VRCodes: Embedding Unobtrusive Data for New Devices in Visible Light

by

Grace Woo

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2012

Author .....
Department of Electrical Engineering and Computer Science
August 31st, 2012

Certified by .....
Andy Lippman
Senior Research Scientist
Thesis Supervisor

Accepted by .....
Leslie Kolodziejski
Chairman, Department Committee on Graduate Theses

VRCodes: Embedding Unobtrusive Data for New Devices in Visible Light

by

Grace Woo

Submitted to the Department of Electrical Engineering and Computer Science
on August 31st, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering

Abstract

This thesis envisions a public space populated with active visible surfaces which appear different to a camera than to the human eye. Thus, they can act as general digital interfaces that transmit machine-compatible data as well as provide relative orientation without being obtrusive. We introduce a personal transceiver peripheral, and demonstrate this visual environment enables human participants to hear sound only from the location they are looking in, authenticate with proximal surfaces, and gather otherwise imperceptible data from an object in sight.

We present a design methodology that assumes the availability of many independent and controllable light transmitters where each individual transmitter produces light at different color wavelengths. Today, controllable light transmitters take the form of digital billboards, signage and overhead lighting built for human use; light-capturing receivers take the form of mobile cameras and personal video camcorders.

Following the software-defined approach, we leverage screens and cameras as parameterized hardware peripherals thus allowing flexibility and development of the proposed framework on general-purpose computers in a manner that is unobtrusive to humans. We develop *VRCodes* which display spatio-temporally modulated metamers on active screens thus conveying digital and positional information to a rolling-shutter camera; and physically-modified optical setups which encode data in a point-spread function thus exploiting the camera's wide-aperture. These techniques exploit how the camera sees something different from the human.

We quantify the full potential of the system by characterizing basic bounds of a parameterized transceiver hardware along with the medium in which it operates. Evaluating performance highlights the underutilized temporal, spatial and frequency dimensions available to the interaction

designer concerned with human perception. Results suggest that the one-way point-to-point transmission is good enough for extending the techniques toward a two-way bidirectional model with realizable hardware devices. The new visual environment contains a second data layer for machines that is synthetic and quantifiable; human interactions serve as the context.

Thesis Supervisor: Andy Lippman
Title: Senior Research Scientist

Acknowledgments

This work exists due to the support of those who speak openly, move quietly and gamble generously on suspiciously-quiet children such as myself.

to Andy Lippman, who in 1990, barked: "Television is after all one of those things that everyone watches and nobody does something about. -But we do do something about it upstairs".

to Ramesh Raskar, who allowed me to join during his own first few days at the Media Lab and on his own quest to define a new camera for the future.

to Gerald Sussman, Vincent Chan, and Joi Ito whose wisdoms became more obvious close-up than from afar.

to Douglas L. Jones at the University of Illinois Urbana-Champaign, who provided guidance for only a couple of years starting in my undergraduate years but laid the foundations for a philosophical degree.

to my family that supports me in all the crazy ways you could possibly imagine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Imagine a synthetic physical world where visible surfaces appear unsuspicious but act as general digital interfaces which transmit machine-compatible data and provide relative orientation and positioning. How can we engineer this visible light environment to provide new interactions and remain human-compatible?

This dissertation presents the design, implementation and evaluation of a novel visible light-based communications architecture based on the wide availability of devices which can transmit and receive light. The goal of the software-defined interface is to create an interactive system for which any aspect of the signal processing can be dynamically modified to fit the changing hardware peripherals and well as the demands of desired human interaction through the visible light medium.

This approach to the design of a synthetic visual environment transforms ordinarily passive pixels into interactive environments that allow machine-to-machine communications and overcome many of the limitations imposed by radio frequency (RF) interfaces. It allows ordinary transmission of visual light to convey proximal data that is understandable by a machine but not obtrusive to the human.

Flexibility and experimentation using this framework allows new techniques to be evaluated on screens and cameras that already exist without the requirement for special-purpose devices. A wide range of techniques that deliver relative positioning and digital data may be quickly tested using this approach. The results can be adapted for readily available hardware. The dense population of public screens can embed information for the dense population of personal cameras while remaining unobtrusive to the human.

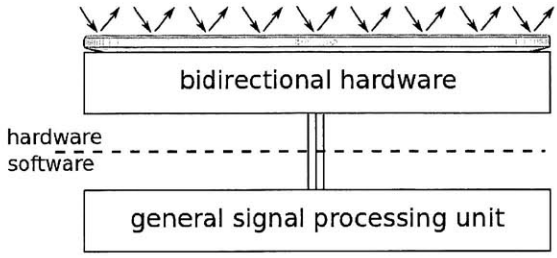

Figure 1.1: An ideal software-defined testing system with pixels on the surface and a signal processing platform that can be developed in parallel. The goal is to be mindful of what our eyes see while making our devices less blind to the rich information in the physical world.

The primary design goal is to create an interactive system with the flexibility to make tradeoffs and optimizations. The ideal system in Figure 1.1 depicts a general-purpose physical frontend which delivers digitized data and makes a clear distinction between the pixelated surface and the underlying system. This architecture makes the pixel surface analogous to an antenna which emits light that is at the same time observable by the human.

A clear separation between a hardware and software development; and an abstracted universal hardware peripheral allows new signal processing techniques to be developed according to the desired goals for human compatibility without building custom prototype hardware. The architecture for a general testing platform presented consists of three parts:

- Pixelated frontend hardware visible to both humans and machines (both surface and captured imagery)
- I/O system for transporting digital data to and from memory. We later describe these mechanisms in terms of specific hardware
- A programming environment to support the implementation of computationally intensive, high data rate, accurate positioning and real-time signal processing.

Through the description of this system, we argue the power of creating this basic platform for testing and inventing new transmission techniques. We use this framework to demonstrate a few novel techniques of our own. Specifically, we present VRCodes (Woo et al., 2012) which are implemented using displays and cameras already widely available today.

1.1 The Need to Move Beyond Radio Frequencies

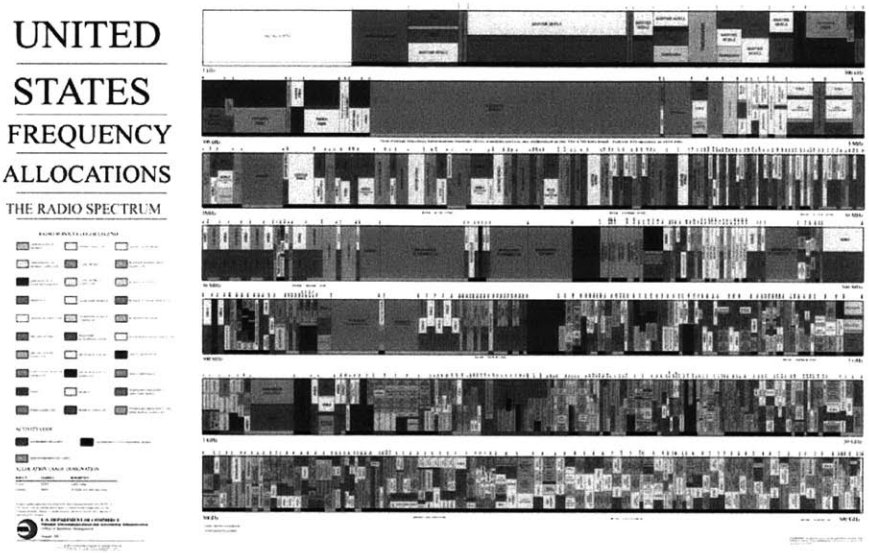

Figure 1.2: Radio frequencies are divided up and auctioned off like real estate thus making it difficult for newcomers to innovate and leverage EM spectrum.

Historically, radio spectrum has been used as the medium for communicating between devices. Today, that spectrum is controlled by a centralized entity, cut up and auctioned off like real estate to large organizations interested in providing long-range communication services (Galbi, 2003). Due to regulation, it is now difficult to innovate using the RF medium. The EM space is seemingly crowded. Figure 1.2 depicts federal regulations which ultimately limit human interactions to those provided by cellular networks, one-way paging services and WiFi protocols.

Directed proximal interactions in the environment dictates the need for utilizing a new medium which is perceivable by humans and supportive of machine-to-machine communications. Light-based communications encompasses elements of directionality and high-capacity datarates, both properties of the short wavelengths unique to the visible light range. The medium is minimally regulated and also controlled by diverse entities. Widely deployed hardware which produce artificial lighting are already widely available due to the human-needs they serve.

The potential of light-based systems comes from taking into account how

the human eye and brain perceive the visual environment. While the human visual system (HVS) is complex and can be unpredictable in what it sees, the camera can be described concisely with parametrized variables. Based on the definition of a universal peripheral device and the currently observed properties of the human visual system, there are clear and quantifiable metrics for the maximum amount of information in an information theoretic sense that can be embedded in the synthetic visual environment.

In comparison to RF, however, inventing new transmission techniques for interaction in the public visual space is challenging. Urban planners have long addressed the problem of excessive or obtrusive artificial light by defining common metrics for light pollution. In addition to the classical challenges present in radio frequency (RF) spectrum regulation, sloppy guidelines for the visual environment can result in a distraction that impact cultural norms.

