PUPIL

Constructing the Space of Visual Attention

by

Moritz Philipp Kassner
Pre-Diploma, Product Design
University of the Arts (UDK)
Berlin, Germany, 2009

William Rhoades Patera
Bachelor of Architecture
Cornell University
Ithaca, New York, 2007

SUBMITTED TO THE DEPARTMENT OF ARCHITECTURE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ARCHITECTURE STUDIES
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 2012

Signature of Author: **Moritz Phillip Kassner**
Department of Architecture, May 24, 2012

Signature of Author: **William Rhoades Patera**
Department of Architecture, May 24, 2012

Certified by: **Terry Knight**
Professor of Design and Computation,
Department of Architecture, Thesis Adviser

Certified by: **Patrick Winston**
Ford Professor of Artificial Intelligence and Computer Science, EECS
Thesis Adviser

Certified by: **Takehiko Nagakura**
Associate Professor of Design and Computation
Chair of the Department Committee on Graduate Students

PUPIL

Constructing the Space of Visual Attention

Committee

Thesis Adviser
Terry Knight
Professor of Design and Computation,
Department of Architecture

Thesis Adviser
Patrick Winston
Ford Professor of Artificial Intelligence
and Computer Science, EECS

PUPIL

Constructing the Space of Visual Attention

by

Moritz Philipp Kassner | **William Rhoades Patera**

SUBMITTED TO THE DEPARTMENT OF ARCHITECTURE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ARCHITECTURE STUDIES
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 2012

Abstract

This thesis explores the nature of a human experience in space through a primary inquiry into vision. This inquiry begins by questioning the existing methods and instruments employed to capture and represent a human experience of space. While existing qualitative and quantitative methods and instruments – from "subjective" interviews to "objective" photographic documentation – may lead to insight in the study of a human experience in space, we argue that they are inherently limited with respect to physiological realities.

As one moves about the world, one believes to see the world as continuous and fully resolved. However, this is not how human vision is currently understood to function on a physiological level.

If we want to understand how humans visually construct a space, then we must examine patterns of visual attention on a physiological level. In order to inquire into patterns of visual attention in three dimensional space, we need to develop new instruments and new methods of representation. The instruments we require, directly address the physiological realities of vision, and the methods of representation seek to situate the human subject within a space of their own construction. In order to achieve this goal we have developed **PUPIL**, a custom set of hardware and software instruments, that capture the subject's eye movements. Using **PUPIL**, we have conducted a series of trials from proof of concept – demonstrating the capabilities of our instruments – to critical inquiry of the relationship between a human subject and a space. We have developed software to visualize this unique spatial experience, and have posed open questions based on the initial findings of our trials.

This thesis aims to contribute to spatial design disciplines, by providing a new way to capture and represent a human experience of space.

Thesis Supervisors

Terry Knight	**Patrick Winston**
Professor of Design and Computation, Department of Architecture	Ford Professor of Artificial Intelligence and Computer Science, EECS

Acknowledgements

First and foremost, we would like to thank our advisers, Professor Terry Knight and Professor Patrick Winston, who both have significantly shaped our experience at MIT. Both Professor Knight and Professor Winston were incredibly supportive of our curiosities and endeavors. They have, and will, continue to serve as inspirational figures both in academics and spirit.

We would also like to thank all of our colleagues in the Design and Computation Group.

Moritz Kassner

Thanks to the German National Academic Foundation, the German Academic Exchange Service, and MIT for financially supporting my studies at MIT.

To family and friends for their support and encouragement to take this journey.

Axel Kufus and Frank Spenling for their guidance, support and friendship.

Yeon, for Cambridge Wonderland.

Will, for being the global optimum in thesis collaborator space.

Finally, I would like to thank Elias, my little, great son, who has been so patient during these two years. I look forward to spending my days with you again.

William Patera

Thanks to family and friends for their continual patience, support, and advice. Nina for encouraging me to expand my curiosities; your support, guidance, and love.

Professor David Mindell, Yanni Loukissas, and those in the LARS group for giving me critical feedback and structure in the early stages of my research.

Moritz, from and with whom I continue to learn.

Table of Contents

Introduction

This thesis explores the nature of a human experience in space through a primary inquiry into vision. This inquiry begins by questioning the existing methods and instruments employed to capture and represent a human experience of space. While existing qualitative and quantitative methods and instruments – from "subjective" interviews to "objective" photographic documentation – may lead to insight in the study of a human experience in space, we argue that they are inherently limited with respect to physiological realities.

As one moves about the world, one believe to see the world as continuous and fully resolved. The proliferation of images, photographic instruments, and screens in everyday routines only contribute to the misguided paradigm that equates a visual and embodied experience with a mechanical and objective reproduction of the world. In this paradigm, vision is understood as *image formation*. However, this is not how human vision is currently understood to function on a physiological level. One does *not* reconstruct images in their brains with a one to one correspondence to the world. While the human field of vision is expansive, the area that can be resolved at high resolution and in color is tiny – about the area of one's thumbnail held at arms length. This small area of high resolution physiologically corresponds directly to the area in the human retinal surface called the fovea. Due to the fact that one only sees but a fraction of the world in high resolution, and the rest in diminishing acuity, one must move their eyes rapidly in order to process the world.

Figure 1: Vision as image formation paradigm. Sculpture by Dimitri Hadzi, "Elmo-MIT," 1963. Location Hayden Library Courtyard, MIT, Cambridge. Photograph by authors, extracted from a film.

In this thesis we argue in support of an alternative paradigm, where vision is understood as *information processing*.[2] In this way of thinking we consider the physiological realities of human vision

[1] Margaret Livingston, "What Art Can Tell Us About the Brain," Design and Computation Group Lecture, MIT, April 13, 2012. We borrow these concepts from Livingston. However the use of the word "paradigm" in this context is precarious as it infers an anti-positivist epistemological model. Instead of adopting this Kuhnian word, we seek an alternative epistemology that is based on the relationship between the interleaved development of instruments, experiments, and theory as written about by Thomas S. Kuhn, *The Structure of Scientific Revolutions*, 3rd ed. (University Of Chicago Press, 1996); Peter Galison,, *Image and Logic: A Material Culture of Microphysics* (Chicago: University Of Chicago Press, 1997), Chapter 9, "The Trading Zone: Coordinating Action and Belief".

as an active, highly selective, intent driven, and fully unconscious process, where the patterns of visual attention are a unique and intimate artifact of a human experience in space. In order to process the world, the human eye must remain relatively still or in order to attend to – fixate – on an area of interest.[3] These patterns of rapid movement and fixation are necessary processes of human vision. Studying these patterns allows us to learn about the relationship between a human subject and a space. We call this subjective relationship a visual construction of space.

If we want to understand how humans visually construct a space, then we must examine patterns of visual attention. In order to inquire into patterns of visual attention in three dimensional space, we need to develop new instruments and new methods of representation. The instruments we require, directly address the physiological realities of vision, and the methods of representation seek to situate the human subject within a space of their own construction. In order to achieve this goal we have developed **PUPIL**, a custom set of hardware and software instruments, that capture the subject's eye movements. While commercial eye tracking systems exist, they are prohibitively expensive and closed for development in both their hardware and software. We created **PUPIL** from raw silicon to high level Python control libraries, primarily because we wanted to inquire into the low level processes of human vision that are highly unconscious. But our development is also motivated by the desire to make an area of inquiry accessible and to make an instrument that would enable future research for a broad range of disciplinary interests.

Using **PUPIL**, we have conducted a series of trials from proof of concept – demonstrating the capabilities of our instruments – to critical inquiry

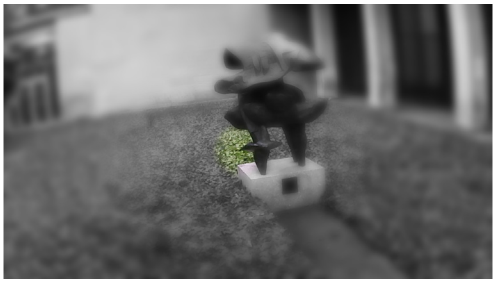


Figure 2: Vision information processing paradigm. Foveal view is simulated as colored ellipse in the center, peripherial vision is gray and increasingly blurred as distance increases from the center of visual attention. Sculpture by Dimitri Hadzi, "Elmo-MIT," 1963. Location Hayden Library Courtyard, MIT, Cambridge. Photograph by authors, extracted from a film.

[2] In fact, one is effectively blind during these rapid eye movements, as the vision system is blocked. You can discover this by looking into a mirror and trying to see your eyes move. You will not see your own eyes move, due to this temporary blindness. This was well known by experimental psychologists and physiologists , but curiously remains an

of the relationship between a human subject and a space.

After conducting trials we were faced with the challenge of representing human experience in space. Understanding vision as information processing, we sought to reveal the unique artifacts of human vision. The representational methods we have developed act as a critique of existing representations of vision, specifically in the associated fields of spatial design. In existing practices, "objective" representations of vision are based on a mechanical or geometric understanding – image formation. The camera obscura, cinematography are instruments employed to represent visions of space. However, these methods of representing vision are limited, both spatially and temporally, by reducing the experience of human subject to a single vantage point locked in time. Even the most "truthful," "objective," or "realistic" representations of human vision are full of treason.[4] But why? From studies of physiology, we know that humans can not resolve but a tiny fragment of our surrounding environment. But those who represent visions of space continue to produce images based on the vision as image formation understanding.

We are in search of an alternative way to represent the human experience of space. There is no way to "objectively" represent this unique relationship. Nor do we propose a solution. We are attempting to represent vision in a way that reveals some of the physiological challenges of human visual attention and simultaneously the reflexive bias of our own methods. This experience, as we have argued, is fragmented and partial. But the amount of information that humans are able to process from this seemingly limited experience is extraordinary. In order to understand how humans visually attend

[3] Lorraine J. Daston and Peter Galison, *Objectivity* (Zone, 2010).

to a space, we have created a collection of tools and representational methods — that situate the human subject as the constructor of representations.

While the retinal surface may capture something analogous to “images” of the world, we believe that there are no representations of space in the human brain. Following literature in cognitive science and physiology of vision, we assume that the process of vision does not reconstruct an image in the brain of the human subject. Rather, vision is information processing – a constructive process. Our tools and sensors, analogs to the retinal surface of the human eye, capture images of the world, but to our computer these images are just information that must be processed. Our computer knows nothing of space, nor how to represent it given a two dimensional stream of information. The representational methods we have developed rely on a series of software libraries that enable the construction of a three dimensional representation based on this information stream. This process is called Structure From Motion (SfM), where a stream of images from a moving subject are used to recover the pose of the subject’s head and three dimensional model of the world as triangulated points in space.[5] We merge the physiological reality of human visual attention with a mechanical construction of vision. Two modes of thinking about vision – mechanical and physiological or image and information – are both partial and fragmented in their own right. But by leveraging one against the other, the nature of their own mechanisms of construction are revealed.

We believe that these instruments and methods of representation allow one to inquire into the relationship of a human experience of space, and pose questions that challenge our assumptions as

[4] The history of SfM can be traced back to the study of optical illusions in experimental psychology, where a human subject is able to construct an understanding of a three dimensional solid by observing the relative movement of two dimensional points. In computer vision, this problem is studied in order to compute three dimensions given only relative movement between two dimensional points or features in images. Shimon Ullman posed the problem for computer vision in 1979.

designers. For example, we believe **PUPIL** may allow us to capture the richness of a human experience even in what may be considered a "complex" space. Or, to even reconsider what we mean when we say that a space is "complex".

Chapter Summary

In the beginning of the first chapter we provide definitions in an overview of the physiology of the human eye. The background chapter continues in a series of essays concentrating on the history of physiological and psychological studies of vision, within the context of eye movement studies in experimental psychology in the nineteenth and twentieth centuries. Through a historical review we identify a shift from qualitative or introspective methods to examine human vision to quantitative or proto-computational understanding. This shift aligns with the development of new instruments and metaphors for vision – vision as image formation. Within this historical context we will outline a number of case studies of visual attention experiments conducted by psychologists that seek quantitative and statistical explanations for literacy in areas from reading to art critique. In the final section we move towards a contemporary literature review that is supported with a discussion on the shifting status of vision in modernity and the formation of "objective" representations.

In the second chapter we introduce our platform of instruments and representational methods – **PUPIL**. The first section of this chapter contains a detailed discussion of the hardware apparatus and the associated technological and physiological constraints that were considered in its design and development. In addition to a hardware system, we introduce novel representation methods, that employ a structure from motion (SfM) pipeline to *construct* a the space of a human experience, and to situate the human subject within that space. The final section of this chapter documents a series of trials conducted with **PUPIL** with different human subjects in a variety of environments and scenarios. We conclude this section with an analysis of our initial findings from our trials, and pose open questions.

Background

Section Overview

The background of this thesis is divided into two major sections. The first section describes the physiology of the human eye, outlines constraints of human vision, and introduces key terms and assumptions central to the thesis. The second section provides a historical review of experiments on human vision in experimental psychology and concludes with a discussion of scientific representations of objective vision.

Physiology of Vision — The Human Eye

Through an understanding of the optics and sensing capabilities of the human eye reveal we reveal constraints specific to human vision, that sponsor the question of why do we move our eyes to see the world? We then proceed to discuss how the eye is moved through a discussion of central motifs — saccades and fixation — and why these movements are necessary on a physiological level. Furthermore, we question how movements of the eye are motivated.

Through a description of physiology of the human eye and by outlining questions based on physiological understandings we will expose the underlying assumptions of our research. One of the central assumptions is that what we fixate on, on the physiological level of eye movements, is what we are attending to. In the next section we will demonstrate how the correlation of saccade-fixation and visual attention evolved in a specific disciplinary context.

History of Eye Movement Studies and Representations of Objective Vision

In this section we introduce a theoretical framework for the thesis by reviewing late nineteenth and early twentieth century psychology experiments on human vision. Within this context our research concentrates on reviewing psychology experiments where new instruments were developed and employed to measure physiological movements – specifically eye movements – and correlated with cognitive response.

Historically, this is an interesting period for the (then, young) field of psychology as it struggled to gain acceptance as a "hard science" in the sub domain of experimental psychology. We will demonstrate through a review of experiments , that it is through physiology that experimental psychology makes its claim to empirical truths, where vision and eye movements played a central part in closing the gap between physiological behaviors and cognitive activity. We will describe how eye movements became an object of inquiry in the space of the experimental psychology laboratory through the study of reading and examining images. We will show how the development of instruments in the laboratory of experimental psychologists brought about new ways of theorizing and representing human vision. We will make connections between research on human vision in experimental psychology with early research on computer vision, which develops from physiological understandings of the human visual system. The connections and intersections between human vision experimental psychology and machine vision will be elaborated upon, providing a foundation for the research conducted in this thesis. Further connections between human physiology and machine vision will be discussed in detail in the essay on physiology.

In the final subsection on vision, we will examine representations of vision specifically in the study of eye movements and computer vision. The representation of vision is always contested territory in terms to claims of reality in the arts and claims of objectivity or truth in the sciences.

The Human Eye

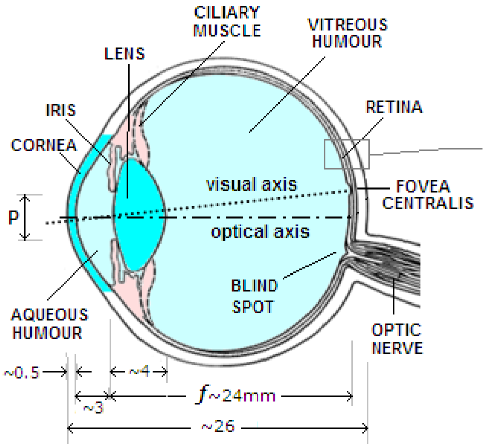


Figure 3: Human eye cross section. Reprinted from: “12. The Telescopic Eye “ www.telescope-optics.net/eye.htm

In this section we will provide an overview of the optical functions of the human eye, the composition and function of the sensory apparatus, and the performance and constraints these systems put on the motor apparatus that moves the eye.

Optics and Sensors of the Eye

The human eye uses a two lens system in a fluid called the vitreous humour to project incoming rays of light from the world onto the retinal surface that is located at the back of the eye.

The first optical element is the Cornea, to be precise, a thin layer of tear fluid that covers the curved corneal surface. Once the rays pass through the cornea they continue onwards to the pupil. The pupil acts like an aperture in a camera. One function of the aperture is to control the amount of light that passes into the lens, or camera. This aperture can change in size, growing larger – dilation in low light conditions to allow more light, or smaller – constricting – to allow less light. The rays of light then pass through the eye lens. The eye lens can change in shape, or deform in order to focus on light coming from different depths. Once past the lens, the rays of light travel through the vitreous humour and are cast on the retinal surface. The retinal surface is covered with photoreceptor cells. In the camera analogy this surface would be equivalent to a CMOS sensor or photographic emulsion.

The retinal surface contains two types of photo receptors, rods and cones. There are twenty million rods and six to seven million cones. The rods are more sensitive than the cones, but are not sensitive to color. The cones are responsible for color sensitivity in the retina.

On the retinal surface there is an area that

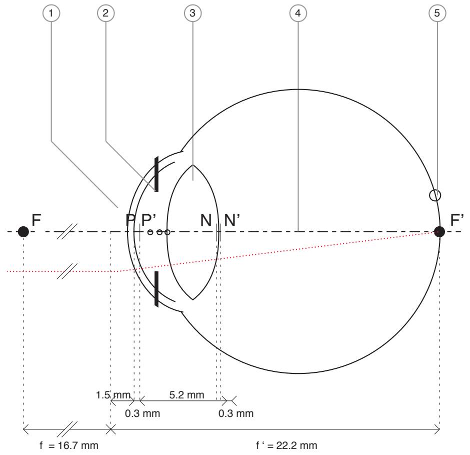


Figure 4: Human eye as optical system. Redrawn by authors based on original from: “12. The Telescopic Eye “ http://www.telescope-optics.net/eye_aberrations.htm

is densely packed with cones in proportion to rods. This area is called the fovea, and appears as a small yellow spot on the retinal surface. The fovea is measured in angular diameter of between 0.3 degrees and 2 degrees, or 1/4000th of the retinal surface area (Steinman 2003). The remainder of the retina is not blind, as the distance increases from the fovea the density of cones and optical acuity is greatly reduced. At twenty degrees from the fovea, visual acuity is down to ten percent. The fovea is densely packed with cones, approximately 161,900 per square millimeter, allowing for high resolution color vision. The surrounding area is populated by rods, densely packed around the fovea. The rods decrease in density relative to distance from the fovea. The physiology of the retinal surface shows us that here is only a small portion of our visual field that we can resolve in high resolution.

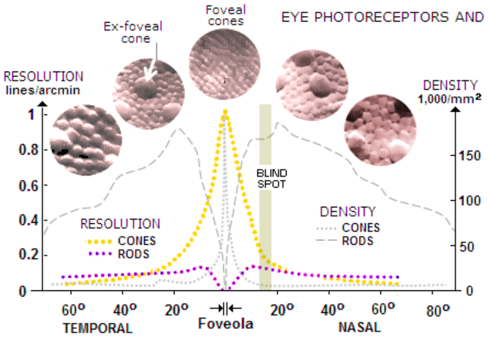


Figure 5: Graph of the eye photo receptors and acuity. Reprinted from: “12. The Telescopic Eye “ http://www.telescope-optics.net/eye_aberrations.htm

Field of View

The human monocular field of vision, without eye movement, is one hundred and sixty degrees in width and one hundred and seventy five degrees in height. While this may seem like an incredibly large field of vision, this field of vision is not all sharp and in color. The visual angle foveal vision is only approximately one degree. As a practical example, the area that a human eye resolves in color and high resolution is approximately equivalent to the area of one’s thumbnail held out at arms length. Due to this limitation we have to rotate our eyes in their sockets, positioning the eye such that the area of interest in the world is projected onto the fovea. This movement called a *saccade* and is usually very fast, up to seven hundred degrees per second and typically lasting for thirty milliseconds. The saccade is a major motif of visual movements.

Not only is the retinal area of high acuity small, but the cones are also relatively slow. The step response time for cones is approximately twenty milliseconds. For an image to fully resolve with all its high frequency details its projection on the fovea need to be motionless. Any movement faster than three degrees per second would result in a blurry image. During fast movements like a saccade the image would be blurred. However, the visual system is blocked during saccades and therefore the brain does not receive this useless information. During a saccade we are effectively blind. This can be easily demonstrated by attempting to look in a mirror and observe your own eyes moving.

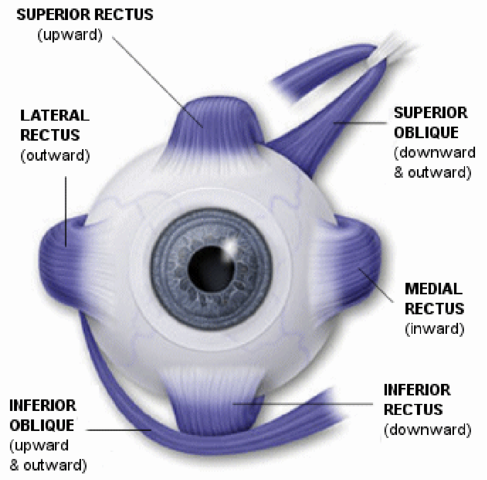


Figure 6: Human eye musculature. Image taken from Wikipedia.

Eye Musculature and Stabilization

For successful resolution of a detail, one's gaze needs to rest on the area of interest, this is called fixation another motif of eye movements. Fixations typically last for three hundred milliseconds. During a fixation two powerful image stabilization mechanisms keep the projection in place. Even if one moves their head while fixating on an object, or the object of interest moves relative to the viewer, the image stays in place.

This stabilization is achieved by the vestibulo-ocular reflex (VOR) and the optiokinetic reflex (OKR). Both reflexes actuate a three pairs of muscles that rotate the eye around its three axis: the lateral rectus, the medial rectus, the inferior rectus, the superior rectus, the inferior oblique, and the superior oblique. When the muscles contract and their counterpart relaxes accordingly, the resulting torque moves the eye in almost pure rotation.

The eye can be rotated voluntarily to yaw and shift, allowing any part of the field of view to be projected onto the fovea. Movement around the optical axis, rolling, can not be triggered voluntarily and is not very noticeable due to the rotational symmetry of the eye, but it is nonetheless frequently used by the compensatory reflexes.

VOR is a reflex that compensates for rotation and translation of head movements. These movements are sensed by the vestibular apparatus in the inner ear. Here the inertia of a fluid is sensed by small hairs in the inner ear. Rotational and translational change is then passed on to a short and very fast neuron network called the three neuron arc, that stimulates the eye muscles to compensate for the movement.

OKR is triggered by the assessment of change in the foveal image. As the projected image starts to drift the angle and velocity of this change is measured and the eye motor muscles are stimulated to compensate for this change. The processes used to evaluate angle and magnitude of change is analogous to a process in computer vision called optical flow. The analogs between human vision and computer vision will be discussed elsewhere in this thesis.

The three main motifs of eye movements are fixation, saccades and pursuit. Saccades are a fast repositioning of the eye that alternate with fixations, a pause between saccades during which the projection on the fovea is processed. For moving targets OKR is utilized to keep the target projection in place, this smooth motion of the eye is called pursuit.

Looking at smaller temporal and spatial scales, other motifs emerge: drift, tremor and micro-saccades.

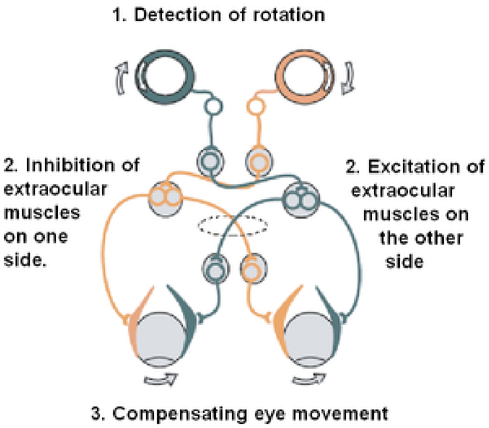

1. Detection of rotation
2. Inhibition of extraocular muscles on one side.
2. Excitation of extraocular muscles on the other side
3. Compensating eye movement

Figure 7: Functional diagram of the vestibulo-ocular reflex (VOR). Image taken from Wikipedia.

Direction of Visual Attention

Beyond these two reflexes, VOR and OKR, the visual system stimulates the eye muscles to rotate in order to inspect areas of interest. The visual cognitive pathways that control and decide what area in the field of vision to evaluate in greater detail is a subject of ongoing inquiry. Current research suggest that two schemes can be used the describe this behavior. The first is typically called "bottom up control," where salient features determine what is to be attended to. The second is called "top down control," where attention is focused by a predetermined cognitive task.

In the bottom up control schemes, the area in our field of vision that is resolved in foveal vision is not only uncontrollable by conscious thought, but it is also completely opaque to it. Salient features of a scene prompt our visual system to saccade the gaze point to this area of interest and fixate on it.

An example for this is a sudden movement in the peripheral vision or a contrasting detail like a blossom on a background of green leaves.

Top down control is controlled by high level motivations, like a verbal cue that lets our visual system shift foveal vision towards areas that we deem as potentially informative for the given visual task. This could for example be a search task: "find the red bottle", or a motor task like opening a drawer.

This control scheme is very interesting, but at the same time opaque to us as it implicates the combined motivations of the viewer, their prior experience, training, and a myriad of other influences.

While neither the neurological implementation nor more detailed rules of interaction are understood, it appears that both schemes are weighted into the decision what to fixate upon next.

Is the Point of Fixation Equal to the Point of Attention?

The community has agreed upon the assumption that the area that we fixate upon is the area we are visually attending to and chosen to look at by either visual saliency (bottom up), potential usefulness for a task (top down) or both. However it should be noted that the converse may not always be true. Our peripheral vision may resolve features sufficiently enough, that our visual system does not need to direct the visual center on them.

However, the implementation of our software and visualization only concerns itself with the point of visual attention as with our hardware and software we have not developed means of observing states of visual cognitive activity, other than pupil position.

Vision as information Processing

What should become apparent from this overview is that the physiological realities of seeing are by no means in accord with how we think we are perceiving the world. While our visual world appears to us as always focused and complete, the physiological reality is constructed from fragments of a resolved field. The area we are resolving at full resolution at any given time is surprisingly small. Very sophisticated machinery has evolved to efficiently use the limited visual cognitive capacity and extract the contextually important visual information from a the world around us. Seeing should not be understood as full resolution of a scene by means of capturing images in a still or video camera, but as an act of information processing. This act is guided by a goal or motivation, is highly selective, and starts at the very beginning of the visual system, the eye.

While the exact mechanisms of vision are still unknown we believe that the choices this machinery makes, in the fixation points of a human subject moving through space, allow for low level introspection. This introspection can include but is not limited to the of the visual perception of a space, the motive of the subject, the influence of subjects background and training on his perception.

Furthermore it can potentially give answers to the impact of a space and spacial design decisions on the human subject. This latter category might allow us to use the insights gained from tracking visual attention is space not only as a mere analytical tool but include it in a design and evaluation feedback loop, ultimately making it an instrument that allows for more appropriate spatial design.

History of Eye Movement Studies

Part 1: "The Best Possible Reader"

Try reading this silently.

The two men were seated at a table upon which many books and papers were scattered. The older man turned to a page in a large book and began to read. The subject of the chapter was something about hypnagogic hallucinations and hyperaesthesia. A few pages further on he came to a sentence which read, "One thing, however, is obvious, namely that the manner in which we now become acquainted with complex objects need not in the least resemble the manner in which the original elements of our consciousness grew up."[5]

This paragraph was composed by Guy T. Buswell as a selection for "advanced readers" in his experimental study of reading conducted as a PhD student at the University of Chicago between 1917 and 1920.[6] The ultimate goal of Buswell's research was to produce the best possible reader.[7]

In order to take steps toward achieving this goal, Buswell followed the framework of applied psychology established by Hugo Münsterberg in his book, *Psychology and Industrial Efficiency*. The idealized relationship between applied psychology and society can be generalized in a three step process with a feedback loop. First, societal problems, economic or pedagogical, are isolated and brought forward to the psychologist in the laboratory. Second, the experimental psychologist in the laboratory extricates the problem from societal constraints, creating an isolated and abstracted object of inquiry. The abstract object of inquiry is studied in relation to an individual subject. Third, results from the laboratory are applied in the field as "psychotechnics." The application of "psychotechnics" can then feed back into the theoretical knowledge of psychology and potentially reinitiate the process from new societal concerns.[8]

How does reading become an object of inquiry in the psychology laboratory? Within the framework of applied psychology, reading enters the laboratory from society as industrial, economic, and pedagogical concern. The importance of reading can be directly connected to a historical narrative of industrialization. Lewis Mumford situated the printing press as one of the most influential machines, "second only to the clock in order if not perhaps in importance..."[9] The mechanical clock transformed time into an abstract

[5] Guy T. Buswell, *An Experimental Study of the Eye-Voice Span in Reading* (Chicago: University of Chicago Press, 1920), 8; William James, *The Principles of Psychology: Volume Two* (New York, Henry Holt and Company, 1890), 630; Buswell selects the last sentence in his reading sample for "advanced readers" (marked above as in Buswell's text sample with double quotation marks) from William James's monumental text. However, italics that begin at "...the manner," in James's original text are omitted in Buswell's citation. This citation is taken from the last chapter of James's second volume in the section titled, "Necessary Truths And The Effects of Experience" and within the further subheading, "The Genesis of the Elementary Mental Categories." Within context of this chapter the quotation can be read as a critique of associationist thinking, where a complex object or experience is based on a combination of prior experiences. The associationist concept of experience clashes with James's spiritual and holistic understanding of the mind. This can be captured best in another quote from the same chapter, "The way of 'experience' proper is the front door, the door of the five senses. The agents which affect the brain in this way immediately become the mind's objects. The other agents do not. It would be simply silly to say of two men with perhaps equal effective skill in drawing, one an untaught natural genius, the other a mere obstinate plodder in the studio, that both alike owe their skill to their 'experience.' The reasons of their several skills lie in wholly disparate natural cycles of causation." (William James, *The Principles*, 628)

[6] Buswell, *An Experimental Study;* The dissertation was published by the University of Chicago Press in conjunction with two educational journals: *The School Review* and *The Elementary School Journal.*

[7] Hugo Münsterberg, *Psychology and Industrial Efficiency* (Boston: Houghton Mifflin Company, 1913), chaps. 4-12. My use of "the best possible reader" is a direct reference to Münsterberg's title for the first section of his book, "The Best Possible Man."

[8] Münsterberg, *Psychology and Industrial Efficiency*; "Only slowly did the pedagogical problems themselves begin to determine the experimental investigation. The methods of laboratory psychology were applied for the solving of those problems which originated in the school experience, and only when this point was reached could a truly experimental pedagogy be built on a psychological foundation. We stand in the midst of this vigorous and healthy movement, which has had a stimulating effect on theoretical psychology itself." (Münsterberg, *Psychology and Industrial Efficiency*, 12); Also, the definition of psychotechnics is the application of experimental psychology back to society: "The task of psychotechnics is accordingly to determine by exact psychological experiments how this mental effect, the satisfaction of economic desires, can be secured in the quickest, in the easiest, in the safest, in the most enduring , and in the most satisfactory way." (Münsterberg, *Psychology and Industrial Efficiency*, 242).

[9] Lewis Mumford, *Technics and Civilization* (Chicago: University of Chicago Press, 2010 [1934]), 134.

object, that was no longer connected to biological or organic rhythms, where time "...could be divided, it could be filled up, it could even be expanded by the invention of labor saving instruments."[10] For Mumford, the printing press was a quintessential example of a labor saving instrument that merged with mechanized time-keeping in the material product of the periodical publication.[11]

However, increased production volume and mobility of inscriptions was not always labor *saving* for the individual psyche.

In the late nineteenth and early twentieth century reading was held accountable for disastrous physiological and psychological affects in the form of myopia and fatigue.[12] In the service of a capitalist society, the applied psychologist was employed to provide a training regime that could standardize the production of the best possible reader. Alternatively, the applied psychologist could be tasked to fit the best possible reader with the best possible work. An exercise in vocational fitness. While applied psychology produced a program that attempted to reintegrate experimental psychology into society, the importance of reading in psychology had its roots in experimental psychology.

In attempts to become recognized as an exact science, psychology developed a new program of experimental psychology. Today, the new program of experimental psychology is attributed to Wilhelm Wundt who developed the first psychology laboratory, *de jure,* at the University of Leipzig in 1879.[13] The *new* science of Experimental psychology marked a definitive shift in the relationship between observer and object, where the object observed was independent of the observer.[14] Hugo Münsterberg traced parallel lines of development between early

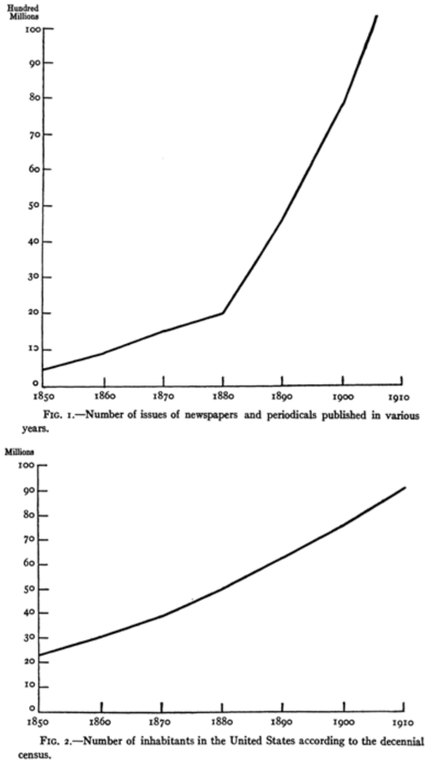


Fig. 1.—Number of issues of newspapers and periodicals published in various years.

Fig. 2.—Number of inhabitants in the United States according to the decennial census.

Figure 8: Graph of publications between the years 1850-1910, and graph of inhabitants in the United States for the same period. Judd uses these graphs to demonstrate the overwhelming volume of reading relative to population and to make an argument for the radical reform of reading pedagogy in elementary education. Also of note are the graphs themselves. The graph for publications on the left is dramatically stretched in the vertical axis within the original publication. Judd, while a professor at Yale, writes about the importance of graphic representation and scale, "In general it will be found advantageous to bring together all the results of an investigation whenever possible in a graphic representation... It is sometimes advantageous to emphasize one characteristic or the other of a given curve."[9] Reproduced from Charles H. Judd, "Relation of School Expansion to Reading," The Elementary School Journal 23 (4 December, 1922), 253-266 on 254, 255.

[10] Mumford, *Technics and Civilization, 17.*

[11] Mumford, *Technics and Civilization, 136.*

[12] Edmund Huey, "On the Psychology and Physiology of Reading: I," *The American Journal of Psychology* **11** (3 April, 1900), 283-302 on 283; "Some general account of what we do in reading seems to be much needed in view of the fact that reading is one of the most frequently performed psycho-physiological operation, and is fatiguing, often disastrously so. What are the conditions of this fatigue?… Moreover, it has been believed, though less strongly at present than a few years ago, that the reading in schools is the cause of the tremendous progress of myopia."

[13] On "Early Psychology Labs: Founding of Laboratories of Psychology (From 1875-1890)," see http://www3.niu.edu/acad/psych/Millis/wundtslab/epl.htm (accessed 15 December 2011).

experimental psychology and the natural sciences, where he claimed that the aims of the discipline were to derive general or constant laws based on the exact observation of individuals.[15] No longer would psychology rely on self observation. Instead, "experimental study was possible only when external manipulation of conditions was possible–that is, it was restricted to relations between stimulus and consciousness in the simplest sense."[16]

For the experimental psychologist, reading formed an essential link between the seemingly unconscious responses of the eye to a stimulus–text on a page–and the conscious activities of the mind–measured in verbal feedback. The physiological realities of the moving eye in connection with verbal reading aloud, or interpretation of text, produced a connection between artifact, subject, and psychologist. In order to assert a position as an exact science, this interpretive relationship required exact methods and instrumentation to measure the eye, voice, and text. The text artifact as stimulus, movements of the eye as physiological response, and verbal feedback as a window into conscious activity needed to be correlated.

The first half of the solution was found in the clock, "... for the clock is not merely a means of keeping track of the hours, but of synchronizing the actions of men."[17] It was in the precise timing of stimulus and response that an individual's actions could be measured and tabulated. By correlating the times in a table, individual responses could be compared, sorted, and evaluated. The second half of the solution was found in photographic medium and apparatus.

A novel instrument to measure the eye movements during reading was developed by an

[14] George Mandler, *A History of Modern Experimental Psychology: from James and Wundt to Cognitive Science,* (Cambridge: MA, MIT Press, 2007) 60: "It was Wundt in 1874 who marked out the 'new domain of science' and who made the break with self-observation by insisting that 'all accurate observation implies… that the observed object is independent of the observer.'"

[15] Münsterberg, *Psychology and Industrial Efficiency,* 5: "Their aim was no longer to speculate about the soul, but to find the psychical elements and the constant laws which control their connections. Psychology became experimental and physiological."

[16] Mandler, *History of Modern Experimental Psychology*, 60.

[17] Mumford, *Technics and Civilization, 14.*

American named Raymond R. Dodge while studying philosophy at the University of Halle in Germany in 1894.[18] Even though Dodge was determined to become a philosopher, he enrolled in a seminar on the psychology of reading taught by Professor Benno Erdmann. During this seminar Professor Erdmann discussed the philosophical need for a special apparatus to study reading. Erdmann was able to describe the apparatus in detail but postulated that this ideal apparatus could not be built. The "ideal tachistoscope" would be able to expose words to both eyes simultaneously.[19] Dodge took his professor's description and skepticism as a personal challenge. His first year as a graduate student at Halle was devoted to the construction of the tachistoscope. The project sponsored a close working relationship with Professor Erdmann and a shift in Dodge's disciplinary interests; from philosophy to psychology, resulting in a dissertation on the kinesthetics, or motor functions, in connection with verbal imagery.[20] After defending his dissertation, Dodge remained in Germany for a year to complete the publication of his research conducted with Erdmann. The Erdmann-Dodge tachistoscope enabled unprecedented precision in the measurement and correlation of a subject's psycho-physiological response time to varying projected verbal imagery.[21]

A diagram of the Erdmann-Dodge tachistoscope portrays a laboratory workbench with projection stimulus instruments, and a timing devices affixed to the surface of a workbench along a central axis. In this representation the subject would be seated at the left end of the lab table. The subject, once seated, would bite down on the bar in order to simultaneously restrict movement of the head and initiate the projection of words or word fragments

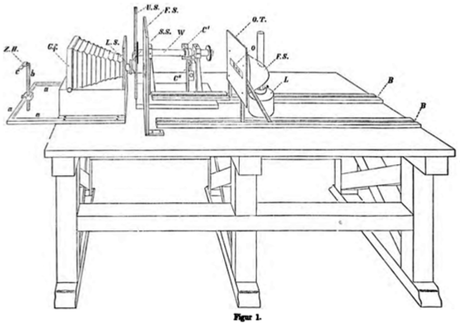


Figure 9: Erdmann-Dodge Tachistoscope, a schematic representation. The subject would be seated at the left with head stabilized by the "Helmholtzer Zahnhalter" (Helmholtz bite bar marked Z.H. above figure). The subject would be presented with "verbal imagery" (words and characters marked as 'o' the plate marked 'G.T.') projected onto ground glass surface marked (G.f.) for incredibly precise durations. Short exposure from with a maximum of 0.00025" of precision. Reproduced from Benno Erdmann, Raymond Dodge, Psychologoische Untersuchungen Über Das Lesen auf Experimenteller Grundlage (Halle: Max Niemeyer, 1898), 99.

