increase. When camera movement is not present, but the pupil moves, then display-parallax conflict should be maximum, similarly lack of motion parallax also increases. It is expected that disparity-parallax conflict and lack of motion parallax are highly correlated and unexpected parallax should be their complement.

For different camera motions, while all other parameters kept constant, obtaining perceived depth scores and taking viewers movements into account, disparity-parallax conflict, unexpected parallax and lack of motion parallax can be estimated and this estimations can be related to the rate of distortion in depth perception.

B.3 Experimental Setup

B.3.1 Setup

Experiment should aim to refine effect of four errors (see Table B.1) on perceived depth: accommodation-convergence conflict, disparity-parallax conflict, lack of motion parallax and unexpected parallax. To refine effect of the errors different scenes causing different errors should be generated. Considering performance of the proposed gaze tracker and limits of a 3-D TV, accommodation convergence conflict will be quantized at 16 levels with 5cm step size. Similarly, disparity-parallax conflict will be quantized at 16 levels with 5cm step size. Lack of motion parallax and unexpected parallax will also be quantized at 16 levels but at 5mm step size. To reduce effects caused by other cues, experiments will be repeated with different levels of depth resolution (disparity), different amount of depth and color contrast and different textures of background.

B.3.2 Measurements

Following measurements will be taken throughout the experiment.

Perceived Depth : Perceived depth will be obtained from subjects by using SSCQE as mentioned in Section B.1.

Convergence : Gaze point of the subject will be measured via the method proposed in Chapter III. Gaze point in 3-D will be the estimate for convergence.

Accommodation : Projection of gaze point on the screen plane will be calculated and distance between pupil and this point will be used as accommodation estimate.

Accommodation-Convergence Conflict : Projection of distance between convergence and accommodation estimates on to z-axis of head coordinate system will be the accommodation-convergence conflict estimate.

Disparity : Depth map given to the display will be used as disparity estimate.

Received Motion Parallax : Depth map produced by depth estimation from camera motion will be used as received motion parallax.

Unexpected Motion Parallax : It is defined when camera movement is present and represented as difference between pupil motion vector and camera motion vector.

Lack of Motion Parallax : It is defined when camera movement is not present and represented as pupil motion vector.

Disparity-Parallax Conflict : When neither camera nor pupil movement is present then this conflict will be zero since HVS neither expects nor receives motion parallax. When pupil or camera movement is present, mean of squared difference between disparity and received motion parallax will be the estimate for disparity-parallax conflict. This conflict will be calculated in Panum's Fusional Area only.

B.3.3 Evaluation of Results

As previous researches show, accommodation-convergence conflict may have a effect on eye strain [1] and possibly headaches, however, considering efferent inputs and ambiguity resolution capability of HVS, this conflict is not expected to have a significant effect on perceived depth. A similar inference may be valid for unexpected parallax. For long video sequences relation between experienced headaches and these ambiguous input is expected to be related. Different from these ambiguous inputs, lack of motion parallax and disparity-parallax conflict is expected to have a significant effect on perceived depth.

APPENDIX-C

CAMERA GEOMETRY, CALIBRATION and POSE ESTIMATION

In order to determine gaze direction and 3-D coordinates of gaze point, the geometry and calibration of the imaging system should be understood. Hartley and Zisserman [22] present camera geometry, calibration and pose estimation issues and their possible solutions in detail. Projective camera model, pose estimation and calibration issues are also discussed in detail. A point X in 3-D and its projection on to image plane is related with following equation:

$$k\begin{bmatrix} x \\ 1 \end{bmatrix} = \boldsymbol{P}\begin{bmatrix} X \\ 1 \end{bmatrix} \tag{C.1}$$

where $x = \begin{bmatrix} u & v \end{bmatrix}$ are the observed point coordinates of X on image plane, k is the unknown scale vector and P is the 3x4 projection matrix. Projection matrix is formed by a calibration and transformation matrix

$$P = K\ {}^{C}T_{W} \tag{C.2}$$

where K is the calibration matrix and ${}^{C}T_{W}$ is the transformation from world coordinate system (in which X is defined) to the camera coordinate system.

Calibration matrix and the transformation matrix have the following form

$$K = \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{C.3}$$

$$^{C}T_{W} = \begin{bmatrix} R \mid -RC \end{bmatrix} \tag{C.4}$$

where R is a 3x3 rotation matrix and C is coordinates of the optical center of the camera in world coordinate system. For a known point on a 3-D structure and its projection on to the image plane, projection matrix can be computed. Since calibration matrix is a upper-triangular matrix and rotation matrix is an orthogonal matrix, projection matrix can be decomposed into the calibration matrix and transformation matrix by computing right-triangular and orthogonal parts (RQ decomposition).