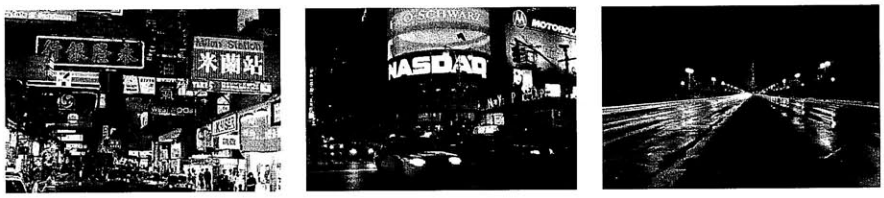

Figure 1.3: Regulation of visible light spaces differ according to culture. From left to right are urban environments found in Hong Kong, New York and Germany.

Figure 1.3 shows how diverse societies treat light in different ways. In addition to the familiar RF guidelines which limit power emission and frequency allocation, Figure 1.3 shows how regulation must also be mindful of design, human perception and cultural traditions. Acceptable techniques leveraging the visible light medium must not only deliver fast datarates and precise positioning to a peripheral device, but also remain considerate toward existing cultural norms. On the left in Figure 1.3, is Hong Kong where busy signage decorate busy streets. In the center, New York Times Square boasts high-priced real estate allowing for a few iconic displays. On the right, Germany deploys new lights which can only be activated using a personal mobile phone.

Digital displays, signage and street lamps are man-made hardware peripherals that serve artificial light for humans. In contrast to the radio which requires deployment of hardware that can specifically serve the purpose of communicating amongst machines, a wide variety of synthetic devices which are capable of both transmitting and receiving light already exist in our

environment. The visual sense is the strongest sensory input of the human [1]. Nearly all of the visual objects in Figure 1.3 are artificial sources of light which serve this sense.

The techniques presented in this thesis propose to add interaction to these visual environments by first refreshing what we do know about the human visual system (HVS) and then parameterizing a personal device with a display and camera peripheral. The methods leverage the difference between how our eyes perceive these synthetic visual environments and how this personal device composed of at least a camera and light emitter sees the same controllable light. In other words, we will explain how to make many of our communicating devices which currently leverage the radio frequency medium less blind to our immediate visual context.

In order to realize and evaluate these methods, we develop a framework which can test these new methods in software. The proposed techniques for a second synthetic layer of unobtrusive visual data takes advantage of the three dimensions that a shutter-based camera which is fundamentally a sampling-based machine can process data: time, space and frequency. This is in contrast to how our eyes perceive the visual environment which is complex and multi dimensional.

As a result,this platform is used to present a method for designing systems which can embed a lot of information in our visual environment through the use of electronic devices already available to us today. It makes an argument for the potential of pixel-based systems as a complement to RF. This is possible by taking into account how our human visual system experiences the visual environment and how a parametrized model for a personal transceiver peripheral views the same artificial light.

1.2 A Second Data Layer in the Visual Space

Recently, a framework for designing radio system in the communications field has emerged which assumes a generic universal software radio peripheral (USRP) which serves as a prototyping frontend to test out transmission schemes and network protocols (Bose et al., 1999). Prior to the software-defined approach, testing a concept required custom hardware. The clear abstraction and division between the sampling frontend and backend resulted

[1] The homunculus is a diagram mapping of the sensory input of our brain. Visual processing actually takes up the most amount of processing in comparison to the other senses like audio, smell and touch.

in a platform which made it possible to test new ideas using an ordinary personal computer.

Observing the advantages and downsides of adopting software radios for the RF spectrum serves as inspiration for predicting the benefits of using a software radio approach for the visible light spectrum. In particular, visibly transmitting devices in the environment already widely exist for human information consumption.

Above all, the availability of general purpose analog-to-digital (A/D) devices and digital-to-analog (D/A) devices in the RF range have allowed for quick reconfigurability to test ideas and concepts in the research environment. When an algorithm has been tested and trusted in a real environment, it can then be used to optimize and produce that method in hardware. For developing new ideas, however, it is useful to have a platform that is flexible and can be easily reconfigured to test out a new algorithm.

Displays and cameras are very quickly being upgraded and getting packaged with other devices which may do similar things. However, they can still be parametrized into a block diagram of a sampling device, a processing device and then a computing device. A software-radio approach can result in algorithms that are optimal and can actually be realized in currently available hardware.

In addition to all the benefits that software radio architectures provide for RF communication design, there is an additional trend that display and cameras enjoy. This is the existing software trend for consumer displays and cameras. Whereas advanced algorithms for communications are currently optimized in hardware on our consumer devices, advanced algorithms for display and vision processing are still being optimized in software.

Similar to the software-defined interface, a well-tested protocol is eventually implemented in hardware. MJPEG encoding/decoding is often hardware accelerated in devices like photoframes, but the majority of development for such schemes are done in separate software. Hence, there is still the wide availability of software programming libraries.

Finally, the most exciting thing about a software-based approach is dissemination. Given the current trends of establishing application markets, a tested display and receive method can almost immediately be disseminated into the hands of consumers in the physical world using intermediate devices like projectors and camera phones. If the mission is to create a more beautiful environment quickly with the wide variety of display hardware, then it is very important that there is a fast method to disseminate solutions quickly into the hands of participants in a visual environment.

Figure 1.4 depicts the envisioned environments enabled by truly unob-

Figure 1.4: Pixelated and controllable surfaces include storefronts, dishware, furniture and security-conscious displays

trusive displays which convey machine information. These interactions have been demonstrated in various forms without a human-compatible digital data-bearing interface. We claim that for these interactions to scale and grow in the environment around us, the underlying technology can be unified with a common testing platform. It is natural for us to look to the camera to help us do this since it most closely maps to how we see the world.

The interaction on the far left currently involves the use of widely-available 2D barcodes. The interaction in the center involves the use of projectors and cameras in the environment. The interaction on the far right requires the use of specialized glasses. We claim that a common software testing platform may be used to turn existing hardware such as mobile cameras, tablets and screens into support for these already demonstrated interactions.

To ease exposition throughout this thesis, we reiterate a few definitions:

- A **pixel** is a single point in a rendered scene and it is the smallest unit that can be represented and controlled both as an input device and also as an output device. Further, each pixel has it own address. Without this address, it would not be able to be controlled. In this work, pixel refers to something that is active so that the temporal behavior may be modulated.

- A **sensor** is a collection of receiving pixels which do not remember the bits passing through. They are arranged on a 2D surface and are composed of independently addressed pixels. In most cases, information captured on the sensor does not correspond to what is actually seen.

- A **display** is a collection of transmitting pixels which do not remember

the bits passing through. They are arranged on a 2D surface and are composed of independently addressed pixels.

- The **camera** in this work refers to the hardware optics and dark chamber which records visible light on an array of pixels. It is passive in that it allows light to pass through freely with no processing.

- The **projector** in this work refers to the hardware optics and additional back illumination which amplifies what is conveyed on the display. It is passive in that it allows light to pass from the sensor with no processing.

All of the form factors shown in Figure 1.5 and Figure 1.6 consist of a *sensor* and display. In some cases, it also consists of additional optics thus turning them into procam (Raskar et al., 2004) systems. We later use these form factors to describe a general parametrized description and even our innovations in form factor (see Audio Headlight (Woo et al., 2010)) and hardware optical setup (see Bokode (Mohan et al., 2009)). These form factors hint towards how a method of embedding visual information might enable very new forms of interaction.

A Personal Transceiver Peripheral - We define a personal transceiver peripheral which is capable of both emitting and capturing optical light. We consider this personal transceiver to be carried by the user. One form of this widely disseminated device is the projector-camera. Raskar et al first described one such form factor in RFIG (Raskar et al., 2004).

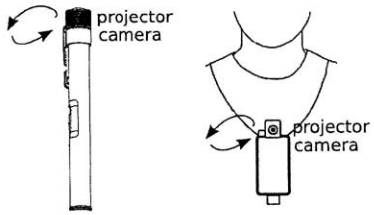

Figure 1.5: Personal peripherals can take on many different personal form factors including some that already exist today.

Controllable Luminous Surfaces - Together, these devices represent the components of a system but they do not describe the processing behind them. First, the way that a human sees the pixels and display is different from how the camera sees the pixels. In order to design an I/O system, we

need observational models for both the camera and the human and how they see the already ubiquitous surfaces such as those in Figure 1.6

One such device deployable in the environment, called the I/O bulb, was first conceptualized in context of the Luminous Room by Underkoffler (Underkoffler et al., 1999). We show in Figure 1.5 that these hardware peripherals today can take on many different form factors including screens, televisions, projectors and everday lighting.

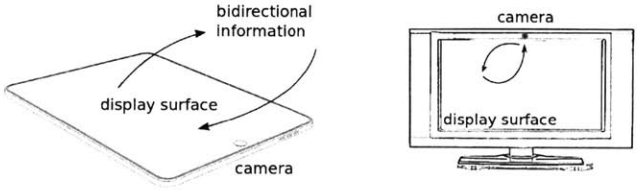

Figure 1.6: Public surfaces can take on many different public form factors including some that already exist today.

Contributions The major contributions of this thesis are:

- A demonstration that it is possible to implement a high data-rate, computational intensive real-time signal processing application for human interactions on a general purpose platform in the visible light medium

- The design of a system architecture for pixelated surfaces that can be subsequently deployed in a visual medium. The flexibility of this system including the evolution of hardware form factors bypasses existing regulation of proximal mediums such as RF.

- A specific technique, called *VRCodes* which take advantage of how the camera sees something different from the human. This technique is an example of how this platform allows a new paradigm of embedding data in the visual environment for human interaction.