[18] Walter R. Miles, *Raymond Dodge 1871-1942: A Biographical Memoir* (New York: Columbia University Press for the National Academy of Sciences, 1956), 70: The story of Dodge enrolling at University of Halle resulted from his rejection from Harvard and Columbia. His choice to study in Germany was based on a conviction that he would become a philosopher, and in order to become a philosopher, he would have to master the German language. Dodge's professor at Williams College (Professor Russell) gave him a copy of Kant's *Kritik des reinen Vernunfts*, which was edited by Professor Benno Erdmann. His biographer speculates, "It is hard to believe that Dodge had no other information about Erdmann than the fact that he edited Kant's *Kritik*. Perhaps Professor Russell was behind the choice in other ways." (70)

[19] Miles, *Raymond R. Dodge*, 70-71: The professor in a seminar on the psychology of reading discussed the need for a special piece of apparatus which could serve to exhibit a word or diagram all at once an in clear view for binocular reading or perception. The desirable features of this ideal tachistoscope the professor could enumerate but he could not picture what the apparatus would look like, and he expressed to his seminar the opinion that it would not be possible to build such a piece of equipment."

[20] Miles, *Raymond R. Dodge*, 73-74: It seems that Dodge did not willingly shift to psychology, but was discouraged in his study of philosophy by professor Erdmann, based on his lack of mastery of the German language; Raymond R. Dodge, *Die motorischen Wortvorstellungen* (Halle: M. Niemeyer, 1896): The title of Dodge's thesis translates literally as "The Motor Verbal Images." A more nuanced understanding of the title allows one to understand the importance of eye movements, where verbal imagery is kinaesthetic/motor.

[21] Examination of the tables in Erdman, Dodge, *Psychologische Untersuchugen Über das Lesen,* reveals Erdmann and Dodge as primary subjects in their own experimental trials. According to Miles, and Münsterberg this is a typical practice of early experimental psychology, where few individuals (even the psychologists themselves) could be used as subjects from which general laws could be proposed.

onto the screen. These devices used analog electronic circuits to control the exposure times of characters on the screen, producing an early cinematographic experience. This instrument enabled Erdmann and Dodge to not only to measure the speed of human visual response times through verbal feedback but also to observe and quantify the physiological behavior of the human eye, enabling new claims about saccadic eye movement.[22]

After his studies and post doctorate research in Germany, Dodge returned to the United States and continued his research at Wesleyan University. Facing critique of his experimental results from forefathers in the field of experimental psychology – specifically Wilhelm Wundt – Dodge set out to defend his research by providing increasingly precise and rigorous experimental proof. With this mission, Dodge along with his students, constructed new instruments and built up a laboratory at Wesleyan University.

In 1901 Dodge and his student Thomas Cline published a paper that introduced the first non invasive eye tracking instrument. The aptly named, Dodge Cline "photochronograph," is able to capture eye movements of a subject by shining a light on the subject's eye. The light that reflects off of the cornea surface is reflected into the camera lens and captured onto a falling plate. This instrument used a falling photographic plate, whose rate of fall was "governed by the escape of air from a cylinder into which the falling plate presses a closely fitting piston. The cylinder and piston are an ordinary bicycle pump."[23] The precise time of the falling plate was correlated with a pendulum that began swinging as the plate dropped and periodically obstructed the falling plate camera aperture.

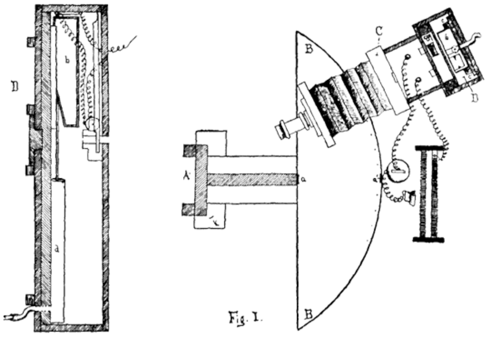


Figure 10: The Dodge-Cline Photochronograph plan diagram (left) with detail of the photographic plate holder (far left). Reproduced From Raymond Dodge, Thomas Sparks Cline, "The Angle Velocity of Eye Movements," Psychological Review 8(2) (March 1901), 145-157 on 150, 152.

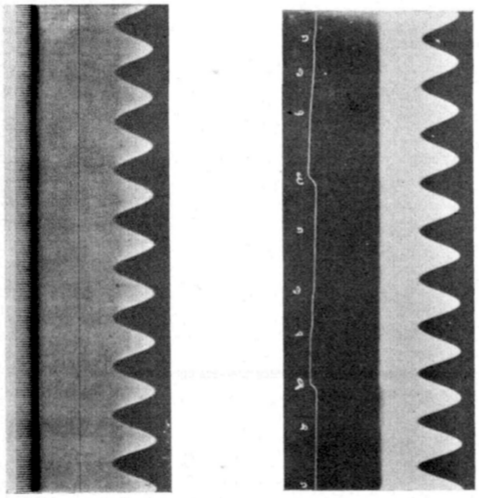
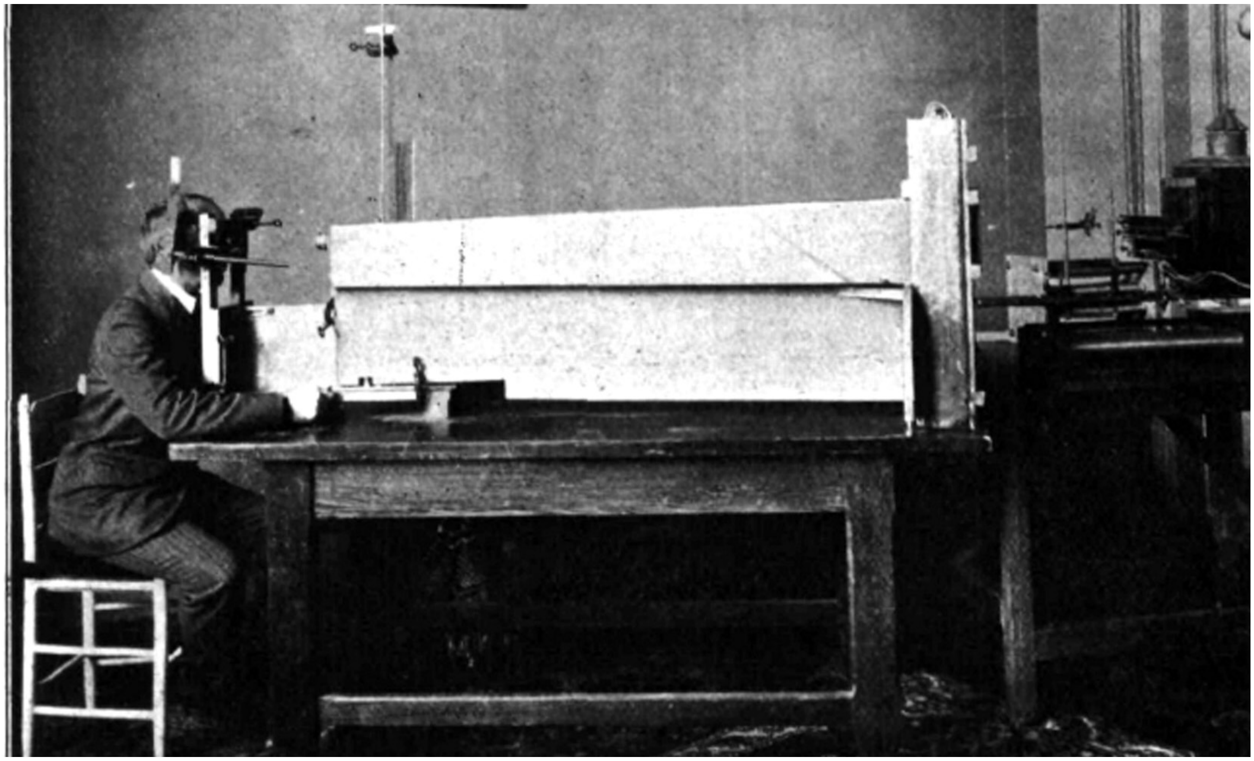


Figure 11: Record of eye movements (far right). The eye movements of a subject are seen as the labeled white line in the right of figure 2.2. The regular waveform represents the time interval. Reproduced From Raymond Dodge, Thomas Sparks Cline, "The Angle Velocity of Eye Movements," Psychological Review 8(2) (March 1901), 145-157 on 150, 152.

[22] Benno Erdmann, Raymond Dodge, *Psychologoische Untersuchungen Über Das Lesen auf Experimenteller Grundlage* (Halle: Max Niemeyer, 1898), Walter R. Miles, *Raymond Dodge 1871-1942: A Biographical Memoir* (New York: Columbia University Press for the National Academy of Sciences, 1956), 82. The Erdmann and Dodge publication was criticized by Wundt

[23] Raymond Dodge, Thomas Sparks Cline, "The Angle Velocity of Eye Movements," Psychological Review 8(2) (March 1901)



The constant time of the pendulum can be seen as the waveform in the photographic record. The fine black lines on the left extreme of the record are produced from the vibrations of a tuning fork periodically obscure another aperture. Together the tuning fork and the pendulum provide major and minor intervals of time that are directly correlated with the movements of the eye on the same photographic negative.

Timing of stimulus and response in the study of eye movements during reading sponsored a great variety new instruments.[24] Many researchers sought to attach cups directly to the corneal surface of the eye and used mechanical levers as a stylus to inscribe movements on smoked paper using a kymograph. In the mechanical process, the eye would have to be anesthetized, by the use of

[24] Nicholas Wade, Benjamin Tatler, *The Moving Tablet of the Eye: The Origins of Modern Eye Movement Research* (New York: Oxford University Press, 2005); Nicholas Wade, "Pioneers of Eye Movement Research," *i-Perception* **1** (5 November 2010), 33-68. Timing of stimulus and response was not only limited to eye movement studies. In fact, many timing instruments were developed in early experimental psychology laboratories to facilitate experiments involving timed exposure to visual stimuli. The key difference between eye movement studies and other timed visual exposure apparatus was in the recording of the physiological response in conjunction with 'conscious' activity.

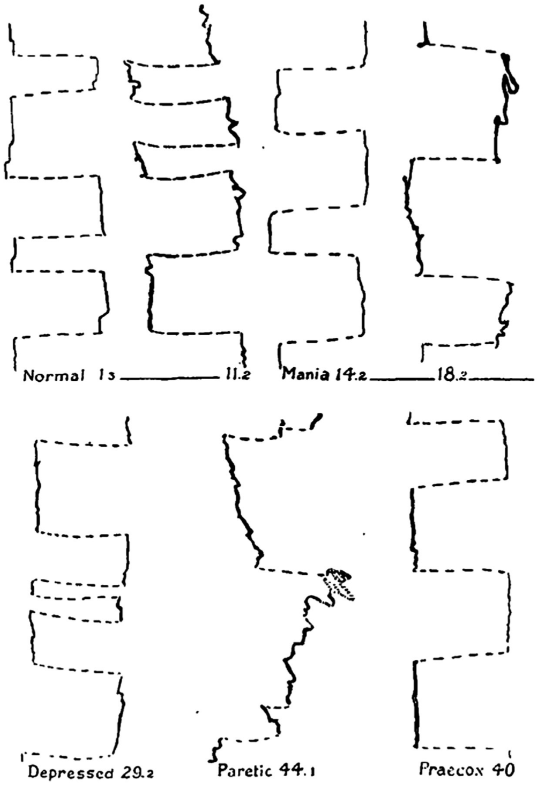


Figure 12: [Facing Page] Diefendorf-Dodge photomicrograph (1908) developed from the Dodge-Cline photochronograph. "The recording apparatus is the Dodge photo chronograph fitted with an enlarging camera, the Bausch and Lomb convertible protar, and the Dodge–Cline plate holder." Reproduced From Diefendorf and Dodge, "An Experimental Study of the Ocular Reactions of the Insane from Photographic Records," Brain 31, 451-489.

Figure 13: Sample records from a Diefendorf-Dodge photomicrograph showing eye movement records of various subjects. "Reproduction of typical records of eye-movements. The records were projected by lantern and drawn from the projected image on a much larger scale. The resulting lines reproduce the original records very well, save that ... the exact shape of each dash is not accurately reproduced. The photographic plate was moving so slowly that the dots run together in the vertical line appearing as dashes only during eye movement. The dashes represent flashes of light succeeding each other every 0.01 s. The paretic line, No. 44.1, is an extreme case of head movement and broken lines. The broken movements are typical, the head movements less so. (The velocity of the movement is given by the number of dashes which represent flashes of light succeeding each other every 0.01 s.) " Reproduced From Diefendorf and Dodge, "An Experimental Study of the Ocular Reactions of the Insane from Photographic Records," Brain 31, 451-489.

cocaine or holocaine, while a lever transferred and amplified the spatial movements of the reading eye.[25] However, with the introduction of the Dodge Cline photochronograph the, often painful, mechanical process and metaphor for timing, observation, and capture of eye movements eventually evolved as an inferior alternative. It is important to note here that the "objectivity" of photographic instrumentation and representation of scientific observations fall are aligned with the scientific zeitgeist of the period – mechanical objectivity as truth. A detailed discussion of this scientific ethos will be discussed in regards to representation later in this thesis.

As the photographic process for recording eye movements developed, the mechanical process was eventually edged out of the newly forming discipline of "exact" experimental psychology.[26]

[25] Edmund Huey, "On the Psychology and Physiology of Reading," 288: "The eye was rendered anaesthetic by the use of cocaine or holocaine. The latter was found most satisfactory and was used in most experiments. The cocaine usually interfered with the accommodation, the holocaine probably never did so. The eye felt fairly comfortable during the experiment, and the reading proceeded normally."

[26] Earlier physiologies of the eye and optics relied on mechanical metaphors. For a discussion of the history of vision see: Wade, Tatler, *The Moving Tablet of the Eye*, chap. 2; Nicholas Wade, *A Natural History of Vision* (Cambridge: MIT Press, 2000); Jonathan Crary, *Techniques of the Observer: On Vision and Modernity in the Nineteenth Century* (Cambridge: MIT Press, 1990).

Experimental psychology found its roots in exact science through physiology.[27] Through the development of new physiological instruments, new observations could be made and theories developed through experimentation.[28] As a recording device, photography and photographic media served as a homeomorphic transformation for the physiological function of the retinal surface. In regards to time, what could be more precise than the speed of light as a constant? It was on strips of photographic emulsion and in the electronic modulation of the photographic aperture that a subject's eye movements could become a quantifiable matter. Delineated space, as printed letters on a page, could be correlated with the measurable movements of the eye.[29] With these new instruments and reading entered the world of experimental psychology as a psycho-physiological object, could be studied as an abstract entity, and tabulated in the laboratory.

[27] Hugo Münsterberg, *Psychology and Industrial Efficiency*; James McKeen Cattell, "Homo Scientificus Americanus," *Science* **17** (10 April, 1903), 561-570 on 564: In Cattell's diagram of the disciplines of science, psychology as a discipline is connected directly to Physiology and Anthropology. In Münsterberg's writing experimental psychology is rooted in physiology, "Psychology became experimental and physiological." (Münsterberg, *Psychology and Industrial Efficiency,* 5). It would be interesting to compare Münsterberg and Cattell's disciplinary diagrams and relationships between science and society. Cattell's diagram is topographical, where theoretical sciences are situated on one "plenum" where affinities and connections can develop, while applied sciences are on another "plenum" situated atop the theoretical sciences. On the surface, science is applied science and acts as a buffer (or derrière garde) for the (avant garde) of experimental or theoretical science.

[28] Peter Galison,, *Image and Logic: A Material Culture of Microphysics* (Chicago: University Of Chicago Press, 1997), Chapter 9, "The Trading Zone: Coordinating Action and Belief". Of interest to this essay is Galison's argument for a more flexible relationship between instrumentation, observation, and theory in opposition to traditional views of positivists and anti-positivists. In Galison's argument, he demonstrates that a new instrument can lead to a revision of theory just as a new observation may lead to a revision in instrument and visa versa.

[29] There were intense debates about the merits of different methods and instrumentation. Raymond Dodge, as a professor at Wesleyan University, was a staunch advocate for the photographic method in eye movement studies–often arguing against invasive procedures for experimental studies of reading. Charles H. Judd–a first generation Wundtian and Buswell's professor at the University of Chicago–at first opposed Dodge's methods of coronal reflection, and opted to place "China White" directly on the cornea of a subject. However, he too eventually adopted Dodge's method, as evidence of his student's work at the University of Chicago in both Gray and Buswell. For a discussion of eye movement instrument development see: Wade & Tatler, *Moving Tablet of the Eye,* chap. 4.5.

[30]Guy T. Buswell, *An Experimental Study.* The exact instructions that Buswell gave to the subjects of the experiment were, "... read the paragraph naturally, just as you would a newspaper. If you meet any new or difficult words, pronounce them the best you can and go on. Try to remember the thought well enough so you could tell what you have read if asked to do so." (Guy Buswell, *Eye-Voice Span*, 10)

Try reading this aloud.[30]

The two men were seated at a table upon which many books and papers were scattered. The older man turned to a page in a large book and began to read. The subject of the chapter was something about hypnagogic hallucinations and hyperaesthesia. A few pages further on he came to a sentence which read, “One thing, however, is obvious, namely that the manner in which we now become acquainted with complex objects need not in the least resemble the manner in which the original elements of our consciousness grew up.”

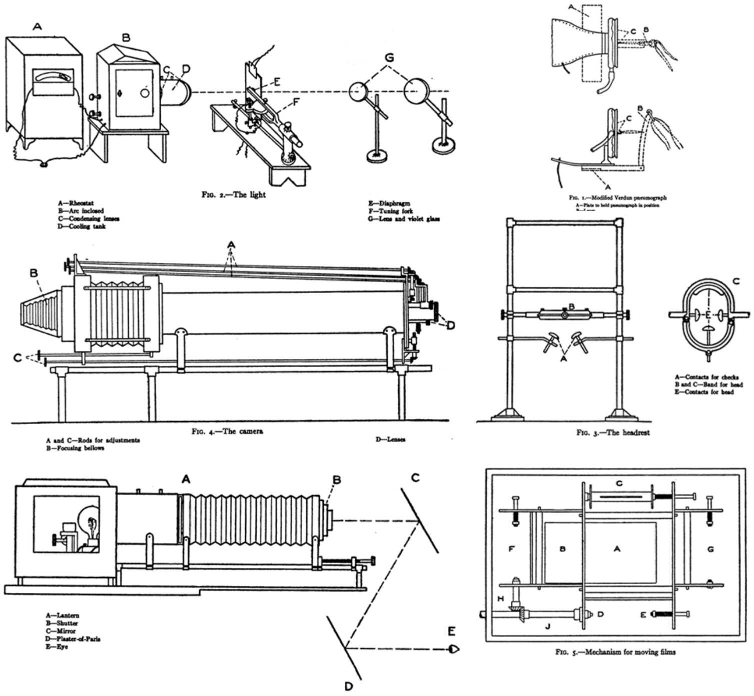


Figure 14: Clarence T. Gray's photochronographic and pneumograph for measuring the eye and the voice for subjects reading aloud. Clockwise from top left: the light uses electric pulses to vibrate a tuning fork which periodically obscures the aperture from the arc lamp. This light is reflected into the subject's cornea; The pneumograph to measure the changes in a subject's breathing pattern or–"breath curve"–from changes in abdominal or breast expansion. The breathing pattern is inscribed on a kymograph; The headrest constrains the movements of the subject's head during reading; The mechanism for moving films is a rotating spool that moves film past the camera aperture -- similar to a cinema camera; The apparatus for exposing reading material uses an lantern or arc light to project passages of text onto the surface 'D.' The shutter of the projector can be precisely controlled in order to limit projection time; The camera is a normal bellows camera connected to a four foot long brass tube that acts like a telescope. The rods are used for adjustment of the camera's height and orientation as well as focus plane. Reproduced From Clarence T. Gray, Types of Reading Ability as Exhibited Through Tests and Laboratory Experiments (Chicago: University of Chicago Press for Supplementary Educational Monographs, 1917), 71, 84, 85, 87, 88, 89.

In 1917, fresh off duty from wartime service in the Signal Corps, Guy Buswell returned to the University of Chicago to complete his dissertation.[31] At the University of Chicago School of Education, Buswell worked closely with Professor Charles H. Judd, the second American to receive a PhD in psychology from Wundt's laboratory in Leipzig.[32] Guy Buswell and Clarance T. Gray, both under Judd's academic tutelage, extended Dodge's photochronograph technically by simultaneously recording the speaking voice with the movements of the eye, leading to new experiments and eventually theories.

Prior to Buswell and Gray's developments, experimental psychologists relied on verbal feedback from the subjects in interviews and written tests as methods to access the mental aspects of reading.

[31] John M. Robertson, "Guy Thomas Buswell (1891-1994)," *The American Psychologist* **51** (2 February 1996), 152: Robertson claims that Buswell's wartime experience in the Signal Corps laid the foundations for his research in reading and education. For example, "During World War I, he was assigned to the signal corps training detachment housed at Kansas State Agricultural College (now Kansas State University). Learning how to communicate with symbols hinted at what would become a major contribution in his academic career-- how the eye moves while interpreting incoming information." However, the use of the term "information" in Robertson's obituary seems to be a much more modern word–taken from information theory and early cybernetics–being applied retrospectively to Buswell's work. For Buswell the correct word would likely have been visual stimulus, objects, or maybe signals.

[32] Arthur R. Jensen, Robert B. Ruddell, "Guy Thomas Buswell, Education: Berkley," *University of California. In Memoriam, 1994* (Berkley: University of California (System) Academic Senate, 1994), 47-49.

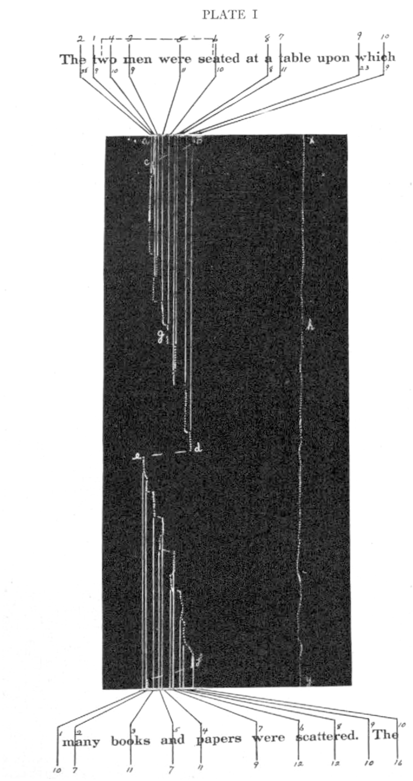


Figure 15: An example photographic record of a reader's eye movements as white dotted lines (in step pattern) within in the black area of the photographic negative (on the left top/bottom) juxtaposed with text that the reader was reading. The eye movements are connected to the text by long solid lines. Each dot represents 1/50th of a second. On the right of the photographic negative is the movement of the head, reflected from a silver bead on the head mount. Reproduced from Buswell, An Experimental Study, 5.

Access to the meaning, or semantics, of the printed word was used as a way to test comprehension and to correlate a physiological stimulus time into an abstract unit of meaning and comprehension. Buswell and Gray still employed prior methods in written comprehension tests and follow up interviews, but extended their technique with new instrumentation. They were armed with a new form of physiological and mechanical objectivity in the speaking voice of the subject recorded by phonograph in combination with falling plate records of eye movements.[33] Movements of the eye and voice could be captured as finite movements in the surface of both photographic and phonographic medium. Points of light reflected from the cornea in photographic medium, jumping curves of vocal amplitude on the smoked drum of on a kymograph, and later as incisions in the surface of a

[33] Loraine Daston, Peter Galison. *Objectivity* (New York: Zone, 2010). In many ways, experimental or applied psychology developed recently in historic record relative to the natural sciences and did not go through the same steps as Daston and Galison's "truth to nature", "mechanical objectivity", and "trained judgement." One could argue that experimental psychology begins with the instruments of "mechanical objectivity," but with the ethos of "truth to nature." The experimental psychologists, like Buswell, were not looking to capture every unique variation of a water droplet, but to develop, as Münsterberg says, "general laws." These general laws could then be developed into a training program, and reintroduced as "psychotechnics." How do Daston and Galison's categories start to break down when applied/imported into the field of psychology?

wax phonographic record.

The apparatus, originally developed by Gray was presented as six separate elements: the pneumograph, the light, the headrest, the camera, the mechanism for moving films, and the apparatus for exposing reading material. Gray should be given credit for his work on developing an apparatus for the University of Chicago Laboratory. But it was Buswell who introduced the phonograph into the loop, conducted rigorous experiments, and analysis.[34]

With this new apparatus, Buswell launched an ambitious experimental research program, building from Gray's work. He secured a wide range of subjects, from elementary school students to university level students. Subjects would read aloud and their eye movements would be recorded on forty two inch strips of photographic medium.[35] Their voices were recorded onto wax phonographic discs.

In terms of analysis, Buswell measured reading ability as the distance between the fixations of the eye–time when the eye is not moving (one is blind during movements of the eye)–and the voice–the annunciation of words. A surprising and complex physiological fact is embedded in Buswell's research on eye movement. It was not until the nineteenth century that psychologists discovered and proved that the eye does not glide across printed words on a page smoothly as one might perceive though introspection, but rather in rapid jerks.[36] These rapid jerks are called saccades.[37] During a saccade the visual system is blocked. The reader only "reads," when the eye fixates. In Buswell's research, mature readers were identified by a large eye voice span. What this means, physiologically, is that the eye is spatially and temporally ahead of the voice. For immature readers the voice and the eye were considered as a tightly

[34] Fred C. Ayer, O.B. Douglas, Frederick Elby, B.F. Pittenger, "In Memoriam: Clarence Truman Gray," (Austin: University of Texas, Austin: Internal Faculty Document, University Stenographic Bureau, November 16, 1951) 1-2: "His doctoral thesis, an experimental study of reading habits, was written under the direction of Professors C. H. Judd and W. F. Dearborn. In this connection he invented and built an ingenious apparatus for studying eye-movements in reading. Later he reproduced this in the laboratory of the University of Texas where it is still being used in experimental work."

[35] Buswell, *An Experimental Study,* 2-3.

[36] Wade & Tatler, *Moving Tablet of the Eye,* viii: "Thus, nystagmus was the initial class of eye movement that rocked the age-old belief in their smooth progress, and introduced the idea that we are not aware of the details of our own eye movements… "It was followed, almost a century later, by the demonstration of discontinuous eye movements in the same direction when reading text. These short jerks became known as saccades, and it was the instrumental ingenuity of Raymond Dodge that rendered them traceable photographically."

[37] There is still–to this day–much debate over who coined the term "saccade," and who discovered the movement empirically. Many give Javal credit for the term, but not for the physiological/psychological discovery. Wade & Tatler argue against Javal in their text.

coupled pair.

Returning now to the passage for "advanced readers," that has been slipped into this text twice already. The first four lines are intended to be, "easy, normal reading-matter."[38] Lines five and six contain three difficult words, and the last sentence is "made up of easy words but containing a difficult thought."[39] A record of a "good reader from the freshman class," reveals the temporal relationship between eye and voice within a single graphic representation. The top line represents the position of the visual fixations, measured in fiftieths of a second, and the lower line represents the annunciation time of specific words. The visual fixations are connected to verbal annunciation by lines. The three "difficult words," that Buswell inserted into the paragraph resulted in a decrease in eye voice span and then a moment of rereading at the word "hyperaesthesia." The "complex thought," also caused the subject to reread. The result of Buswell's studies can be summarized in a single diagram, that reveal how the eye and the voice can be spatially and temporally disciplined to construct best possible reader.

But what does it mean? There are three elements considered in the eye voice span diagram: the eye, the voice, and meaning. Buswell situated "meaning" as always between the eye and the voice. For advanced readers meaning was closer to the eye than to the voice. Buswell admitted that there is no objective way to locate meaning or to the unit size of meaning in his research. Meaning could have been considered as a single word, a segment of a word, a whole word, or even an entire sentence. In Buswell's words, "...the recognition of the complete meaning must be in a liquid state during the reading process, being subject to continual change and being held in

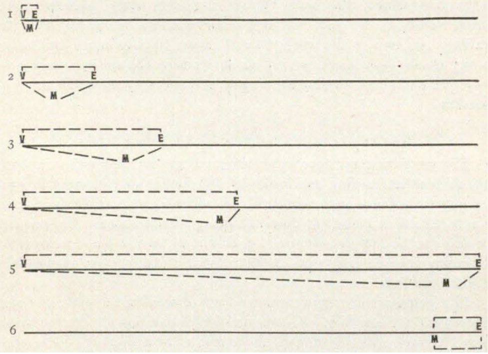


Figure 16: The development of the attention span in reading. The mature reader is represented in the fifth line, wile the immature reader is at the top. The last line shows a process of silent reading. Reproduced from Buswell, An Experimental Study, 100.

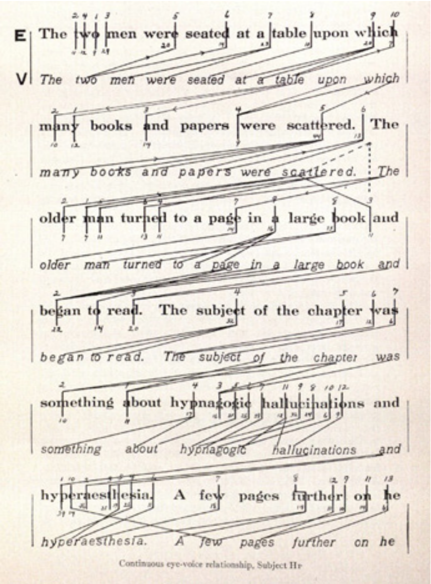
[38] Buswell, *An Experimental Study,* 10.

[39] Buswell, *An Experimental Study,* 10.

the mind in a tentative fashion until the end of the unit of thought is reached... a location for the recognition of such a developing meaning as this would probably refer to the focal point in the moving focal point... nearer [to] the eye than the voice..."[40]

How does the psychological study of reading in the laboratory become a way of discipling the eye outside of the laboratory? By understanding the relationship between the eye and the voice, Buswell–along with his academic mentor Judd–sought to develop a new program of education, to develop the best possible reader. However, this was no small task. The construction of the best possible reader also required the best possible educational system supplied with the best possible printed learning materials.[41]

The loop between laboratory psychology and society could be closed in the classroom. However, there seems to be a somewhat darker side to the project of applied psychology. The keyword here, although unstated, is optimization–if not objectification. While Münsterberg claimed that the aims of applied psychology were primarily humanitarian in nature, he also supported the goals of industrial optimization. "We must not forget that increase of industrial efficiency by future psychological adaptation and by improvement of the psychophysical conditions is not only in the interest of the employers, but still more of the employees; their working time can be reduced, their wages increased, their level of life raised."[42] The eye and the voice, synchronized in time , captured on photographic medium, and impressed in phonographic wax were no longer subjective qualities. They became abstract quantities subject to centralized control. (How would Marx respond?) In Buswell's world, meaning reproduced from the



The two men were seated at a table upon which
many books and papers were scattered. The
older man turned to a page in a large book and
began to read. The subject of the chapter was
something about hypnagogic hallucinations and
hyperaesthesia. A few pages further on he

Continuous eye-voice relationship, Subject Hr

Figure 17: The relationship between the eye and the voice. Notice how the eye and the voice are tripped up with the word "hyperaesthesia." Reproduced from Buswell, An Experimental Study, 66

[40] Buswell, *An Experimental Study,* 101.

[41] Buswell, *An Experimental Study*; Charles H. Judd, "Relation of School Expansion to Reading,": Both texts were published by the University of Chicago Press in collaboration with the *The School Review* and *The Elementary School Journal.*

[42] Münsterberg, *Psychology and Industrial Efficiency,* 308-309.

printed word, may have escaped measure, but perhaps not for long.

The printed word may have seemed like a labor saving device, for Mumford, but in the hands of applied psychology it was transformed. By submitting the subjective acts of reading to the exact measure and mechanically objective record, reading became an abstract object, not unlike Mumford's mechanical time itself.

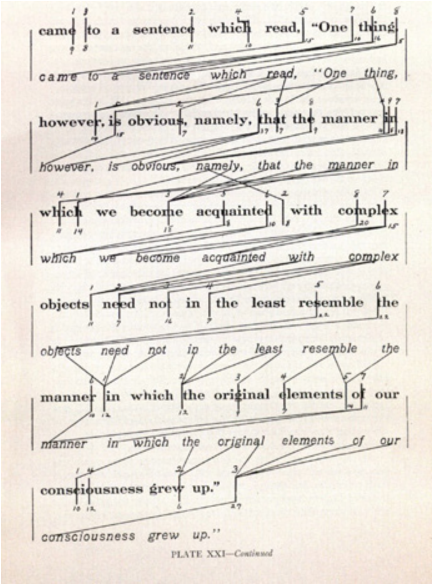

came to a sentence which read, "One thing,
came to a sentence which read, "One thing,
however, is obvious, namely, that the manner in
however, is obvious, namely, that the manner in
which we become acquainted with complex
which we become acquainted with complex
objects need not in the least resemble the
objects need not in the least resemble the
manner in which the original elements of our
manner in which the original elements of our
consciousness grew up."
consciousness grew up."
PLATE XXI—Continued

Figure 18: The relationship between the eye and the voice. Reproduced from Buswell, An Experimental Study, 67.

History of Eye Movement Studies

Part 2: Section Overview

In the previous section we demonstrated how reading and text, became an object of inquiry in the laboratory of experimental psychologists, through the development of instruments and in laboratory research. We argued that eye movement studies formed a crucial link between theoretical psychology in the laboratory and applied psychology (as psychotechnics) in the workplace. Furthermore we suggested how the development of new physiological instruments using photographic techniques and precise timing devices contributed to the sociocultural admittance of psychology into the fold of empirical sciences.

In this essay we explore another category of experiments in the history of eye movement studies in psychology where human subjects are asked to visually evaluate still images – paintings, drawings, photographs. The study of human vision in the act of examining pictures provide a much more challenging problem for experimental psychologists. While written text is predominantly linear, paintings and photographs share no such inherent structural constraints. We examine the challenges faced by experimental psychologists in attempts to evaluate fixation patterns of subjects looking at pictures.

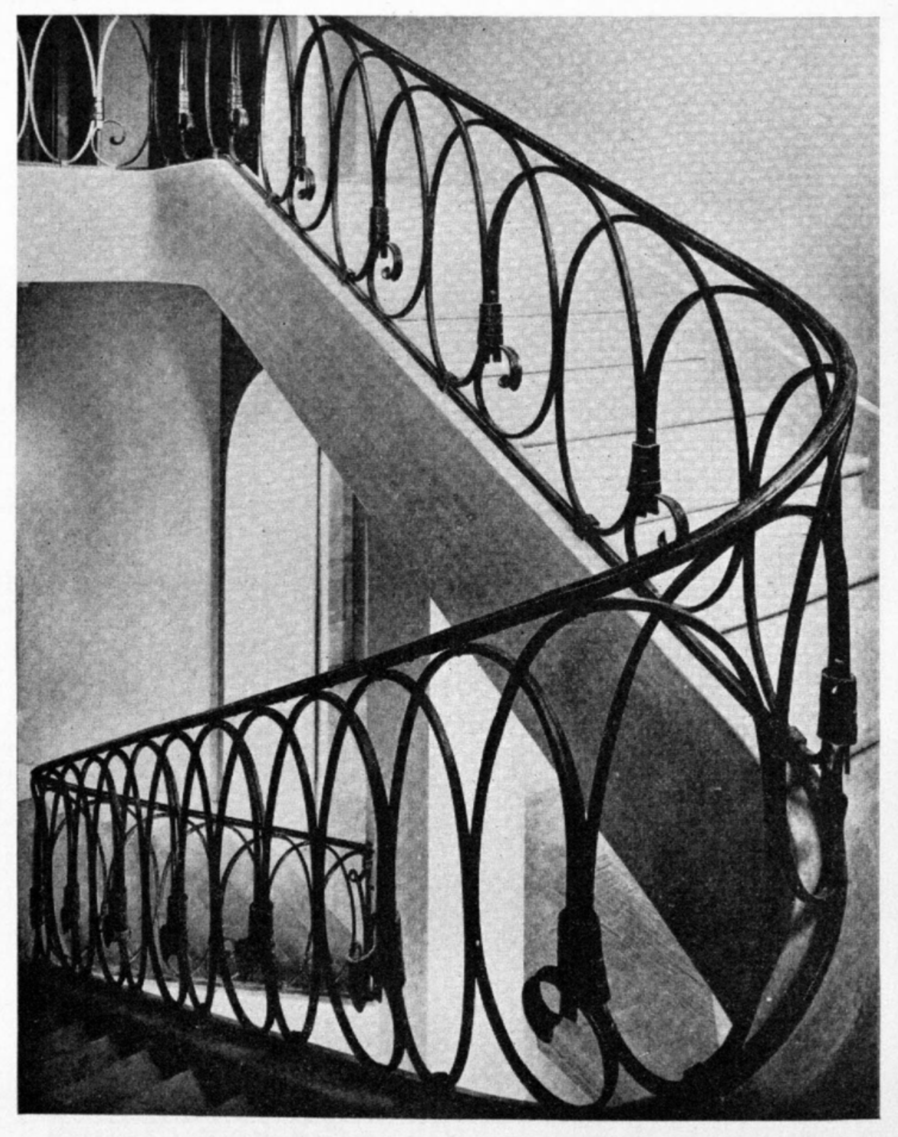
This essay revolves around two historical actors and experiments they conducted on humans evaluating two dimensional images. The first actor is Guy Buswell, introduced in the previous section in his research on reading and eye movements, who initiated and conducted an extensive experiment on human eye movements and two dimensional images in the mid 1930's at the University of Chicago. The second actor is Alfred Yarbus who conducted similar experiments in the mid 1960's in Soviet Russia. Both actors developed novel — but very different —

instruments and methods to measure and represent eye movements using photographic techniques.

Through an examination of these two experiments we will be able to concretely address challenges faced by the two psychologists and introduce questions of agency between human vision and visual stimulus media.

“How People Look at Pictures”

Study this image.



This image was selected by Guy T. Buswell to be used in his experimental study on the human perception of art conducted at the University of Chicago in the 1930's. In the introduction to his publication Buswell opens with broad question, "What does a person do when he looks at a picture?"[43] The answer to this broad question, Buswell posited, could be approached through a study of the processes of human perception and attention while one examines a two dimensional picture. The motivations for Buswell's research on the human perception of pictures stems from — and extends —his prior studies of eye movements and reading. The shift in the object of inquiry, from printed text to images on a two dimensional surface, can be characterized as the introduction of a new variable to the experimental process. With the introduction of the image (the new variable) the problem of correlating physiological and psychological processes increases in dimensionality. While the act of reading text is primarily a linear process, the act of viewing a painting shares no inherent structural and temporal constraints. Not only does the image viewing require new instrumentation to be developed to measure the physiological responses of a human subject, but also new methods of analysis.

A secondary motivation stems from Buswell's skepticism towards claims made by literature on the psychology of art and history or art. In the beginning of his book, Buswell cites many quotations that make claims on the relationship between artworks or designs and eye movements. For example:

> "A more or less differentiated pattern, gradually lengthened in its design and intensified in its color, will draw the eye from the lightly

[43] Guy T. Buswell, How People Look at Pictures, 4.

Figure 19: [Facing Page] Guy T. Buswell, How People Look at Pictures a Study of the Psychology of Perception in Art. (Chicago, IL.: The University of Chicago Press, 1935). Appendix, picture 25. Original from: Charles Moreau, "Wrought Iron Stairway ", La Ferronnerie Moderne (Paris: 1930). Size 20.8 x 26.6 cm.

> developed part toward the more expressive... Since a picture is something different from a section cut out from nature, it must provide a means of allowing the eye to travel through all parts associated within the frame."[44]

Buswell argues that these claims are based on subjective and introspective evidence of perceptual and psychological processes. Buswell sets out to corroborate or refute these claims with data from his research that are "entirely objective."[45] Armed with "objective" instruments, methods of experimental psychology, and prior results on eye movements and external two dimensional stimuli, Buswell sets out to establish the ground truths of how people look at pictures. Following the tenets of applied psychology, Buswell sought to make a contribution not only to the scientific community in an understanding of how people look at pictures, but also as a contribution back to society. In this case, this feedback loop would go back to the creator – artist – and critic – art historian.