Camera model presented in (C.1) is a linear model, however projection of a point on to the image plane includes non-linear distortions, namely radial and tangential distortions. Calibration toolbox [31] includes such distortions and camera parameters are defined as follows.

For a point $X = \begin{bmatrix} x_c & y_c & z_c \end{bmatrix}$ given in CCS, it's normalized projection is defined as follows

$$\begin{bmatrix} x_n \\ y_n \end{bmatrix} = \begin{bmatrix} x_c / z_c \\ y_c / z_c \end{bmatrix} \tag{C.5}$$

When radial and tangential distortions are applied, new normalized coordinate of the point is defined as follows:

$$\begin{bmatrix} x_d \\ y_d \end{bmatrix} = \left(1 + k_1 r^2 + k_2 r^4 + k_5 r^6\right) \begin{bmatrix} x_n \\ y_n \end{bmatrix} + \begin{bmatrix} 2k_3 x_n y_n + k_4 (r^2 + 2x_n^2) \\ k_3 (r^2 + 2y_n^2) + 2k_4 x_n y_n \end{bmatrix} \tag{C.6}$$

where $r^2 = x_n^2 + y_n^2$ and k_i's are lens distortion parameters.

Once the distorted coordinates are found, image coordinates of the point is obtained with the following relation:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & \alpha & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_d \\ y_d \\ 1 \end{bmatrix} \tag{C.7}$$

where f_x and f_y are scaling factors on x- and y axis, c_x and c_y are principal point and α is the skew coefficient.

When 3-D coordinates of the points in world coordinate system and their projections on to the image plane is provided, toolbox solves the internal camera parameters and external pose throughout a non-linear optimization by minimizing reprojection errors [33].

When internal calibration parameters, 3-D coordinates of the points in world coordinate system and their projections on to the image plane is provided, toolbox first removes the lens distortions, then finds the transformation between world coordinate system and camera coordinate system using (C.1) and known camera calibration matrix.

APPENDIX-D

TRIANGULATION METHODS

In order to obtain gaze point in 3-D, intersection of right and left LoG/LoS is used. Intersection of two lines in space is referred as triangulation problem. Hartley and Sturm [21] present a review of triangulation methods for 3-D reconstruction and propose an optimal solution for triangulation named *polynomial triangulation*.

Midpoint Triangulation: Midpoint method is a popular approach for triangulation. In this method, the midpoint at which the lines are closest to each other are taken as the solution. Midpoint triangulation with left and right LoG is presented in Figure D.1.

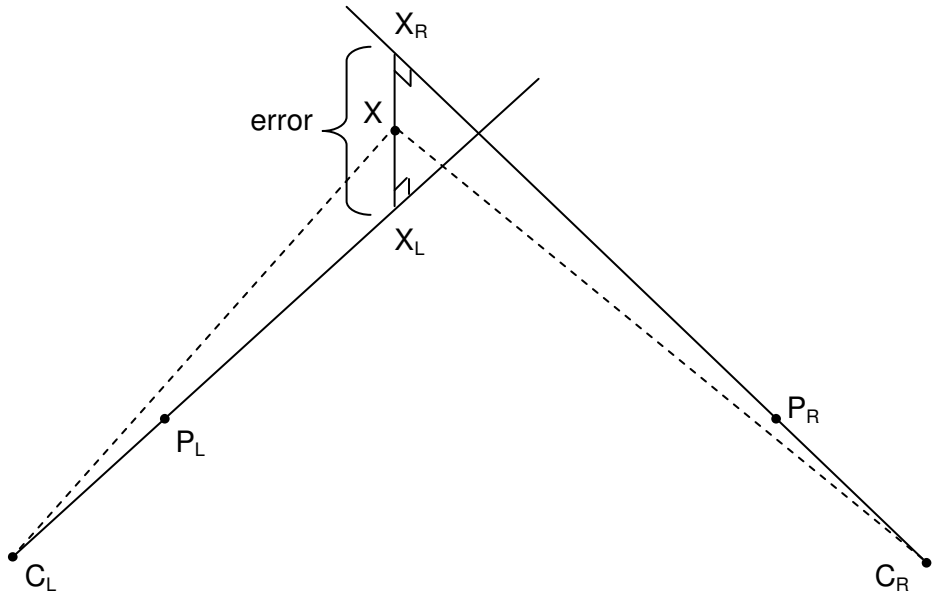


Figure D.1 : Midpoint triangulation with left and right LoG

Least Squares Triangulation: This method is a linear triangulation method. Projection of a point X in 3-D on to image plane (C.1) can be written in following form:

$$\begin{aligned} ku &= p_1^T \begin{bmatrix} X & 1 \end{bmatrix}^T \\ kv &= p_2^T \begin{bmatrix} X & 1 \end{bmatrix}^T \\ k &= p_3^T \begin{bmatrix} X & 1 \end{bmatrix}^T \end{aligned} \tag{D.1}$$

where p_i^T is the ith row of the projection matrix. When we rearrange (D.1),

$$\begin{aligned} \left(up_3^T - p_1^T \right) \begin{bmatrix} X & 1 \end{bmatrix}^T &= 0 \\ \left(vp_3^T - p_2^T \right) \begin{bmatrix} X & 1 \end{bmatrix}^T &= 0 \end{aligned} \tag{D.2}$$

can be obtained for a point X and it's projection $\begin{bmatrix} u & v \end{bmatrix}$. If projections of the same point X on different images are known one can write following equation

$$\begin{bmatrix} u_1 p_{1,3}^T - p_{1,1}^T \\ v_1 p_{1,3}^T - p_{1,2}^T \\ \vdots \\ u_N p_{N,3}^T - p_{N,1}^T \\ u_N p_{N,3}^T - p_{N,2}^T \end{bmatrix} \begin{bmatrix} X \\ 1 \end{bmatrix} = 0 \tag{D.3}$$

where $\begin{bmatrix} u_i & v_i \end{bmatrix}$ are the observed point coordinates of X on the ith image and $p_{i,j}^T$ is the jth row of the ith projection matrix. For N different images, (D.3) states a 2N equations, 3 unknowns problem. Least squares solution of this over-determined problem can be found by pseudo-inverse or SVD.

When right and left LoG (LoS) is given, gaze point should satisfy following set of equations

$$\begin{aligned} X_i &= k_{L,i} G_{L,i} + C_{L,i} \\ X_i &= k_{R,i} G_{R,i} + C_{R,i} \end{aligned} \tag{D.4}$$

where G_i is the unit vector in the direction of LoG and C_i is the eyeball center at i[th] instant for left and right eyes. (D.4) can be written in following form

$$k_i G_i = \left[I_{3x3} \mid -C_i \right] \begin{bmatrix} X_i \\ 1 \end{bmatrix} \tag{D.5}$$

Since (D.5) has the same form with (C.1), one can easily apply least squares triangulation with given eyeball centers and LoG (LoS). If left and right LoG at a single frame is utilized, then this solution is referred as the least squares triangulation. If more than one frames are utilized, then this solution is referred as multi-line least squares triangulation.

Polynomial Triangulation: Hartley and Sturm [21] state that an optimal solution to triangulation problem can be found by minimizing reprojection errors. Therefore, cost function should be defined as the reprojection error.

$$J = d(x_1, \hat{x}_1)^2 + d(x_2, \hat{x}_2)^2 \tag{D.6}$$

where $d(*,*)$ represents Euclidian distance, subject to the epipolar constraint

$$\hat{x}_1^T F \, \hat{x}_2 = 0 \tag{D.7}$$

and F is the Fundamental matrix. (D.7) states that pairs on two images should line on epipolar lines, then (D.6) can be expressed in the following form

$$J = d(x_1, \lambda_1)^2 + d(x_2, \lambda_2)^2 \tag{D.8}$$

where λ_1 and λ_2 range over all choices of corresponding epipolar lines.

To minimize (D.8), following strategy is proposed.

1. Parameterize the pencil of epipolar lines in the first image by a parameter t. Thus an epipolar line in the first image may be written as $\lambda_1(t)$.
2. Using the fundamental matrix F, compute the corresponding epipolar line $\lambda_2(t)$ in the second image.
3. Express the distance function $d(x_1, \lambda_1(t))^2 + d(x_2, \lambda_2(t))^2$ explicitly as a function of t.
4. Find the value of t that minimizes this function.

This solution can be applied to triangulation of two LoG (LoS) with following procedure.

1. Construct two coordinate systems centered at left and right eyeball centers. z-axis is defined by current LoG. Define a common y-axis by cross product of the left and right LoG.
2. Place two pinhole cameras to the origin of the coordinate systems constructed in previous step. Focal length of the camera is equal to the eyeball radius.
3. Optical centers of the cameras are eyeball centers and projection of the 3-D point on to image planes are pupils.
4. Find the transformation between constructed coordinate systems and find the fundamental matrix.
5. Apply polynomial triangulation to find gaze point.