- The characterization of the computational availability of signal processing algorithms on a general purpose processor abstracted from the screens and cameras available today and how the human perceives those same implementations.

- Design guidelines for possible changes in optical setups to fulfill the purpose of a visual environment which conveys a lot of information to our cameras but not intrusive to our eyes.
- Demonstrations of interactions which are enabled using this software development environment.
- A layered model for the specification inspired by those for single-purpose communications systems. This allows new techniques to embed unobtrusive information by also taking into account the physical properties of devices.

Unlike software radio applications for radio frequencies where the perceived bottleneck is the data I/O rate, the bottleneck for enabling proximal interactions in the visible light medium traces back to the power consumption of a personal peripheral. This thesis does not address the availability or potential solutions for power in context of a development environment. Here, the focus is on specific software solutions based on the parametrization of projectors and cameras.

In Chapter 2, we present a generalized development which outlines an approach that allows us to understand the basic setup and how our eyes operate differently from our cameras. In Chapter 3, we introduce a specific technique called *VRCodes* which take advantage of how cameras see differently from our eyes. Following, in Chapter 4, we present alternative optical hardware setups which can still benefit from the development of VRCode software and explain how it fits into this overall approach. Chapter 5 shows upcoming interactions based on this work. Chapter 6 concludes with a systems level argument for light-based proximal interactions.

Chapter 2

A Capacity Gap in the Information Theoretical Sense

This chapter gives some background for how much information content can be embedded in a signal meant for humans. While it is impossible to give a concise argument for the capacity of our human visual system, it is possible to describe the factors that can give us engineering estimates. A large part of this chapter will be devoted towards describing how aspects of the human eye operate as it is relevant to this work. At the end, we will have an understanding which will make it possible to evaluate the techniques designed within this framework. A setup and a binary scheme where temporally alternating colors are used to design pictures that our eyes do not find intrusive but our cameras can interpret information is the goal of this understanding here.

Basic Setup The goal of this thesis is to give our devices some context into the world that we are experiencing ourselves. It becomes very natural for us to look towards cameras since for most people. Sight is already the dominant way that the world is perceived.

The basic setup used is that shown in Figure 2.1. Essentially, we have a display and a camera looking at the same active screen. Even though we have built our cameras to simulate what our human eyes see, the camera still sees something very different from our eyes. The goal is to be able to build a system that leverages the small differences to embed data in active pictures so that our eyes don't notice them but our cameras can pick up the information embedded underneath. Ideally, we can do this without introducing incredibly

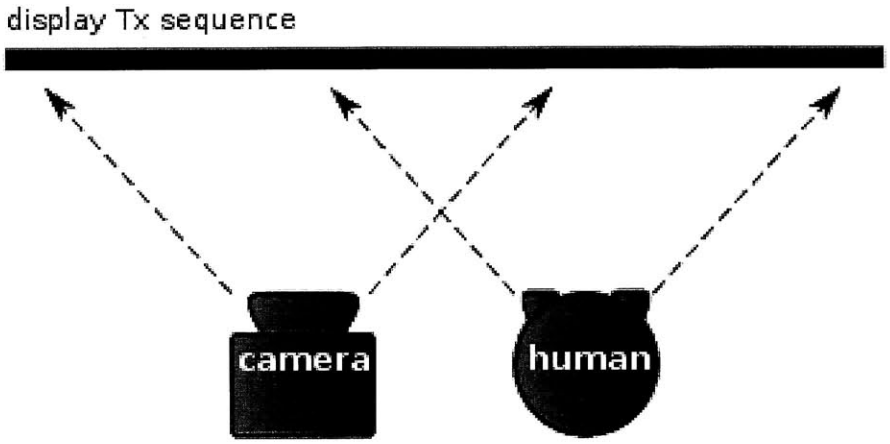

Figure 2.1: Basic setup which consists of an active screen and both a camera and a human looking at it. They see differently!

new hardware designs that require adoption.

2.1 Description of the Human Visual System

The human visual system (HVS) is complex in how it perceives a rich physical world. Unlike cameras, which can be defined as a composition of sensors with no memory and a computing backend, the HVS is composed of short-term memories and long-term memories. When considering how to convey information visually to the human brain, it still makes little sense to use an informational theoretic model.

Minsky describes seeing "red" as a very complex task for a human (Minsky, 1986). In Minsky's words, making a machine see red is an easy task: "start with sensors that respond to different hues of light, and connect the ones most sensitive to red to a central red agent". However, for a human, seeing or conveying the information red could be conjured up using words like "lips", "cherries" or the word "red".

Consideration for the psychological aspects of human vision perception is incomplete. -And, this may be difficult to unfold since information can be stored to various levels of detail in our brain that do not necessarily resemble the sampling machines that we are familiar with. In our own characterization of the HVS, we make a distinction between the sensory input of the eye and the processing of the brain. The basic image we keep in our mind is that

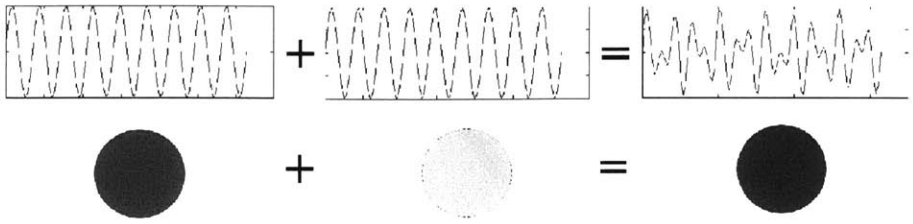

Figure 2.2: Top row shows how two signals with different wavelengths add up in signal space. The bottom row shows how two colors add up according to color perception. The two rows do not seem equivalent since the pure red is not represented by the interference signal. This hints at an alternative model for how our human eyes perceive. One which we may not fully understand yet as a scientific community.

of a camera and a human in front of an active screen. By understanding aspects of how the human sees visible light, we will be able to characterize how we use the differences between the camera and the human to embed information.

Since there is not yet an analytical model of our human eye, we scope our explanation in this work to point out the aspects that are relevant to the techniques presented here. We will address just the following phenomenon:

- Young's tristimulus theory of color perception
- Flicker Fusion Frequency of the eye i.e. how we perceive temporal blending of a flickering light
- Effective aperture of our eyes

Young's tristimulus theory The basic picture which leads the engineer to question how the human eye perceives color differently from a camera is shown in Figure 2.2.

The top row shows how two signals with different wavelengths and how they add up in signal space. When we add two signals $\sin(f_1 x) + \sin(f_2 x)$, we expect a resulting superimposed waveform as shown in Figure 2.2. For a single sampling system such as a camera, the result of these two superimposed signals does not resemble the pure sinusoid that we would expect to be associated with a pure red.

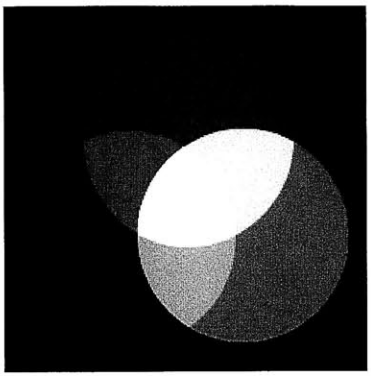

Figure 2.3: Tristimulus theory: Familiar color mixing rules we use to create mixtures of colors

The bottom row shows how two colors with different wavelengths add up and appear to our eyes. Thomas Young first discovered that the *sensation* of "red" can be created by additively mixing "magenta" and "yellow". Thus, without showing a pure red wavelength to our eyes, the *sensation* of red can be created using two combinations (Gregory, 1979) of colors such as magenta and yellow.

Further, Young showed that exactly three colors: red, green and blue can be used to create any perceivable color other than black and brown. This is known as the Tristimulus theory and roughly explains how our eyes perceive color. Today, we know that our eyes are sensitive to more than just the three red, green and blue parameters. Young was not far off in observing that the most sensitive receptors in our eyes actually are red green and blue.

When used together with the properties of the flicker fusion frequency threshold (discussed next), we can see why the same color shown to our eyes can actually be created using combinations of colors called *metamers*. Thus, Young was able to show through experiments using multiple projected colored light that colors could be mixed according to the familiar rules we see in Figure 2.3.

However, the story doesn't actually end here. Young's model only partially explains how our color perception operates. In practice, many researchers eventually showed that other sets of three colors could also be used to create the sensation of perceivable colors. Young's experiments also raised questions about whether the three colors he chose had to be spaced evenly apart as he had chose them or could there also be other picks. Historically, Maxwell revisited Young's work to try and come up with a unified theory for how our

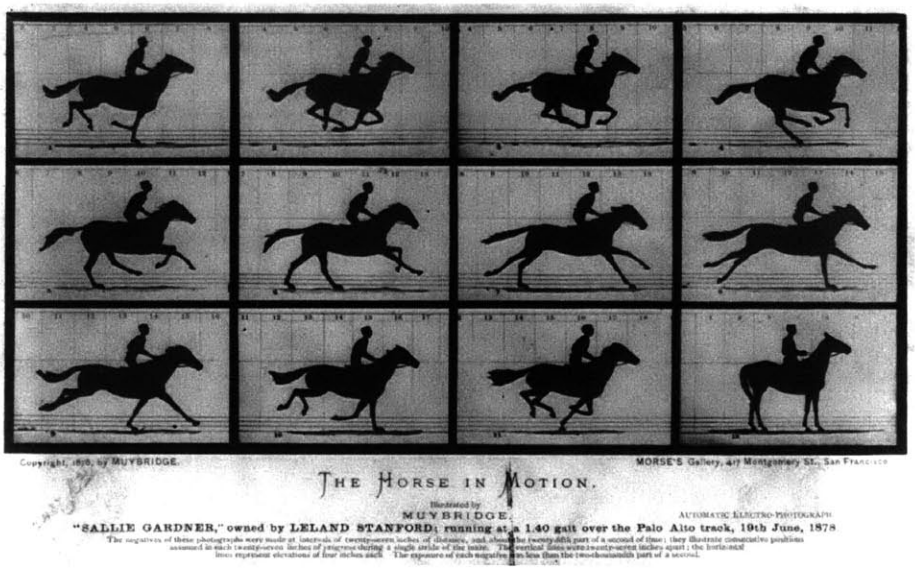

Figure 2.4: Muybridge's famous gallop sequence capturing a horse's movement.

color perception works.