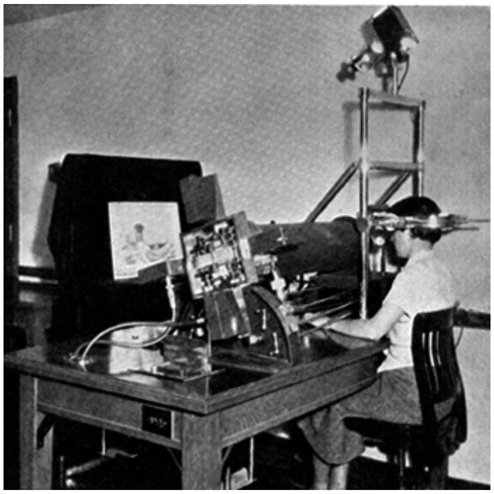


Figure 20: The photographic instrument that Buswell developed in order to measure eye movements of a human subject while examining a two dimensional picture. Here we see a subject seated at a laboratory bench with head stabilized. The photographic apparatus dominates the space of the table (seen to the left of the subject). The image is seen with a black background in the left of the image. The image being shown appears to be a reproduction of Katsushika Hokusai's color woodcut The Great Wave off Kanagawa, c. 1829-1832. Reproduced from Guy T. Buswell, How People Look at Pictures, plate V.

In order to conduct his study, Buswell had to first develop new instruments in order to objectively measure the movements of the human eye. These instruments would need to accommodate a pattern of movement that was not inherently linear like that of reading, requiring Buswell to extend the instrument previously used to measure eye movements in reading. In the study of reading, Buswell and Gray developed an instrument that used a moving strip of film upon which dots of light reflected from the corneas of subjects were registered as precise increments of time and space. The problem of instrumentation is succinctly stated by Buswell:

> "Technically the problem of photographing eye movements in reading is much more simple than

[44] Guy T. Buswell, How People Look at Pictures, 7. Quotation from Eugen Neuhaus. The Appreciation of Art (Boston: Ginn & Co., 1924), 155.

[45] Guy T. Buswell, How People Look at Pictures, 7.

> in the case of looking at pictures, since in reading it is only the horizontal movements of the eyes which are of any great significance. However, in looking at pictures it is necessary to secure simultaneously a record of both the vertical and horizontal eye movements. This involves technical difficulties which were surmounted by the construction of the elaborate apparatus used for the present study."[46]

The major principles of the new photographic instrument were recycled from the previous experiments on reading, where light reflected off of the moving cornea of the subject was registered on the moving photographic negative. In early instruments, such as the Dodge-Cline Photochronograph, the photographic medium was contained on a falling plate, using pressure to allow the plate to fall at a controlled rate. The University of Chicago instrument, originally developed by Clarence Gray and later used by Buswell, used more sophisticated mechanics to move a long strip of film along the focal plane of a camera. The movement of the photographic medium served to created a empty sliver of space on the photographic medium upon which reflected light from the cornea could be registered. But before the beam of reflected light could reach the photographic medium it would be interrupted by a precisely timed vibrating tuning fork or rotating fan blade. Thus, the record of eye movements were transformed into discrete points of light, registered on a dedicated sliver of photosensitive medium. The apparatus Buswell developed employed a fan that revolved at 30 Hertz. Curiously Buswell used the words "moving kinetoscope film" to name his instrument.[47] However the goal of the instrument were not to reproduce movements, as with most kinetoscopes, but to capture

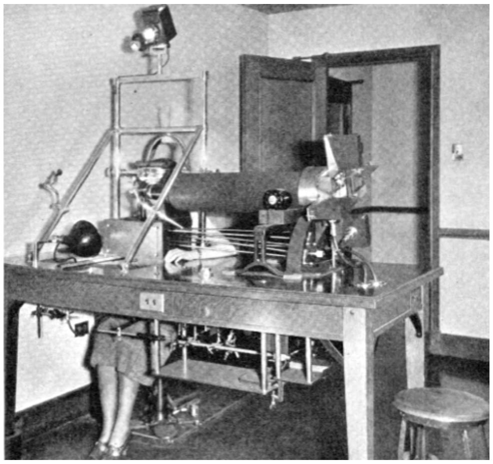


Figure 21: The photographic instrument that Buswell developed in order to measure eye movements of a human subject while examining a two dimensional picture. Reproduced from Guy T. Buswell, How People Look at Pictures, plate VI.

[46] Guy T. Buswell, How People Look at Pictures, 9.

[47] Guy T. Buswell, How People Look at Pictures, 10.

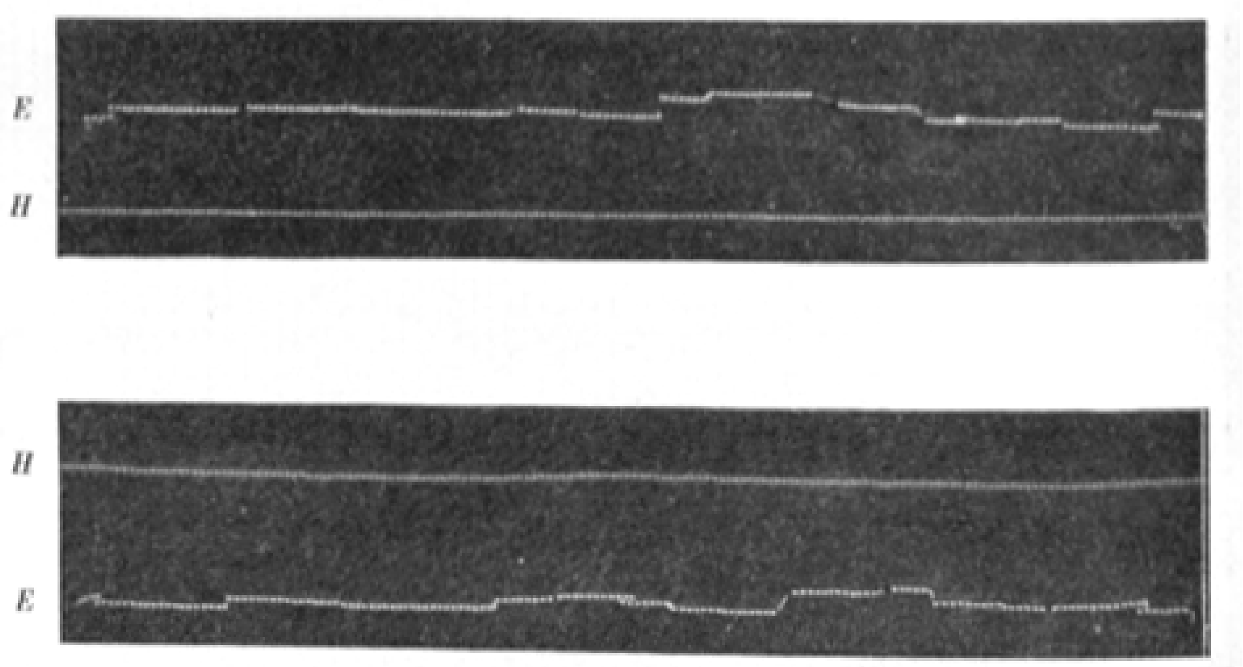


Figure 22: An example of a photographic record produced from Buswell's instruments. The top record shows vertical movements, and the bottom shows horizontal movements of the eye. The dotted line on the top of each record show the movements of the head, as reflected points of light at 1/50th of a second interval. The bottom white dotted lines show the eye movements in vertical and horizontal components respectively. Reproduced from Guy T. Buswell, How People Look at Pictures, plate VII, 13.

isolated movements as discrete points in time and space.

While problems of movement and discretization may have been considered solved in constrained situations, the problem of recording eye movements of subjects looking at pictures remained open. For the purposes of Buswell's experiments, this was solved by literally adding another dimension to the photographic recording instrument. In the new instrument there were not one, but two strips of film. The strips of film were oriented in parallel to one another and were mechanically pulled across the focal plane in synchronous time. Similar to previous instruments, light from a lamp was projected onto the subject's eye and the corneal reflections were captured by the camera lens. However, once the beam entered the lens of the camera, the process differed in how the beam of light was handled. The continuous beam of light was split by prisms into vertical and horizontal components. The horizontal and vertical components were then interrupted by a fan blade within the long cylindrical lens chamber of the camera, and registered separately on two empty areas of the moving strips of photosensitive medium.

The photographic records of reflected light points showed when and where in discrete space-time steps the subject's eye fixated.[48] In his publication, Buswell wrote at length to emphasize the precision of his instruments. For Buswell, the photographic record served as a precise measurement of perception and allowed him to stake a claim for an "objective," albeit mechanical, truth of human vision. The precise labor of recording eye movements, consumed 18,000 linear feet of photographic negatives. These negatives were the ground upon which further analytical representations of vision were based in Buswell's research. A reliance and claim to the objective truths of the instrumentation and photographic process were critical to the truth claims of Buswell's research. While a full discussion of objectivity and representation will be treated elsewhere in this book in full, it is important to introduce it here in order to point out that mechanical objectivity is not necessarily embodied in the final artifacts of Buswell's research.

Before recording eye movements, the subject was directed to fixate upon registration marks at the four corners of the picture. These marks were used as a calibration routine to mathematically correlate the extents of the image, in width and height, with the extents of light-point record. Once the two spaces – light points on film and source image — were spatially correlated Buswell graphed the fixations back onto the surface of the image. The discrete moments of fixation, extracted from the two moving film strips, were plotted as a function of the calibration

[48] Each point represents one thirtieth of a second as well as a discrete position in space.

coefficients precisely back onto the image.

Buswell recorded the eye movements of over two hundred subjects in response to a selection from a set of fifty five pictures. In total, 1,877 records were obtained. Both backgrounds of selected subjects as well as types of pictures covered a broad range. The subjects chosen for the experiments ranged in age, ethnic origin, and expertise.[49] The pictures selected covered a broad range of media – from photographs to paintings to line drawings – and subject matter — from everyday interiors to historical portraits and advertisements. Acclaimed paintings, works by Seurat, Duchamp, and Hokusai were also selected for the study. Aside from the work of Seurat, Duchamp, and perhaps the Japanese prints, all of the pictures selected are either photographic documentations of existing places and artifacts or realist representations.[50] With a varied population of human subjects, pictures, and precise instrumentation Buswell began his experiments.

Buswell's rigorous experimental program was organized in two main sections. The first section establishing "ground truths" of human vision in response examining pictures using his novel instrumentation to measure and record eye movements. The second section examined the variations in human subjects and pictures relative to eye movement response. In order to establish "objective" ground truths of human vision Buswell introduced a statistical method to quantify and compare subjective patterns of fixation. Density scatter plots were used to show the overall distribution of fixations, where each fixation point was plotted onto back onto the source image irrespective of order. For portraits the density plots are revealing, in that human faces dominate fixation patterns, such as in

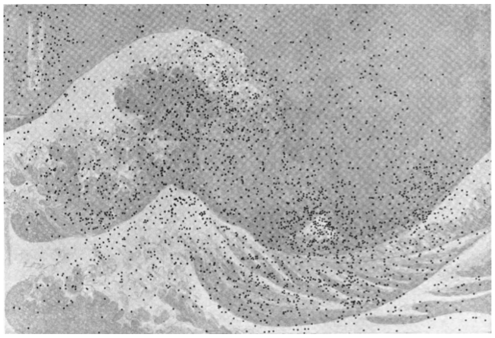
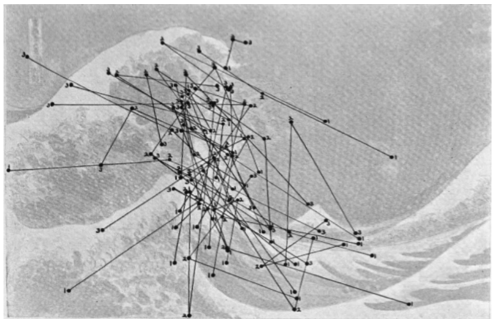


Figure 23: Plot of all fixations points of thirty five different subjects. Reproduction of Katsushika Hokusai's woodcut The Great Wave off Kanagawa, c. 1829-1832. Reproduced from Guy T. Buswell, How People Look at Pictures, plate XI.

Figure 24: The first three fixations of thirty five subjects. Reproduction of Katsushika Hokusai's woodcut The Great Wave off Kanagawa, c. 1829-1832. Reproduced from Guy T. Buswell, How People Look at Pictures, plate XI.



[49] Guy T. Buswell, How People Look at Pictures, 16. "Of this number 12 were elementary grade children, 44 were high-school pupils, and 144 were adult subjects. Of the adult subjects 47 were secured from the Art School of the Art Institute of Chicago and were persons who had from two to five year of special training in the field of art. Fourteen other subjects had made sufficient study of art to be classified as art students. The great majority of the remaining adult subjects were college or graduate students."

[50] The topic of "realism" will be treated in the subsequent section on representations of vision.

Walter Ufer's *Solemn Pledge*. In landscape images and architectural images, where there are no human or animals are clearly present, dominant fixation patterns are not immediately evident such as in Hokusai's *Great Wave*.

In the first section, Buswell established his methods and works toward creating a foundation for the relationship between eye movements and pictures. Buswell introduced a four by four grid to provide structure to the two dimensional field of the picture. This grid served as boundaries of a histogram for statistical analysis and comparison, where fixation points could be tabulated locally. With the four by four histogram, Buswell established categories like "centers of interest" and "patterns of perception." These categories enabled Buswell to extend his line of questioning in the agencies of the human subject versus picture as object, and to speculate on reversals of reversal of agencies between the two. Buswell quantifies centers of interest, by counting the number of fixations that were made in each area of the histogram bins and dividing by the total number of fixations for the image.

What emerges from Buswell's statistical analysis is evidence that points toward an influence on attention from features present in the picture. For example, human visual attention is drawn to faces as salient features in a picture. While this may be a well known observation from a contemporary position, and may have been tacitly known by artists and art critics alike, it was not until Buswell's "objective" methods that one could *see* how faces were attended to. Buswell's statistical analysis also suggests that movements of the eye, and more precisely movement motifs of the first few fixations are influenced by general motifs in the image. As one example, an

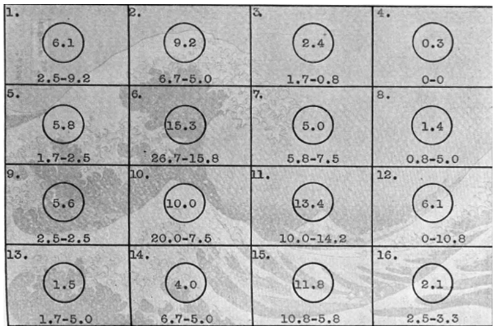


Figure 25: Four by four histogram where numbers in the circle indicate the average number of fixations (out of the first eighteen fixations) that fall into the rectangle for thirty five subjects. The numbers below the circles correspond to the percentage of the first three and last three fixations that are within the area of each rectangle. Reproduction of Katsushika Hokusai's woodcut The Great Wave off Kanagawa, c. 1829-1832. Reproduced from Guy T. Buswell, How People Look at Pictures, plate XI.

Figure 26: The last three fixations of thirty five subjects represented as dots connected by lines to show sequence. Reproduction of Katsushika Hokusai's woodcut The Great Wave off Kanagawa, c. 1829-1832. Reproduced from Guy T. Buswell, How People Look at Pictures, plate XI.

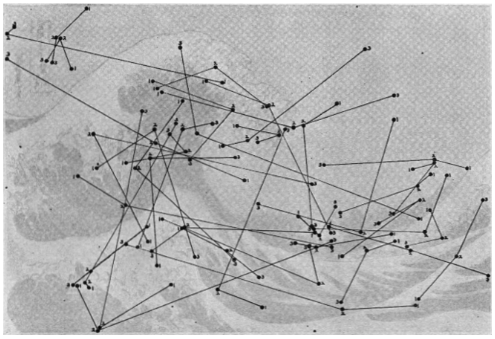
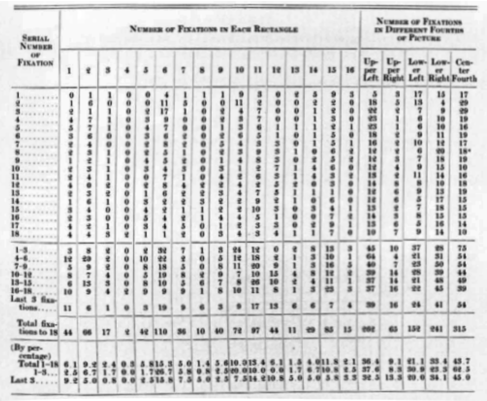


Figure 27: Table of fixation distribution for Hokusai's Great Wave print histogram for thirty five subjects. The table is organized by the number of fixations from top to bottom with a percentage score at the bottom row. Columns are ordered by number of fixation per rectangle as a sum of thirty five subjects, ascending. The last five columns show aggregate scores combinations of the histogram bins, where the image is subdivided into fourths. This allows for a shift in scale when analyzing the fixation patterns.



Serial Number of Fixation	Number of Fixations in Each Rectangle																Number of Fixations in Different Fourths of Picture				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Upper Left	Upper Right	Lower Left	Lower Right	Center Fourth
1..........	0	1	1	0	0	4	1	1	1	9	3	0	2	5	9	3	5	3	17	15	17
2..........	1	6	0	0	0	11	5	0	0	11	2	0	0	2	2	0	18	5	13	4	29
3..........	2	1	1	0	2	17	1	0	2	4	7	0	0	1	2	0	22	2	7	9	29
4..........	4	7	1	0	3	9	0	0	2	3	7	0	0	1	3	0	23	1	6	10	19
5..........	5	7	1	0	4	7	0	0	1	3	6	1	1	1	2	1	23	1	6	10	16
6..........	3	6	0	0	3	6	2	0	2	6	5	1	0	1	5	0	18	2	9	11	19
7..........	2	4	0	0	2	8	2	0	5	4	3	3	0	1	5	1	16	2	10	12	17
8..........	2	3	1	0	2	5	1	0	2	3	9	3	1	0	6	2	12	2	6	20	18*
9..........	1	2	1	0	4	5	2	0	1	4	8	3	0	2	5	2	12	3	7	18	19
10..........	2	3	1	0	3	4	3	0	3	1	2	7	1	4	6	0	12	4	9	15	10
11..........	2	4	1	0	0	7	1	0	4	2	6	3	1	4	3	2	13	2	11	14	16
12..........	4	0	2	0	2	8	4	2	2	4	2	5	2	0	3	0	14	8	8	10	18
13..........	2	3	2	0	1	6	2	2	3	4	7	5	1	1	1	0	12	6	9	13	19
14..........	1	6	1	0	3	2	2	3	2	2	9	2	1	0	6	0	12	6	5	17	15
15..........	3	4	0	0	4	2	1	1	2	2	10	3	0	3	4	1	13	2	7	18	15
16..........	2	3	0	0	5	4	2	1	4	4	5	1	0	0	7	2	14	3	8	15	15
17..........	4	2	1	0	3	4	5	0	1	2	3	3	0	2	9	1	13	6	5	16	14
18..........	4	4	3	2	1	1	2	0	3	4	3	4	1	1	7	0	10	7	9	14	10
1-3........	3	8	2	0	2	32	7	1	3	24	12	0	2	8	13	3	45	10	37	28	75
4-6........	12	20	2	0	10	22	2	0	5	12	18	2	1	3	10	1	64	4	21	31	54
7-9........	5	9	2	0	8	18	5	0	8	11	20	9	1	3	16	5	40	7	23	50	54
10-12......	8	7	4	0	5	19	8	2	9	7	10	15	4	8	12	2	39	14	28	39	44
13-15......	6	13	3	0	8	10	5	6	7	8	26	10	2	4	11	1	37	14	21	48	49
16-18......	10	9	4	2	9	9	9	1	8	10	11	8	1	3	23	3	37	16	22	45	39
Last 3 fixations.....	11	6	1	0	3	19	9	6	3	9	17	13	6	6	7	4	39	16	24	41	54
Total fixations to 18	44	66	17	2	42	110	36	10	40	72	97	44	11	29	85	15	262	65	152	241	315
(By percentage)																					
Total 1-18	6.1	9.2	2.4	0.3	5.8	15.3	5.0	1.4	5.6	10.0	13.4	6.1	1.5	4.0	11.8	2.1	36.4	9.1	21.1	33.4	43.7
1-3...	2.5	6.7	1.7	0.0	1.7	26.7	5.8	0.8	2.5	20.0	10.0	0.0	1.7	6.7	10.8	2.5	37.6	8.3	30.9	23.3	62.5
Last 3......	9.2	5.0	0.8	0.0	2.5	15.8	7.5	5.0	2.5	7.5	14.2	10.8	5.0	5.0	5.8	3.3	32.5	13.3	20.0	34.1	45.0

analysis of fixation patterns for Hokusai's wave print Buswell demonstrated the correlation:

> "The general direction of this major movement will be seen to follow the main direction of the wave, starting at the bottom in Rectangles 15 and 10 and then moving up through Rectangle 10 and 6 to Rectangle 2. The general direction of this movement is even more apparent parent in the individual plotting for this picture..."[51]

This observation seems to provide some objective weight for the subjective claims made by art critics that Buswell introduced in the introduction of his research.[52]

Once the foundations were laid for the relationships between human vision and pictures using statistical methods, Buswell proceeded to isolate and analyze variables of the experiment. First Buswell examined "variations in perception related to characteristics of the picture", such as color, detail, symmetry and balance, outline, profile, and pictures in various stages of completion. Second he examined "variations in perception related to characteristics of individuals", based on differences such as experience or training in viewing art, age, and ethnic background. Third, he examined the role of verbal priming or "variations in perception due to directions for looking at pictures."[53] This isolation seeks to untangle the agencies of the picture versus the agencies of the subject, sponsoring simultaneous questions: What do we attend to when we look at a picture and in what ways do the intrinsic properties of a picture have agency over human attention? In what ways does human attention shape the way in which a picture is attended to or seen?

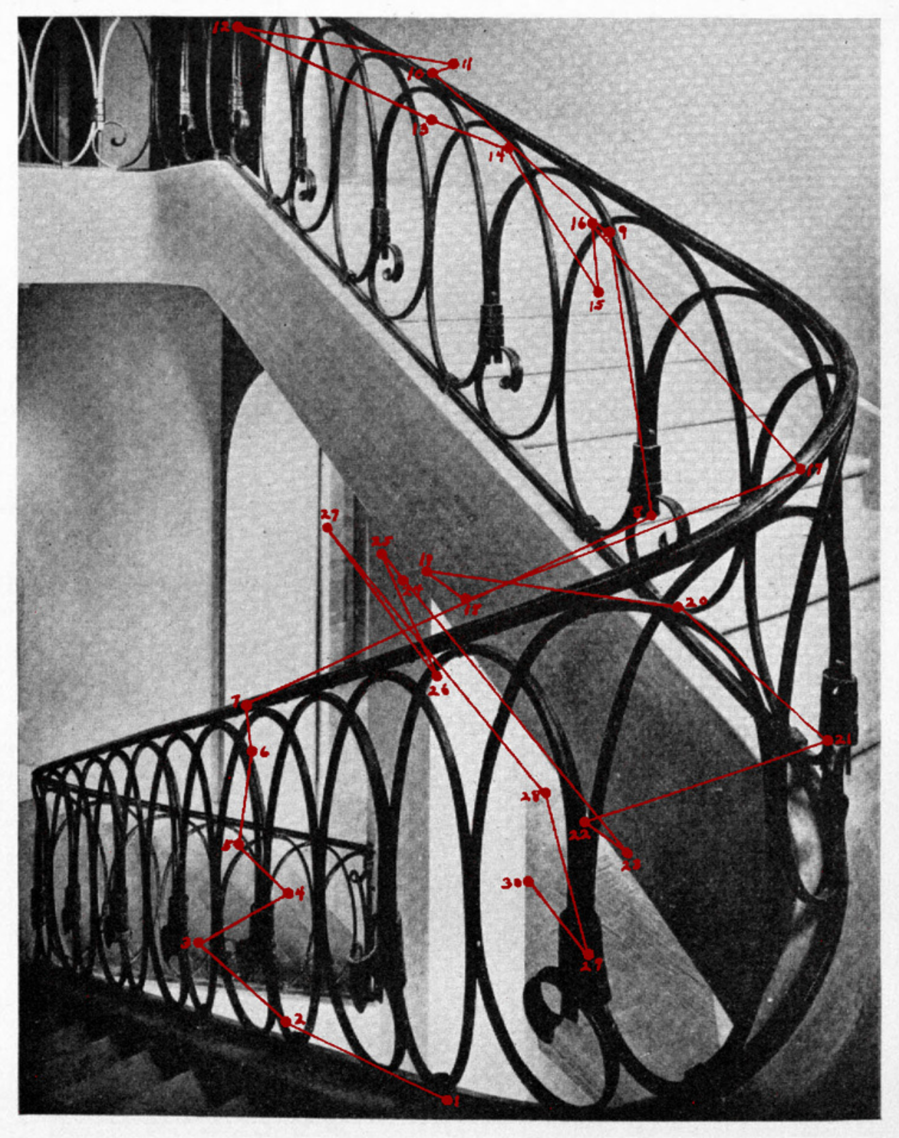
Figure 28: [Facing Page] Guy T. Buswell, How People Look at Pictures, Plate LIV Reproduced from: Original from: Charles Moreau, "Wrought Iron Stairway ", La Ferronnerie Moderne (Paris: 1930). Size 20.8 x 26.6 cm. (Color of the fixation points and poly-line connection originally black in Buswell's publication, modified by the author to emphasize the fixation points and path.)

[51] Guy T. Buswell, *How People Look at Pictures*, 35-36.

[52] A further examination of Buswell's statistical tables also reveals a subtle spike in the center grid cells which, again from a contemporary position may come as no surprise, but reveals a tendency toward the center in human visual fixation patterns. For a more detailed discussion see Tilke Judd, *Understanding and Predicting Where People Look in Images* (Cambridge: MIT, 2011),

[53] Guy T. Buswell, *How People Look at Pictures,* Chapter headings.

What did you see?



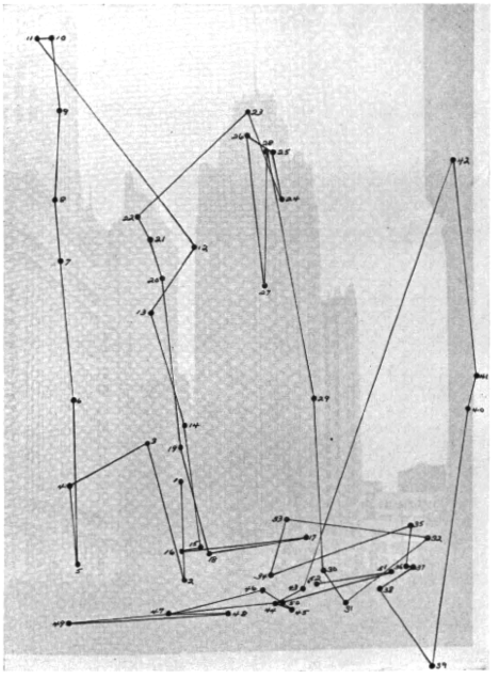


Figure 29: Fixation patterns of a subject freely examining (without instruction) a photograph of the Tribune Tower in Chicago. Reprinted From Guy T. Buswell, How People Look at Pictures, Plates LXV. Original photograph credits: Chicago Architectural Photographing Company (no date given).

For intrinsic variations in images, Buswell observed little difference in fixation patterns between color and black and white images. He speculated that similarity in fixation pattern might not be so aligned "[i]f color had been used in some unexpected way and if the main emphasis in the pictures had been on color rather than on form, the results might have been different."[54] In regards to details in a picture, Buswell speculated that fixation patterns for landscapes would appear as random distributions, but to his surprise he discovered that subjects often attended to "unexpected" elements in the pictures such as the "the crests of waves."[55] Using a set of architectural photographs Buswell demonstrates that his set of human subjects fixated on the architecture and specifically on details around doors and windows, while features in the landscape were virtually ignored. In terms of designs, Buswell demonstrates that during a free examination (without prior instruction) the fixation patterns of the subject do not correspond to the pattern of the design in any unit correspondence, but do follow general motifs within the space of the picture. Buswell remarked:

> "Picture 25, showing the stairway, furnishes another marked illustration of how eye movements follow the general pattern of a design. Here again there is no evidence of a series of rhythmical movements going either up or down the stairs in units corresponding to those shown in the decoration of the design, but there is a marked tendency to follow the general pattern both up and down."

Unfortunately, Buswell was unable to make any further conclusions on the ways in which intrinsic properties of an image affect human vision without a deeper understanding of cognitive responses. While

[54] Guy T. Buswell, *How People Look at Pictures*, 97.

[55] From a contemporary perspective, elements like the crests of waves, do not seem unexpected, in that they represent salient gradient changes within the image. Based on a contemporary understanding of biological processes of vision, especially the center-surround mechanism, leads one toward understanding why the wave crests were fixated upon, as gradient changes excite the optical nerve.

the salience of pictures or types of features in those pictures could not be directly quantified, Buswell was successful in gaining insight into the ways in which human attention shapes visual fixation patterns.

Different people see differently, and people see differently depending on what they are looking for in a picture. While this claim may seem tautological from everyday, tacit, and subjective experience, it is not easy to prove. According to Buswell's research subjects with artistic training do have different fixation patterns when viewing a painting, for example, than those without artistic training. However, the difference is very subtle. Those who have artistic training tend to have shorter fixation durations and cover a broader area of the two dimensional surface. Those without prior training in art, on average, make fixations of longer duration.[56] While difference between individuals may be considerable, Buswell determined that the statistical differences between groups, whether by experience, age, or ethnicity are not statistically significant.[57]

Perhaps the most interesting, yet understated, contribution that results from Buswell's research are found in the observation that fixation patterns vary based on verbal directions or instructions given to the subject. The resulting artifacts show incredible variation between the *same* when looking at a picture with *different* verbal priming.

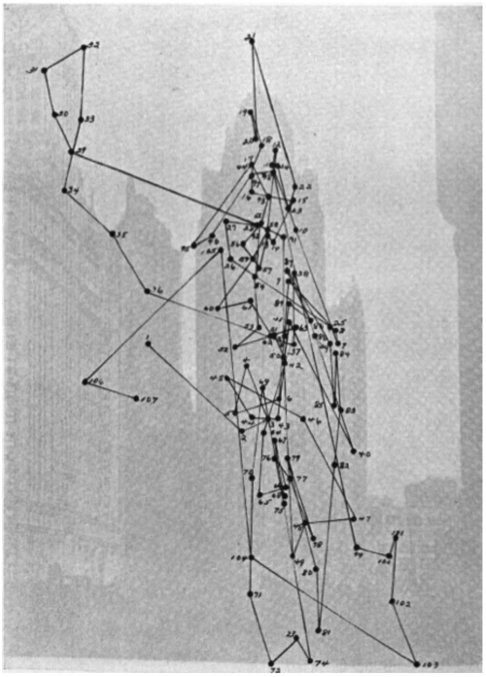


Figure 30: Fixation patterns of the same subject examining a photograph of the Tribune Tower in Chicago given the instruction, "... look at the picture again to see if he could find a person looking out of one of the windows of the tower." Reprinted From Guy T. Buswell, How People Look at Pictures, Plates LXV. Original photograph credits: Chicago Architectural Photographing Company (no date given).

In this case one can clearly see the results of verbal priming, where the pattern of fixation is influenced by an objective that is known, by both subject and researcher in advance. This leads to insight of the importance of attention on fixation patterns and calls into question the role of human agency in attending to a picture versus the agency based on salient features

[56] "In respect to duration of fixations, the group which had studied art made consistently shorter durations, both during the first 25 fixations in looking at seven different pictures and during Fixations 76- 100. On the other hand, ability as measured by the art test apparently has little relation to duration of fixation pauses." Guy T. Buswell, *How People Look at Pictures,* 130.

[57] Minor differences were apparent from picture to picture, but no consistent major differentiation in the patterns of perception could be identified. The average differences between the groups were so much less than the individual differences within each group that the results cannot be considered significant." Guy T. Buswell, *How People Look at Pictures,* 131.

in the picture.

The publication artifacts of Buswell's experiments, much like the final publication artifacts of his reading studies, close the loop between the literal object of inquiry – stimulus – and the physiological response. In the reading studies Buswell re-inscribed the recorded eye and voice movements back into the text, producing an new object of inquiry as an analytical view of what Buswell called the "eye-voice span" in reading. On reading the analysis, a contemporary reader can reenact simulation of the experimental trials, just by reading along, imagining the stutters and stops from the tangled web that connected eye and voice. However, it is much more of a challenge to follow the pattern of fixations over the less structured field of the two dimensional image. We can not see ourselves seeing, nor is it easy to control the movements of our eyes. The subject and object are transposed (or collapsed into one artifact) in Buswell's analytical *re*presentations of vision. The artifacts produced by Buswell are a sort of hybrid between mechanically objective transcription and a trained judgement. Trained judgement, in that assumptions are made by Buswell in the representation, and hybrids of mechanically objective epistemology where movements of the eye are omitted due to the construction of the photographic instrumentation. While fixations are not physiologically dimensionless points, Buswell's instruments captured points, and these discrete moments in time were transcribed back onto the space of the image. In Buswell's records, we witness the collision science and art. From this collision, emerges as a new artifact, bolstered with new information, experiences, and traces of multiple actors and authors.

“Yarbus’s Complex Objects”

In 1965 Alfred Yarbus, working as a biophysicist at the Institute for Problems in Information Transmission in the USSR, published his extensive research titled *Eye Movements and Vision*.[58] Yarbus’s book was translated into English by Basil Haigh and republished in 1967, and has since become one of the most cited books on the subject of human vision and eye movements.[59] *Eye Movements and Vision*, not unlike Buswell’s *How People Look at Pictures*, begins with an elaborate explanation of laboratory techniques. Like Buswell, Yarbus seeks to establish ground truths of vision based on his new instruments. Through the use of new instrumentation Yarbus develops a new way of *observing* vision and is able to posit novel relationships between the vision of a human subject and two dimensional images – or in his words, “complex objects.”[60] In this essay we will first review the instruments developed by Yarbus and then conduct a detailed analysis of the final chapter of Yarbus’s book, “Eye Movements During the Perception of Complex Objects.” The analysis of Yarbus’s work will run parallel to our analysis of Buswell’s research, where we question the relationship between the agency of the human subject and the stimulus-object in the active construction of vision.

In opposition to Buswell’s non invasive photographic instruments for recording eye movements, Yarbus developed eye cups that were fixed directly to the surface of the human eye. These eye cups were hand fabricated from rubber or duralumin in Yarbus’s lab, and employed differential pressure from a suction to affix themselves to the scleral surface of the eye.[61] These instruments and processes were highly invasive, posed health risks

[58] A direct translation from the Russian title would be: *The Role of Eye Motion in Vision Processes*. See: DeAngelus M, Pelz J, “Top-down control of eye movements: Yarbus revisited” (*Visual Cognition 2009*: 17), 790–811.

[59] Tatler et al. “Yarbus, Eye Movements, and Vision” *i-Perception* 1: 2010. 7-27. According to Tatler et al’s research Yarbus’s book had been cited more than 1,659 times from 1967 to 2010 (Source: ISI Web of Knowledge), by a wide range of disciplines in descending order: Neuroscience, Experimental Psychology, General Psychology, Ophtalmology, Artificial Intelligence, Electrical Engineering, Computer Science, Optics, and Behavioral Science.

[60] For Yarbus a “complex object” is a stimulus that has an agency of its own, in that a subject can not voluntarily control his eye movements. The complexity lies in the interaction between the subject and the object.

[61] Duralumin is a trademark name for an aluminum alloy.

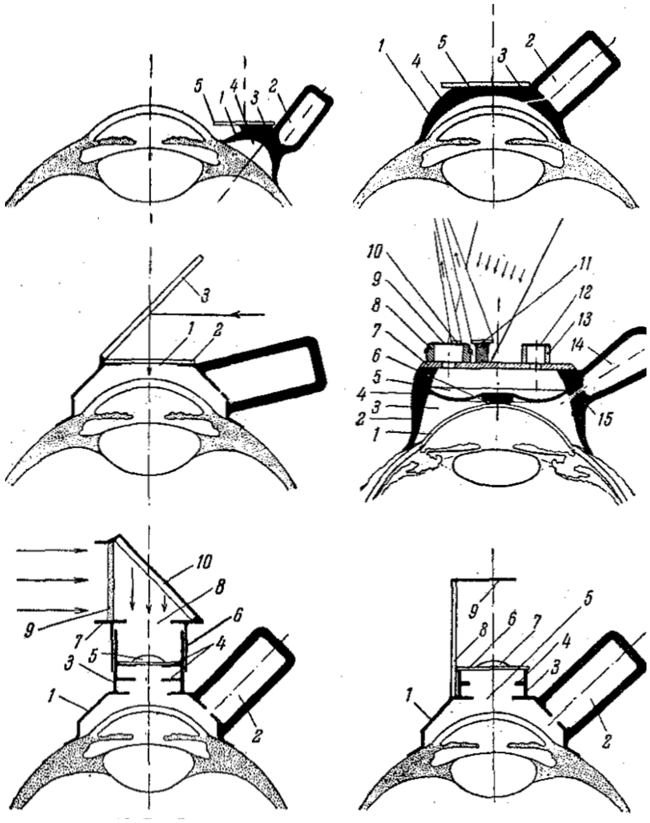


Figure 31: Eye Caps (P1, P2, P4-P7). Eye caps used to Reprinted from Alfred L. Yarubs, Eye Movements and Vision. (Moscow, USSR: Institute for Problems of Information Transmission: 1967), 44.

to the subjects, experiments were inherently limited in duration, and required special preparation of a subject prior to and after an experimental trial. So why would Yarbus, knowledgeable of noninvasive photographic techniques of Buswell and Dodge, invest in developing instruments and laboratory procedures on the physical surface of the eye?

Yarbus's motivation to develop these invasive eye cups was based on a desire to capture very fine movements of the human eye that, at the time, were not able to be captured by photographic instruments such as those developed by Buswell. Additionally Yarbus sought more direct and mechanical records in the form of a photographic process:

> "In evaluating methods of recording eye movements based on still and motion-picture

> photography, it must be remembered that these methods can be used successfully in many cases when the large movements of the eyes have to be recorded. Their main disadvantage is the relatively laborious method of analysis of the records required."[62]

In Buswell's experiments the scientist (or experimental psychologist) was required to interpret the results of the mechanical recording instruments based on a calibration routine. This process involved judgement on the behalf of the scientist in that it required a manual transfer of the points back on to the surface of the stimulus image. Yarbus, on the other hand, sought to develop a procedure that would be more invasive to the subject, but less invasive to the process of translating physiological patterns to photographic record. In Yarbus's decision to choose invasive procedures for higher resolution and more "hands off" attitudes, we can see an example of how manual intervention was minimized in favor of mechanical objectivity.

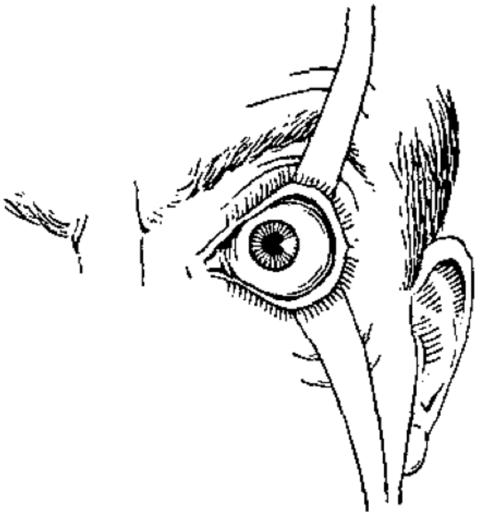


Figure 32: Preparation of the subject for Yarbus's experiments, the eye is exposed and prevented from blinking by taping the eyelids back with plaster tape. Reprinted from Alfred L. Yarubs, Eye Movements and Vision. (Moscow, USSR: Institute for Problems of Information Transmission: 1967), 44.

Yarbus developed eight different types of eye caps that were used in conjunction with a massive lighting and recording device. Most eye caps covered the corneal area completely, constraining the space and exposure of vision to a tiny frosted glass rectangle placed within millimeters from the surface of the cornea. Short focus lenses were incorporated into some of the eye caps to allow the subject to focus on these close projections. Other suction caps were affixed to the scleral surface, adjacent to the corneal area, leaving the eye free to inspect objects in space while reflecting light onto a photosensitive surface. In order to attach these suction caps, the eyelids of the subject would be taped back, the eye anesthetized with amethocaine to minimize ocular discomfort,

[62] Alfred L. Yarubs, *Eye Movements and Vision.* (Moscow, USSR: Institute for Problems of Information Transmission: 1967). 23.

and the suction applied to the eye and adjusted for focus and position.[63] Light projected onto the eye in a convergent beam would be reflected off of the mirrored surfaces of the eye caps and could be recorded on a still or moving sheet of photosensitive paper.

Seeking to establish ground truths of vision Yarbus defines different types of movement for the human eye: fixation, saccades, tremor, drift, pursuit. Buswell and other predecessors in eye movement studies were well able to measure saccades from observation and record them with photographic techniques, however they were unable to precisely measure drift and tremors that occur during the act of fixation. Through alternative, and invasive, methods Yarbus was able to precisely record drift and tremors that occur between major saccadic movement. As the eye drifts during fixation, reflected light recorded on a still photosensitive sheet yields a magnified representation of the foveal area as an ellipse.[64]

After establishing ground truths, Yarbus engages in a discussion of how people look at pictures. But instead of using the word "picture" as a title for the section Yarbus employs the words "complex object." The use of the word "complex object" points to an inherently difficult object of study within his rigorous experimental program – much more complex relative to a black dot on a white ground – but in a theoretical sense, moves away from the idea that we see images, but rather that we obtain *information* in the act of seeing. After all, Yarbus was employed as a researcher at the Institute for Problems in Information Transmission. What information can be constructed through a process of visual attention and what are the factors that shape the way in which humans attend to "complex objects"?

[63] Alfred L. Yarubs, *Eye Movements and Vision,* 43-57. Amethocaine is a local anesthetic.

[64] "In other words, if the axis of the eye is mentally continued to its intersection with the frontal plane, as a result of the tremor it will describe elliptical figures on that plane" Yarbus, *Eye Movements and Vision,* 115.

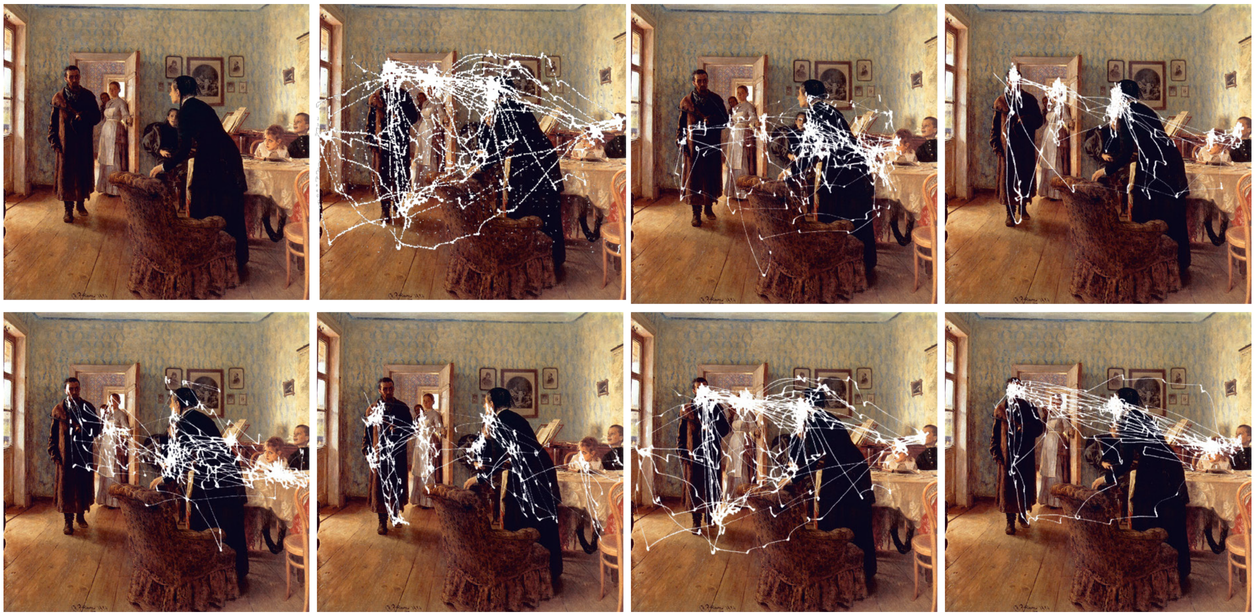


Figure 33: "Seven records of eye movements by the same subject. Each record lasted 3 minutes. The subject examined the reproduction with both eyes. 1)Free examination of the picture. Before the subsequent recording sessions, the subject was asked to: 2) estimate the material circumstances of the family in the picture; 3) give the ages of the people; 4) surmise what the family had been doing before the arrival of the 'unexpected visitor': 5) remember the clothes worn by the people; 6) remember the position of the people and objects in the room; 7) estimate how long the "unexpected visitor' had been away from the family. Text caption from Alfred L. Yarubs, Eye Movements and Vision. (Moscow, USSR: Institute for Problems of Information Transmission: 1967), 44. Images partially reproduced from Sasha Archibald, "Ways of Seeing," Cabinet Magazine: 30, 2008.

Perhaps the most important contribution in his studies of "complex objects" were on the ways in which visual attention is affected by verbal instructions. While Buswell already demonstrated this top-down mode of influence in his 1935 publication, Yarbus's work is now widely cited in papers on active vision. Why is Yarbus's research cited over Buswell's work? This is a question we will return to in the conclusion. But, perhaps one of the answers exists in Yarbus's mechanical "hands-off" methods when it comes to the records and publication of the records. The influence of priming, or instructions given to the human subject before viewing the image are most clearly demonstrated in the records using a realist painting by Ilya Repin titled, "The Unexpected Visitor."[65]

The records show a series of grids with eight cells, the first cell contains a reproduction of the original Repin painting and the remaining seven cells varying recordings of eye movements. It is important to note that the recordings of subject's eye

[65] Ilya Repin, *An Unexpected Visitor,* 1884. (Also referred to as "An Unexpected Return").

movements are not overlaid back onto the original image, but are presented as mechanical records directly from the photographic negative. The first set shows seven different subjects who were allowed to freely examine (without instruction) the Repin image with both eyes for three minutes. These images show variation between the subjects, but also a density of concentration around certain elements such as faces and representations of people in the scene. The second set displays seven eye movement recordings of the same subject while allowed to freely examine the Repin image with both eyes for three minutes. In this series the subject remains the same, and the duration between viewing the images is one to two days. The records are remarkably similar. The third, and perhaps most insightful, set of images show records of eye movements of a single subject who has been instructed with verbal cues (primed) before viewing the picture. Before viewing the Repin painting, the subject was given instructions varying from highly subjective — "surmise what the family had been doing before the arrival of the 'unexpected visitor'" — to objective "remember the clothes worn by the people."[66] The resultant fixation records reveal distinct patterns of eye movements relative to given instructions. This led Yarbus to the conclusion that fixation patterns are a direct channel into the mysterious workings of human thought: "Eye movements reflect the human thought processes; so the observer's thought may be followed to some extent from the records of eye movements (the thought accompanying the examination of the particular object)." By studying the records of eye movements, Yarbus believed that he had closed the gap between the agency of the "complex object" and human agency, "It is easy to determine from these

[66] Alfred L. Yarubs, *Eye Movements and Vision.* 174.

records which elements attract the observer's eye (and, consequently, his thought), in what order, and how often."[67] With a firm handle now on how human thought could be connected to vision, Yarbus extended his analysis:

> "It may be concluded that individual observers differ in the way they think and, therefore, differ also to some extent in the way they look at things... In just the same way it would be natural to assume that a complex object of perception understood by a physicist but unfamiliar to a biologist (or vice versa) will be examined quite differently by a physicist and a biologist." [68]

What may seem like tacit knowledge can now be defended with empirical record. Different people see the same object differently. And furthermore, the same person can also see an object differently depending on their motivation. From Yarbus's position this difference in human visual fixation patterns was a difference in *information* obtained by the eye.

[67] Alfred L. Yarubs, *Eye Movements and Vision.* 190.

[68] Alfred L. Yarubs, *Eye Movements and Vision.* 192.

Representations of Vision, Constructing Objectivity

In the first two sections of this chapter we wrote about the history of eye movement studies in the examination of two types of experiments, the first on studies of how humans read text and the second on studies of how humans look at pictures. These experiments were conducted within the disciplinary context of experimental psychology, and were representative of a period in time when psychology was gaining ground in its social acceptance as an empirical science through experimental methods rooted in physiology. Knowingly or unknowingly in support of their field's trajectory, the researchers we examined – Dodge, Münsterberg, Buswell, and even later Yarbus — sought to produce "objective" records and representations of human experiences of vision. Their records were representations of vision, both as theoretical claims toward understanding human perception, and as records of a physiological reality.

We framed these experiments in a discussion on the relationships between theory, instrumentation, and experimental observation. This framework is based on concepts developed by scholars in the history of science, Peter Galison and Loraine Daston.[69] In both our own research, and historical survey, we have found that claims to objectivity or reality are enmeshed with the development of instrumentation and provide a foundation upon which one constructs a vision of the world. From these studies we learn that seeing is far from a passive act. Seeing is an act of construction, where a vision of the world is formed based on a collision of theory, instruments, and experimental observation at any given moment.[70] It is through representations that these world views are captured, constructed, and communicated.

Historians of science, bound by a disciplinary

[69] Lorraine J. Daston and Peter Galison, Objectivity (Zone, 2010). Peter Galison, *Image and Logic: A Material Culture of Microphysics*, 1st ed. (University Of Chicago Press, 1997).

[70] We abolish the word "spectator" from our text specifically for the reason that it is a passive way of thinking about seeing. Seeing is an active construction.

methodology, periodize shifts in disciplinary praxis. Often imposing divisions between theory, observation, and instrumentation. In a positivist model, experimental observation is treated as a constant while theories change over time to accommodate what the trajectory of observation.[71] In an antipositivist, or Kuhnian, model theory and experimental observation are bound together in lock step, continuing in incremental progression except in the case of a disruption or "revolution," provoking what Kuhn calls a "paradigm shift."[72] In the Kuhnian methodology, theory frames possible visions of a world – whether enabling or inhibiting – and therefore informs the instruments of observation.[73] Galison eloquently characterizes the two methods of periodization:

> "The positivist and antipositivist periodizations have a grandeur to them: they both sought and found a single narrative line that would sustain the whole of science – in observation for the positivists and theory for the antipositivists. Both agreed that language was the linchpin of science – though the positivists looked for a language of experience, and the antipositivists located the key terms in theory."[74]

Shifts in theoretical world views or experimental observations are accompanied by claims to truth, objectivity, or access to an underlying reality. In either mode of thinking, there must be a circulation of knowledge in for ideas to gain acceptance as a *social* reality. [75] The necessity to circulate ideas requires representation. While there are many forms of representation – from written text, spoken word, drawing, models, simulations – we are primarily concerned with visual representations of human vision.

We have demonstrated how representations

[71] Galsion, *Image and Logic,* 784-785.

[72] Thomas S. Kuhn, The Structure of Scientific Revolutions, 3rd ed. (University Of Chicago Press, 1996); Galison, *Image and Logic,* 793-797.

[73] Galison, *Image and Logic,* "The Trading Zone: Coordinating Action and Belief," 781-844.

[74] Galison, *Image and Logic,* 793.

[75] Bruno Latour, "Visualisation and Cognition: Drawing Things Together," *Knowledge and Society Studies in the Sociology of Culture Past and Present* **6** (1986): 1-40. See Latour's concept of "immutible mobiles" is salient here.

of human vision are constructed in the domain of experimental psychology, in late nineteenth and early twentieth century studies of eye movements with specific focus on the experiments of Guy Buswell and Alfred Yarbus. Buswell and Yarbus's representations of vision were, and remain, artifacts that are inscribed with an epistemology. To borrow Bruno Latour's term, these representations of vision are "immutable mobiles," objects that are optically consistent and can be transported – literally or metaphorically – without corruption. Furthermore, these representations are inscribed within an epistemology of objectivity and make claim to revealing an underlying truths of nature. In these cases, the nature of a human experience is an experience of vision. While powerful representations, we argue that they can not be studied in isolation. Without the development of instrumentation, these representations of vision would not exist, nor would they be motivated to be constructed if it were not for earlier theories of human vision, optics, and so forth. Which leads us, in accord with Galison, to question both the previously discussed methods of periodization that purport temporal unities and hierarchy within fields of scientific study and the construction of objectivity in representing visions of a world.

Galison's concept of intercalated periodization is insightful for our research, serving as a critique of positivist and antipositivist methods of periodization and hierarchy. Underlying positivist and antipositivist methods is the idea of a unified scientific field. Galison argues, along the same lines as Feyerabend, although with less anarchic bent, that science is *not* unified. Shifts in epistemology are fragmented, not sea changes or paradigms, and can develop from a variety different sources and interactions between

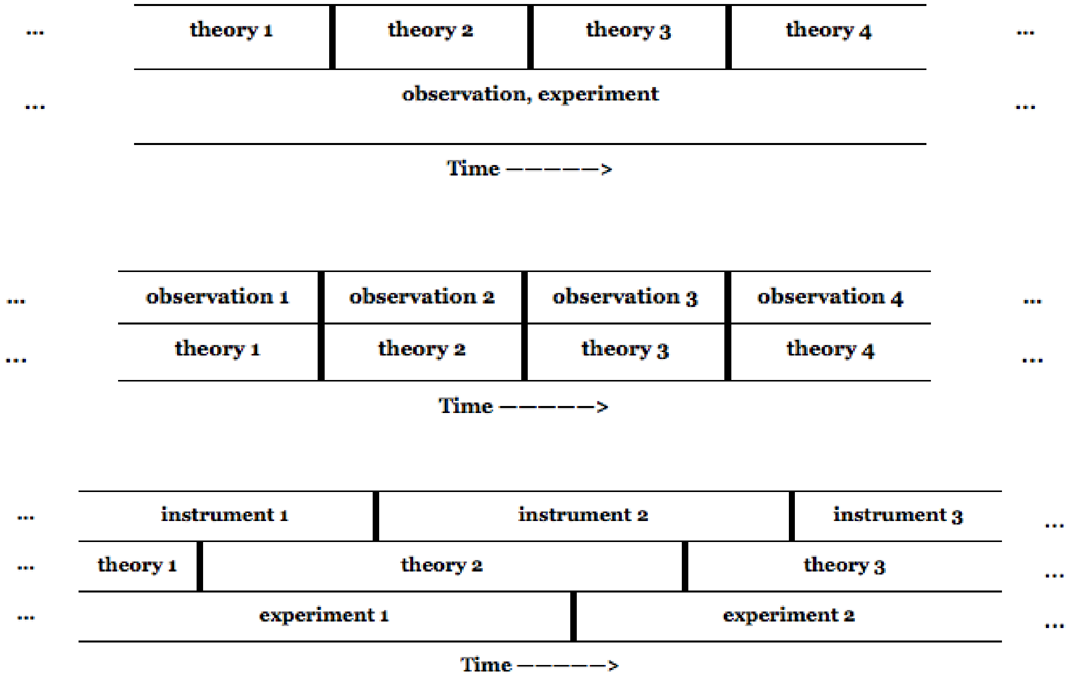


Figure 34: Galison's diagrams of periodization. [1] Positivist periodization. [2] Antipositivist periodization. [3] Galision's Intercalated periodization. Redrawn by the authors. Originally in Galison, Image and Logic, 785, 794, 799.

those sources.[76] In his research on the development of the field of microphysics, Galison proposes an intercalated model of interaction. Intercalated, literally defined as an interpolation in a temporal period or the insertion of something between layers in a crystalline lattice or structure, introduces a flexibility into the traditionally rigid historical structures. The three tables suggest diagrams for positivist, antipositivist, and intercalated methods.

This model is a useful way of thinking about how the development of instrumentation, experimentation, and theorization can develop independently without an inherent hierarchy. The interactions between strata in Galison's model can lead to local or systemic shifts in knowledge. We argue that at every stage of development there are representations of knowledge put forth as an argument or claim to a "reality." In our research we have learned that developments in instrumentation

[76] Paul Feyerabend, *Against Method: Outline of an Anarchistic Theory of Knowledge*, 3rd ed. (Verso, 1993). "Are we really to believe that naive and simple-minded rules which methodologists take as their guide are capable of accounting for such a 'maze of interaction'?", 10.

can lead to novel experiments and ultimately a shift in the understanding of the human visual system. While it appears that instrumentation took the lead role in Dodge's inquiry into the psychology of human vision, it is difficult to speak with confidence on the order of causality of each area on one another. And determining causality is not an argument that need be pursued in our research; we will leave it to the historians to argue about the order of causality – instrumentation or theory. However, we do have some confidence in discussing the representations of vision and developing an understanding of how objectivity is constructed in scientific practice.

But what exactly is an objective representation of reality, is it even possible? And, in what ways has our understanding of human vision changed over time, and how can we begin to study these changes? The answer to both of these questions lies in the relationships between these quasi-categories of instrument, theory, and experiment. Our argument parallels Daston and Galison's argument in that we want to show how, "epistemic virtues [are] inscribed in images, in the ways they are made, used, and defended against rivals."[77] Our argument diverges from Daston and Galison in the material artifacts studied. In their research, the objects of inquiry were primarily images from seventeenth century scientific atlases. In our research we are concerned with representations of vision (visions of vision), primarily from the nineteenth century to present. Our study also requires us to answer questions about how epistemological virtues shifted with respect to human vision. Is vision situated outside of the subject or is it something that is rooted within the subject? Studies of vision, not unlike objectivity, have a "summersault history," where human vision shifts

[77] Lorraine J. Daston and Peter Galison, *Objectivity* (Zone, 2010). 42.

from something "out there" to be obtained from the world to an embodied phenomena.[78] To fully address shifts in historical conceptions of vision we will rely on frameworks established in the work of Jonathan Carry.[79]

Crary postulates shifts in the history of vision in modernity through a discussion of the relationship between an observer and optical devices – notably the camera obscura and the stereoscope. The optical devices in Crary's research serve a purpose that parallel the scientific atlases in Daston and Galison's research. For Crary, optical devices are, "points of intersection where philosophical, scientific, and aesthetic discourses overlap with mechanical techniques, institutional requirements, and socioeconomic forces."[80] The camera obscura optically situates the observer as outside of vision, in that they do not actively construct vision but observe the world from inside a camera. The stereoscope and the phenomena of afterimages, on the other hand, situate vision as an embodied experience; within the observer's body. A stereoscope presents two images with disparity, that when presented to two human eyes, can be fused into a three dimensional understanding of the world. For Crary, the stereoscope marks a shift in vision in relation to the observer from objective – being outside – to subjective – being inside the observer's body.[81] However, the use of the terminology relative to an observer's body as, objective and subjective here is misleading relative to representations of vision. Within each optical device was an inscribed claim and representation of the realities of human vision. In essence both technologies were argued to be capable of producing objective representations. So, how can both an externalized and internalized experience be objective? In order to clarify some of the arguments

[78] Daston and Galison, *Objectivity*, 29.

[79] Jonathan Crary, *Techniques of the observer : on vision and modernity in the nineteenth century* (Cambridge Mass.: MIT Press, 1990).

[80] Crary, *Techniques of the Observer,* 8.

[81] This shift begins in the study of after images for Goethe and Purkinje.

it is necessary to first review the origins of the word objectivity itself.

Linguistically the words objectivity and subjectivity themselves have a complicated history. While they have always been a dipole pair (since the fourteenth century), in the course of recent history they have undergone complete reversals in their polarity. According to Daston and Galison, the word objectivity has a "summersault history." Originally, "'[o]bjective' referred to things as they are presented to consciousness, whereas 'subjective' referred to things in themselves."[82] The contemporary reversal in the meaning of objectivity and subjectivity can be attributed to Immanuel Kant, where the objective is paired with the *universal* (preconditions of experience: time, space, causality) and the subjective with the *particular* (merely empirical sensations).[83]

The framework developed by Daston and Galison, in their research on the history of objectivity through a close study of scientific atlases serves as precedent framework for our research. Daston and Galison outline four major stages, or epistemologies, in the development of scientific objectivity that follow a historical progression from the 18th to 21st century: truth-to-nature, mechanical objectivity, structural objectivity, and trained judgement. At each stage, there is a new way of seeing the world and describe to represent natural phenomenon.

Truth-to-nature, describes a process where abstractions are produced by an examination of a collection of individual species. In this stage, the scientist, artist, and engraver work together to study tangible specimens and natural phenomenon, and create generalizations or "types." The truth-to-nature artist, draws not what is seen, but an idealization of a universal type.

[82] Daston and Galison, *Objectivity* 29. Etymologically, the word derives from the Latin *obiectivus/obiective* and *subiectivus/subiective* introduced in the fourteenth century scholastic philosophers.

[83] Daston and Galison, *Objectivity*, 30.

Mechanical objectivity, is a stage that Daston and Galison argue as the moment where objectivity emerges in its modern form. The artist and the scientist produced images through machines, usually photographic media, with a drive toward automation. In an ideal mechanically objective observation, neither the hand of the artist, scientist, or machine maker would be present in the rendering of the natural phenomenon. While mechanical objectivity seemed to distance the scientist from the observation, it also proved difficult as a method of guiding a community — the purpose of making representation in the form of a scientific atlas — due to a sacrificed lack of abstraction.

The third practice Daston and Galison define is structural objectivity, or a desire to abolish images all together from scientific practice, in favor of equations, logical relationships, or sequences of signs.

If mechanical objectivity trained the scientist not to trust his eyes, the stage of trained judgement saw the integration of subjective abstraction and mechanical objectivity. The scientist knows more than his tools are able to tell (tacit knowledge) and therefore is given ethical jurisdiction to "smooth out," or make abstractions from mechanical representations of the world.

The fourth category is trained judgement. In this method of representation, mechanical objectivity is questioned and the tacit knowledge of the scientist is called upon to rectify the inconsistencies of a mechanically captured image.

In the experiments we surveyed in the previous section, we argue that the experimental psychologists studying vision occupy an epistemological position between mechanical objectivity and trained judgement. In these experiments, representations of human vision were created predominately with photographic techniques and custom instruments to capture eye movements.

Employing Daston and Galison's framework, Buswell's representations and scientific attitude can be situated as trained judgement. Buswell's hands are directly present in his published representations of vision. Movements of the eye on Buswell's mechanically objective film records, always appear as perturbations. There are no stationary points, and therefore foveal fixations must be inferred from the mechanical capture. Once the data has been interpreted by Buswell, fixations are mapped back onto the stimulus representations as points and connected by a poly-line. In Buswell's case, an attitude of pure mechanical objectivity would have only published the photographic records, despite their legibility. Or, one can imagine a scenario where an instrument could be developed, like a capture instrument in reverse, where the movement records, once captured and developed, are back projected onto the space of the original representation such that all movements would be preserved. However, it is not so easy to sever the instrument from representation from theory. Embedded in instrumentation are theories and experience of others.[84]

Yarbus, on the other hand, puts forth an attitude of pure mechanical objectivity, where representations are either presented as graphs or as direct records from mechanical devices he constructed. It is curious that Yarbus's attitude towards scientific representation, characterized by Daston and Galison as a predominant attitude in the mid nineteenth century would persist into the middle of the twenty first century. As new stages are defined

[84]. This mode of thinking is well elaborated upon in a discussion of navigation technology in: Edwin Hutchins, *Cognition in the wild* (Cambridge Mass.: MIT Press, 1995).

through the development of new instruments and shifts in scientific epistemologies, prior ways of image making are not eclipsed or overwritten. Instead, the ways representing the world are accumulated, and are “all still available as ways of image making and ways of life in the sciences today.”[85] In the original publication of his book, now heavily cited, images of a subject viewing Repin’s “Unexpected Visitor,” were presented in a grid with no attempt to superimpose the artwork stimulus with the human response.[86] In this book, we present Yarbus’s records as superimpositions of the photographic record and the original. The human observer may serve as a subject and the artwork as object during the experiment, but afterwards the human observer’s eye movements become the objective representation and the artwork a subjective stimulus.

It is in these experiments that we find the collision between art and science and reversals in the subjectivity/objectivity dipole. By using the frameworks discussed above, we have examined representations of vision and attempt to reveal how objectivity is constructed through different attitudes and practices in representation and instrumentation. Furthermore, the discussion of an intercalated periodization allows us to better situate not only the work of these historical actors, but our own research, where we are developing new instruments that may enable us to see the world differently – to see how we see. What should be taken away from this short sketch is that objectivity is not fixed, it is a construction that is manifest in representations. This process of constructing objectivity is always a negotiation or series of exchanges instrumentation, theory, and experimentation.

[85] Daston and Galison, *Objectivity,* “Epistemologies of the Eye”

[86] In our reproduction of these images, we have superimposed the movements of the subject’s eye on top of the images.

Pupil

Introduction

In the previous chapter we developed the background of this thesis in three parts with a physiological overview of human vision, a historical review of eye movement studies, and a discussion of shifting objectivity in representations of human vision. The historical survey of eye movement studies we conducted concentrated primarily on the relationship between a subject and a two dimensional object in the world. The background provides a strong theoretical and historical framework within which we situate our work. While majority of the historical literature review concentrates on the relationship between a human subject and a two dimensional object, the next section launches our inquiry into a human visual experience of three dimensions.

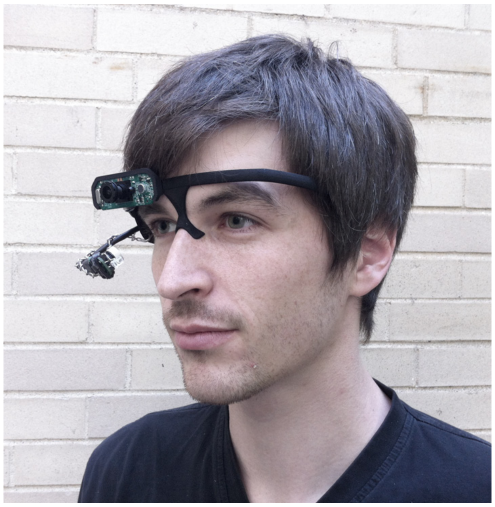
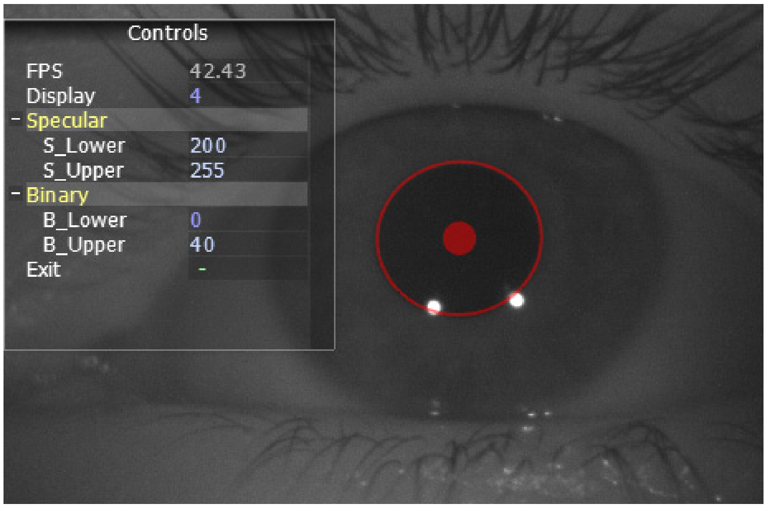


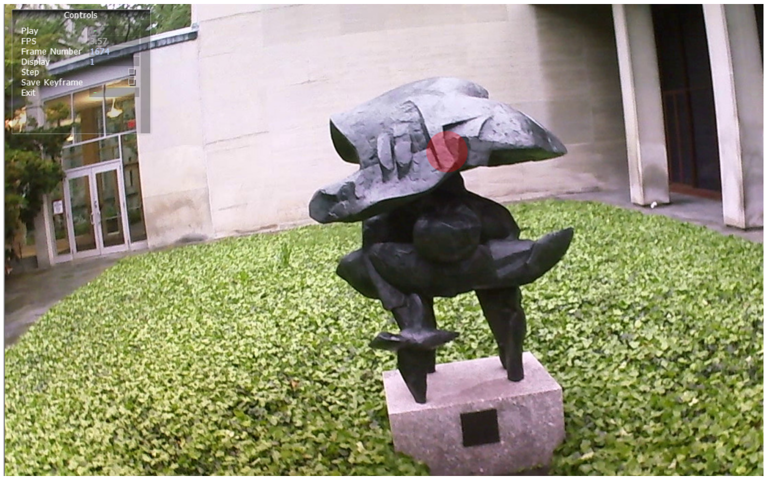
Figure 35: Author wearing the third revision of the headset prototype.

In this section we introduce **PUPIL**, a hardware and software prototype that we have developed in order to study the relationship between a human subject and a space. The first section of this chapter provides a detailed discussion of hardware development, from a low level discussion of real time image processing, camera control libraries, and the manufacturing of custom camera boards to a high level discussion of prototype design based on physiological constraints. In the second section we introduce a representation method that constructs a space of visual attention as a three dimensional point cloud from a sequence of two dimensional images captured by our hardware. It is within this space that we are able to situate the subject, and the area of their visual attention. The implementation of open source computer vision and Structure from Motion (SfM) algorithms used in this platform will be discussed in detail in this section. In the third section we conduct a series of trials that range from tests of instrumentation to inquiries into the relationship between a human and a space. The findings of the trials are analyzed and discussed in the final section of this chapter.



```
Controls
FPS          42.43
Display      4
- Specular
  S_Lower    200
  S_Upper    255
- Binary
  B_Lower    0
  B_Upper    40
Exit         -
```

Figure 36: Pupil center detection in the capture routine. Still image of eye taken as screen capture of capture software.



```
Controls
Play            -
FPS             3.57
Frame Number    1674
Display         1
Step
Save Keyframe
Exit            -
```

Figure 37: Two dimensional browser screen capture. Red dot shows the subject's area of visual attention. This is a still image from a trial conducted at the courtyard outside of Hayden Library at MIT.

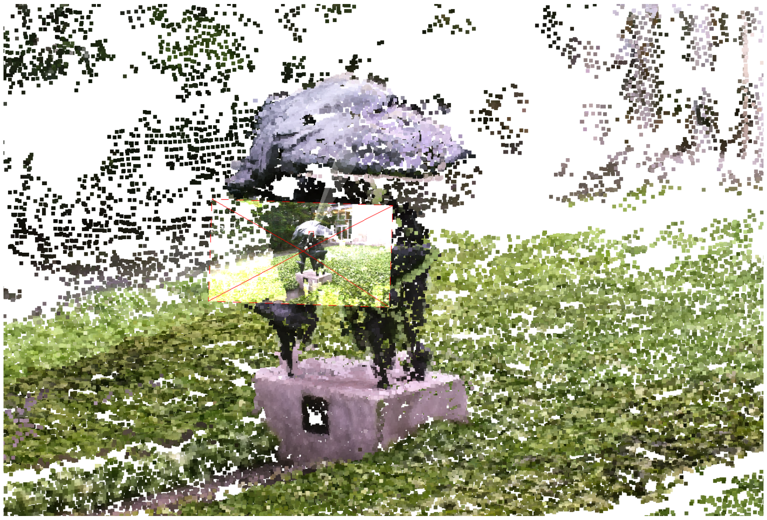


Figure 38: Three dimensional browser screen capture. Red pyramid shows the subject's pose in the three dimensional model, reconstructed from the trial shown in the previous figure. The white cone represents the three dimensional space of visual attention (shown as red dot in previous figure).

Hardware Development

Section Overview

In the introductory essay to this section we argue for the development of custom hardware due to desires of higher spatial and temporal resolution, cost, open accessibility, and future development. We present two prototypes and compare two prototypes, the first iteration using consumer grade web cameras and the second using our custom built cameras and software interface. This section provides a detailed description of the development of hardware — cameras, parallel processing XMOS micro-controllers, USB communication, head mounted prototypes – and software used to control/drive the software

Physiological understanding of human vision, as addressed in earlier section, is critical to the design of the prototype. These physiological constraints are incorporated into the design constraints of both prototypes in formal and electronic design. The last section puts bounds on tracking accuracy, due to sensor performance the physiological considerations.

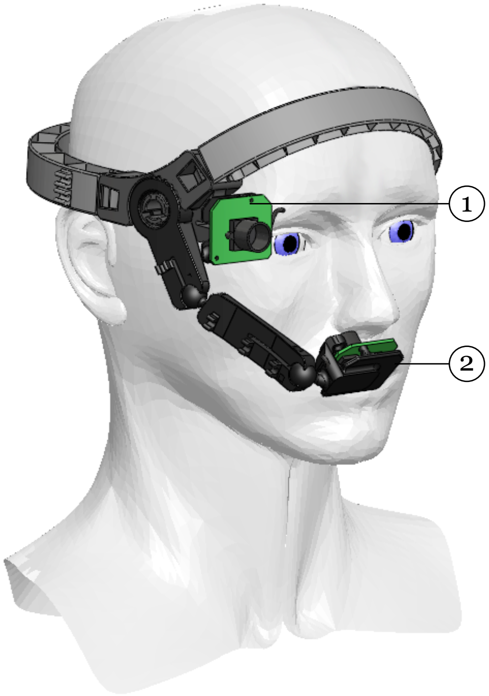


Figure 39: 3D-Cad model of the first prototype. Two "XboxLive" cameras facing outward [1] and looking at the eye [2].

First Prototype

The first prototype we developed used off the shelf consumer webcams, more specifically XBOX-Live webcams which sold for twelve US Dollars at the time of writing.

The headset was designed to be modular, and adjustable. It was fabricated using ABS on a Fused Deposition 3D-Printer. The headset is composed of seven modular parts that are connected at a "hub." The camera arm that holds the "eye camera" has two ball joints and one rotary joint that connects the arm to the hub element. The freedom of movement allowed for adjustment of the eye facing camera to accommodate different subjects. The mount plate for the world camera was attached via a ball joint to the hub, such that the world cam could be aligned to accommodate the subject's field of view.

The cameras adhered to the UVC camera standard and thus were accessible though OpenCV's camera capture module. The only modification that was made to the consumer cameras were the addition of an infrared band pass filter and matched infrared LEDs to the Eye camera. The setup for the first prototype was a good proof of concept in that it was accessible, easy to use, and affordable. However, the prototype had some disadvantages and limitations:

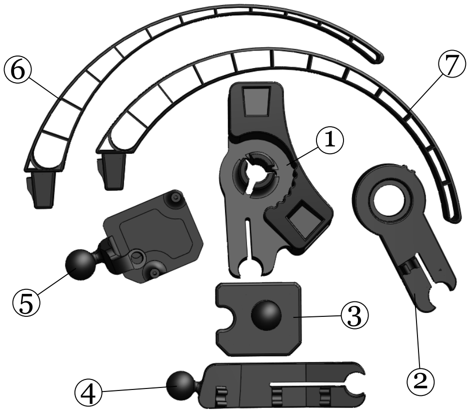


Figure 40: All components of the first headset are snap fit assembled. Centerpiece is the hub [1] it connects the head straps [6,7] to the camera mounts. The world-cam mount plate [3] and the arm elements [2,4] and eye-cam mount plate [5] are connected using a rotary joint and ball-joints. The joints allow the cameras to be adjusted to the user.

Size

Like most web-cameras the sensor chip is located on the same PCB as all other components. This makes the camera relatively large, especially when placed in front of the subject's eye, obstructing the field of vision.

Control

The camera was designed to be a easy-to-use consumer webcam. Therefore, low level control of the capture parameters such as exposure, frame size or raw pixel access is not possible.

Resolution

The Xbox Live camera provides a VGA image at 30 frames per second. The results we achieved with this camera where acceptable, but we wanted to investigate whether higher spatial — more pixels per frame —and temporal — more frames per second — resolution could lead to more robust and better pupil tracking.

Ergonomics

The first headset was modeled based on a idealized model of a human head. Due to the mass of the eye camera and its outward leaning arm the headset had to fit tightly on the head in order to resist any torque exerted by the inertia of the camera arm mount. The tight fit resulted in discomfort when the headset was worn for prolonged periods of time.

Obstructions:

The eye camera obstructs the subject's field of vision. In most cases one will adapt to small obstruction within the field of vision. In other cases obstructions may lead to dizziness or unnatural viewing patterns.

Appearance:

The headset is rather voluminous and its shape and color is reminiscent of lab and medical equipment. For experiments in natural environments with social interactions a less visible shape that is more associated with everyday headgear like eye glasses or sunglasses was determined to be more favorable.

New Prototype

In the iteration of the headset, we attempt to address the above issues. We addressed the problems we had with the camera by building a capture system from the silicon up and redesigned the headset based on the new constraints established in a review of the first prototype.

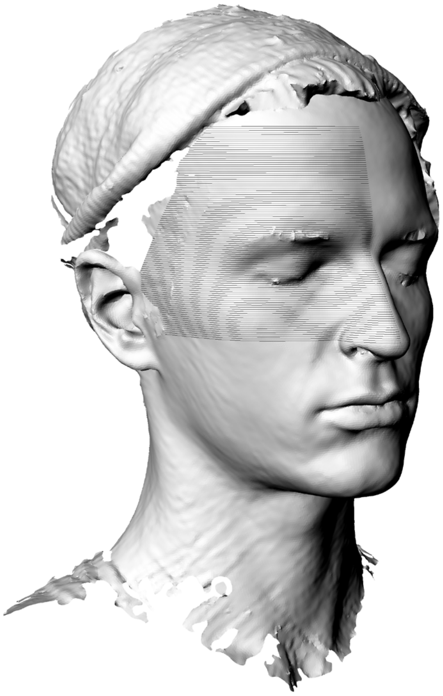


Figure 41: Mesh based 3D representation of one Authors face and head. The shaded area was chosen to be remodeled using NURBS surfaces to serve as a geometrical reference for the headset.

Anatomical Reference

In order to design an ergonomically precise headset, we required an anatomical reference that was more accurate than the idealized head we had used for the first prototype. We made scans of our heads using a portable three dimensional scanner. These scans provided high resolution meshes that were used as initial ergonomic design references. Second Prototype

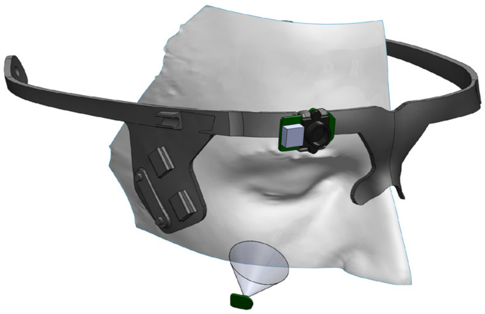


Figure 42: Ergonomic study. The shape of the headset is a precise geometric offset of the authors face and head shape.

Second Prototype

The first ergonomic study we developed was created as a mere offset of one author's head and face mesh from the three dimensional scan. The eye camera arm mount hub detail was significantly refined. In order to reduce camera sizes and weights, parts of the circuit boards would be mounted on the hub, in front of the subject's ear. The hub was designed to accept a two and a half dimensional snap fit arm that could be formed and modified during calibration.

As expected the prototype was overly specific and only suitable to be worn by the reference head. The prototype was comfortable to wear and light weight, but it was determined to be undesirable to have the headset touch moving parts of the face. Movements of the eyebrows, as the subject talked and gestured, caused the headset to move. Movements of the headset could cause the calibration and tracking of the eye to return corrupted data, therefore movements should be minimized.