However, as we know today, there are more than just a few parameters that will determine how we perceive the mixing of colors. We start by describing another dimension of how our eyes perceive before explaining how this particular method for perception of color operates.

Critical Flicker Fusion Frequency To understand how we choose to blend colors using VRCodes, we introduce a temporal dimension to how our eyes perceive. The minimum frame rate necessary for a picture to appear as smooth motion is a phenomenon that is very familiar to us. It is the basis for how we create television, motion pictures and computer graphics. Figure 2.4 shows Muybridge's famous shot capturing a horse running [1]. This animation sequence, along with many years of research (Gregory, 1979), gives rise to a phenomena which we are very familiar with: the minimum frames per second required for humans so that we may experience smooth motion.

This perception of motion phenomena is known as the ϕ phenomena. It is distinct from that of the critical flicker fusion threshold in that the

[1] This sequence of images was originally taken as a part of a bet to see whether all four legs of the horse actually left the ground when galloping

threshold for triggering ϕ phenomena is significantly lower. Imagine viewing a blinking light flashing at around 24fps. As a viewer, we would certainly notice the obtrusiveness of this blinking light at this frequency. In comparison to today's displays which are able to show consecutive scenes far above 15 frames per second, the ϕ phenomena is fairly unimpressive.

Instead, we can understand the rate at which a blinking light appears entirely solid. This experiment is conducted by using a flickering light appearing on and off. When the light flickers on and off beyond a particular speed, then the viewer will see the wavelengths blend into one another thus appearing as a solid color.

The critical flicker fusion threshold or the frequency at which light stimulus appears to be completely steady to the observer and is more complex than understanding the minimum framerate required for smooth motion. The actual critical threshold values for the critical flicker fusion rate relies on at least six parameters:

- windowing of modulation. For example, the duty cycle of the emitted sequence may effect the perception of these colors
- amplitude or depth of the modulation
- average (or maximum) illumination intensity
- wavelength or color frequency of the illumination: a parameter knows as luminous flux
- position on the retina where the stimulation occurs i.e. the region of interest. In particular, we have evolved to view flickering of light on the peripheral of our vision with more sensitivity than in the center
- degree of light and dark adaptation prior to the flicker fusion experiments

Most recently, Bimler (Bimler, 2012), shows a fairly extensive study which shows how the eye perceives based on a series human subject studies and a flickering CRT screen. His findings show significant perception patterns based on the color patterns. The conclusion from these studies is that 120Hz is a rather high flicker fusion rate beyond which individual colors cannot be distinguished by the eye.

Assuming that we have a setup where both the viewer and the camera are in front of an active display screen, we can describe the resulting mixing of color using the vectorscope images as shown in Figure 2.5. Here, each

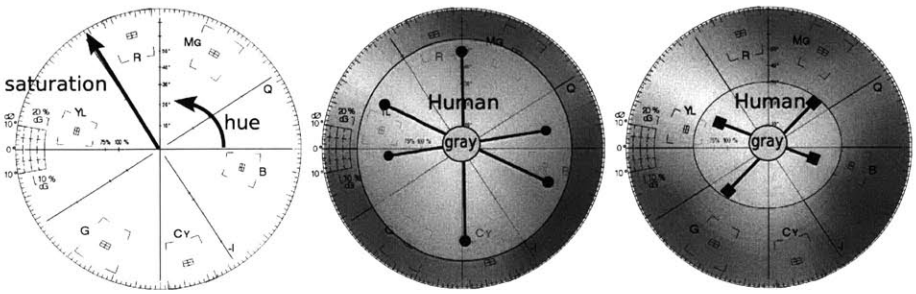

Figure 2.5: Simplified graph assuming only a binary scheme used to explain how multiple colors can be used to mix together

color is decomposed into a hue, saturation and brightness. We assume for now that the brightness is held constant so we will not plot this in Figure 2.5.

On the left in Figure 2.5, a color is mapped using an angle to represent hue and a radius to represent saturation. For the center figure, it now becomes clear how all colors of the screen could be plotted on this vectorscope plot using this representation. For a target color, the middle figure in Figure 2.5 now shows how multiple pairs that can mix together create this gray.

The middle figure shows that if the designer wishes to display a target color gray, there is at least one way to mix that color. We assume there is a camera with frame rate high enough to distinguish between two alternating colors. With this assumption, we can see from the vectorscope plot that there is a way to show a color such that the human only sees the blending target color gray. At the same time, a camera which is sufficiently high frame rate would be able to detect the flickering sequences! This will be the basic concept behind how we develop VRCodes.

The region outlined in the middle of Figure 2.5 shows that we can impose constraints which includes pairs (through experimental measurement) for what the human observer perceives based on the speed of the emitting device. If the emitting devices move slower, then there exists a region for these pairs of colors that can produce a color without the eye noticing. On the far right, as the emitting devices flash faster, the region where these pairs can be chosen for the color gray increases.

When we change the target color to a more saturated color, for example red, then we draw a smaller region for the space where we can pick candidate pairs and still not have our eyes be able to distinguish the mixed colors. This leads us to the conclusion that more saturated colors are more difficult to create using this method.

By combining our understanding of linear color mixing and the critical flicker fusion frequency, we now have a better understanding of how to pick pairs on the vectorscope graph to create the desired target color.

CIE Color Chart Up to this point, it may seem clear how we can go design the perfect encoding scheme based on what the designer wants. In practice, however, color mixing doesn't necessary occur according to the vectorscope representation. The CIE color chart is a color mixing chart that was generated by the International Commission on Illumination (CIE) in 1931. Although it is a mathematically defined color space, the data points themselves are gathered experimentally by researchers David Wright and John Guild.

We note that the CIE color chart doesn't actually represent the perfectly circular representation depicted by the vectorscope.

The premise for defining this color chart is to form an experimental table for looking up how colors mix in the human visual system. The notion that different colors can mix in a particular manner is known as metamerism (as we discussed in Young's experiment). Metamerism is the term used to describe the effect of mixing colors. The same color can be produced using two different combinations of basic colors. Yet, at the same time, these colors are sufficient to uniquely describe each color combination.

There are several reasons why the color chart is difficult to produce. Among the obvious reasons are that color perception varies greatly from individual to individual. Ambient light conditions may, themselves, blend colors in ways that we can't predict. When a single color is shown on a surface, it might get perceived very differently when placed in an alternative environment.

Much like pixel burn-in on a screen, the sensors in our eyes can get overstimulated resulting in less predictable color blending properties. Our eyes are complex sensors that contain rods and cones. When the rods are exposed to light, the rhodospin molecules can become "bleached". This is referred to as our visual short term memory.

Also, as Edwin Land of Polaroid discovered, the spatial positioning of colors relative to a background also alters our perception of them (Gregory, 1979). We understand the CIE graph itself to only be an experimental result representing how we see.

Finally, we point out here that in addition to the imperfections of our eyes, our transmitting screen devices also cannot emit all of the colors in the CIE graph. In fact, the available colors that can be represented are bounded by the independent and pure emitters that bound a specific region as shown

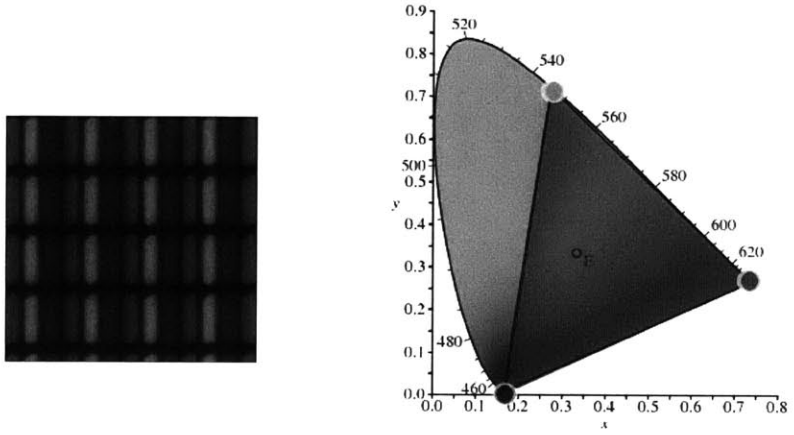

Figure 2.6: The range of producible colors is determined by the region defined by available wavelength emitters.

in Figure 2.6.

On the left in Figure 2.6, we show a microscopic shot of a pixel and we see the subpixels underneath which are capable of emitting independent wavelengths. The corresponding region of representable hues is shown on the right. This indicates that the range of available pairs for creating a desired target color is further limited using this model where only a triangular range of colors may be represented. We note this characterization results from the assumption of available emitting devices and how they operate.