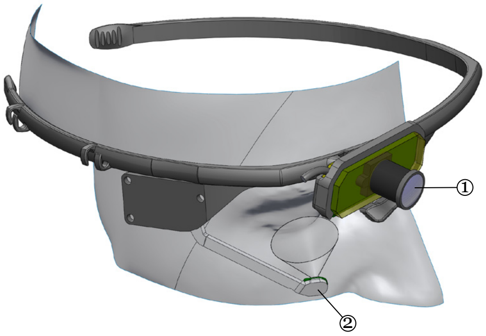


Figure 43: Second headset prototype. Pictured here with a mount accommodating a Logitech C525 Webcam [1] as the world facing camera, and a Omnivision Camera-Cube as the eye facing camera [2].

Third Prototype

Learning from the first revision of the prototype we created a second iteration with more defined contact points and a slightly more generalized in geometry in order to accommodate a wider anatomical range. In order to do this, we first refined the anatomical mesh in order to provide a continuous surface that could be used as a reference for the entire prototype. In order to hold the headset in place, the arms where pre tensioned,much like eye wear made for active sports.

Aspects of modularity were preserved in this prototype, where the camera mounts are designed to be snap fit onto a joint with one degree of freedom and can be replaced to accommodate many different revisions and models.

The third revision uses four times less material than the first revision, is more comfortable, less obtrusive for the subject's field of vision, and less conspicuous. Due to the reduced material volume, this prototype was cheaper to print than the old one. At the time of writing, the cost for printing one headset is twenty five US dollars, compared to the approximate price of one hundred US dollars for the first prototype.

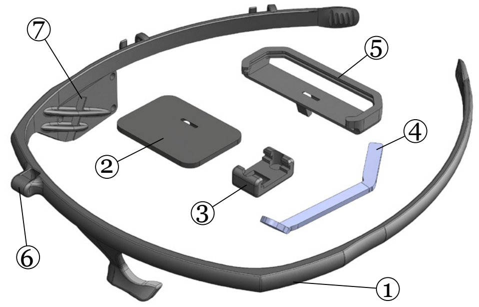


Figure 44: The disassembled headset: The Frame [1] has a rotary joint connector [6] that can be outfitted with different camera types: Generic camera mount plate [2], small camera mount clip [3], mount clip for Logitech C525 [5]. The eye-facing camera is attached to a aluminum arm that slides into the side hub [7].

Capture System Hardware

Capture System

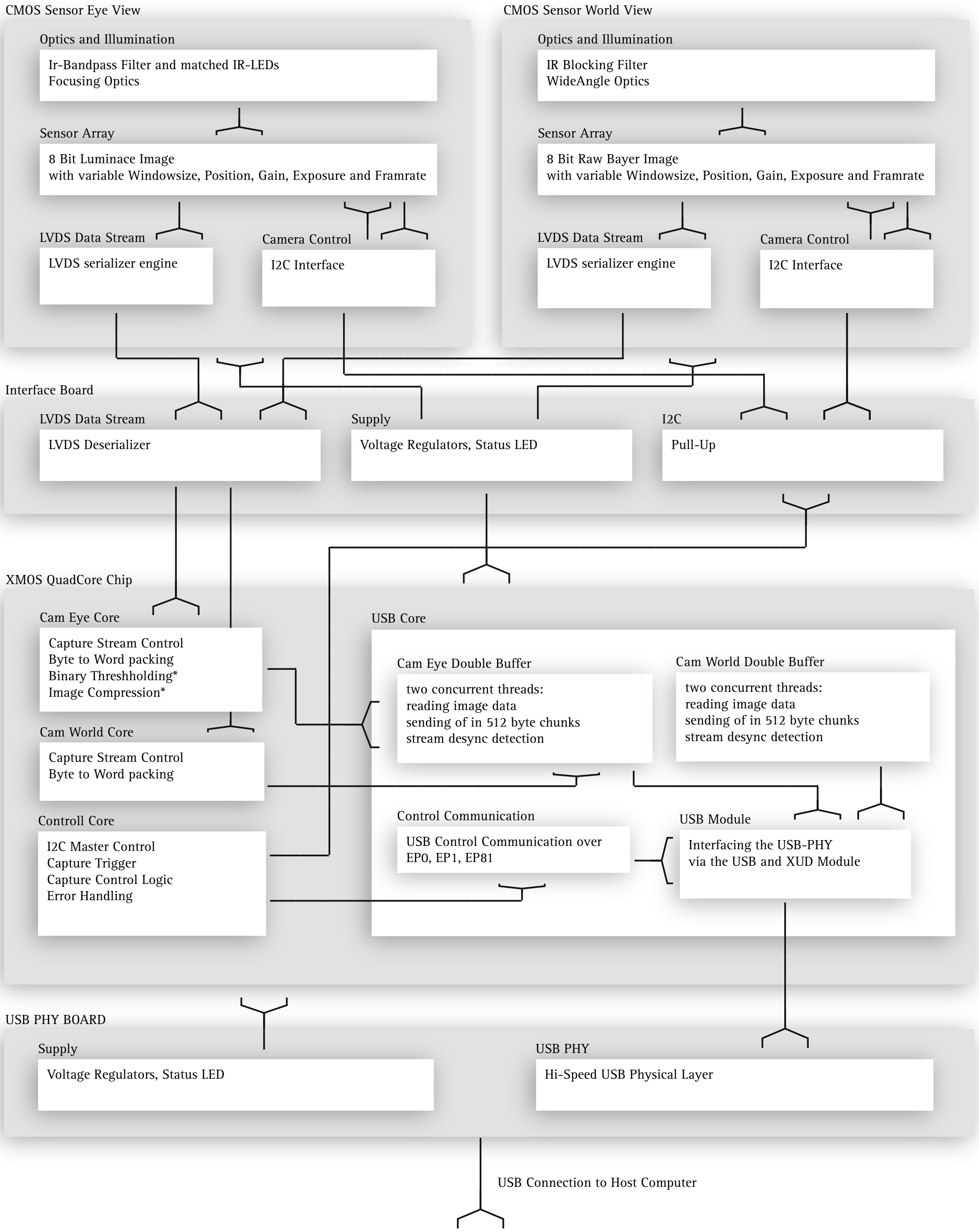
Designing our own capture system posed many challenges. Image sensors are high bandwidth data sources that generate byte streams at their own pace. If any component in the process chain is slow or unreliable the data stream looses synchronization and the image becomes garbled.

Designing our own capture system provides us with the benefit of low level control of the capture pipeline as well a flexibility in choosing the components, layout, and signal processing strategies.

CMOS Camera Chip

The CMOS sensor is the interface between optics and digital electronics. Normally invisible, it sits under the camera optics, where you would find

Figure 45: [Facing Page] Functional component overview and data-flow visualization of capture hardware.


CMOS Sensor Eye View
Optics and Illumination
Ir-Bandpass Filter and matched IR-LEDs
Focusing Optics
Sensor Array
8 Bit Luminace Image
with variable Windowsize, Position, Gain, Exposure and Framrate
LVDS Data Stream
LVDS serializer engine
Camera Control
I2C Interface
CMOS Sensor World View
Optics and Illumination
IR Blocking Filter
WideAngle Optics
Sensor Array
8 Bit Raw Bayer Image
with variable Windowsize, Position, Gain, Exposure and Framrate
LVDS Data Stream
LVDS serializer engine
Camera Control
I2C Interface
Interface Board
LVDS Data Stream
LVDS Deserializer
Supply
Voltage Regulators, Status LED
I2C
Pull-Up
XMOS QuadCore Chip
Cam Eye Core
Capture Stream Control
Byte to Word packing
Binary Threshholding*
Image Compression*
Cam World Core
Capture Stream Control
Byte to Word packing
Controll Core
I2C Master Control
Capture Trigger
Capture Control Logic
Error Handling
USB Core
Cam Eye Double Buffer
two concurrent threads:
reading image data
sending of in 512 byte chunks
stream desync detection
Cam World Double Buffer
two concurrent threads:
reading image data
sending of in 512 byte chunks
stream desync detection
Control Communication
USB Control Communication over
EP0, EP1, EP81
USB Module
Interfacing the USB-PHY
via the USB and XUD Module
USB PHY BOARD
Supply
Voltage Regulators, Status LED
USB PHY
Hi-Speed USB Physical Layer
USB Connection to Host Computer

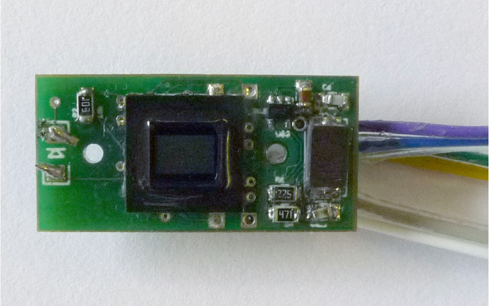


Figure 46: Aptina CMOS sensor chip on custom designed and fabricated PCB.

the film in an analog camera. Before continuing, we provide a very brief overview of the working principle of a CMOS camera sensor.

Every time a photon impacts a special semiconductor called a photo transistor, it makes this piece of silicon conductive, allowing a bit of charge to run into a reservoir. The greater the amount of light, the more photons impact, the greater the charge that flows into the reservoir and thus, the higher the potential (voltage) of this reservoir. On a WVGA camera chip, there are 480 rows by 752 columns of such light sensitive transistors, each with its own reservoir.

An analog digital converter (ADC) reads the voltage of each reservoir and sends out a result coded in 8 bit resolution. These 8 bits are presented on 8 data lines that are accompanied by a Line Valid, Frame Valid, and Data Clock Signal coming from the CMOS sensor. For one WVGA frame this makes 360,960 8 bit values. After readout, the reservoirs are dumped and the process repeats to display the next frame.

Microcontroller

The CMOS chip cannot operate on its own, it has the be set up and started using a external control chip. The pixel stream coming from the CMOS chip has to be converted and put into packets and transmitted to the host via a communication interface like Firewire, USB or LAN. Until recently this was the exclusive domain of Field programmable Arrays (FPGA) or specially designed silicon chips. While FPGAs are a little bit more flexible, they are notoriously difficult as a development platform. Specifically designed silicon chips do not leave room for changes and during the process of building a

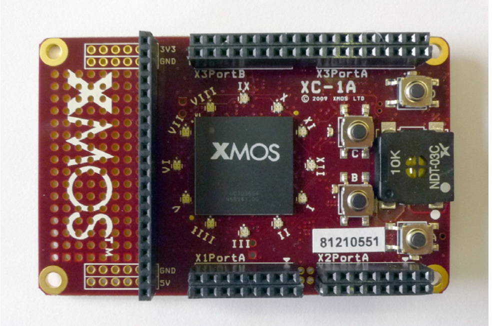


Figure 47: XMOS Development board XC-1A. Quad core processor 1,600 MIPS.

camera capture device it becomes a laborious exercise of checking data sheets and connecting matching components.

We have investigated two other hardware options: An ARM-cortex M4 based Microcontroller with a dedicated camera interface, USB-controller-interface and multiple Direct Memory Access (DMA) streams to shuffle the data from A to B. The ARM-cortex development board seemed to be a valuable option at first, but the process of developing firmware code resulted in scouring over 3000 plus pages of data sheets and configuring dedicated hardware. This approach felt unrewarding and risky as there was a set limit to the bandwidth of the controller and it was neither explicitly stated nor researchable for our application scenario.

The second option was a new hardware architecture that followed a diametrically opposite approach to the previously discussed microcontrollers. Instead of creating more and more dedicated silicon to perform special tasks, this architecture performs all tasks in software. This is accomplished in using up to 4 cores, running 8 threads on each microprocessor, providing up to 1600 million instructions per second as well as a single cycle Port I/O at 100 MHz sampling rate.

XMOS Embedded Software

In order to accommodate the concurrency of tasks running on a multi threaded platform as well as low level access to ports, the Microcontroller manufacturer developed a language based on C called XC. XC serves as the primary language for the XMOS Microcontroller architecture. Using this language, development is made possible by using an IDE or a set of command line tools, both are available for the major operating systems (Windows, Linux, Macintosh).

Unlike most microcontrollers XMOS does not use interrupts, instead it uses timing sensitive operations that can run concurrently. Threads have exclusive access to ports and single cycle access. Saving a port-state to RAM can be done in two cycles.

In the XMOS architecture, ports have built in buffering and can be clocked from various sources. This buffer enables the development of simple but flexible serializer and de-serializer applications which increase port bandwidth and simplify the development process.

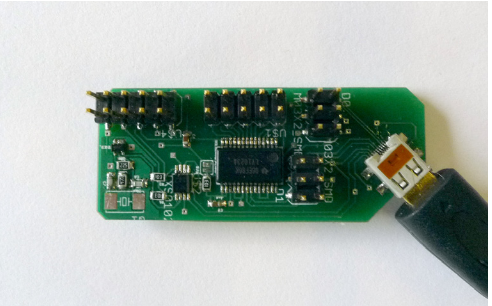


Figure 48: LVDS Serializer and HDMI interface board. Designed to fit on the headset. Custom designed PCB.

Low Voltage Vintage Differential Signal Transport

The data clock of the CMOS chip we use is 26 MHz, at these rates, special care has to be put into the routing and wiring of data busses. To allow for reliable transmission from the CMOS sensor to the Microcontroller, we used a Low Voltage Differential Signaling (LVDS) module on the CMOS chip and a Random Lock De-serializer added to the Microcontroller board to Serialize the 8 data lines with their Line and Frame Valid Signals into a single differential pair.

Using LVDS the data bus frequency is increased 12 fold but all 13 data lines are reduced to single line. This signal along with its complement is transported on a pair of wires. Any noise the signal is exposed to on its way to the de-serializer is added as common mode and can be filtered out by the de-serializer. The de-serializer presets the camera data stream to a 8 bit XMOS port along with Frame Valid, Line Valid, and Data Clock which are fed into three 1 bit ports.

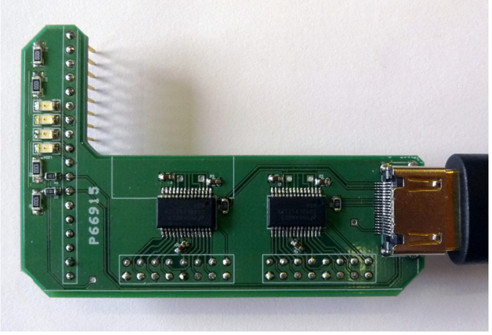


Figure 49: Dual LVDS De-serializer and HDMI interface board. Add on board for the XMOS XC-1A Development Board. Custom designed PCB.

Inter-integrated Circuit Embedded Control Bus

Control commands and status reports are communicated via the inter-integrated circuit (I2C) bus. This two wire interface allows low bandwidth multipoint communication. In our case the XMOS control chip talks to two sensor chips via the I2C bus.

Communication between Capture System and Host Computer

We use High Speed USB 2.0 to stream the data from the Capture System to the Host Computer. Choosing the Interface was aided by a small survey of possible communication Interfaces:

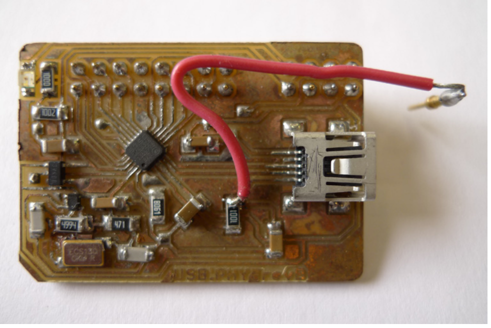


Figure 50: USB PHY add on board for hi-speed USB communication.

Bandwidth requirements: Device to Host

The CMOS sensor we use can read WVGA (752x480) pixel images at up to 60 frames per second, the Pixel Depth is 8 bit leading the following bandwidth requirements:

$$30fps*752*460*8bit = 10,828,800\ \frac{byte}{\text{sec}}$$

$$60fps*752*460*8bit = 21,657,600\ \frac{byte}{\text{sec}}$$

These numbers present a theoretical bandwidth requirement. Because the Microcontroller does not have enough Ram to buffer a full image and the image stream coming from the CMOS sensor is not fully homogenous in timing (the actual bandwidth of the chip is higher and image information is transmitted only during line valid high states), our actual bandwidth requirement is approximately 2 times greater. In addition, the protocol overhead has to be taken into account on USB. Achievable rates that are usually 80% of their theoretical maximum.

Interface Survey

LAN, easy to access from the host side , UDP would be a suitable protocol.

100BaseT = 12.5 Mega Byte/sec.; Dedicated controllers as well as Integrated solutions are widely available.

1000baseT = 125 Mega Byte/sec. ; Controllers and PHYs for embedded systems are rare.

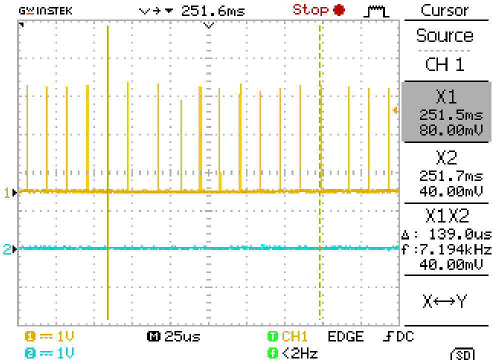


Figure 51: The Oscilloscope screenshot shows the state of a output pin that is flashed every time 512byte are sent through the USB interface.

USB, wide spread interface of host-centric, short distance communication.
Low-Speed: 1.25 Mega Byte/sec.; Slow and deprecated

Full-Speed: 12.5 Mega Byte/sec.; Wide spread, PHY often integrated in the uC (STM32F4 and many other Cortex M3+ devices), dedicated uC with parallel interfaces exists (FTDI).

Hi-Speed: 60 Mega Byte/sec.; Uc's with USB controllers either in silicon (STM32F4) or software (XMOS) exist.

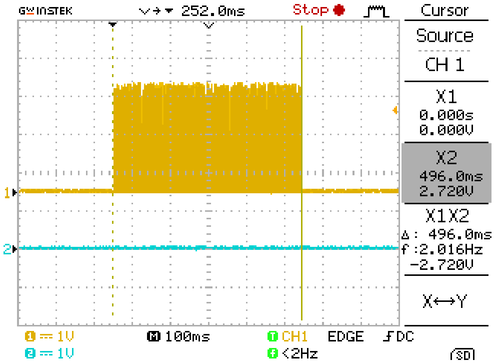


Figure 52: This screenshot shows a cluster of "flashes" that represent data worth 60 vga frames send in .5 seconds. This data shows that the XMOS chips is able to reach USB data rates of ≈40Mbyte/sec.

XMOS USB Bandwidth

XMOS USB performance was a major point of uncertainty. XMOS does support Hi-speed USB, the question was whether the software module was fast enough to sustain data rates that would saturate the USB interface. We bench marked by sending a endless stream of 512 byte chunks of test data. The oscilloscope screen capture shows the state of a output pin that is flashed every time 512 bytes are sent through the USB interface.

USB Host Interface

While hi-speed USB is fast enough for our purposes, USB can be tricky to implement on the host side. Sustaining continuous streams of high data rates can become a demanding task. We used a USB library that allows access to USB devices attached to the computer called libUSB.

libUSB is a USB Library written in C. We created a capture function that is compiled as a shared module and can be called from a Python environment. The capture function looks for the right PID/VID signature, opens the device and claims the appropriate interface. The Python module we wrote exposes functions to read and write control registers and grab image frames.

Hardware Components

As a development platform we choose the XC-1A development board. The additional PCBs we designed and fabricated were: USB PHY add-on, HDMI interface and de-serializer board, micro HDMI breakout and power supply board, and a CMOS camera chip board. Note the additional serializer IC on the micro HDMI breakout board. It allows one to interface with a camera chip that does not have LVDS output. The HDMI cable was used because it has multiple twisted pairs for LVDS signal transport, additional single wires for I2C and power, and good shielding.

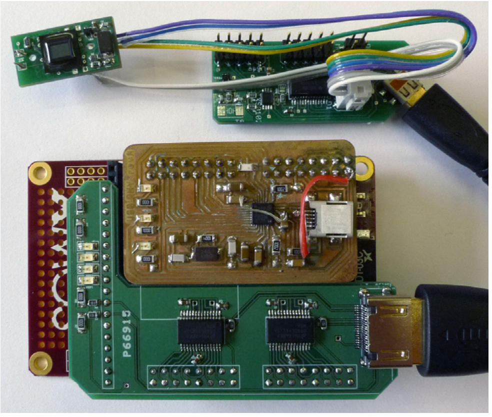


Figure 53: Assembled components, CMOS censor, Interface Boards, XMOS XC-1A Development Board, and USB PHY.

Conclusion

In summary, the development of the third headset prototype as been a success. The new headset is more comfortable, cheaper to produce, and less conspicuous in that it more closely resembles everyday eye wear.

Developing the capture system turned out to be more work than initially anticipated. The current stage of development for the capture system, can be described as a first working prototype. Future work will enable the capture system to become a mature image capture tool chain that could become a contribution to the computer vision community in the realm of eye tracking and beyond.

Capture System Software

Capture Software

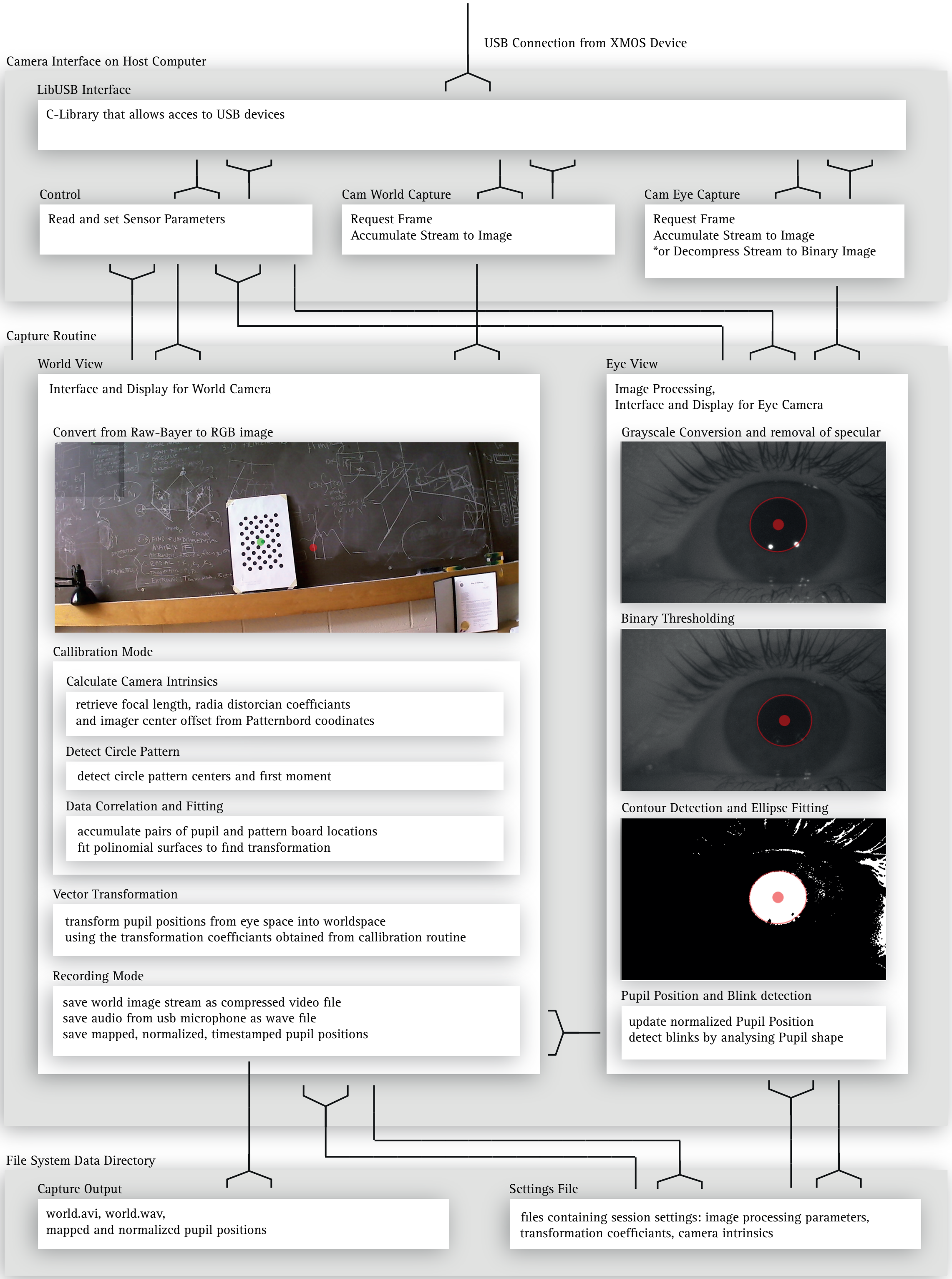
The video stream from the eye and world cameras alone do not lead to much insight into the nature of human visual perception of his surrounding environment. In order to build an understanding of how the subject's eye attends to the surrounding world, we need to process the image captured from the eye camera. In our capture routine we track the movements of the eye by isolating and extracting the location of the pupil centroid.

Given that the eye and world cameras do not change their positions relative to the subjects tracked eye – as they are securely mounted on the subject's head – we make the assumption that the position of the pupil centroid directly correlates to the position of the subject's space of foveal vision, or gaze, within the field of view of the world camera. After performing a calibration process, the pupil position can be mapped into the space of the world camera. Based on a history of physiological precedents, we make the assumption that the center of the pupil in a subject with "normal" vision equates to the center of visual attention – or foveal vision. As the capture routine runs at speeds higher than the human limits of perception ,we can assume that we are able to capture all relevant physiological movements of the eye.

Eye Process

Image processing, and position mapping has to be at least as fast as the image refresh rate. In our capture routine, the eye image and the world image are processed in separate threads to allow for more CPU time and independent frame rates. Most image processing operations are implemented using the Open Computer Vision Library (OpenCV) for its ease of use, accessibility, and speed.

Figure 54: [Facing Page] Functional component overview and data-flow visualization of capture software.



USB Connection from XMOS Device

Camera Interface on Host Computer

LibUSB Interface

C-Library that allows acces to USB devices

Control

Read and set Sensor Parameters

Cam World Capture

Request Frame
Accumulate Stream to Image

Cam Eye Capture

Request Frame
Accumulate Stream to Image
*or Decompress Stream to Binary Image

Capture Routine

World View

Interface and Display for World Camera

Convert from Raw-Bayer to RGB image

Callibration Mode

Calculate Camera Intrinsics

retrieve focal length, radia distorcian coefficiants
and imager center offset from Patternbord coodinates

Detect Circle Pattern

detect circle pattern centers and first moment

Data Correlation and Fitting

accumulate pairs of pupil and pattern board locations
fit polinomial surfaces to find transformation

Vector Transformation

transform pupil positions from eye space into worldspace
using the transformation coefficiants obtained from callibration routine

Recording Mode

save world image stream as compressed video file
save audio from usb microphone as wave file
save mapped, normalized, timestamped pupil positions

Eye View

Image Processing,
Interface and Display for Eye Camera

Grayscale Conversion and removal of specular

Binary Thresholding

Contour Detection and Ellipse Fitting

Pupil Position and Blink detection

update normalized Pupil Position
detect blinks by analysing Pupil shape

File System Data Directory

Capture Output

world.avi, world.wav,
mapped and normalized pupil positions

Settings File

files containing session settings: image processing parameters,
transformation coefficiants, camera intrinsics

The following steps are taken for each frame of the eye camera stream to extract the pupil center:

The image is preprocessed by removing the spectral reflections of the IR-LEDs.

Binary thresholding is used to convert the gray scale image into a binary image. Where the dark pupil becomes white, with a value of 255, and black everywhere else.

Calculating the x and y spatial image derivatives we obtain region boarders.

Ellipses are then fitted around these regions and the biggest ellipse is taken to be the one describing the pupil.

This method reliably describes the pupil position, which the eye image process reports to the world camera stream process.

World Process

The world process receives the pupil position, it then calls a mapping function which projects the pupil position into the world image space using transformation coefficients obtained from a calibration routine described later in the chapter. The world process concurrently receives world image frames and converts them to the appropriate color format.

The world view process has the capability to detect the centers of a circle or chessboard pattern grid. We use this functionality for two purposes: to find a transformation from the space of the eye view to the space of the world view, and to estimate the camera intrinsics.[1] The pupil transformation coefficients are used to describe two surfaces that are used to project the pupil position from eye space into world space.

87. Camera intrinsics are discussed in the chapter on Software.

Calibration and Transformation

During the calibration routine the subject is instructed to keep their gaze fixed on the center of the circle gird pattern, while moving his head from side to side and up and down. The world process detects the moving pattern in the world view and stores the centroid of the pattern in a list together with the pupil positions, correlated in time. After collecting enough points to cover the extrema of the world view boundaries, the two resulting lists of points are saved and passed to a fitting function. Two point clouds from each list of points are constructed that each represent a surface which transforms the pupil centroid coordinate value into the space of the world view. This surface is then approximated by a polynomial surface using Singular Value Decomposition. The number of coefficients is important for a transformation that describes the system well. Choosing to few coefficients reduces the transformation surfaces into planes, resulting in a poor fit. The following graphs show the calibration point could and the respective surfaces. The x and y dimensions are the position of the pupil in pupil space. The Z. dimension is the x (for the blue plot) and y (for the red plot) position of the circle pattern.

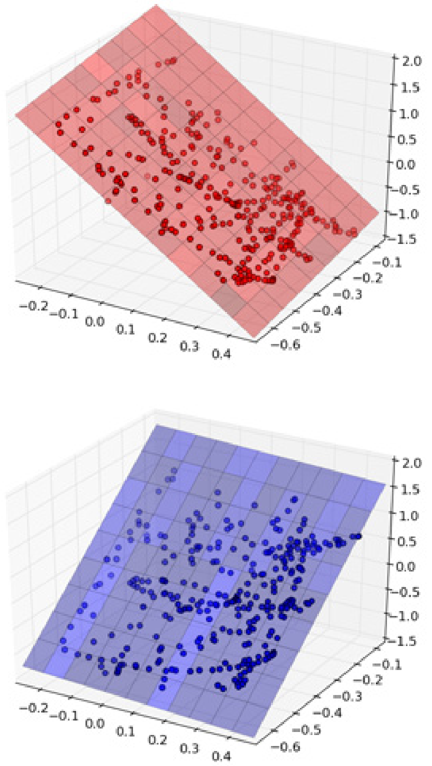


Figure 55: First order multivariate polynomials result in simple planes which poorly fit the data.

In this first case following equations are used, note that the coefficients for World$_x$ and World$_y$ are not identical:

$$world_x = c_x x + c_y y + c$$

$$world_y = c_x x + c_y y + c$$

Due to the rotation of the eye-ball that is projected onto the camera as well as radial distortion of the lenses in both eye and world camera, the eye space does not linearly translate into the world space. Using planes therefore results in a bad fit.

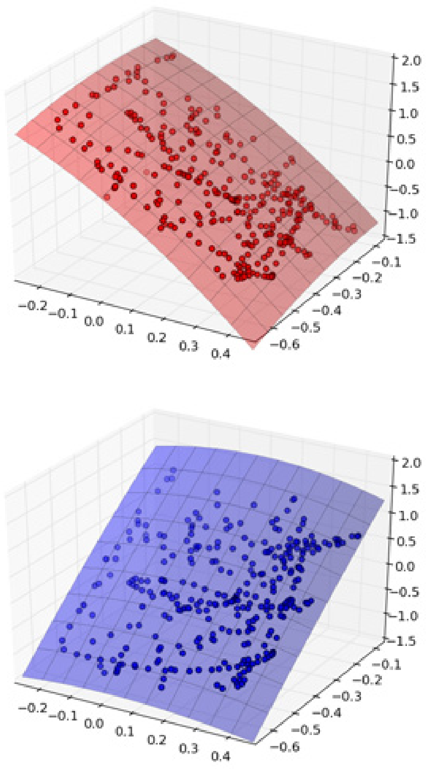


Figure 56: Fitting a second-order multivariate polynomial using singular value decomposition gives a better estimation of the transform.

In this case the following equations was used:

$$world_x = c_{x2}x^2 + c_{y2}y^2 + c_{xy}xy + c_x x + c_y y + c$$

$$world_y = c_{x2}x^2 + c_{y2}y^2 + c_{xy}xy + c_x x + c_y y + c$$

This system of second order bi-variate equations describes the system better while remaining simplistic. Using third or fourth order equations did not improve accuracy.

Next steps of improvement are using more advanced fitting algorithms like RANSAC and selecting a model tailored to the geometric constraints of the system.

After calibration is complete, the position of the pupil is mapped into the world space. This transformation is made possible because of the calibration process. Now, instead of knowing the point the subject is looking at, as we did during calibration, we use the transformation plane and are able to calculate the point the subject is looking at. This works well only when the calibration routine results in meaningful transformation coefficients. The calibration results can be quickly verified through a simple verbal check to see if what the subject is looking at aligns with the point in the world view.

The secondary purpose of the circle pattern detection method, is to obtain the camera calibration matrix, for three dimensional scene reconstruction. This matrix captures the intrinsic properties of a camera within a three by three matrix, typically notated as K. The calibration matrix is an essential element used in constructing a three dimensional scene from the world view video. The specifics of the calibration matrix are discussed in the chapter on software.

After calibrating we can conduct a trial with a subject moving through a space. In order to capture the results for analysis we will need to save audio, video, and pupil positions. The world view process contains functionality to save the world video stream alongside a audio recording from a USB microphone and a list that contains time intervals and pupil positions that are each associated with a video frame.

Bounds on Tracking Accuracy

While tracking is quite good, bounds on accuracy exist. These bounds originate from physiological constraints and can be further worsened by sensor and sensor evaluation process noise as well a conceptual flaws.

Physiological constraints
Why is the eye not a good pointer?
Even with the limited overview of the physiology of the human eye and human vision in Chapter 3, important conclusions can be drawn. With ~1° FOV the area of acute vision is small but still much bigger than the features we can resolve in the center of our field of vision. When we fixate upon an area, the human eye does not require the object of interest to be perfectly centered on the fovea, instead just being within it suffices. Since the positioning of the gaze is involuntary we can not force our eye to do better than that. Additionally small movements within fixation like micro saccades introduce further variance. The exact function of these small saccades is a highly debated topic. One theory is that These small movements ensure that the individual photo receptors are continuously stimulated. Another is that they are a corrective measure to compensate for drift. Micro saccades usually do not move the eye more than .2°. This limits the accuracy to which the

point of visual interest can be determined is 1° in the visual field of view.

Sensor and calibration accuracy

The pupil position is estimated from the image of the eye. As with all sensor systems measurement noise can be reduced but is never fully eradicated and will be present in the readings. More sophisticated image processing and evaluation strategies promise to reduce error and will be incorporated in the future.

As mentioned earlier the gaze tracking system requires calibration before each run, process variance due to the physiological constraints is therefore already present in the calibration data. Variance in the calibration data results in variance in the transformation coefficients which in the worst case has to be added to the already present variance.

This resulting inaccuracy determines the minimum size of the target object used in our visualizer, as anything that is smaller than approximately 1.5 degrees would not reflect the physiological realities of human vision.

Software Development Overview

In this section we introduce a software component of **PUPIL** that produces a representation of visuospatial attention patterns within a three dimensional space. In this method of representation a subject's patterns of visual fixation — eye movements captured with our custom hardware and software platform — are integrated and used to construct a three dimensional representation of a space.

The three dimensional model produced by our software serves as a representation of the space of a subject's visual attention. We argue hat this method of representation is insightful and novel for the disciplines of spatial design and those studying relationships between humans and constructed environments, as it situates a human subject as a central actor in the development of spatial representations, not as a product or result of representation.

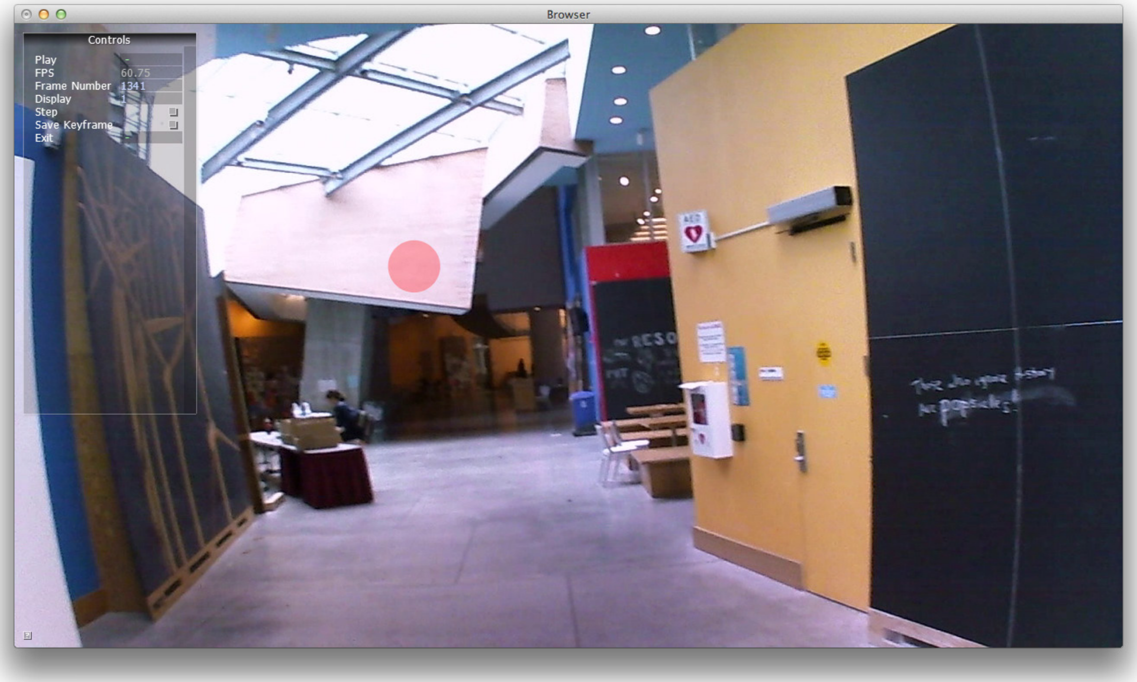
This method of representation while highly subjective — in that it is unique to one subject's relationship with a specific space and time – is also precise and can be used to provide quantitative information on the space of visual attention. This way of seeing, or representing vision, acts as a critique of existing representations of vision specifically in the associated fields of spatial design. In existing practices, "objective" representations of vision are based on a mechanical or geometric analogies. Photography and cinematography and photorealistic renderings are methods used to represent visions of space, or vision in space. However, these methods of representing vision are limited, both spatially and temporally, reducing the experience of human subject to a single vantage point locked in time. Even the most "truthful," "objective," or "realistic" representations of human vision are full of treason. But why? From studies of physiology, we know that humans can not resolve but a tiny fragment of our surrounding environment. But those who represent visions of space continue to produce images based on the paradigms inherited from the camera obscura, stereoscope, and linear perspective.

We are in search of an alternative way to represent the nature of human experience. There is no way to "objectively" represent the space of a human experience. Nor do we propose such solution. We are attempting to represent vision in a way that reveals some of the physiological challenges of human visual attention and simultaneously the reflexive bias of our own methods. The visual experience of space, is fragmented and partial, but the amount of information that humans are able to process from this fragmented view is extraordinary. In order to understand how humans visually attend to a space, we have created **PUPIL** — a collection of tools and representational methods — that situate the human subject as the constructor of representations. In these constructions conscious and unconscious aspects of experience in a space are combined. We believe that this mode of representation will lead to insight into a human experience of space and as the first step toward a new way of thinking about how we represent the world.

While the retinal surface may capture something analogous to "images" of the world, we believe that there are no representations of space in the human brain. Following literature in cognitive science and physiology of vision, we assume that the process of vision does not reconstruct an image in the brain of the human subject. Rather, we support the argument that vision is information processing. Our tools and sensors, analogs to the retinal surface of the human eye, capture images of the world, but to a computer these images are just information that must be processed. Our computer knows nothing of space, nor how to represent it given a two dimensional stream of information. The tools we have developed rely on a series of software libraries that enable the construction of a three dimensional representation based on this information stream. This process is called Structure From Motion (SfM), where a sequential stream of images, extracted from a video captured with our hardware, are used to calculate the pose of the subject and their point of visual attention in a three dimensional space. With these methods we merge the physiological reality of human visual attention with a mechanical construction of vision. Two modes of vision, both partial and fragmented in their own right, reflexively revealing the nature of their own mechanisms of construction.

These next series of components of **PUPIL** that we will describe constitute the representational tools of our platform of software tools, taking information gathered during the capture routine and post processing the videos, audio, and **PUPIL** positions in order to construct a representation of a human experience. We will begin this section by describing the development of a two dimensional browser that is used to visualize the results of a trial by representing fixation points and synchronizing audio with a video feed. The two dimensional browser is also used to select keyframes from the video that will be used in the reconstruction pipeline. Once a trial has been verified and keyframes are stored to a list, we extract the frames from the video and write meta-data to the extracted images based on our calculations of camera intrinsics from our calibration routine. With these selected keyframes and extracted still frames, we then proceed to our Structure from Motion Pipeline. The Structure from Motion Pipeline takes images, extracts features from the images, matches features, and calculates three dimensional locations of matched features. Once we have calculated a construction with the SfM Pipeline we are able to merge results from our trials into a single representation of a three dimensional space in our three dimensional browser.

Two Dimensional Browser



Play	-
FPS	60.75
Frame Number	1341
Display	1
Step	□
Save Keyframe	□
Exit	-

Figure 57: Screen Capture of the Two Dimensional Browser recorded with our capture hardware (Logitech C510 Camera). Video recorded in the Stata center, MIT, Cambridge, on May 06, 2012. The red dot represents the fixation point of the subject at keyframe 1341.

The two dimensional browser has two main stages and purposes. First, to visualize the results of a trial by playing back synchronized audio and video and rendering the calibrated pupil positions on each frame of the video. Second, the two dimensional browser is used to prepare the video for the SfM Pipeline.

The first stage of the two dimensional browser can be used solely for the visualization of results from a trial, where one wants to verify that the calibration was successful or for the presentation of results. The browser coordinates the playback or examination of individual frames by loading three files captured during the trial: video from the "world camera," audio from the head mounted microphone, and the pupil positions. The pupil position is represented either as a red dot, superimposed on the image stream, or as a mask that isolates only the space of visual attention. The red dot represents the space of foveal resolution

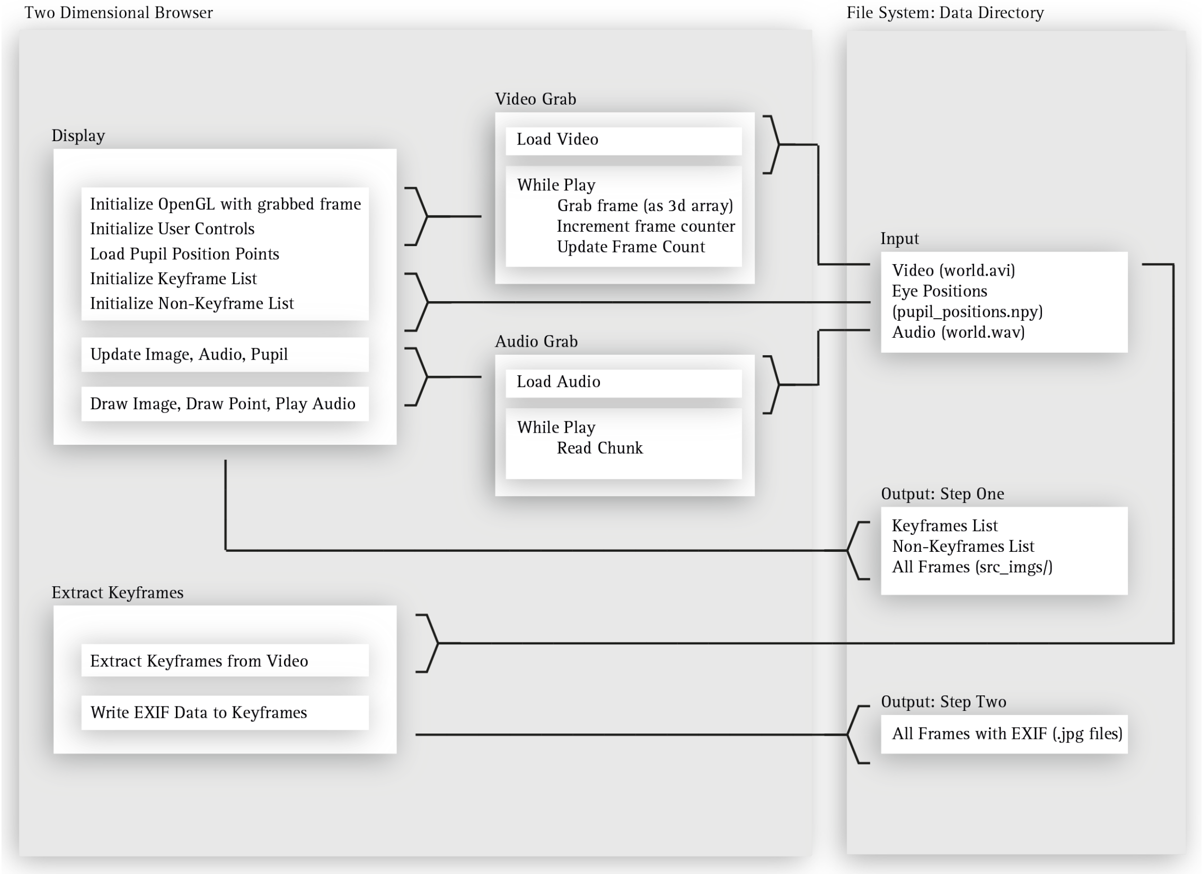


Figure 58: Diagram of the Two Dimensional Browser software and file system. The the gray box on the left shows the program flow of the two dimensional browser. The gray box on the right shows the input and output files of the browser at different stages in the process.

of the subject. This means that everything outside the diameter of dot is viewed by the subject with diminishing resolution and in gray scale. The size of the dot is taken as an approximation of the area of foveal acuity. In order to play the video we created a lightweight frame grab routine that depends on the open source computer vision library (OpenCV) and video back-end (FFMPEG) to decode the compressed video file. Once the video is successfully loaded, a single frame is extracted, stored as a three dimensional array, and put into a queue that is then read by OpenGL display loop that runs as its own process. The first frame of the video is passed to the open OpenGL display loop in order to set the size of the display window and to initialize the OpenGL texture

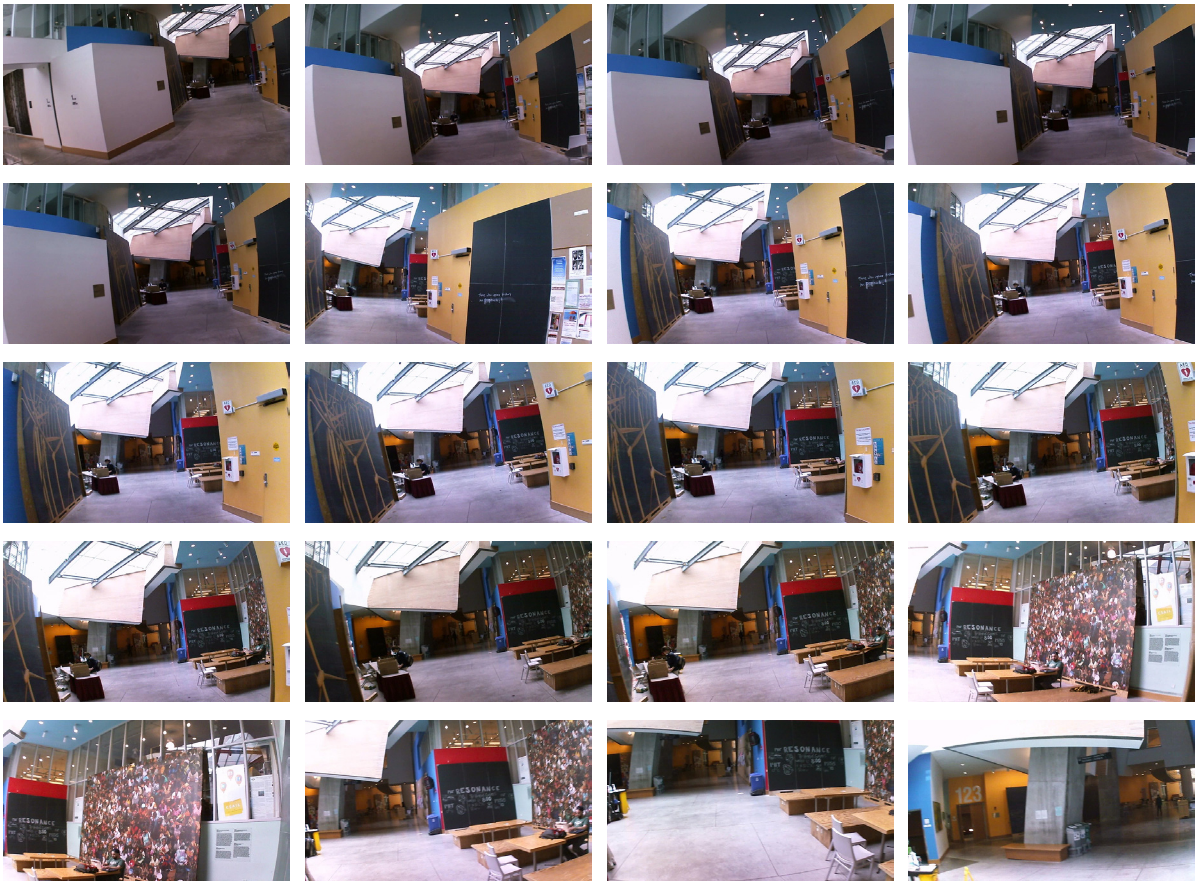


Figure 59: Example set of twenty keyframes extracted from a video file recorded with our capture hardware (Logitech C510 Camera). Video recorded in the Stata center, MIT, Cambridge, on 05/06/2012.

object that will be used to render the array in the window. The grab routine and browser now waits for the user to play the video or step through the frames sequentially before reading any further frames. If the user plays the video from the user interface control panel, then the grab frame routine continues to grab frames and add them as arrays to the queue that is sent to the display process.

In our capture routine, we record audio and video files separately. This means that we have to create a separate method to play back the audio and insure that the audio is synchronized with the video. The audio grab routine follows a similar procedure as the video grab routine. The audio files are recorded and encoded as .wav files. This simplistic encoding format samples an audio signal at a given rate and saves the signal as a string of sixteen bit values. In order to sample an audio stream we use PyAudio, a Python language wrapper for PortAudio. The number of audio channels, rate, and sample width is encoded into the file format. In order to play the files, we simply open the wave file and start an audio stream with the encoded meta-data in the .wav file. Just as we write the file we can read the string, or play the audio, by reading a specified segment of the string or chunk of bytes in succession until the there are no more bytes to read or the user pauses the video.

Alternatively, one can step through the frames one by one in order to check the quality of the frames and the pupil position within the frame. Frequently frames will be blurry or distorted, due to rapid head movements, rolling shutter behavior from the "world camera," or contain artifacts from video compression. Distorted or out of focus images are not useful, and often detrimental to the success of our SfM reconstruction.[87] Therefore, we have devised a routine that allows one to browse through the video and hand pick frames to use for reconstruction. Frames that a user selects are saved to a list of what we have called "keyframes." All frames between the beginning and end of the keyframe selection that are excluded from selection are saved to a list named "other frames."

Once the desired amount of keyframes are selected from the video and the "keyframe" and "other frame" files are written the browser exits and calls the frame extraction subprocess. Frame extraction is performed for all frames in the range of keyframes at the maximum frame rate by FFMPEG. The frame rate may vary depending on the recording process and the computer used with our hardware, but generally it is around thirty frames per second.

[88] The precise reasons why these frames are detrimental will be discussed in when we introduce the SfM pipeline.

Once all of the frames are extracted from the video, the intrinsic information of our “world camera” is written into each frame in the Exchangeable Image File Format (EXIF). Almost all contemporary cameras, and many other recording formats, save files like .jpgs with EXIF headers that encode the intrinsics of the camera like focal length, time of photo, image width and height, camera model, aperture size, GPS coordinates, and so forth. Most video files also contain metadata related to the frame rate, duration, camera model, and image size, and so forth. However, each individual frame of a video, once extracted from the video, is devoid of meta information which is essential for our SfM pipeline. So, using an open source tool (exiftool) we can write the known and calculated camera intrinsic information for each frame of the video.

If keyframes from the video are not selected, then all frames of the video will be extracted and saved to a folder in the file system. Once the frames are extracted from the video and have been given proper EXIF tags, the SfM Pipeline can be initiated.

The two dimensional browser currently requires human intervention in the selection of keyframes and verification of data. However, we have taken steps toward implementing a blur detection algorithm so that the frames can be automatically chosen. The blur detection algorithm does spectral analysis on the images, using a discrete Fourier transformation, and compares metrics between frames to determine blur. Blurred images lack sharp edges and therefore high frequency components in their spectral space. We can compare the spectral signature of images to a baseline reading generated from a temporal neighborhood and discard the blurrier frames in favor for their sharper neighbors. This simplistic scheme for detecting blurred images can later be replaced by more sophisticated algorithms. However, detecting “good” images is only one part of the problem, as a good keyframe must satisfy a series of constraints, such as: continuity from one frame to the next, consistency in exposure and gain, quantity of "features," and non-moving elements in the scene. In the next section we will define what makes a "good" image of the world, from the perspective of computer vision and why an undistorted and non-blurry image is important for feature detection and patch expansion in the SfM pipeline.

As an aside, it would be interesting to compare what makes a "good," or salient, image from the perspective of computer vision (in attending to, or in anticipation of constructing a space) with what makes for a "good," or salient, image from a human perspective. A further question along these lines could consider what makes a for a salient space from both human and computational perspectives.

Structure From Motion

If the Structure from Motion (SfM) Pipeline were an opaque process, you would input a series of two dimensional photos, wait a while, and receive a three dimensional point cloud reconstruction of the space pictured in the image. An understanding of all the processes involved in producing a photogrammetric reconstruction from a series of two dimensional images, however, is much more involved and requires an understanding of the basic principles and assumptions in binocular vision, optics, and computer vision.

For computer vision — analogous to and in many ways inspired by physiological understandings of human vision — images captured of the world by a photosensor are not representations of a world, but information that must be processed in order for meaning to emerge. Sophisticated methods have been developed in the field of computer vision in order to make sense out of the massive quantities of information from the world. SfM is one such way of structuring a stream of information, and relies on the constraints that the image sensor, or subject, is moving through a space and that there will be points – or features – that can be tracked from one image to the next.

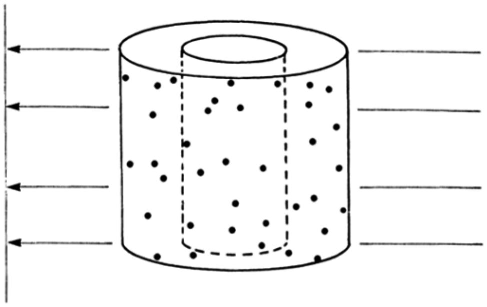


Figure 60: Ullman's Caption: "The interpretation of structure from motion. The dots comprising the two cylinders are projected onto the screen (the outline of the cylinders are not shown in the actual presentation). The 3-D structure of the cylinders can be recovered from the motion of dots across the screen. (See Ullman, 1979)." Reprinted from Shimon Ullman, "Computation Studies in the Interpretation of Structure and Motion", (A.I. Memo: 706, March, 1983), 2.

The history of SfM can be traced back to the study of optical illusions in experimental psychology, where a human subject is able to construct an understanding of a three dimensional solid by observing the relative movement of two dimensional points.[88] For a human, tracking the movement of a point in a two dimensional field or between two images may seem like a trivial task. But for a computer, locating and tracking a moving point or points requires sophisticated mechanisms and

[89] The origins of SfM are discussed in the background of this thesis.

is still an area of active development in computer vision today. It is much more difficult to explain how a human perceives a rotating cube from only a collection of three dimensional points, let alone to make a computer program that can construct a three dimensional solid from these points. The subtle difference between this historical example and a generalized SfM pipeline, is in the movement of points in space. In our research and in the typical case of SfM algorithms, points are considered static features of a scene and the subject or imaging sensor as moving agent.

While humans with unimpaired vision are able to gauge depth from binocular disparity, a computer can not inherently do so with a single image sensor (or even with two calibrated cameras). SfM routines leverage the movement of a monocular sensor and assume that the world is relatively stable, such that two images that share enough features can be considered for binocular vision – seeing the same things from slightly different vantage points. Once a series of mathematical correspondences between these points are established and if the intrinsics of the imaging sensor are known, the images can be situated in space relative to one another. The assumption that is made is that matching points in two dimensional space will intersect in three dimensional space. Thus, these points can be triangulated and form a three dimensional representation of the two dimensional correspondence. This process can be extended to accommodate a potentially infinite number of images from an infinite number of cameras, but is not without difficulty, requiring exponentially greater computational power. Resulting in tens or hundreds of hours to construct three dimensional models from large sets of images.

While the two dimensional browser enables the observation of visual attention fixation areas at any given instant in time, it is difficult to analyze the pattern of visual attention in this format. To put it another way, it is easy to conduct a comparative analysis of visual fixation patterns if the object of inquiry and the subject are static. Recall Buswell's early studies of eye movements for viewing pictures, where the fixations of a human subject can be compared over time and different subjects can be compared, because both space and the subject are constrained, static. These constrained experiments and methods of representation are still conducted today, resulting in statistical heat maps of a subject's cumulative fixations. But, when a subject is moving freely through a space, one can not simply accumulate the points of fixation over time, as the world appears to move relative to the subject's field of view. One method of representation we considered was a flattened representation, like the "joiners" of the artist David Hockney.

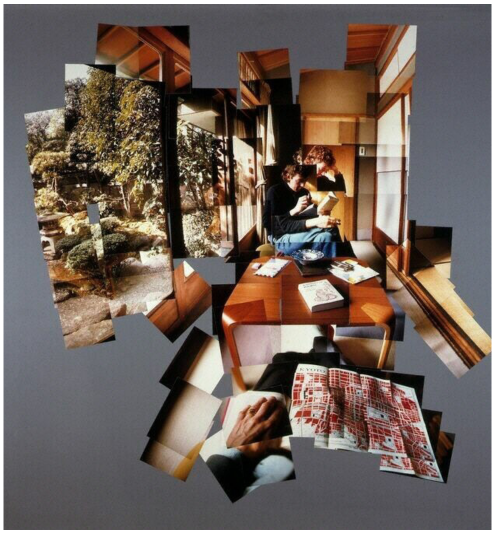


Figure 61: In David Hockney's "joiners" series he takes a series of photographs from different perspectives and the collages them together, resulting in a multi-perspectival construction of the world where time and space are collapsed. Reprinted from David Hockney Gregory Reading in Kyoto, 1983.

In this method of representation, time and space are flattened in a multi-perspectival photographic collage. This method of representation is undoubtably elegant, and would enable a comparative study of fixations over time, but is limited spatially. Allow us to consider the scenario of walking down a long straight hallway and taking a photograph at each step. If one were to collage these images together by roughly aligning features in the photographs, without scaling, the sequence would overlap and obscure each other after only a couple of images, resulting in spatiotemporal incoherence. If images were scaled relative to the first image, temporally earlier images would be highly visible while, later images smaller, and fixation patterns would either become too low in resolution. For our purposes, a Hockney*esque* method of spatiotemporal representation fails, as in a general case, it is impossible to orient and unfold images of movements through space within a two dimensional space. With these considerations in mind, we decided to work within a three dimensional space in order to achieve greater spatiotemporal coherence.

SfM is a well known problem in the field of Computer Vision, and continues to receive attention in a diverse range of applications navigation, geographic surveys, and robotics (to name a few). For the purpose of this thesis project, we are interested in the application of SfM to construct a subjective space. The spaces produced from our SfM Pipeline are inherently incomplete three dimensional models of a place, based on the subject's field of view in a specific time. The space of visual fixation occupies only a subset of this three dimensional construction, while the remainder provides contextual information that would otherwise be processed in the peripheral field of human vision. Beyond precisely locating a moving human subject in space, this method enables a closer approximation to the perceived continuity human spatial experience.

If we want take steps towards understanding human visual experience in space, based on real time information collected from a subject moving through space, then we need to develop representational methods that can capture the richness of this interaction through computational mechanisms. The SfM Pipeline begins by looking for features in each photo. After features are found in each photo, they are matched between each photo. At the heart of the SfM process is an algorithm called "bundler" developed at the University of Washington, in Seattle, by Noah Snavely, Steven Seitz, and Richard Szeliski. This algorithm automates calculation and optimization of spatial relationships between two dimensional clusters of matched points.[89] Finally, we use another algorithm, also developed at the University of Washington by Yasutaka Furukawa. This algorithm is responsible for matching and projecting two dimensional patches of points, given an existing three dimensional SfM construction. We developed a program that tied all of these algorithms together and into one process. The results of our SfM pipeline provide us with all of the information we need in order to construct a three dimensional space and precisely situate a human subject within that space.

In order to make this process transparent for the reader, we will provide an overview of methods we have implemented, starting from a given set of images, moving on to feature detection in images and feature matching, to an overview of camera models and optics, and concluding with multiple view geometry and bundle adjustment. We have developed our own algorithms for all stages of the process, up to, but not including bundle adjustment. While the ultimate implementation of our SfM Pipeline depends on Snavely's bundler, the development of our own algorithms was critical in developing an understanding of the process.

[90] Citation needed.

Computer Vision Methods

Feature Detection and Feature Descriptors

In our SfM Pipeline, the first step is to provide a way for the computer to relate two or more images. As we discussed previously in this section, from the perspective of a computer, images are really only numerical arrays. So, the first step that we take is to find salient local features in an image.

Feature detection algorithms generally work by first transforming or filtering an image and then searching through pixels of the transformed image for salient features, or differences, within a neighborhood around each pixel. One of the most well known feature detection schemes is the Harris corner detection algorithm. This algorithm locates interest points, or features, where the surrounding neighborhood of pixels shows edges in more than one direction, these are then defined as corners in the image. Depending on the characteristics of the scene captured and the transformation kernels applied to the image, many pixels may contain no salient features within their neighborhood.

For example, imagine standing in the center of an entirely white room that is diffusely illuminated such that there are no sharp shadows. For a feature detection algorithm, the diffuse all-white room would be very a challenging space. In this scenario there might not be many differences in the pixels across the image. However, if one were to look at the image at a different scale of neighborhood (either by broadening the search area or by blurring the image) salient features might emerge for that same pixel in question.

The methods of transformation, detection, and extraction are specific to each feature detection algorithm. There are many algorithms available for feature detection, written in low level code (C or C++) and optimized for multicore processors. Most of these

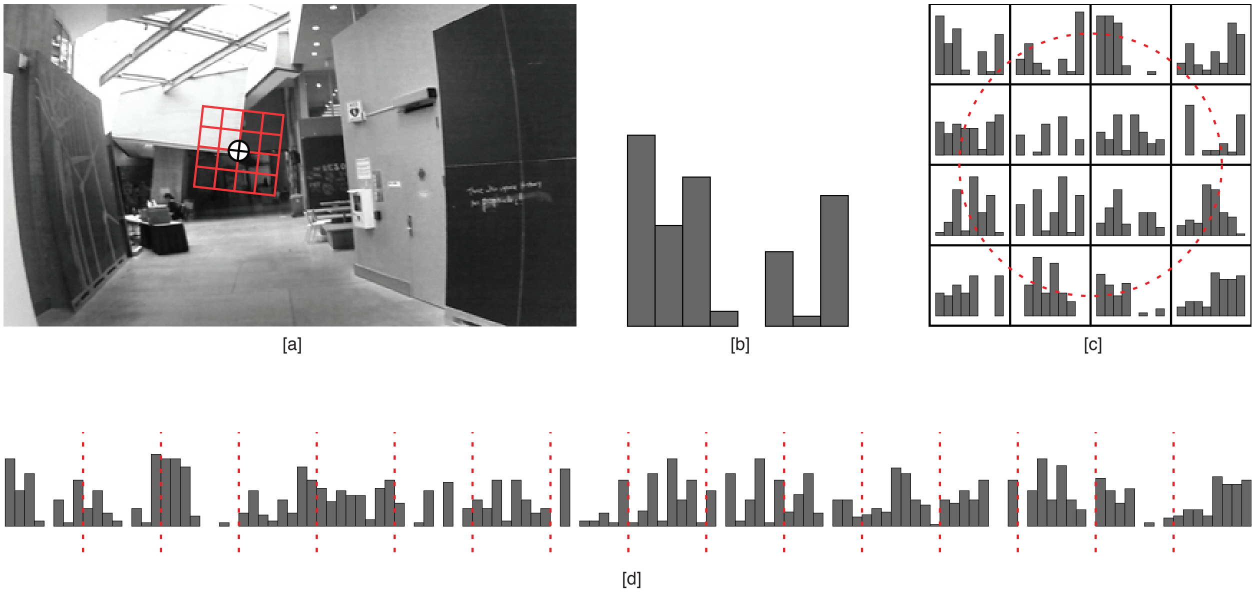


Figure 62: Illustration of a SIFT key point and the construction of a descriptor. [a] The key point is marked by a white circle in the grayscale image above, and the neighborhood is the red grid surrounding the point, oriented along the dominant gradient direction. [b] 8 bin histogram from a part of the neighborhood grid. [c] histograms from each area of the neighborhood about the point. [d] concatenated histogram forming a 128 element long descriptor vector. Drawn by the authors, based on diagrams from Solem, Programming Computer Vision with Python, 54.

algorithms are open source and are implemented or accessed by different computer vision software libraries, such as Open Computer Vision (OpenCV). Some feature detection algorithms will require an image to be transformed into a threshold image and then will begin to search for edges or corners, gradient differences between dark and light. While other algorithms transform the source image into a gradient intensity image or integral image before searching for features. In this section we will explain how feature detection algorithms work.

Detected features in an image are simply points of interest within the space of the image. In order to compare these points of interest across images to find correspondences, we need a way to describe the feature. In computer vision, descriptions of features are defined by a multidimensional vector. This vector describes the image appearance around the point of interest or "key point." Each algorithm typically uses a different way of describing the neighborhood about the key point, but in general the better the description of the neighborhood the better the correspondence between images.

In our research we have evaluated a large range of feature detection algorithms available in OpenCV directly from authors of algorithms. In our evaluation of algorithms we have found that David Lowe's Scale Invariant Features Transform (SIFT) algorithm, while not the fastest, provides the greatest number of key points with a strong description of neighborhoods. The combination of a strong description and large number of key points is beneficial for finding pairwise correspondences, which we discuss in the next step of the Pipeline. Our SfM Pipeline is capable of using other feature detection algorithms available through OpenCV, but for sake of clarity we will limit our discussion here to the SIFT algorithm.

Scale Invariant Feature Transform (SIFT)

SIFT key points are found by using difference of Gaussian (DoG) function. Candidate key point locations are maxima and minima of the DoG functions across both image and scale. Stable key points are defined by criteria of high contrast, points, and edges. Unstable key points are removed from the candidate pool based on a specified threshold.[90] As the detection algorithm is scale invariant, the key point is defined as the x,y position in the image and scale of the key point. Invariance to rotation is achieved by defining a direction and magnitude of the image gradient within a neighborhood that surrounds each key point.

The sift descriptor is computed based on the position, scale, and rotation of the key point using image gradients. "The descriptor takes a grid of

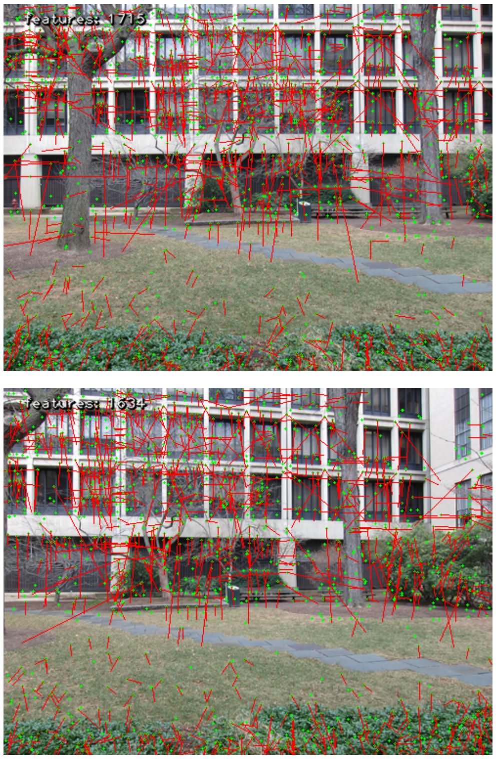
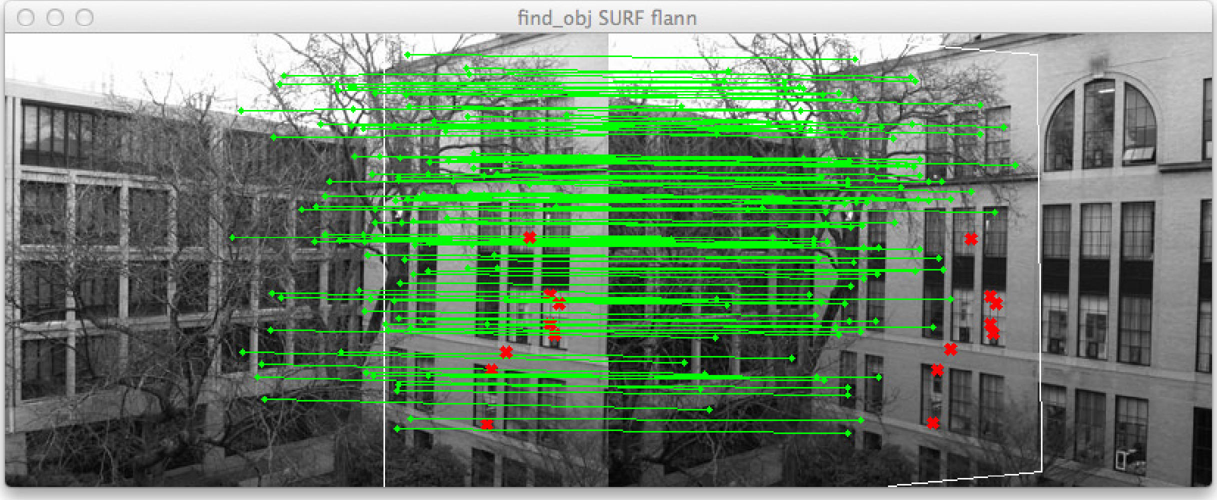


Figure 63: Two images that show SIFT key points as green dots for two views of a courtyard at MIT, images are separated by ten degrees in rotation about the Z axis. The red lines represent the scale and orientation of each key point. The number of key points found in each image is listed in the top left corner of each image. It is possible with visual inspection alone to locate some matched features.

[91] David G. Lowe, "Object Recognition from Local Scale-Invariant Features," Proceedings of the International Conference on Computer Vision, September, 1999.



subregions around the point and for each subregion computes an image gradient orientation histogram. The histograms are concatenated to form a descriptor vector. The standard setting uses four by four subregions with eight bin orientation histograms resulting in a one hundred and twenty eight bin histogram…" which is the length of the feature vector.[91]

Typically feature detection algorithms are not visualized, however we have found it useful for didactic purposes to represent the results of these algorithms. By doing so, we can begin to gain a visual understanding of how these algorithms process an image, and structure an image of the world provided to the computer.

Matching Feature Descriptors

Once feature key points and description vectors have been established for each image in our set of images, we attempt to find matches between

[92] Jan Solem. *Programming Computer Vision with Python*. (Draft: March 18, 2012), 51.

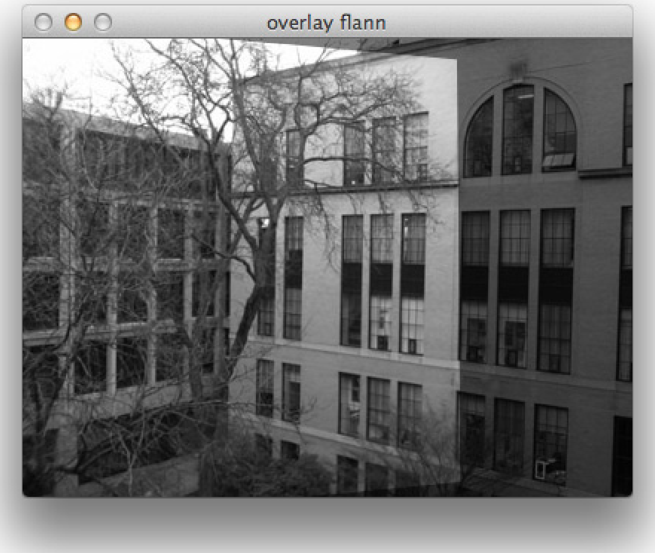


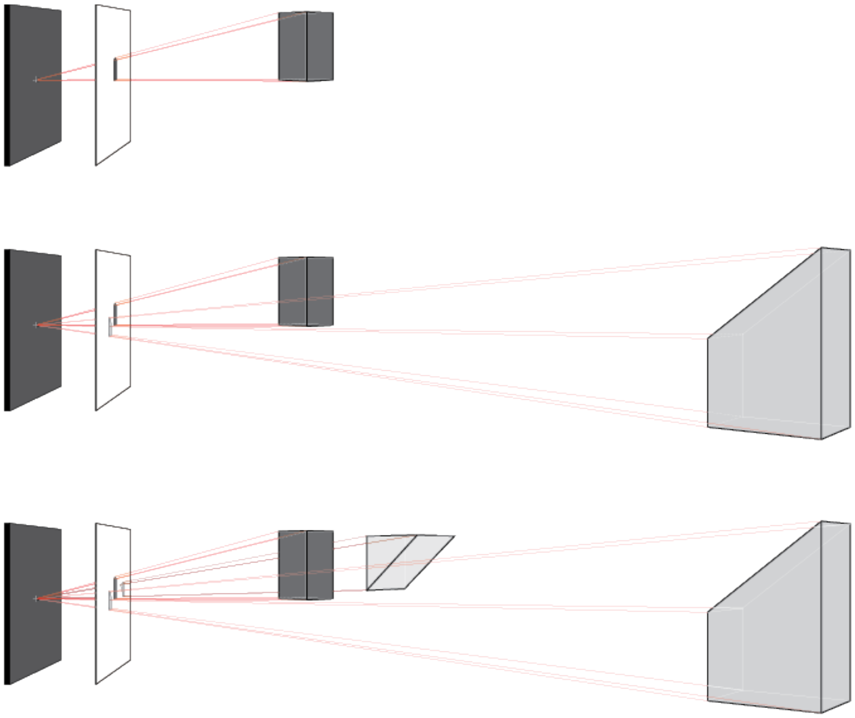
Figure 64: [Facing Page] Two images of buildings in an MIT courtyard from subtly different angles. The green dots show the location of feature key point in each image and the green lines represent matches between each feature descriptor. Matches calculated as outliers are indicated with a red 'x.' A white rectangle is drawn on top of the right image which represents the perspective transformation of the left image onto the right image using the Homography matrix computed from the matches between the two images.

Figure 65: Two images of the courtyard (as seen in preceeding Figure). Matched features from the images are used to compute a homographic transformation between the two sets of key points. After the homography is found, the image on the left is warped with a perspective transformation (a type of affine transformation) within the space of the right image with transparency.

images. A simple example of matching feature descriptors between two images can be accomplished by using the ratio of the distance of the two closest matching features – or a specified number of nearest neighbors. As locations, orientations, and scale of features is likely to change dependent on the movement of the subject or image sensor from one image to the next, robust matching algorithms like RANSAC (Random Sample Consensus) or FLANN (Fast Approximate Nearest Neighbor Search) can be used to estimate best fits between two sets of features. Features outside a specified threshold will be ignored, and treated as outliers, allowing for a potentially better fit between the two sets of points. However, the specification of the threshold must be tuned in order to avoid finding local maxima or minima. An example of feature matching between two sets of feature descriptors is shown in the following image.

Transformations: Image to Image Mappings

Once we know how images are related, based



on a set of features, we can find a two dimensional projective transformation between the two images. In an analog example, this transformation would be similar to aligning two images by hand as in the Hockney photo collage. In mathematical terms, this is type of two dimensional projective transformation is called a homography and is defined by a three by three matrix. The minimum number of points needed to find a homography between two images is four, but judging from the number of SIFT features we find, this will likely never a problem. Using iterative methods to fit data that may contain outliers, like RANSAC, we can find the best fit between the two sets of points and therefore a transformation between the two points. With a known homography matrix, we can perform perspective transformations in order to align two images, which can be used for creating panoramic images.

Figure 66: [Facing Page] Pinhole Camera Diagrams: Different geometric objects being captured on the image plane of a pinhole camera. Note, the image plane is shifted in front of the optic in order to emphasize the geometric relationships between the optic and the image sensor. In a real camera the image sensor is behind the optic, but this causes the images to be inverted and convolutes the diagrammatic clarity. In this representation all the geometric principles in optics are preserved despite the shift in the image sensor plane.

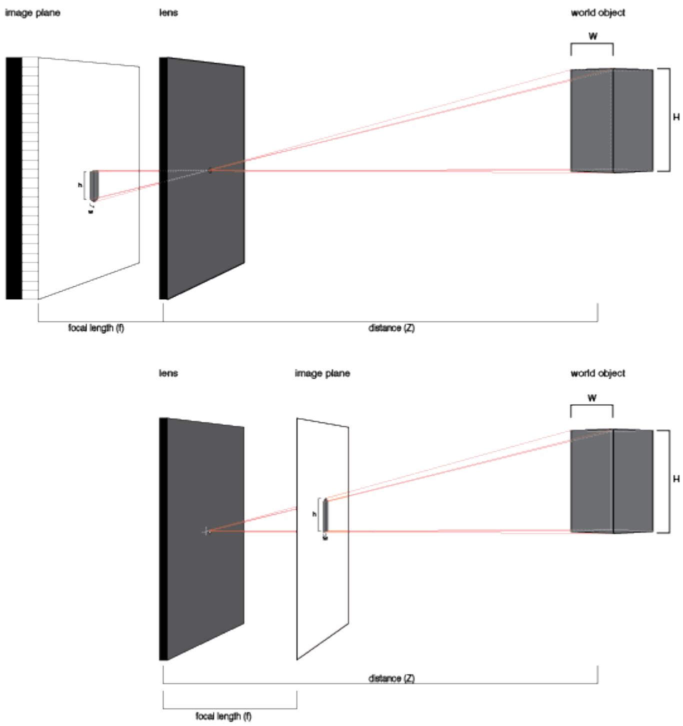


Figure 67: [top] A three dimensional rectangular prism in the worldview is captured on the image sensor. The image of the prism is inverted as the rays from the world pass through the single point of the pinhole camera. [bottom] The sensor plane is shifted in front of the lens for geometric clarity.

Computational Camera Modeling

Before delving further into the mechanics of the SfM Pipeline, it is necessary to provide an overview of optics and a simplified computational camera model.

The pinhole camera model is standard approach to computationally modeling a camera. This abstract model of a camera is composed of two elements, the lens and the imaging plane. The lens for a pinhole camera is nothing more than a tiny aperture that permits light from the world to pass into the chamber of the camera and onto the surface of the imaging plane. The imaging plane is located at a variable distance away from the lens, and serves as a planar surface for a photosensitive material or device that records the rays of light passing through the lens. While the pinhole model may appear overly simplified, it is extended with further numerical parameters to account for the distortion caused by optics and relationships between optics and the image sensor. The lens distortion parameters will be covered in the discussion on camera calibration. The pinhole model captures fundamental geometric relationships between intrinsic properties of the camera and extrinsic objects in the world.