An Information Capacity Gap For each encoding scheme that is proposed using this method, we can use a simple sphere-packing argument for explaining the associated probabilities of errors. We adopt the terminology here from our characterization of radio frequency transmission systems. The basic model here is that of a transmitter-to-receiver system that has both an encoding and decoding scheme associated with it as shown in Figure 2.7.

From here, we can use the encoding scheme described by the bars shown in Figure 2.7 and the decision boundary outlined in red. In particular, we can show that the overall bit error rate can be quantified using a simple Gaussian error model.

For example, in the center, we are using a binary scheme where one of two colors are encoded i.e. either a solid violet or a mixed violet. Here, the

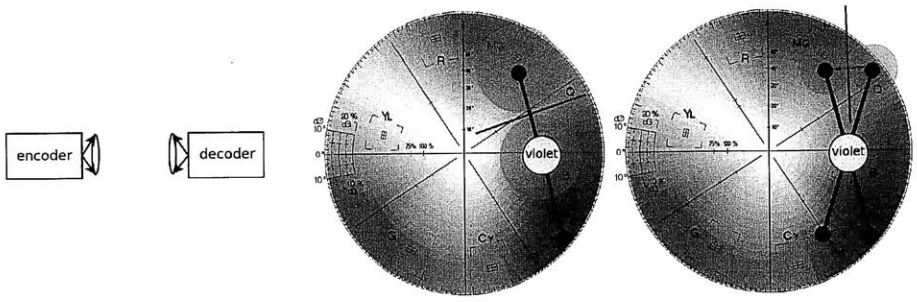

Figure 2.7: Two possible encoding schemes and the corresponding decision boundary for achieving a target color violet

decision boundary lies between the solid color and one of the hues. As a result, the probability of error may be computed assuming a Gaussian noise model where we assume that the receiver makes an observation according to:

$$y_i = hx + n \tag{2.1}$$

Here, we assume that n is Gaussian noise, h is an estimable and invertible channel estimation, and x is a value chosen according to the encoding schemes depicted in Figure 2.7. With this, we can draw a circular region in each of the schemes presented in Figure 2.7 to show the maximum allowable area that an observed symbol is allowed to deviate before the receiver makes an incorrect estimate for what the original symbol was.

The decision boundary is drawn to give the decoder a rule for estimating what the original symbol x was based on the observation y. In this sense, if the landed signal falls on the wrong side of the decision boundary, then the decoder will make an incorrect estimate for what the original symbol was. Here, the decision boundary determines the probability of error that will be used to characterize the performance of this particular system. At this point, we can do an analysis of the data that the camera can interpret if we can come up with an appropriate encoding and decoding scheme.

Chapter 3

VRCodes for Cameras with Shutters

This chapter presents how the presented material can be made useful today. For example, 2D barcodes today are often printed very small in the corner in the subway. Once the viewer takes out his mobile phone device to scan the code, a URL is decoded. However, most of the time, there is no Internet connectivity in the subway to allow the user to go to that website! What we would like to do is bring the concepts presented in the previous chapter into today's world without having to introduce new devices.

Today, many of the limitations that are in today's proximal and vision-based communications systems stem from a lack of flexibility since they rely on specific hardware features. For example, visual barcodes require standardization across all devices. At the same time, they are unsightly and work poorly from afar. The need for a flexible and camera-based framework is best argued through a solution which satisfies desired aspects of a vision-based code.

We present a novel hue-based barcode design called VRCode (Figure 3.1), which may be read by a rolling-shutter camera with consumer-grade capture speeds but remains unobtrusive to the human eye. VRCodes display active video patterns which can only be revealed by a rolling-shutter architecture with an effectively high capture speed. It is a method of embedding data in a commodity color-display that takes into account how our human-visual-system operates differently from a camera. We temporally switch complementary hues at 60Hz (beyond the typical fusion-frequency of the human eye) on a commodity active screen. What the human sees may just

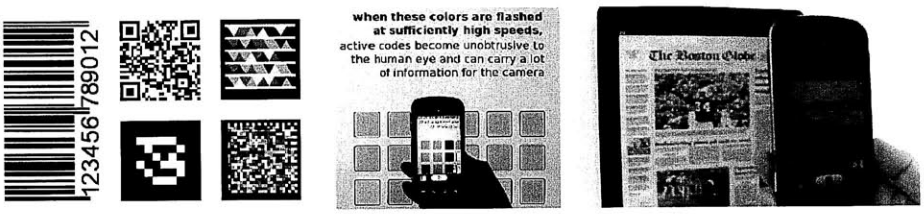

Figure 3.1: How do we use today's devices to embed more information for our camera but make it not unobtrusive to the human eye?

be a pure gray-colored screen whereas an ordinary consumer-grade mobile camera with a 30fps capture rate will see the decomposed aliasing effects of the changing hues. We use many combinations of hues distinguishable by a wide class of cameras and encode data by assigning a unique symbol to each combination. We use the VRCodes not only for embedding large amounts of data but also for embedding relative positioning and orientation.

The goal is to describe a usable and practical system that can be built with ordinary displays and off-the-shelf rolling shutter cameras found in the majority of current mobile devices can be summarized with the following technical contributions:

- Method to spatio-temporally embed digital data in an active screen that is unobtrusive to the human eye
- A technique for encoding that allows us to recover data, as well as relative distance and angle of the camera without noticeable artifacts
- Prototype design which shows the use of a regular screen used to encode digital data for a rolling-shutter camera
- Approach to encode a pattern which gives additional features to the camera while at the same appearing unobtrusive to the eye
- Working systems and performance analysis to demonstrate the concepts for several plausible interactions

Thus far, we have explained how we can use the alternating set of colors to be able to encode a solid "target" color by understanding how a particular sensation of color is achieved. However, we have assumed that the cameras available today are high speed and able to distinguish the alternating encoded sequences. We all know that the most widely deployed cameras today *are not high speed cameras.*

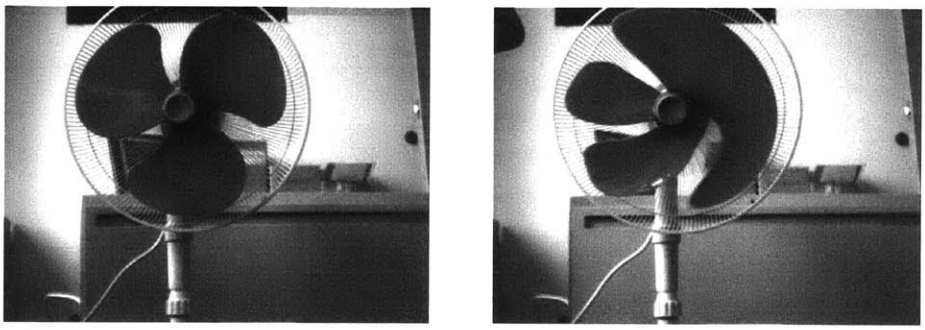

Figure 3.2: On the left, is the image of a fan taken by a higher-end camera. On the right, is the image of a fan taken by a rolling-shutter camera.

The key observation thus far is that unlike humans, cameras are sampling machines that have shutters to begin with! Unlike the human eye which can thus far only be modeled in terms of individual characteristics, the camera has a shutter which is mechanically controlled. When Muybridge took images of the galloping horse, he lined up a few cameras next to the race track and threaded a string through all of the shutters. When the horse took off, he pulled the string and all the shutters were triggered sequentially. Thus, even his setup was a line by line sampling.

We now look to how commodity cameras today work. Figure 3.2 shows two images of a moving fan that is being captured by a camera. Both of these images are taken from Point Grey, a high end camera manufacturer specializing in high-grade cameras. On the left, is a snapshot of the moving fan. Here, we see what might be expected i.e. a still shot of three blades. On the right, is what we see instead when using this rolling-shutter camera. We see instead the blades of the fan being smeared out on the right hand side. On the left, the blades seem to get chopped up as they are imaged on the sensor.

This begins to hint at why we might be able to use ordinary cameras to pick up information that our eyes don't find obtrusive. The image on the right in Figure 3.2 shows something that we would never perceive ourselves. If we look back to Muybridge's horse in the previous chapter, we can now pick up other subtle details. The horse in each frame appear different lengths!

VRCodes operate by replacing the fan in Figure 3.2 with an active screen, we can now use a setup that exploits the same effect as that shown with the fan. For example, what may just appear as an ordinary snapshot in time for us may actually embed more information for the camera.

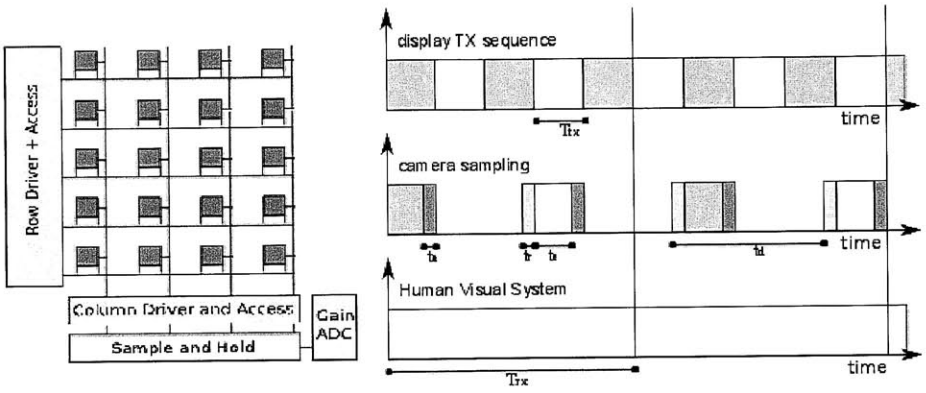

Figure 3.3: Architecture for a rolling shutter camera that simulates an analog global shutter. Typically, the effect results in artifacts known as wobble and skew.

Techniques that involve spatio-temporal data modulation on an active screen are known. But as far as we know, ours is the first system designed to be compatible with widely-deployed active screens and rolling-shutter cameras while at the same time satisfying the primary goal of producing an unobtrusive visual environment. This is completed with the prototyping methods using ordinary screens and cameras. The next section explains why this effect happens when we use the sensors we have today.

3.1 Unobtrusive Optical Communication with Rolling Shutter

CMOS image sensors are rapidly becoming commonplace in surveillance video cameras and mobile cameras. These sensors are digital light collectors that are designed to behave similar to an analog camera. In other words, they are meant to try and simulate how we humans perceive. Unlike the Human Visual System (HVS), they can be concisely described as readout circuits which simultaneously read pixels into a line-memory and get exposed to incoming light rays. We first introduce this common CMOS architecture and then give a brief explanation of how the HVS behaves differently. Then, we describe the specific setup for VRCodes.

3.1.1 Effective Rolling Shutter Speed

Most consumer-grade camera shutters implement line-scan frame acquisition as shown in Figure 3.3. This is called a rolling-shutter. Even though each individual pixel can be controlled, the design of a rolling-shutter camera grabs each line individually or in blocks of lines.

The motivation for this type of design in consumer electronics is two-fold. First, it is employed because the read speed of each line is not sufficiently fast in circuitry (especially on a mobile-phone) to achieve the effect of a global shutter. Second, it is very difficult to design a CMOS sensor to simulate a true optic shutter because this requires a storage element within each pixel that can hold the charge before it is time to be read out.

We start with a simple setup and use the example of a rolling-shutter camera pointed at a single point source which is blinking on and off with a period of T_{tx} as shown in Figure 3.3. From Figure 3.3, the rolling-shutter camera system can be described with simple abstracted parameters. The line scan reset time t_r is the time needed for the access driver to query for another line or batch of lines. Correspondingly, the line scan acquisition time t_a is the time for the access driver to dump the collected data. The actual amount of time that each sensor is exposed to light is t_e. The total acquisition time for a line in practice is:

$$t_d = t_r + t_c + t_a \tag{3.1}$$

and the reported frame rate for the camera is:

$$fps_{rx} = \frac{1}{T_{rx}} \tag{3.2}$$

Here, each frame can contain many batches of line scans each of time t_d which occur sequentially n times and also overlapping in parallel as shown in Figure 3.4.

In other words, a typical rolling shutter camera which has a reported frame of T_{rx} actually has n number of t_d length readouts for a blinking point source. Depending on the exact number of readouts n for the camera architecture, the effective time that a pixel in a line scan is exposed to light is now:

$$T_s = \frac{T_{rx}}{n} = \frac{1}{fps_{rx} * n} \tag{3.3}$$

For $n = 5$ on a 15fps off-the-shelf camera, the effective fps is now 75fps since the effective "shutter" is now on for $0.0666/5 = 0.013.8$ seconds. For

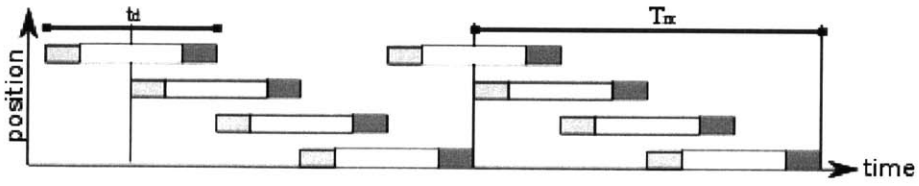

Figure 3.4: There are many reads per frame of captured data which may occur in parallel. As a result, the displayed frame contains "bands" from more than one transmitting frame.

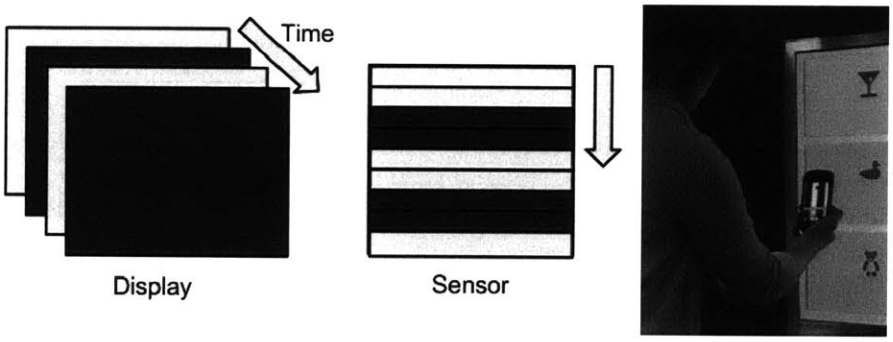

Figure 3.5: Basic setup using an active screen that uses a sequence of colors to mix the target color.

a single point source with a period of $T_{tx} = 0.02$ seconds, this means a low-end rolling-shutter 15fps camera can resolve the blinking light with post-processing that undoes the rolling effect. In practice, we see this to be a reasonable approximation for the effective shutter speed.

Now, we see the effect of replacing the fan with an active screen using the basic setup shown in Figure 3.5. Here, for a blended target color of gray, a sequence of images can be shown on the active screen which will be picked up by the rolling shutter camera but perceived. Flashing colors from different points in time are shown on a single frame imaged by the camera. Figure 3.5 shows this is not just in theory but also in practice.

Given everything we now know, the image in Figure 3.5 itself is taken using a high end camera with a long exposure time. The image of the screen captured in front simulates what we might see and the stripes captured by the camera are the bars as seen by the mobile phone camera which is an ordinary CMOS sensor. The rest of the images used to describe how VRCodes operate are also taken using this third person high-end camera.

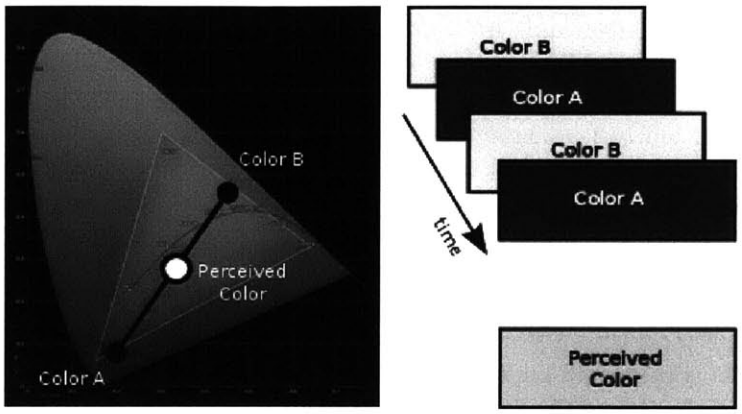

Figure 3.6: International Commission on Illumination (CIE) generates a color chart based on experimental visual data. For example, if we assign Color A as a solid blue and Color B as a solid yellow and alternate them at 120Hz far beyond the crtical fusion-flicker threshold of our HVS, the perceived color will be gray.

3.2 Metamerism

By now, it becomes clear that the HVS can not be modeled in the same way as the camera. As a heuristic, the frequency of visible light colors is extraordinarily high on the order of terahertz whereas the fastest nervous impulses in our brain are only on the order of kilohertz (Gregory, 1979). It would be naive to think of the HVS as a sampling system in the same manner as the camera.

In the early 1800s, Thomas Young first recorded and conjectured that it may be possible that humans generally have three color receptors and perhaps all colors may be created from by mixing these combinations. Young's experiment projects three distinct colored-lights spaced apart from one another to create any spectral hue. Three primary colors such as red, green and blue which can be mixed to produce any new hue have since been called "tristimulus" values and the entire school of thought is considered the "tri-chromatic" additive color model.

In 1931, the International Commission on Illumination (CIE) created the CIE color space as shown in Figure 3.6. The CIE RGB color space is one of many candidate color spaces but is of note because it is generated from user observations. It is done using a circular split screen 2 degrees in size modeled after the angular size of the human fovea. On one side, a testing color is projected. On the other side, three orthogonal colors red, green and

blue which are adjustable in brightness are exposed and users are asked to match the two sides. As a result, the CIE color space shows how each hue can be broken down based on real experimental data.

In addition to leveraging the extensive color-space research, VRCodes rely on research in the psychophysics world to consider the flicker fusion threshold. This is the temporal frequency at which an intermittent light appears steady to the human observer. This is generally understood to be around 90Hz (Gregory, 1979) and can be lower or higher depending on amplitude or depth of the modulation and the color wavelength. For example, when high-contrast pairs of colors are flickered alternatively, the flicker fusion threshold is relatively high. When low-contrast pairs of colors are flickered alternatively, the flicker fusion threshold is lower.

In the context of VR Codes, a low-refresh screen must show low-contrast colors in order for the HVS to not perceive any flickering. On a high-refresh screen, higher-contrast colors can be shown alternatively thus directly corresponding to higher data capacity. VR Codes are tested at both 60Hz and 120Hz and we must use different amplitudes for each of the frequencies in order to present an unobtrusive image to the user.