The pinhole camera model is usually shown as a diagram with the image plane in front of the lens. Of course placing the image sensor between the object in the world and the lens would not function as an actual camera, but this representation serves to clarify geometric relationships between the lens, image coordinates, and real world coordinates and also serves as a mathematical convenience. By looking at any two rays (red lines) in the ideal pinhole camera model we can form two similar triangles. The smaller triangle, between the lens and the image plane, is similar to the triangle that is formed between the lens and the object in the world. Similarity between the two triangles, intrinsic and extrinsic, becomes even more explicit when examining a cross section of the modified pinhole camera model, with the image plane displaced in front of the lens.

The ratio of these two triangles allows one to begin calibrating the camera, given known extrinsic measurements, or to recover extrinsic measurements given known camera intrinsics. The intrinsic properties referred to in this simplified model are the camera focal length and the information on the image plane. The simplified ratio for calibration can then be written as:

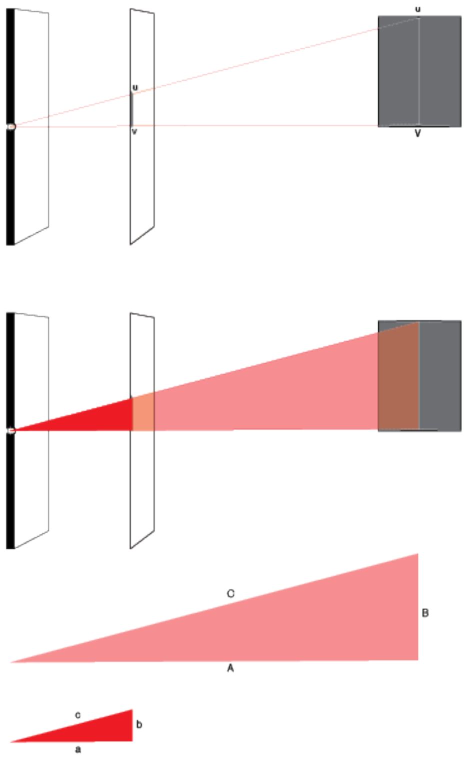


Figure 68: Pinhole Camera Cross Section: [top] The three dimensional points U, V in object in the world are projected to the image plane as the points u, v. [middle] Two triangles can be formed between the points u,v and the optical center and U, V and the optical center. [bottom] Similar triangles ABC and abc.

$$u = f_x \cdot \left(\frac{U}{Z}\right)$$
$$v = f_y \cdot \left(\frac{V}{Z}\right)$$

This equation can be reconfigured so that one is solving for an unknown distance or world coordinate.

Calibration Matrix:

The calibration matrix captures the intrinsic properties of a camera within a three by three matrix, typically notated as K. The calibration matrix is not affected by extrinsic variables, or things that happen in the world. In general form the calibration matrix has three parameters, one of which is often ignored or insignificant:

1. The focal length *f* as the distance between the camera center and the image plane. The focal length is often expressed in two dimensions as *fx* and *fy* due to the fact that most pixel arrays are not necessarily square.
2. The skew *s*, almost always ignored, is used if the pixel array in the sensor is skewed.
3. The optical center coordinates, in pixels. This is often called the camera center, where the principal ray intersects the image plane. The optical center can be roughly estimated to be half of the height and width of a full resolution image produced by the sensor.

The general form of the calibration matrix, including skew would look like this:

$$K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

And, since skew can safely be ignored the majority of the time, the calibration matrix then takes on the form:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

If the aspect ratio of the pixels is equal to 1 (i.e. square pixels) then the calibration matrix is further defined:

$$K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

One can calculate the calibration matrix for any camera by taking a photograph of a planar object that is parallel to the camera imaging plane at a known distance. Knowing the height and width of the planar object, and the height and width of the planar object in the coordinates of the imager (pixels) one can calculate the focal distortions and estimate the center-x and center-y parameters to obtain the calibration matrix. This matrix can be saved and reused whenever using the same camera. However, when processing an image that has been scaled one should also scale the calibration matrix. This can be done by multiplying fx by the ratio of the new image width/original image width and likewise with fy.

$$f_{x\,scaled} = f_x \cdot \left(\frac{w_{new}}{w_{old}}\right)$$

$$f_{y\,scaled} = f_y \cdot \left(\frac{w_{new}}{w_{old}}\right)$$

Also, the center-x and center-y parameters must be adjusted for new image sizes. Typically these values can be multiplied by the same scaling ratio.

Using libraries like OpenCV, one automate

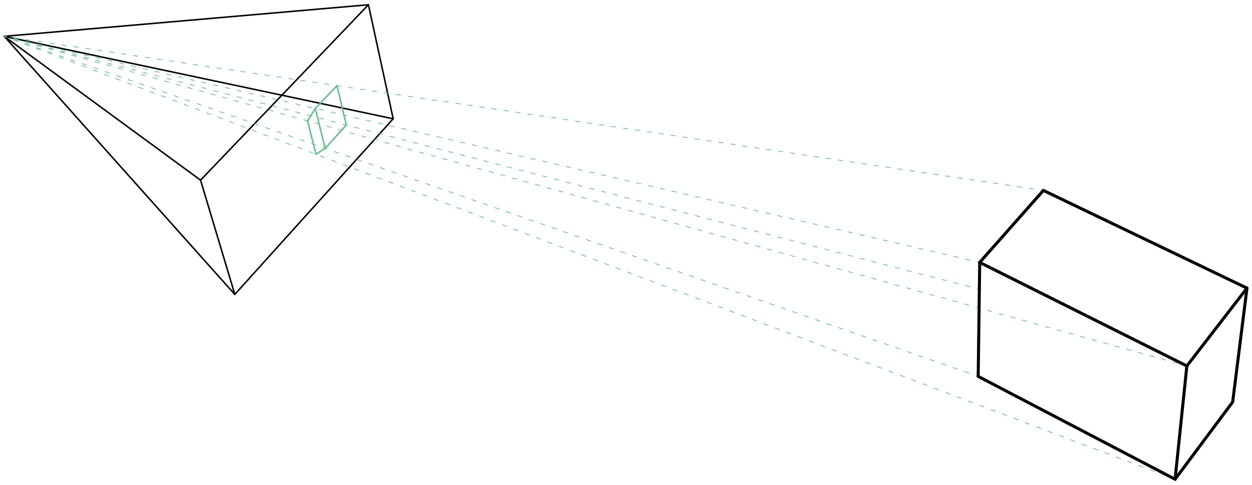


Figure 69: Projection model diagram. Black edges of the pyramid connect the extents of a rectangular image sensor or image plane to an apex or camera center. Green dashed lines are imaginary rays that relate real world coordinates to coordinates in the image plane. We use this pyramidal diagram to graphically represent the attributes of the projection matrix.

the calibration process by automatically detecting the location of points in the world on a planar object and solving for the calibration of a camera by taking many photographs of that planar object. Typically a black and white chessboard or dot pattern is used in computer vision calibration routines. This pattern is used because it is easy to construct (can be printed or drawn), is easily measured for size of the square or dot diameter, and has strong contrast providing corners or blobs of defined diameter for feature detection algorithms. In this pattern corners are defined as the location where two black squares meet and these can be automatically located in the space of the imager (pixel coordinates) with sub pixel accuracy. Knowing the real world size and number of the checkerboard square intersections along with their corresponding corner locations in the image plane, one can calculate a calibration matrix.

Projection Matrix:

The Projection matrix typically notated as P, is a three by four matrix, that captures all information about a camera with respect to its intrinsics — the relation between optics and sensor — and the extrinsic properties — the camera's pose or global translation and rotation in the world. The general form of the projection matrix takes the form:

$$P = K\,[R|t|]$$

Where the K is the three by three calibration matrix of the camera, R is a three by three rotation matrix and t is a one by three translation vector. A projection located at the origin will look like this prior to multiplying by the calibration matrix:

$$\begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \end{bmatrix}$$

The last column represents the translation vector and the three by three identity matrix represents the rotation matrix R.

For the purposes of this project, the projection matrix is important because it allows us to compute the relationship between points on the image plane, their projected point equivalents in three dimensional world space, and to robustly estimate the origin of that projection which is the equivalent to the camera pose. For example, if we have two images of the same object from two different vantage points, we can think of these images corresponding to two projection matrices. These two projection matrices, if taken by the same camera with invariant intrinsic parameters, vary only in pose. Furthermore, if the two projection matrices are known, we will be able to reconstruct a three dimensional triangulation of the space defined by the two projection matrices.

By determining correspondences in features between the two sets matched features in each image plane, we can project a line from each camera origin through each feature key point (or each pixel) and estimate a triangulated intersection point in world coordinates. However, this scenario is highly unlikely, because it requires two known projection matrices and therefore the knowledge of the poses of both cameras. This could be accomplished, by measuring the relative movement in translation and rotation of the actual camera while taking photographs, but this would be a very tedious process prone to error and specific only to the two images in question. So, we must seek a general solution that allows us to recover projection matrices between an arbitrary number of potentially different cameras, given relationships between matched features in sets of two dimensional images.

In order to achieve this goal we need to model the spatial relationship between matched features in image coordinates, use the relationship of matched points to estimate possible projection matrices, project and triangulate points from each camera into world space, reproject the triangulated points onto the imaging plane, and iteratively minimize errors in reprojection in order to estimate an optimal projection matrix. This process is referred to as Structure From Motion.

Structure From Motion Pipeline

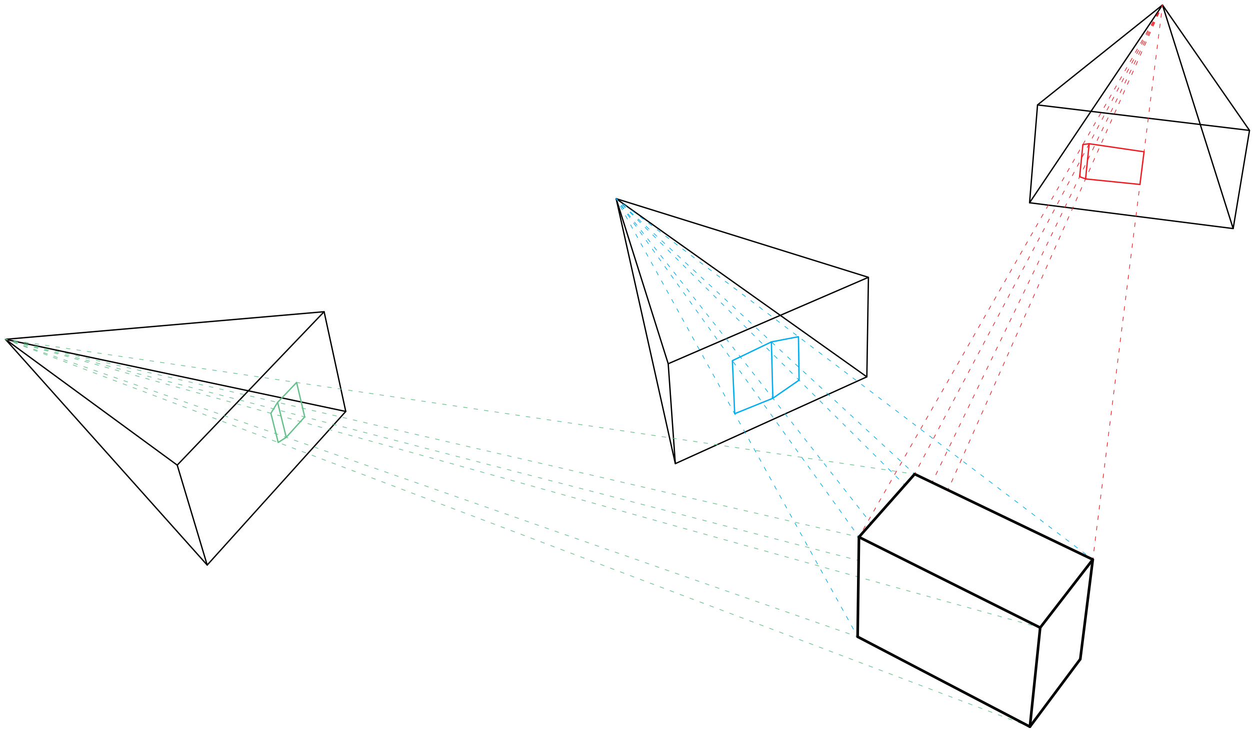


Figure 70: Structure from motion Diagram. Three images and their associated image pyramids that represent the pose and focal length of the camera. The dashed lines show how the object in the scene maps to the two dimensional planar surface.

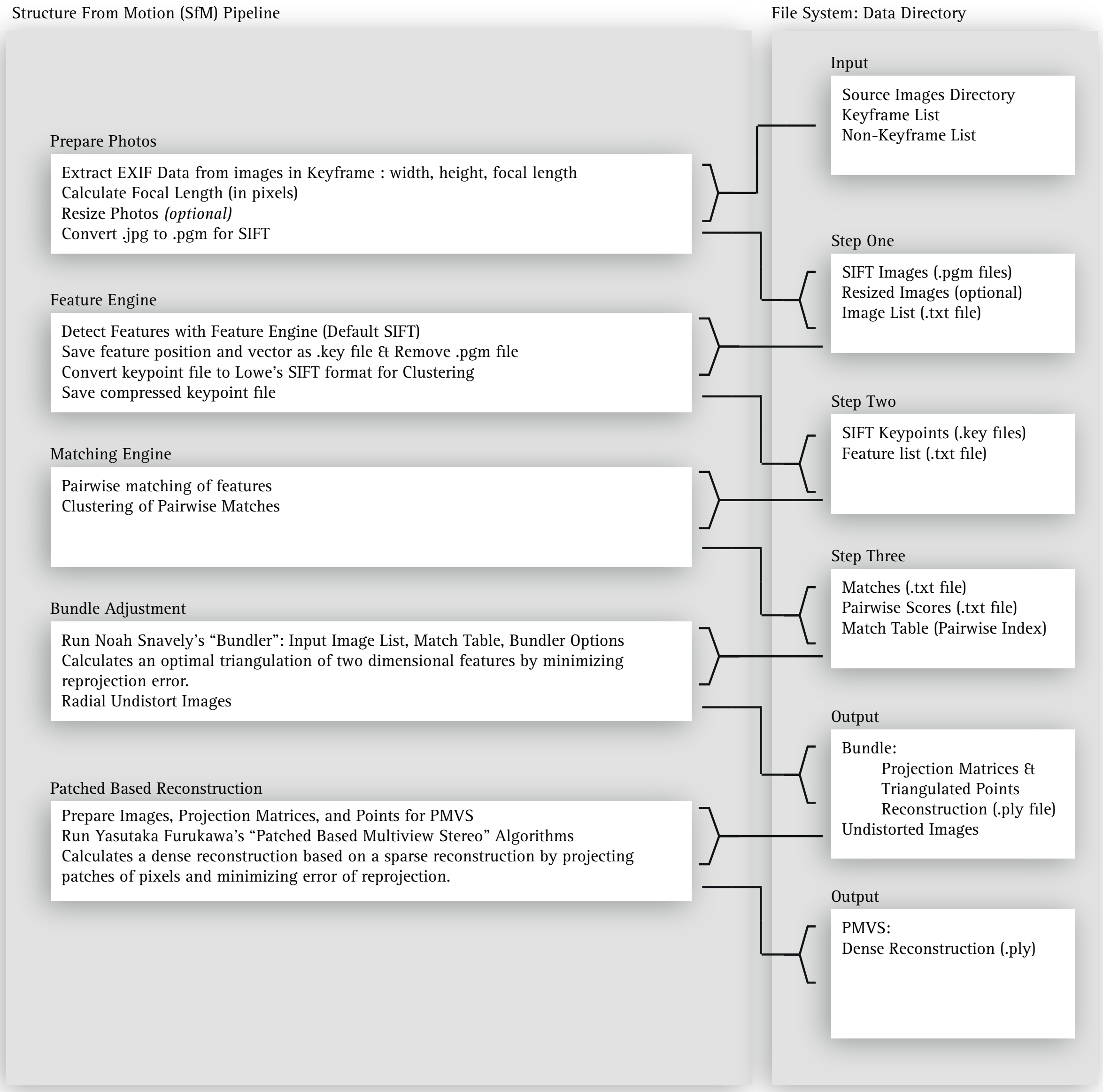
In the SfM Pipeline that we have implemented, we start with a list of "keyframes" and a directory of source images extracted from a video captured from a head mounted camera.

The first step in the SfM pipeline is to process images. Metadata is extracted from each image, images are scaled, focal lengths are calculated, and copies of the images made in a format that can be used for feature detection. Metadata is extracted by examining EXIF tags, yielding intrinsic information about the camera – make, model, focal length in millimeters, sensor dimensions – and the intrinsics of the image produced – width and height. If the images are determined to be too large they are determined to be too large. Next, the images can be scaled. While larger images may result in a greater number of features detected, and therefore a higher potential for pairwise matches, larger images will drastically reduce the speed of reconstruction. If the images are scaled, the scale factor is saved as a local variable and a copy of the scaled images are saved to the SfM folder in the data directory. After optional scaling, we must then recalculate the focal length, in pixel units. The names of each image file and the focal lengths are written as entries in a list and saved to the file system. After images are processed, SIFT key points and descriptors for each image are extracted and saved to a file.

The second step is to match features. This process is similar to the description we gave above, but differs mainly because of the number of features that are being matched. This processes uses a K-Dimensional Tree to structure the high dimensional nearest neighbor searches between feature descriptors.

Once features have been matched for each image pair, bundle adjustment can begin. This is a process of optimization that results in the calculation of an optimal solution to an under constrained problem. This process can be described in the following series of steps:

1. Calculation of Essential Matrices, or the relationship between key points in world coordinate space.
2. Estimation of Projection Matrices from computed Essential Matrices
3. Triangulation from estimated Projection Matrices
4. Reprojection of triangulated points to image plane
5. Error Calculation from difference between matched features on image plane and reprojected points
6. Bundle Adjustment: The simultaneous optimization of triangulated points and projection matrices in order to minimize error in reprojection.



Structure From Motion (SfM) Pipeline

Prepare Photos

Extract EXIF Data from images in Keyframe : width, height, focal length
Calculate Focal Length (in pixels)
Resize Photos *(optional)*
Convert .jpg to .pgm for SIFT

Feature Engine

Detect Features with Feature Engine (Default SIFT)
Save feature position and vector as .key file & Remove .pgm file
Convert keypoint file to Lowe's SIFT format for Clustering
Save compressed keypoint file

Matching Engine

Pairwise matching of features
Clustering of Pairwise Matches

Bundle Adjustment

Run Noah Snavely's "Bundler": Input Image List, Match Table, Bundler Options
Calculates an optimal triangulation of two dimensional features by minimizing reprojection error.
Radial Undistort Images

Patched Based Reconstruction

Prepare Images, Projection Matrices, and Points for PMVS
Run Yasutaka Furukawa's "Patched Based Multiview Stereo" Algorithms
Calculates a dense reconstruction based on a sparse reconstruction by projecting patches of pixels and minimizing error of reprojection.

File System: Data Directory

Input

Source Images Directory
Keyframe List
Non-Keyframe List

Step One

SIFT Images (.pgm files)
Resized Images (optional)
Image List (.txt file)

Step Two

SIFT Keypoints (.key files)
Feature list (.txt file)

Step Three

Matches (.txt file)
Pairwise Scores (.txt file)
Match Table (Pairwise Index)

Output

Bundle:
Projection Matrices &
Triangulated Points
Reconstruction (.ply file)
Undistorted Images

Output

PMVS:
Dense Reconstruction (.ply)

Figure 71: Diagram of the Structure from Motion Pipeline. The the gray box on the left shows the program flow of the SfM Pipeline. The gray box on the right shows the input and output files of the SfM Pipeline at various stages in the process. This process begins with input from the Two Dimensional Browser.

The results of bundle adjustment are estimates for the Projection Matrices of each camera and a sparse cloud of triangulated feature key points with colors inherited from the pixel where the key point is located on the image plane.

If the sparse reconstruction is successful, then a dense reconstruction will be attempted using the same series of photographs and the projection matrices calculated from bundle adjustment. Our pipeline currently uses Yasutaka Furukawa and Jean Ponce's "Patch-based Multiview Stereo" algorithm, which takes the set of images and projection matrices calculated in our sparse reconstruction routine and then attempts to expand surface patches between the already constrained images and projection matrices (using Difference of Gaussian and Harris operators), match features by triangulating, expand patches, and filter patches (optimize for minimal projection and reprojection errors). A detailed description of the algorithms can be found in Furukawa and Ponce's Accurate, Dense, and Robust Multi-View Stereopsis.

The final output if the SfM pipeline is a set of projection matrices for each image frame along with a point cloud that represents the constructed three dimensional space.

Three Dimensional Browser

This final piece in our software toolchain allows the user to view and analyze the subject's patterns of visual attention as they moved through a space. The three dimensional visualizer constructs a representation of the three dimensional space by merging data from the SfM Pipeline into a single representation. This representation reveals: the subject's movements as a path through a space, his field of view as recorded by the world camera in the capture routine, a three dimensional point cloud construction as calculated by the SfM Pipeline, and the patterns of visual attention as three dimensional projections.

The three dimensional is primarily written in Python and uses multiple external library most notably an OpenGL library named GlumPy and OpenCV to allow access to OpenGL functionality for rendering and display and loading of images. The visualizer aims to be an easily modifiable platform to coalesce data from trials and visualize results.

Using the data generated by the SfM Pipeline the three dimensional browser generates a colored point cloud that represents the image pixels projected into three dimensional space. By loading the video capture stills and their corresponding projection matrices the space of the three dimensional construction is populated with each frame at the approximate location they were captured. This part is a crucial contribution to the structure from motion pipeline as it allows us to determine the exact movements of the subject in space along with the orientation of their field of view. Using the pupil positions, that are associated with frames of the video, we can now re-associate the point of visual fixation, the subject's gaze, with each image. Superimposing

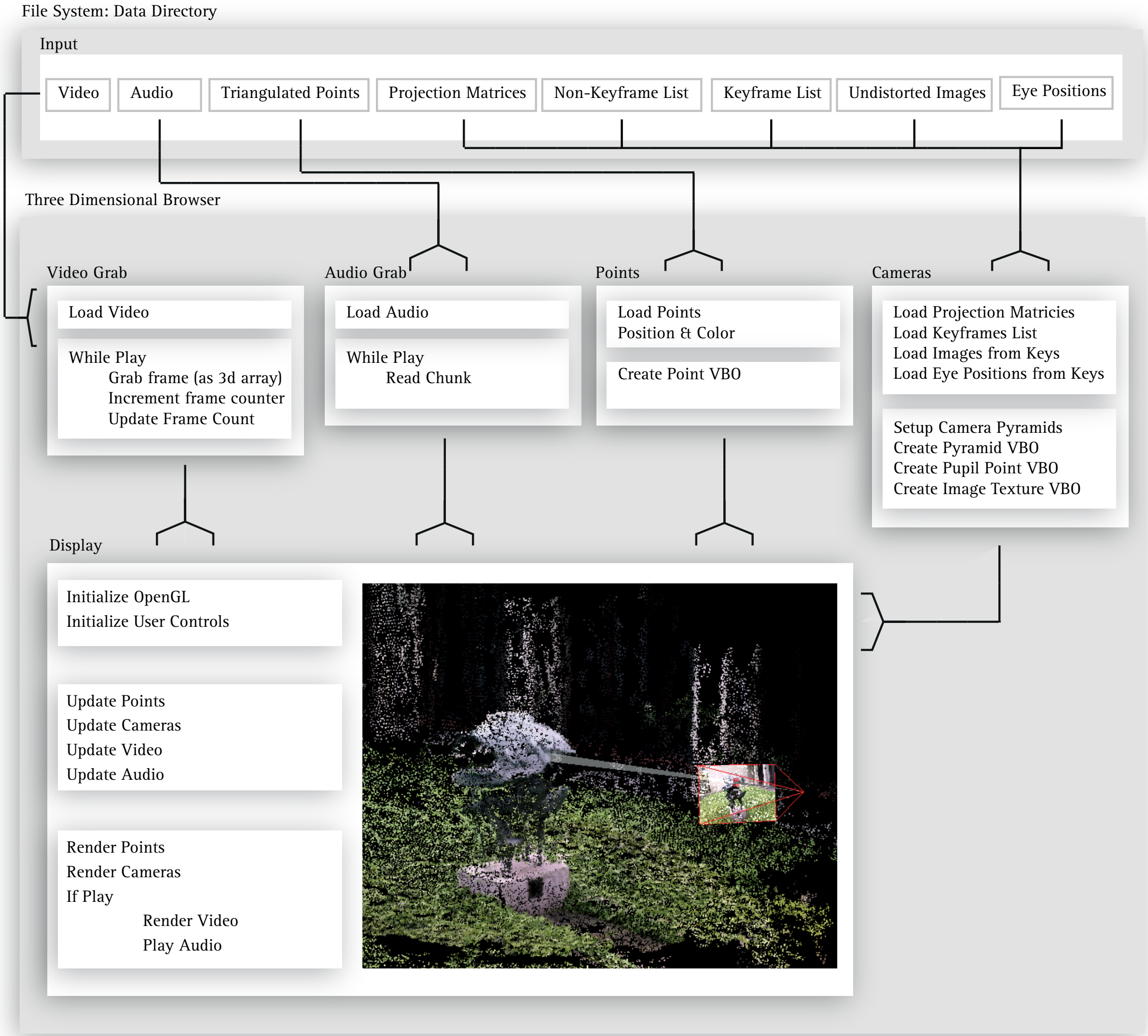


Figure 72: Diagram of the Three Dimensional Browser program flow. The the gray box on the top shows the file system and the files used to construct the visualization in the browser. The gray box on the bottom shows the modules and processes used to prepare different media for the browser and synchronize streams in time. The display function spawns an OpenGL window to visualize vertex buffer objects: image textures, points, cones of visual fixation.

the position of the gaze onto the surface of the image, now oriented in space, we obtain a cone of cone of the subject's gaze within the constructed space. The cone represents the space of acute vision of the subject at that instant in time. If the eye is fixating, then the cone represents the space of visual attention.

The only information that is not currently available is the precise depth of the cone. The depth of the cone relates to the visual depth plane the subject had fixated upon while the frame was captured. At this point one can either choose to set the length of the visual cones globally or let them end at the first intersection with the point cloud. The latter approach assumes that the subject does not look past an object in his field of acute vision. This assumption may not always be true but can serve a good indicator for fixation depth. An alternative approach would be to track both pupils and triangulate the visual depth of the gaze point. We have not yet implemented this method, but it would serve as a promising addition to the entire toolchain.

Trials Overview

We define trials as small test runs that are used to demonstrate the capabilities of Pupil in both its two dimensional and three dimensional constructions. The trials begin as simplistic studies that are primarily concerned with testing certain aspects of the pipeline and move toward more sophisticated studies that introduce questions about a human experience of space. However the trials are not quite experiments, as they lack proper preparation and rigor to serve as a basis for further claims or proofs. At this stage we hope to be able to show what is possible and probe into possible areas of future inquiry.

Trial Setup

The trials we present were performed with two people, however it is not necessary for trials to be conducted with two people. In our trials, one person who is wearing the headset and whose eye movements will be recorded, will be referred to as the subject. The second person will assist the subject in the calibration routine and will often ask questions or give prompts to the subject. In some cases the subject and the operator will conduct a running discussion.

Trial 1: Two Dimensional Operation

The first trial is concerned with the accuracy of the pupil detection and vector transformation coefficients obtained from a calibration run. We will show the necessary setup steps for this trial, but documentation of this type will be omitted from all other trials as it is an analogous procedure.

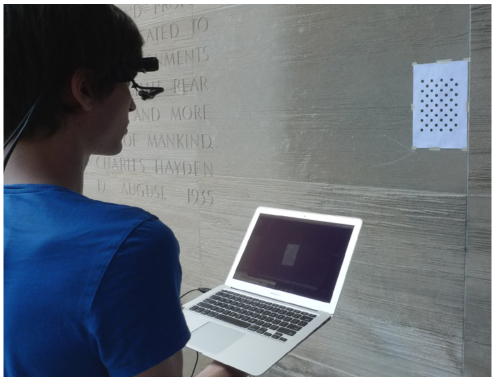


Figure 73: Subject performing calibration routine.

The Calibration Process:

Calibration is a process where the movements of the eye and the movements of the head are correlated such that the position of the eye can be

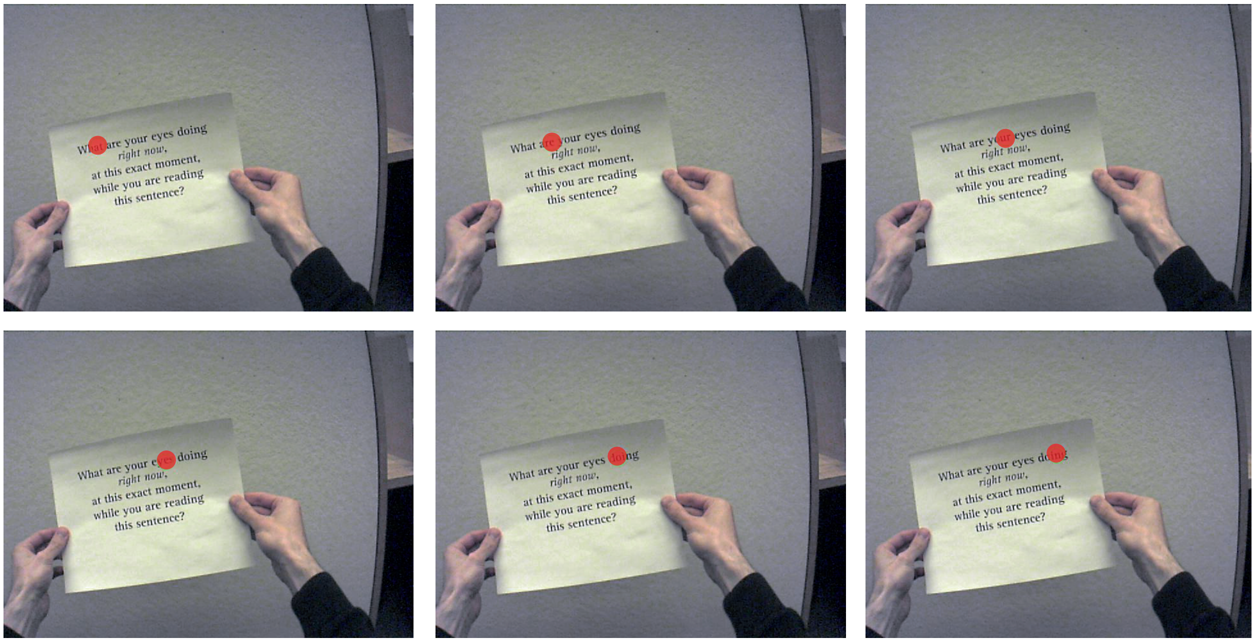


Figure 74: Reading a text as a simple test. Red dot shows the subject's point of visual attention overlaid onto a text that the subject is reading.

precisely mapped into the coordinate space of the subject's field of vision camera (or "world camera"). In practice this means that the subject has to look at a fixed point — the center of the circle pattern — while moving the head left and right and up and down for approximately one minute. Data fitting of the pattern centroid and the position occurs immediately after the calibration, and the trial can begin. During calibration the capture routine also saves information about the pattern for later calculation of the intrinsic camera parameters, such as focal length and radial distortion. Calculating camera calibration requires no action from the user as uses the same pattern for camera calibration as is used in correlating the pupil position with the world camera.

Once calibration has been verified, the trial session can begin. The operator can start and stop recording of audio, pupil position and video as desired.

In this trial the subject follows a speak aloud protocol, describing the object or detail in the environment to which they are attending.

Reading text is another good method that can be used to test the accuracy of calibration. It should be noted that voice audio and eye movements as raw data may not necessarily be synchronized due to, the eye is typically ahead of the voice (see chapter "Best possible reader").

Trial 2: Proof of concept Three Dimensional Pipeline and Browser

The primary purpose of this trial can be considered as a proof of concept of the pupil software pipeline and the three dimensional visualization environment. The location of this recording is a small courtyard between building 2 and 14 on the MIT campus. The subject's task was to visually trace a horizontal contour located on the upper quarter of the sculpture, while moving freely through the environment. Due to the physiological behaviors of human vision, not all misplacements of the eye position can be attributed to process errors. As discussed in the beginning of this book, eye movements are not directly controllable. Perfectly tracing a line with one's gaze is impossible. Nonetheless, this artificial and very constrained viewing behavior ensures that gross misplacements of the area of visual attention in the final visualization must stem from process and display errors and do not represent uncontrolled eye movements of the subject. Looking at smaller length scales this assumption does not hold true.

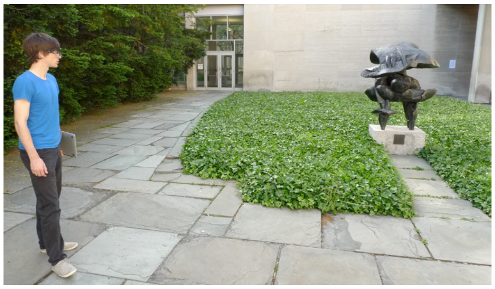


Figure 75: Subject observing and moving about a sculpture outside of Hayden Library at MIT. Sculpture by Dimitri Hadzi, "Elmo-MIT," 1963.

After partial circumambulation of the sculpture with a short interruption to examine a detail at the sculpture’s front side , the recording is stopped and the passed into the software pipeline.

Using the two dimensional browser, 90 keyframes where selected based on the criteria (introduced in section Software Development) to be passed into the structure from motion pipeline. Extracting features, key point matching, bundle adjustment took approximately 100 minutes. Due to the feature rich space even the sparse point cloud yielded 92,767 vertices. Using Patch based multi view stereo extraction to extend patches we only saw a mild increase to 125,052 points. This may be due to the highly fragmented appearance of most surfaces (ivy on the ground and trees in the background).
In the future we plan on adding the remaining frames in order to fill in 100% of the recorded frames, leading to a dense camera path of an average 30 fps temporal density in the constructed space.

The output data was then loaded into the three dimensional visualizer to visually verify the result. Now we will give a tour of the three dimensional browser for the Hayden Library trial.

Figure 76: Looking through the image pyramid of a single key frame. The red point shows the center of the subject's visual attention. Here the subject is attending to the contour of the sculpture.

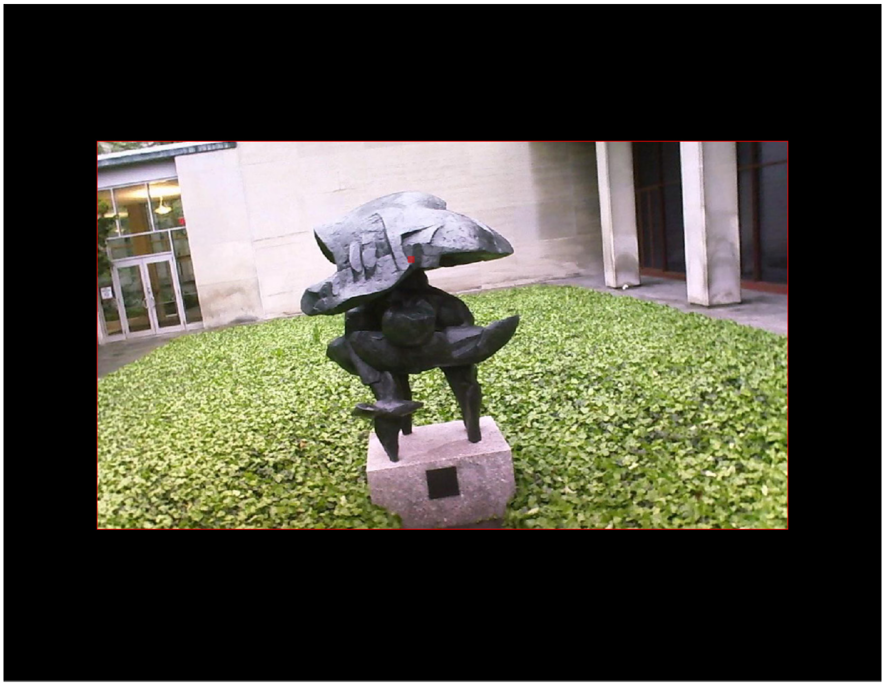
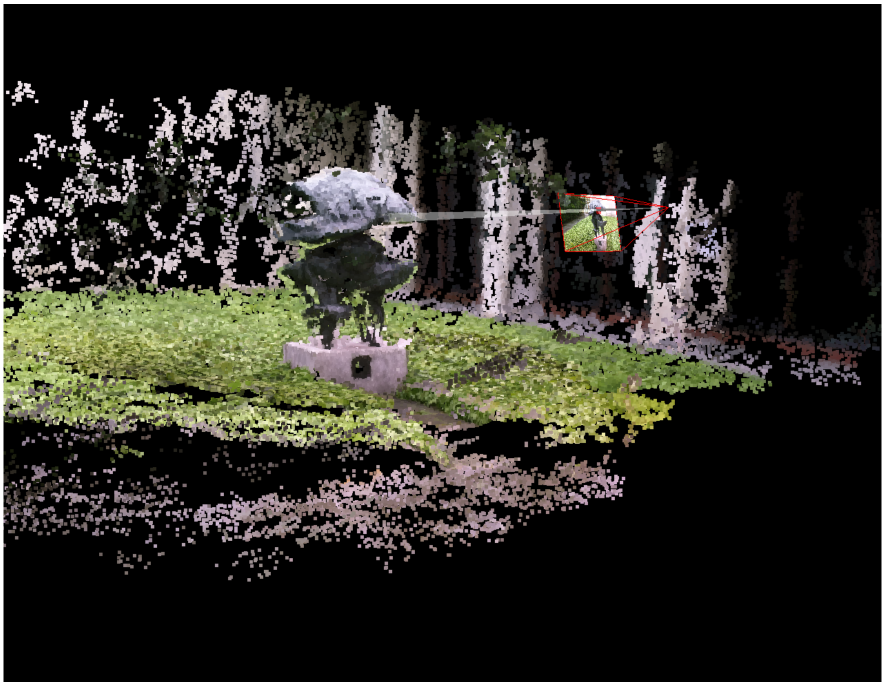


Figure 77: The same frame as the previous figure, now observed from a simulated third person perspective. The red pyramid represents the projection matrix corresponding to that image and the subject's pose in the world at that instant in time. The space constructed using SfM is shown here as a dense point cloud. The white cone describes the subject's volume of visual attention. This cone intersects with the three dimensional points, revealing the relationship between the subject's gaze and the world.



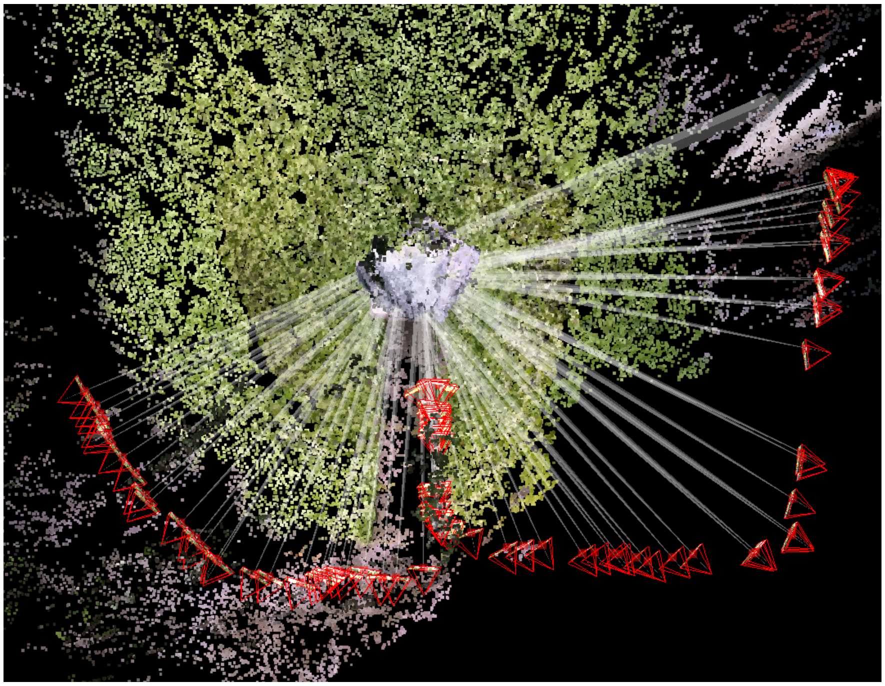


Figure 78: Birds eye perspective that shows 90 keyframes captured from a single subject as they moved about the sculpture. The visual "tracing" of the a sculpture's horizontal contour is well visible in this image.

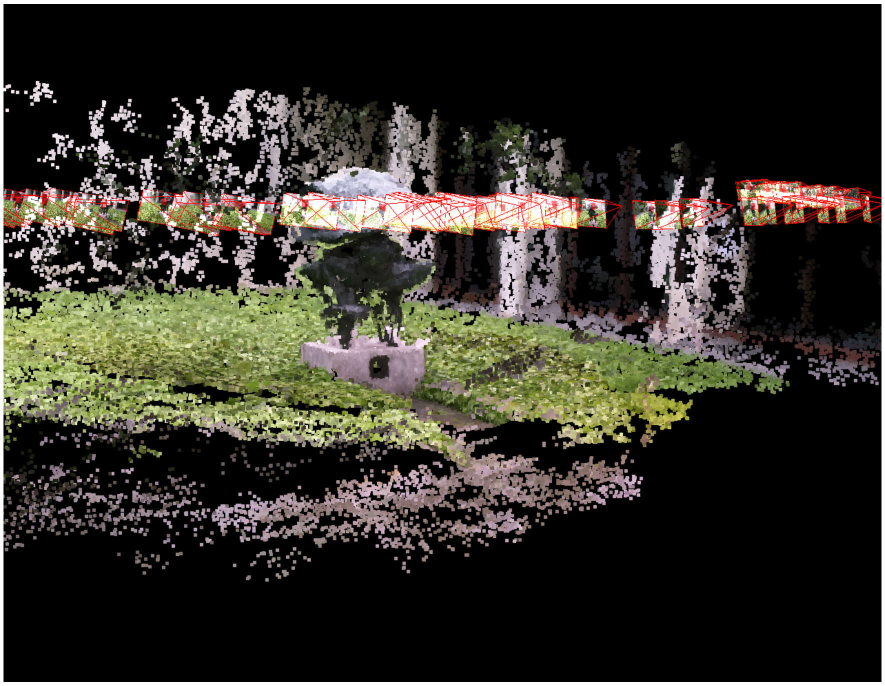


Figure 79: Elevation perspective that shows 90 keyframes captured from a single subject as they moved about the sculpture. Looking from a side view, the subject's physical height becomes visible. All frames are constrained to a small band of variance in the vertical dimension.

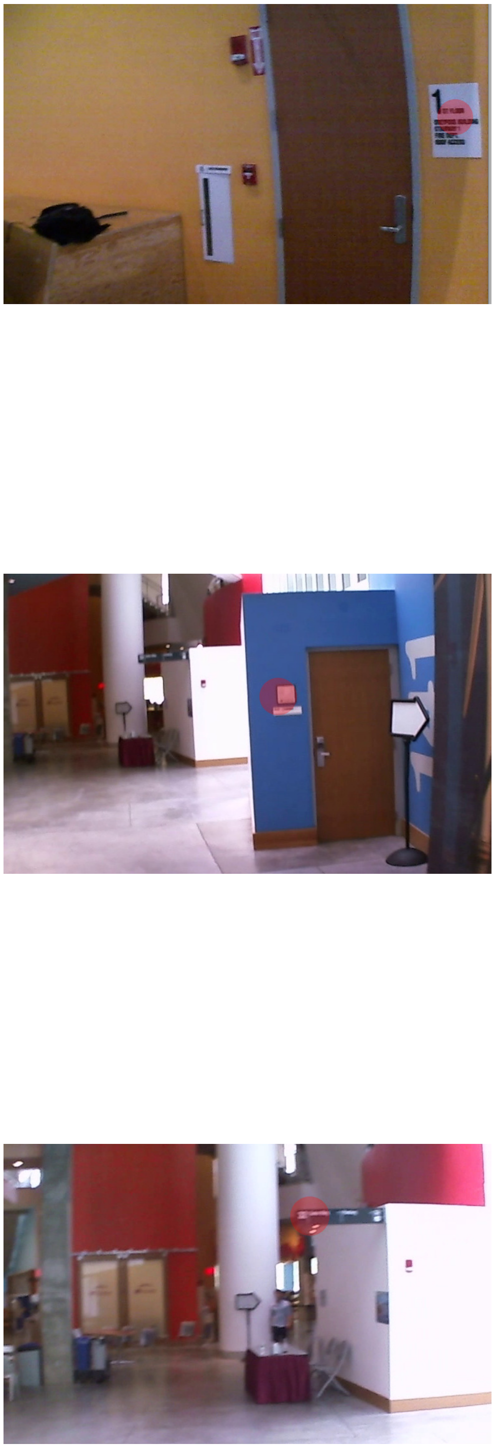


Figure 80: Three key frame images extracted from the two dimensional browser where the red dot shows the subject's area of visual attention. Each key frame shows the subject reading room numbers as they moved through the student street in the Stata center at MIT.

Trial 3: Search in three dimensions

This recording session took place in the Stata center on the MIT campus. Walking down the main corridor the subject was asked look for the highest room number, and read them out aloud as they moved through the space. Here we present only a two dimensional series of images extracted from our two dimensional browser.

What becomes apparent in these trials is that visual tasks, like searching for a room number, form highly recognizable patterns of visual attention. This trial confirms previous research of top-down motivation in human vision. For example, the search space is pre constrained to plausible locations that the subject believes he will find room numbers. In this trial, the subject did not attend to the ceiling or floor at all. Furthermore, in this trial the eye always precedes head movements. This observation is supported by all of our trials.

The frequency and speed at which the eye processes information from the world is astounding. In order to properly capture a task of search with a moving subject, one would need capture the world in higher temporal resolution than our current method.

Trial 4: Critique of a space

Using the same location as the previous trial, the subject, a designer by training, was asked to critique the space as they moved through the main corridor. Here we present a series of two dimensional images extracted from our browser.

In this trial we find that there is much less constraint in the area of visual fixation. The subject examines a much wider range of the space, frequently attending to the ceiling. Each one of these trials were conducted with an audio recorder. We found that verbal descriptions, in this trial, were always preceded by the fixation pattern of the eye. During critique the area of visual attention remained fixated on the feature being discussed or critiqued. By closely examining the visual fixation patterns in our two dimensional browser, we find that the fixations jump from one area to another when discussing a comparison between features in the world.

In the second half of this trial, we observed a predominant pattern where the subject fixated upon edges and corners. In reviewing the audio recordings, the subject did not verbally address the topic of edges, but did discuss broader observations about the geometries of the space. The three figures to the right, exemplify the subject's fixations on edges. Some of these "edges" are defined by geometry and others by texture.

We believe that this observation is interesting as a parallel to computer vision, where edge detection is a method for making sense of the information embedded in images. For computer vision, analogous to human vision, an edge signifies discontinuity and therefore information.

Trial 5: Navigation and intrigue

This trial took place in the STATA center, beginning at a set of stairs that connect the second floor to the third floor. The subject had never visited this space before, so he did not know what to expect spatially. A virgin spatial experience. Not only was this a special experience for the subject in its novelty, it also introduced noticeable navigational challenges, as the subject had to coordinate his ascent while inspecting the new environment.

Even though the capture only elapsed 30

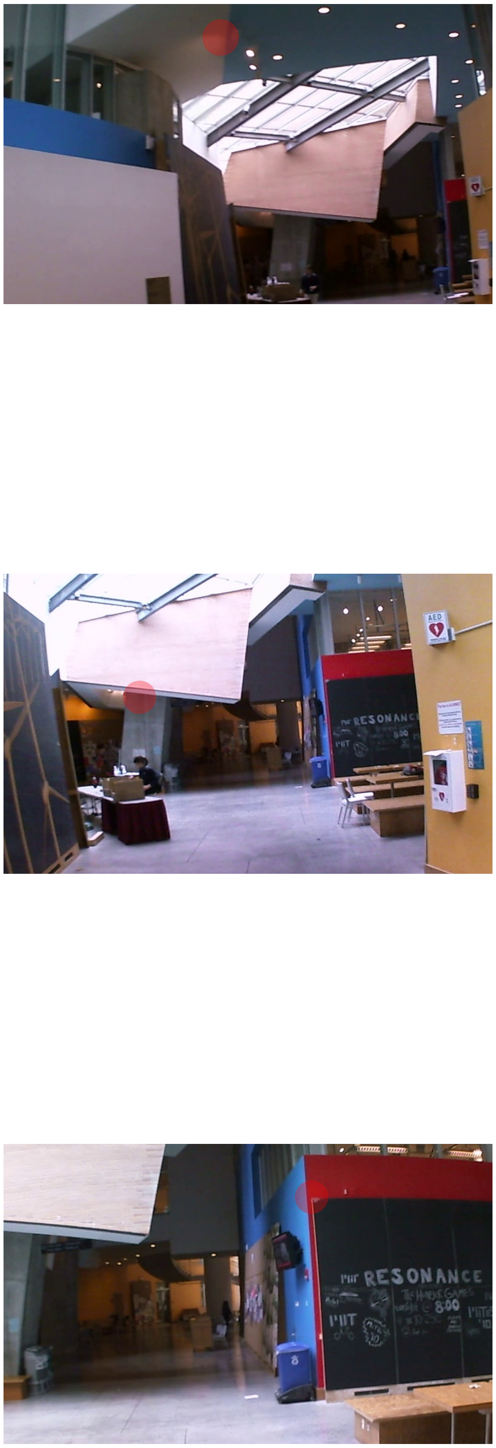


Figure 81: Three key frame images extracted from the two dimensional browser where the red dot shows the subject's area of visual attention. Each key frame shows the subject fixating on "edges" or salient differences in the space.

Figure 82: The subject walked up a flight of stairs from the second to the third floor of the Stata center at MIT. This is the subject's first encounter with this space. The red pyramids are the pose of the subject in the space. This capture lasted 30 seconds.

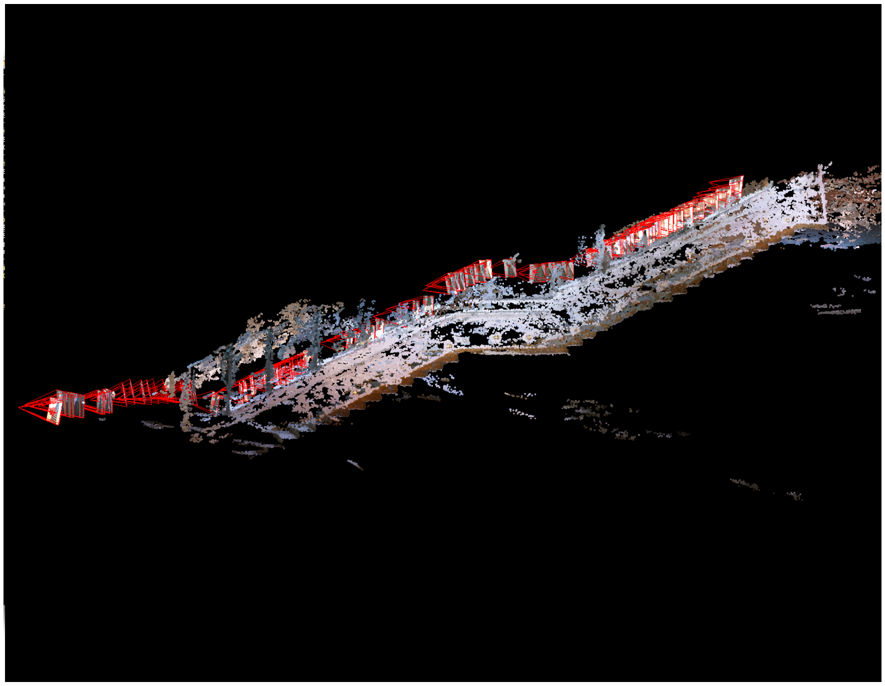
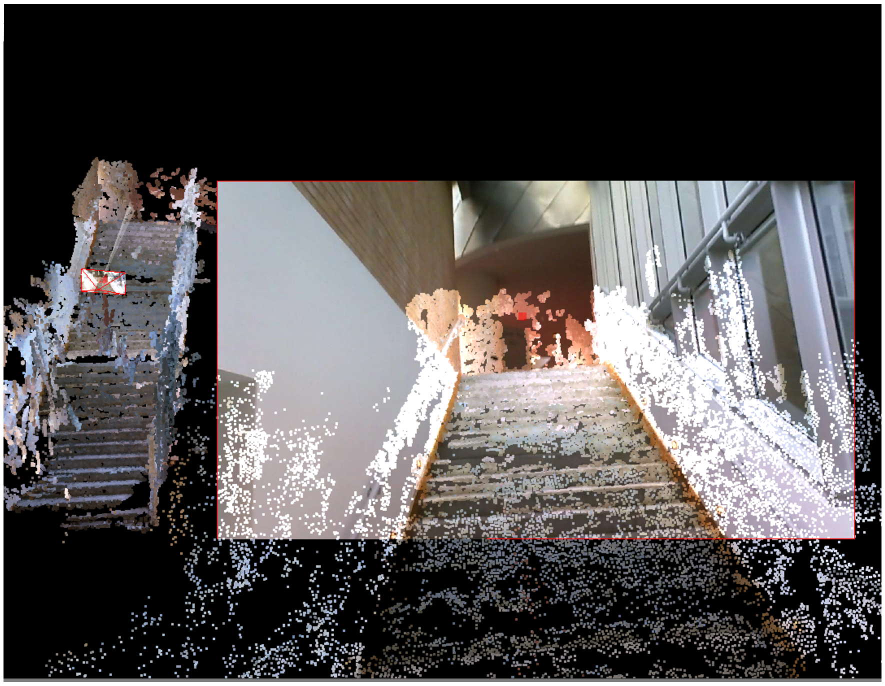


Figure 83: This screen capture from our three dimensional browser shows both the view of the subject on the right as a two dimensional image with the subject's point of visual attention as a red dot, with the three dimensional point cloud aligned with the image. On the left, the subject situated in the three dimensional construction as viewed from a simulated third person perspective. Here we see the subject on the landing, midway up the stairs, inspecting the space he anticipates to arrive at, as it becomes visible from his current elevation.



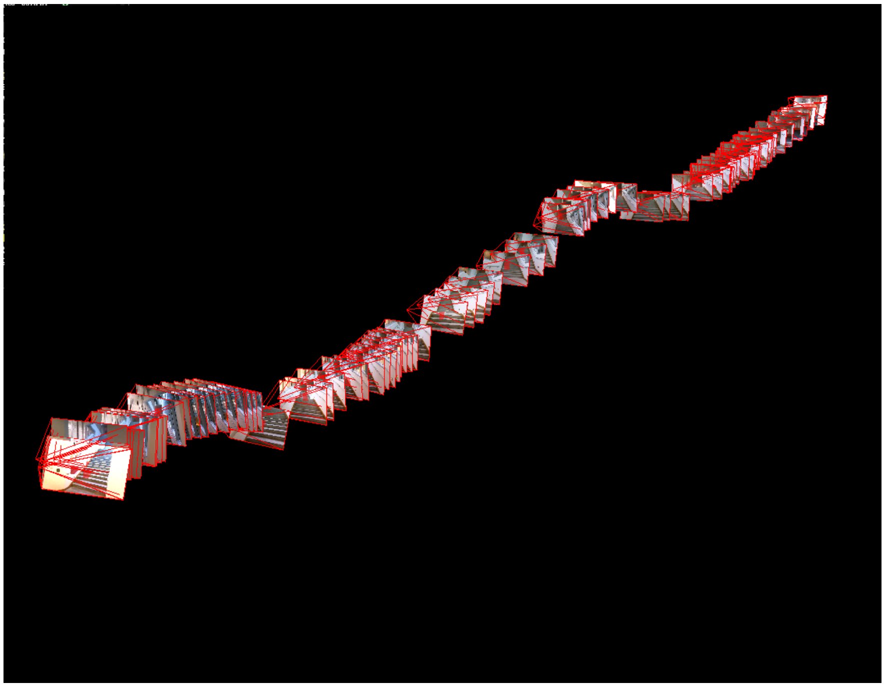


Figure 84: This image is a screen capture from the three dimensional browser that shows 87 keyframes that were used in the SfM construction. From these images, even without the point cloud, we can make out the typology of the space.

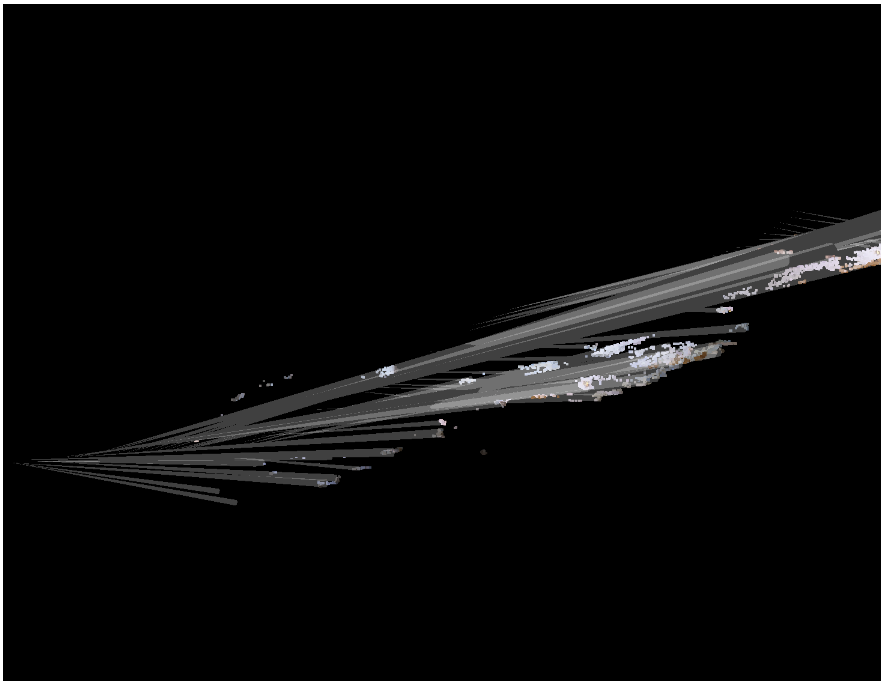


Figure 85: Side elevation view that shows the cones of visual attention and the points in the three dimensional model that intersect the cones.

seconds, the density of fixations reveal two main motifs, that of motor planning and curiosity driven inspection.

In motor planning, the subject fixates on the edges of the stair treads of each flight prior to ascending. This is visible in the three dimensional browser, where the cones of visual attention show a distinctive pattern as the subject begins ascending the first flight of stairs, and again after the first landing.

The second motif becomes apparent as soon as motor planning does not seem to require visual attention. Here the patterns of visual attention can be characterized by an alternation of fixations between features in close proximity, such as the handrail, and distant fixations toward the unfolding space at the top of the stairs.

Discussion

In conducting these trials we have accumulated a wealth of information, only a fraction of which we have been able to analyze at the time of writing. As we have demonstrated, the amount of information the human eyes, and brain, are able to process even within a short duration of 10 to 30 seconds is astounding. Furthermore, what we as observers can learn about a human experience — even in what may seem like incredibly brief temporal slices — not only confirms prior research but, we believe, sets the foundations for future work in studying spatial experiences.

The first two trials we present serve two main purposes. First, they support prior research human vision on a physiological and psychological level. For example, we have confirmed that not only do humans need to move their eyes in order to process their surrounding environment, but they also are unconscious of their own eye movements. By recording the subject and employing speak aloud protocols, we gain insight into what the human subject believes that they are doing – or a glimpse of a subject’s cognition. By correlating the audio recordings and the patterns of visual attention in these trials we have confirmed that human vision is motivated both by high level cognitive demands and salient features in the world. Second, these early trials demonstrate the precision and capabilities of **Pupil**. We are certainly not the first to construct a head mounted eye tracking prototype. However, with **Pupil,** we are able to capture the world with higher spatial resolution and movements of the eye at higher temporal resolution than most commercial solutions. Furthermore, to our knowledge, we are the first researchers to situate visual attention in a three dimensional space. The Hayden Library trial serves as a demonstration of the precision of our instruments and the novelty of our representational methods.

Beginning with the third trial, constraints on the subjects are considerably relaxed in comparison with the prior two trials. In the third trial the subject is allowed to freely move through a space while primed to search for "the highest room number." What we have learned from this trial is that verbal priming, or motivation, serves as a constraint in the task of visual search. As the subject was searching for "room numbers," the space of his search was constrained within a horizon area roughly between a door handle and top of a door. The particularities of these constraints leads to questions about the prior experience and background of the subject and the design of the space. Specifically in the second trial, we observe the subject attending to the small placards in search of room numbers and projecting signage, but not to the "super graphics." From a design perspective we could begin to inquire into how the scale of informational graphics affect the experience of a space and attention of a subject.

In the fourth trial a subject, a designer, is asked to critique a space while moving through the space freely. In this trial we observe that the subject's field of vision is much less constrained in comparison to trials that involve visual search tasks. This trial reveals interesting phenomena between vision and speech, where the subject's visual attention precedes verbal description or critique. Furthermore, we observed that the area of visual attention is in direct relationship to the topic of critique. At moments when the subject is making a comparative spatial critique, the area of visual attention can be observed to be jumping back and forth in the space and making cycles, revisiting prior areas of fixation. We also found that the subject attended to edges or corners when discussing geometric relationships. In this trial, the subject was critiquing the collision between geometric primitives that intersect at skewed angles.

In the fifth trial, we investigate what we assume to be different visual tasks and the shift between the two tasks. In this trial the subject was walking up a flight of stairs that he had never experienced prior to this trial. Prior to ascending the stairs we observe the subject fixating on the edges of stair treads, a task that appears to be crucial for motion planning. We believe that the visual planning of motion occurs prior to the physical movement. Interleaved between these planning motifs are patterns of visual attention that seem to be driven by the curiosity of the subject. These fixations are concerned with both close small details and distant investigations into the gradually unfolding space at the top landing of the stairs. This dichotomy becomes apparent in our three dimensional browser. By examining the cones of visual attention alone, the typology of the space and the subject's motivations and actions are revealed.

Future Work

In the process of development of **Pupil**, we have discovered many new areas of inquiry that we would have liked to pursue but were limited by the duration of our thesis. Furthermore, by conducting a series of trials, some of which are included in this book, we have identified deficiencies in our instruments and representational methods. In this section we will outline areas of inquiry for future research and the next steps for the development of our platform.

Three Dimensional Browser

- Play interpolated video. Currently our three dimensional browser uses only a subset of the video stream captured by Pupil. This means that we are only showing a fraction of the information that we have collected. In order to more fully represent the space of human experience, it will be necessary to include a higher temporal resolution. We imagine that this feature could still employ what we call "keyframes" to mark frames used in the SfM pipeline, but would allow the video to play along a Bezier curve between these keyframes.

- Improve 3D navigation. Currently navigation uses a three button mouse and a keyboard as modifier keys for panning. This is enabled by GlumPy's wrapper around OpenGL's GLUT event handlers. As the GlumPy project is rather young the GLUT back end is not fully developed and some keys are not accessible. Minor changes need to be made in order to make the navigation smoother and more friendly to the user by enabling zooming with a scroll wheel, for example. Of course, one could also imagine other modes of navigation that involve eye movements themselves, not as a “pointer” but as an extension of the body in a virtual space.

- Associate points with images. After performing bundle adjustment with Bundler, we have a known correspondence between the three dimensional points and the two dimensional images. With only a few small steps we can read these correspondences from the bundle.out file and associate these points with each image. This would provide a didactic understanding of how SfM works as well as an analytical tool to inspect which points are visible from each image – or each position of the subject in space. This feature would also enable a more precise calculation of intersections between the area of visual attention and the point cloud.

- With video playback we will integrate audio output. The audio signal could be visualized as an abstract two or three dimensional object in the three dimensional browser environment.

- Transcription of the audio track could lead to annotation of the audio object and correlation of audio annotation and points of visual attention.

- Points of visual attention could then be tagged with symbols, either based on audio-visual correlation, computer based object-recognition or human assisted tagging.

- During fixations, the area of visual attention often appears to be glued to a feature of interest, despite rapid head movements. Because such fixation makes heavy use of the visual stabilization and compensation reflexes, raw pupil position data shows movement. If we were to asses the movement of the area of visual attention in the world view using optical flow, we will be able to calculate relative displacements of the center of the pupil on world space in relation to such features on the actual world image. This way we should be able to detect fixation even under head movements.

- The pupil position signal could then be analyzed (filtering, PCA, FFT) and visualized.

- Because the relative movements of fixations have characteristic signal properties, machine learning algorithms could be trained to discriminate between saccades, fixations and drift. Such classification could be added to the pupil positions and used for evaluation and analyses in the two dimensional and thee dimensional browser.

- Improved representation of peripheral vs. foveal vision. Currently we do not represent the area of peripheral vision in the three dimensional browser. We would like to enable a representation where the three dimensional points attended to are illuminated and in color, while the all that is peripheral is gray scale and potentially blurred.

SfM-Pipeline

The three dimensional pipeline worked and produces useful, accurate and rich output, but was rather picky and would occasionally fail. We believe this happened when image quality was insufficient, the image stream had spatial discontinuities, and spaces yielded few useful features.

The structure from motion library we use makes no assumption about the order or spatial continuity of the input image set, without these assumptions many constraints inherent to the captured images are disregarded. In future implementations we plan to address this issue and hope to increase speed and robustness the 3D constructions by exploiting the additional constraints. Ideally we would have time to write our own bundle adjustment algorithms.

- Use of system constraint. Currently our SfM Pipeline assumes an unordered set of images, without spatial and temporal consistency. Our trials are always temporally ordered, therefore we hope to use this constraint to improve the quality of SfM constructions. Due to these known constraints, we believe that we could implement a real time sparse reconstruction that would result in projection matrices from a sparse set of features tracked between sequential frames of a video stream.

- Compare and benchmark different feature detectors in a real life example. We have already implemented other feature detection algorithms in the SfM pipeline from the OpenCV library. However, we have yet to test other feature detection algorithms other than SIFT. In order to determine which feature detection algorithms are best suited for a scene and for SfM bundle adjustment, we would need to first benchmark the performance of each with a real life example. In the future it may be possible to employ more than one feature detection algorithm in order to improve performance.

- Multi thread key point matching. Many elements of our SfM Pipeline are already multi threaded, leveraging the power of contemporary multi core processors. One bottleneck however in our pipeline is the key point matching routine, which uses a k-d tree to cluster and correlate the potential matches between features. This process takes a lot of time and could be improved by parallel processing.

- Multi core bundle adjustment. Currently bundle adjustment algorithms we are using are only running on a single core. Other algorithms, optimized for multi core processors exist and are available as open source.[1]

- Add to existing 3D construction. We have implemented a way to add to the current SfM construction, using features available in the bundler package. This feature allows us to calculate new projection matrices given an existing model. However, this does not allow us to add new three dimensional points to the construction. This feature would be useful not only for the calculation of the points, but on a higher level to inspect the relationship between a subject returning to a space or the overlapping paths of multiple subjects.

Two Dimensional Browser

- Proper timeline scrubbing. Once we have a proper decoder for the video compression we will be able to know exactly where keyframes are located in the video stream and interpolate between these smoothly. This would allow one to scrub through the video instead of stepping and playing as is currently implemented.

- Visualize sound as a spectrogram. By visualizing the sound as a spectrogram, we can

[93] Changchang Wu, “Multicore Bundle Adjustment” http://grail.cs.washington.edu/projects/mcba/, visited May, 2012; Ceres Solver Project Page, “Ceres-Solver: A Nonlinear Least Squares Minimizer,” Visited, May 2012.

identify human speech events.

- Better representation of peripheral vs. foveal vision. Currently we show the area of attention as a dot superimposed on the video stream. This area should serve as an elliptical mask. Outside of this mask we would simulate peripheral vision with a Gaussian blur over a grayscale image. The blur falloff would be calibrated based on physiological constraints of the human eye.

Capture Routine

As future work increases in tracking robustness under varying lighting conditions will be addressed as we would like to be able to fully rely of this part of the system without having to occasionally check the results during the trail runs. Potential strategies for improvement are:

- Using the Canny edge detection algorithms and arc-fitting in combination with a Random Sample Consensus (RANSAC) scheme to find the pupil center.[2]

- Using RANSAC for calibration surface fitting.

- Another potential area of improvement would be speeding up bundle adjustment to allow for real-time 3D pose estimation in three dimensional space. Thus allowing for real-time interaction with a 3D-construction of the scene as experienced by the subject.

Electronics

[94] Martin A. Fischler and Robert C. Bolles (June 1981). "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography". *Comm. of the ACM* **24** (6): 381–395.

Improvements in electronics design of the capture system are driven by afford ability, reduction of system complexity, and reduction of visual and ergonomic obtrusiveness:

- Reduce PCB footprint
- Reduce component count
- On-board image filtering would allow for less load on the data interface and computational processing on the host computer.

Headset

- Find ways to allow it to be used in conjunction with eye-glasses.
- Explore different material options.

Conclusion

In this thesis we have provided a theoretical framework for the study of the nature of a human experience – vision – through a survey of historical precedents and contemporary theory. We have developed our own studies on the relationships between a human subject and a constructed environment using a custom designed hardware and software platform that we developed – **PUPIL**. Trials were conducted using the platform we developed that both reflexively revealed the biases and assumptions of our prototype and took steps toward understanding the complex relationships of a human experience of space. Our inquiry into this relationship began with a simple question, what does one *really* see as they move through a space? While this question may ultimately be unanswerable, we believe that we have taken significant steps towards framing this question and have developed an arsenal of tools that will enable us and others in a diverse range of disciplines to pursue and refine these questions in the near future. The representation of human vision has, and will continue to be contested territory in domains that make claims to quantification, truth, and objective knowledge. In diverse fields, from experimental psychology to architectural design – we find a collision between art and science and reversals in the subjectivity/objectivity dipole. We examine how representations of vision reveal how objectivity is constructed through different attitudes and practices in representation and instrumentation. Furthermore, the discussion of an intercalated v between instrumentation, experimentation, and theory allows us to better situate not only the work of the historical actors we survey, but our own research, where we are developing new instruments that may enable us to see the world differently – to see how we see.

Objectivity is not fixed, it is a construction that is made manifest in representations. This process of constructing objectivity is always a negotiation or series of exchanges instrumentation, theory, and experimentation.

We conceptualize human vision as information processing, rather than image formation, as it allows us to make connections between the physiological realities of human vision and the algorithmic constructions of computer vision, but also because as a metaphor it immediately disassociates representations of vision from the realities of human vision. This position reinforces a theoretical framework that reveals the roots of a malleable and shifting epistemology of objectivity and simultaneously allows one to appreciate the known physiological realities of human vision. As we now understand it, the physiological realities of human vision reveal that what we actually see is a visually fragmented world. The continuous world, that we perceive to "see," is a sophisticated construction of information processed. Our eyes move rapidly to scan through the world and send salient information about the environment around us to the optical nerves and up to the brain to be processed.

We capture the positions of the pupil as an artifact of the selective process of spatial experience, unique to each subject. These physiological artifacts, in combination with film records of the subject's field of view provide us insight into both the way a subject processes information of an environment and the environment as image. We have developed representational methods that coordinate a subjective experience and images of that environment, spatially and temporally. This results in a dense cloud of light points that can be read as analytic records of a specific space and the position of a subject situated within that space at a moment in time. These resulting representations are constructions that were made possible through the development of new instrumentation — **PUPIL**. With these representations we hope to understand how a human attends to a space and perhaps how one constructs a story of their experience in a space.

We have provided the instrumentation, and have established a theoretical framework, and representational methods. Many of the questions we initially posed and developed during the course of this thesis remain open. However, we believe the next steps are incredibly promising and we hope that this line of inquiry will be continued by other researchers in diverse fields.

Bibliography

1. Sameer Agarwal et al., “Building Rome in a Day,” in *International Conference on Computer Vision* (Kyoto, Japan, 2009).

2. Richard Bolt, “Eyes at the Interface,” *Association for Computing Machinery*, Architecture Machine Group (1981): 360-362.

3. Richard Bolt, “Eyes at the Interface,” *Association for Computing Machinery*, Architecture Machine Group (1981): 360-362.

4. Gary Bradski, *OpenCV: Open Source Computer Vision Library*, 2000.

5. Guy T Buswell, *An experimental study of the eye-voice span in reading* (Chicago: University of Chicago, 1920).

6. Guy T Buswell, *An experimental study of the eye-voice span in reading* (Chicago: University of Chicago, 1920).

7. Italo Calvino, *Invisible Cities* (Harcourt Brace Jovanovich, 1978).

8. W Charman, “Optics of the Human Eye,” in *Visual Optics and Instrumentation* (Boca Raton: CRC Press, 1991).

9. Tom N Cornsweet, *Visual perception* (Academic Press, 1974).

10. Jonathan Crary, *Techniques of the observer : on vision and modernity in the nineteenth century* (Cambridge Mass.: MIT Press, 1990).

11. Jonathan Crary, *Techniques of the observer : on vision and modernity in the nineteenth century* (Cambridge Mass.: MIT Press, 1990).

12. Lorraine J. Daston and Peter Galison, *Objectivity* (Zone, 2010).

13. Doug DeCarlo and Anthony Santella, “Stylization and Abstraction of Photographs”, 2002.

14. Raymond Dodge, “Five Types of Eye Movement in the Horizontal Meridian Plane of the Field of Regard,” *American Journaly of Physiology* (1903): 307-329.

15. Marc Downie and Paul Kaiser, *Spatializing Photographic Archives*, White Paper for the National Endowment for the Humanities, 30 2010.

16. Heiko Drewes, *Eye Gaze Tracking for Human Computer Interaction* (München: LFE Medien-Informatik der Ludwig-Maximilians-Universität, 2010).

17. Paul Feyerabend, *Against Method: Outline of an Anarchistic Theory of Knowledge*, 3rd ed. (Verso, 1993).

18. John M. Findlay and Iain D. Gilchrist, *Active Vision: The Psychology of Looking and Seeing*, 1st ed. (Oxford University Press, USA, 2003).

19. Yasutaka Furukawa and Jean Ponce, "Accurate, Dense, and Robust Multi-View Stereopsis," *IEEE* 1, no. 1 (August 2008).

20. Gad Geiger, Jerome Lettvin, and Olga Zegarra-Moran, "Task-determined strategies of visual process," *Cognitive Brain Research* 1 (1992): 39-52.

21. F. Gonzalez-Crussi, *On Seeing: Things Seen, Unseen, and Obscene*, 1st ed. (Overlook Hardcover, 2006).

22. Clarance Truman Gray, *Types of reading ability as exhibited through tests and laboratory experiments : an investigation subsidized by the General education board* (Chicago Ill.: University of Chicago Press, 1917).

23. Richard L. Gregory, *Eye and Brain*, 5th ed. (Princeton University Press, 1997).

24. Richard Hartley and Andrew Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. (Cambridge, England: Cambridge University Press, 2003).

25. Mary Hayhoe and Dana Ballard, "Eye Movements in Natural Behavior," *TRENDS in Cognitive Sciences* 9, no. 4 (2005): 188-194.

26. David H. Hubel, *Eye, Brain, and Vision*, 2nd ed. (W. H. Freeman, 1995).

27. Edwin Hutchins, *Cognition in the wild* (Cambridge Mass.: MIT Press, 1995).

28. Eric Jones, Travis Oliphant, and Pearu Peterson, *NumPy: Open Source Numeric Computing Tools for Python*, 2001.

29. Eric Jones, Travis Oliphant, and Pearu Peterson, *NumPy: Open Source Numeric Computing Tools for Python*, 2001.

30. Tilke Judd, *Understanding and Predicting Where People Look in Images* (Cambridge, Massachusettes: MIT PhD Dissertation, 2011).

31. Brian W. Kernighan and Dennis M. Ritchie, *C Programming Language, 2nd Ed*, 2nd ed. (Prentice Hall, 1989).

32. Thomas S. Kuhn, *The Structure of Scientific Revolutions*,

3rd ed. (University Of Chicago Press, 1996).

33. Michael Land and Benjamin Tatler, *Looking and Acting: Vision and eye movements in natural behaviour*, 1st ed. (Oxford University Press, USA, 2009).

34. Bruno Latour, “Visualisation and Cognition: Drawing Things Together,” *Knowledge and Society Studies in the Sociology of Culture Past and Present* 6 (1986): 1-40.

35. Bruno Latour, “Visualisation and Cognition: Drawing Things Together,” *Knowledge and Society Studies in the Sociology of Culture Past and Present* 6 (1986): 1-40.

36. Gad Geiger, Jerome Lettvin, and Olga Zegarra-Moran, “Task-determined strategies of visual process,” *Cognitive Brain Research* 1 (1992): 39-52.

37. Richard A. Monty and John W. Senders, *Eye Movements and Psychological Processes* (John Wiley & Sons Inc, 1976).

38. David Noton and Lawrence Stark, “Eye Movements and Visual Perception,” *Scientific American* 224 (1971): 34-43.

39. Marc Pollefeys, *Visual 3D Modeling from Images* (Chapel Hill, North Carolina: University of North Carolina, n.d.).

40. Cliodhna Quigley, Selim Onat, and Sue Harding, “Audio-visual integration during overt visual attention,” *Journal of Eye Movement Research* 2, no. 4 (2008): 1-17.

41. Nicolas Rougier, *GlumPy: Fast OpenGL NumPy Visualization*, Macintosh & Linux, 2011.

42. Noah Snavely, Steven Seitz, and Richard Szeliski, “Photo Tourism: Exploring Image Collections in 3D,” in *ACM Transactions on Graphics*, 2006.

43. Noah Snavely, Steven Seitz, and Richard Szeliski, “Photo Tourism: Exploring Image Collections in 3D,” in *ACM Transactions on Graphics*, 2006.

44. Jan Solem, *Programming Computer Vision with Python* (Book Draft, 2012).

45. Noah Snavely, Steven Seitz, and Richard Szeliski, “Photo Tourism: Exploring Image Collections in 3D,” in *ACM Transactions on Graphics*, 2006.

46. Michael Land and Benjamin Tatler, *Looking and Acting: Vision and eye movements in natural behaviour*, 1st ed. (Oxford University Press, USA, 2009).

47. John Tchalenko, “Eye movements in drawing simple lines,” *Perception* 36 (2007): 1152-1167.

48. E. Llewellyn Thomas, “Movements of the Eye,” *Scientific American* 219 (August 1968): 88-95.

49. Michael Land and Benjamin Tatler, *Looking and Acting: Vision and eye movements in natural behaviour*, 1st ed. (Oxford University Press, USA, 2009).

50. Nicholas Wade and Benjamin Tatler, *“The Moving Tablet of the Eye”: The Origins of Modern Eye Movement Research*, 1st ed. (Oxford University Press, USA, 2005).

51. Nicholas Wade and Benjamin Tatler, *“The Moving Tablet of the Eye”: The Origins of Modern Eye Movement Research*, 1st ed. (Oxford University Press, USA, 2005).

52. Bruno Latour and Steve Woolgar, *Laboratory life : the social construction of scientific facts* (Beverly Hills: Sage Publications, 1979).

53. XMOS Corporation, ed., *XC Programming Guide* (United Kingdom: XMOS, 2011).

54. Alfred L. [Basil Haigh, translator] Yarbus, *Eye Movements and Vision* (Plenum Press, 1973).

List of Figures