We expect that these two colors as shown on the CIE chart will average according to a linear model based on our understanding of Young's theory combined with the critical flicker fusion frequency principle. As shown in Figure 3.6, we start from the CIE color space and pick a line segment on the chart. For two colors where Color A is (c_{ar}, c_{ag}, c_{ab}) and Color B is (c_{br}, c_{bg}, c_{bb}), the perceived color when alternating Color A and Color B beyond the appropriate flicker fusion frequency may be predicted using an average:

$$(c_{pr}, c_{pg}, c_{pb}) = \left(\frac{c_{ar} + c_{br}}{2}, \frac{c_{ag} + c_{bg}}{2}, \frac{c_{ab} + c_{bb}}{2} \right) \tag{3.4}$$

For example, from the CIE chart, when Color A is set to (0, 0, 255) and Color B is set to (255, 255, 0), the resulting perceived color is (128, 128, 128). The resulting color is predictably the one in the center and also the average of Color A and Color B. We verify this prediction using our experimental setup made from commodity hardware. One limitation of the VRCodes is that the number of produced perceived colors is determined by this CIE chart and the spectrum-range of the display. Table 3.1 shows how the candidate metamers change for differing flickering frequencies.

It becomes clear from the CIE chart that colors in the center of the graph are easier to produce with different combinations than colors on the

Table 3.1: Color assignments for frequencies close to or beyond flicker fusion frequency.

Frequency	Color A	Color B	Perceived Color
120Hz	(255, 0, 0)	(0, 255, 255)	(128, 128, 128)
60Hz	(255, 217, 217)	(217, 255, 255)	(235, 235, 235)
30Hz	NA	NA	NA

periphery. The strategy here is that for any perceived color, it is desirable to pick the two colors c_a and c_b on a line segment which maximize the Euclidean distance while at the same time average together to create C_p:

$$\max_{(c_{pr},c_{pg},c_{pb})} |C_a - C_p| + |C_b - C_p| \tag{3.5}$$

In practice, a special and uniquely-generated CIE chart must be generated for each display. Together with Table 3.1, the CIE chart will differ according to color, flickering frequency, and make of television. In generating the CIE chart, users vary greatly in color perception and thus reported results are in fact averaged across the user study. Similarly, flicker fusion frequency reported by users can vary by up to 50% (Gregory, 1979). This margin can result in greatly varying results across multiple users. We leave a full study of human perception in this work as future work and only present results for a subset of data. We also point out here the range of colors perceived colors which can be created is limited by consumer-grade displays although it is now clear how to achieve the effect shown in Figure 3.3.

3.3 Encoding Position and Data

The basic setup for VRCodes is shown in Figure 3.7 where an active video screen is shown to both a camera and a human viewer. The active video screen displays VRCodes which are shown at a frame rate higher than the flicker-fusion-frequency for the available set of display colors. Correspondingly, the camera is a rolling-shutter capable of decomposing and subsequently decoding the VRCodes.

Display: With this setup in mind, a VRCode is composed from a sequence of symbols. Each symbol is assigned a pair of colors which produces a metamer which matches the original desired color. As a result, depending on the number of available pairs which can produce the desired perceived

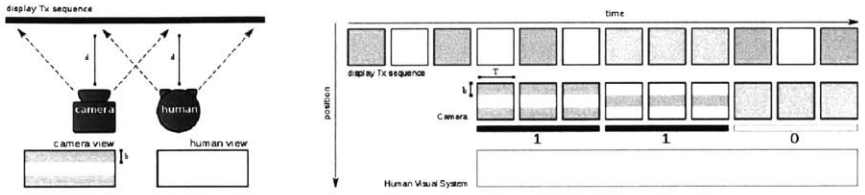

Figure 3.7: Each batched line scans are considered a single read of data and there are many reads per frame of captured data. Our human visual system (HVS) is not a sampling system in the same way a rolling-shutter system is. Each unique combination as perceived by the camera may be assigned a "logic" value for encoding information.

color, the effective data capacity increases. On most off-the-shelf displays, this means a desired gray background can hold a lot of information.

Figure 3.7 shows how a pair of colors is assigned a symbol representing logical "1" and a consecutive solid color is assigned a symbol representing logical "0", b is the spatial width of each color band as it is imaged on the camera sensor. T is the spatial width of each cell. Here, a cell is the defined unit of encoding. Depending on the number of acceptable color combinations which can be distinguished by the camera model, more data capacity can be achieved.

The encoding scheme we demonstrate in this work is very straightforward. To show data, we take an incoming sequence of raw data and use a basic Hamming Code scheme (18004:2006, 2011) for blocks of size 5. Then, we map the resulting encoded sequence to alternating and solid colors. For demonstrating relative positioning, we produce a checkerboard of alternating colors and solid gray colors to measure stability. For demonstrating encoding, we show the encoded sequence embedded in an actual design.

From here, we will show that the decoding process is very similar to that of a typical radio frequency (RF) decoder and has strong parallels to the block diagram abstractions in a RF receiver. We explain each step of the software decoding process. Each captured frame is processed in real-time as a part of a video-processing loop. For each loop that is processed, a chunk of bits are stored and passed up to the application. The basic encoding and decoding approach is that shown in Figure 3.7.

Preprocessing: Each frame is first preprocessed using color equalization. We point out here that in contrast to a decoding chain such as that found in the QR-code (18004:2006, 2011), a binary thresholding step is not sufficient

since there may be two color candidates for a single threshold. Bimber et. al (Langlotz and Bimber, 2007) called this an artifact in their unsynchronized barcodes work. As we described in our system setup, this is actually the feature we rely on from the rolling shutter. In our demonstration we use only a binary scheme so use a shortcut filter which makes a decision for each pixel with hue or no hue.

Natural Marker Assistance: Depending on the application, we use a natural marker to cut down down on processing time. Specifically, in our decoding demo, we use the black edges of the screen to define a search region for the encoded sequence area. Natural markers can also be implemented as in the case of our relative positioning demo where the entire frame is scanned for the pilot sequence.

Homography: Once the VRCodes have been identified, the homography is calculated from the expected shape of the marker. The result is then used to apply a perspective transform to the frame. Once a valid homography is found, it is applied to all subsequent real-time frames until another similar homography is found in the background. Only then is another perspective transform applied to incoming frames. This is done to maintain the real-time processing.

Sampling: After applying the perspective transform, a sampling grid is created where each value can be read out from the 2D frame. These values from the sampling grid are particularly important since they can be also be used as the confidence value and multiple samples from each cell can be used to improve confidence as described in Section 4. Each of these analog values are then assigned a symbol for the assigned threshold value c_{th}.

Decode: Finally, each sequence is decoded using a Reed Solomon decoder. The resulting decoded sequence is passed up to the application for specific use. In the case of positioning, there is no decoding step and the sampled points are directly used for calculating the relative orientation vectors.

The system described is implemented using a Samsung Galaxy S Android phone. The VRCodes are shown and tested on both a Samsung LN46 1920x1080 120Hz 3D television as well as 60Hz Apple I-Pads. The results shown here are taken with a digital SLR camera with an adjustable global shutter speed and are captured from third person view to show off as best possible the difference between how the human sees and the rolling-shutter camera sees.

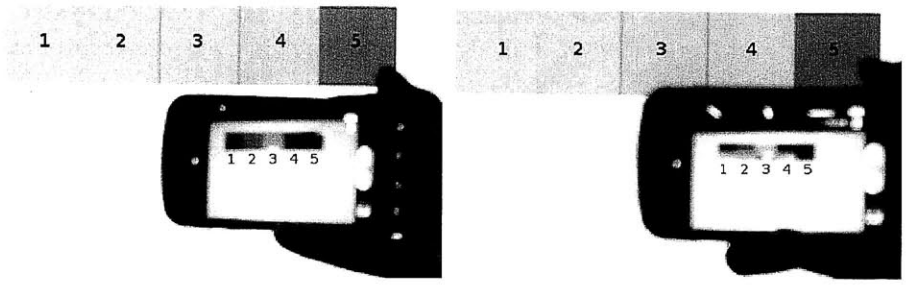

Figure 3.8: For a color towards the center of the CIE color graph (i.e. gray), there are many combinations of colors which can be used to produce the gray.

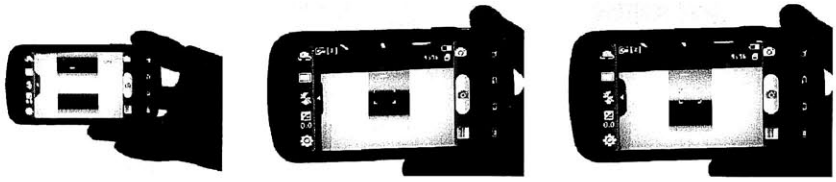

Figure 3.9: The absolute number of pixels in each band decreases as one moves further out.

3.4 Results

Figure 3.8 shows the result of using 4 pairs of colors that show the same perceived gray by the user. On a 120Hz display, higher contrasts are allowed and thus, an encoding scheme with more color pairs may be achieved. In contrast, on a 60Hz display, lower contrast levels must be shown. As a result, fewer combinations of colors are allowed and the decoding sequence must be limited to only a few symbols as seen in Figure 3.14 Also, lighter contrast effectively results in a low signal level which makes decoding more challenging.

Here, we see that the color gray has many combinations i.e. black-white, red-blue, green-magenta, blue-yellow which can create that same mixed gray (although the last one is a pure gray which must be blended to look similar as the other colors). As we described conceptually in the previous chapter, the probability of error regions can be defined for each proposed target color and corresponding encoding scheme. We will see in the next chapter that the overall probability of error also depends on the parameters of the new hardware.

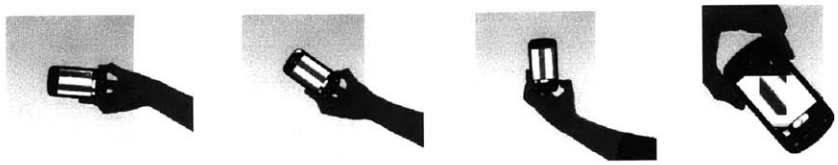

Figure 3.10: Three images show encoding that is invariant to orientation. Instead, by creating a tiling pattern consisting of both alternating images as well as solid images creates a trackable marker that conveys orientation.

Figure 3.11: Encoding pattern used for creating additional features for tracking relative positioning

Figure 3.9 shows how the bands change for $d = 1\text{m}$, $d = 2\text{m}$ and $d = 3\text{m}$ away from the screen. Although the bands appear to be growing more narrow, the absolute number of pixels in the width of each band remains the same. This is due to the sampling size as d increases between the user and the camera. In comparison to existing properties of 2D barcodes, the use of geometry to figure out relative positioning and distance is no different. The only additional benefit here is that the codes themselves appear unobtrusive to the eye.

Figure 3.10 shows how a position marker can be created by tiling several cells that are different adjacent to one another. One cell on its own is not sufficient since the rolling-shutter property of the camera is built into the phone. From left to right, the images show that regardless how the phone is held up against the screen, the bands are always aligned horizontally with the phone.

Here, we use a checkerboard pattern as shown in Figure 3.11 When placed adjacent to a reversed sequence color pattern, the marker can then be used to deduce relative orientation even though the transmitted surface still appears as a single texture to the viewer.

In order to extract the relative orientation, we use the Hough line detection

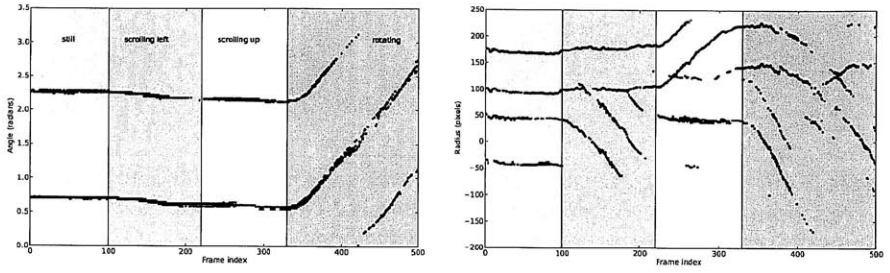

Figure 3.12: Plot of Hough Line Transform stability using VRCodes. Together with the appropriate parameters, sensitive and responsive position tracking may be achieved.

algorithm to find relative edges in the pattern. Each one of the found lines segments are returned as a vector and a relative angle. This collection of vectors and relative angles are then used to get an overall position relative to the active screen by taking the average. There is a small ambiguity in the case of the mobile phone where we cannot distinguish between the phone being upright or upside down. Here, we use the accelerometer to tell the difference.

Figure 3.12 reports a result from an experiment which shows stability of a position tracking data using a modified Hough line transform. The background appears an ordinary green screen. However with the correct rolling-shutter camera, the stunning color underneath may be revealed and in the simplest case show a checkerboard. We show a frame number on the x-axis and show stable tracking of movement going from "still" to "rotate". We see here that there are These results suggest the feasibility of VRCodes as a visual marker. Here we see the position going from still to scrolling upward.

The relative positioning for VRcodes can be used for a variety of applications as we will show in the last chapter of this thesis. At the very least, the personal mobile peripheral can be used as an interface which allows us to control the public environment using subtle motions that are detectable and interpreted by the camera. In many cases, the camera can be used as an alternative to the accelerometer which may give a much coarser estimate.

In addition to the digital positioning readout of the VRCode, we encode data using the same block diagrams. Figure 3.14 shows the use of VRCodes in a low-contrast I-Pad setup where the backgrounds of newspaper images are used to embed data by using the target color gray. By switching subtle colors

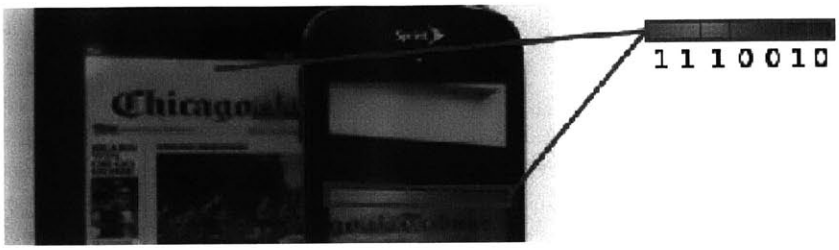

Figure 3.13: Result of capture and decode as shown in the described system. On a 60Hz screen, the color tones must be of particularly low-contrast in order to be unobtrusive to the human eye thus making the decoding more challenging.

of light green and light magenta, the camera can distinguish the underlying bit sequence, decode and map each code to a URL.

3.4.1 Decoding Block Diagram

Improving the decoding of VRCodes largely depends on having a clear understanding of the freespace optical system block diagram. The results here are based on the design, prototype and experimentation with commodity hardware (Woo et al., 2011) where the transmitter is a simple LCD screen and the camera is a consumer 35 mm hardware. There are currently many visible light communication (VLC) broadcast configurations exploring the potentials of a data link established using custom freespace optical hardware equipment. The design presented via VRCode's data transmission differs only in it's use of a consumer hardware setup.

Komine et. al presented prototypes and a fundamental framework for transmission of white LED light in (Komine and Nakagawa, 2002). More recently, Little et. al (Little et al., 2008) build an indoor wireless lighting system also employing OOK time modulation. This class of works represent an effort to demonstrate visible light data transfer systems using classic temporal techniques. QR Codes, Data Matrix Codes, Shot Codes and EZ Codes are examples of popular 2D barcodes which encode information to be transmitted visually. These modes of 2D visual information transmission already conform to ISO standards (18004:2006, 2011) to utilize 2D encoding and decoding of visual information.

We build on a simple understanding of a channel model and propose a communication system that uses a camera and a LCD display to communicate using visible light.

Each reported sample is a (R,G,B) value represented as $c_i = (r_i, g_i, b_i)$.

Camera sensor hardware generally use R, G, B pixel arrays with Bayer tiling (Grey, 2012). As a result, each color value is recovered independently. Due to noise sources such as camera thermal noise, lack of access to raw sensor data and errors in spatial-sampling, a single reported value captured on the camera sensor may not be representative of the actual source. We model the noise level of the camera as Gaussian with σ^2 as the variance to present a decision mechanism for estimating the transmitted color.

We point out here the effect of camera resolution. We define h to be the vertical resolution and w to be the horizontal resolution. The rolling shutter CMOS camera has an orientation and thus we define the horizontal dimension to be aligned with the scanlines. Based on our explanation for the effective shutter speed of the camera, the effective resolution for each linescan readout is:

$$\text{sampling pixels} = h/n * w \tag{3.6}$$

Multiple observations of the same point source due to each line scan batch from the camera may be used to estimate the original value. For a camera with parameters of $h = 1280$, $w = 960$ and $n = 5$, each line scan may have on the order of $1280/5 * 960 \approx 2000$ number of samples of the same source. If we aim to recover the original value of the transmitted source, the effective confidence can be greatly boosted due to the multiple observations.

For example, with a binary scheme and observations $c_1 ... c_k$, we let $x^a = (x_1^a ... x_m^a)$ be the amplitude values corresponding to symbol a and $x^b = (x_1^b ... x_m^b)$ be the amplitude values corresponding to symbol b for a threshold c_{th}. In order to boost the overall confidence, we use the following strategy which maximizes the recovery probability:

$$\text{if } \sum_{1}^{k} \frac{c_i}{\sigma_i^2} \leq c_{th} \text{ then "symbol a", otherwise "symbol b"} \tag{3.7}$$

By using the multiple measurements, we significantly increase the overall confidence of the estimated (R,G,B) value. The number of measurements depends on the range of sensors on the camera. Together with the effective shutter speed of the camera, the rolling-shutter camera model is enough background to build a data transmitting system.

Figure 3.14 shows a basic setup using an ordinary consumer Cannon SLR camera setup to look at the LCD screen. The LCD displays images corresponding to the input binary data; the camera captures a photo of the display and decodes the images to recover the data. We are interested in

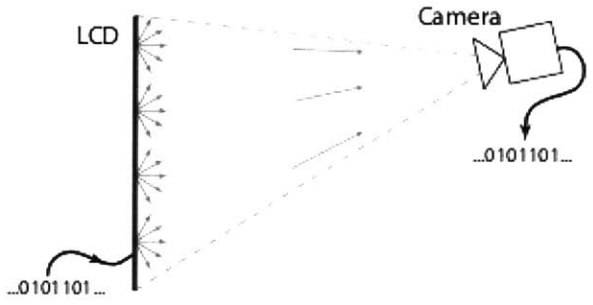

Figure 3.14: LCD and consumer camera system used to test receiver block diagram

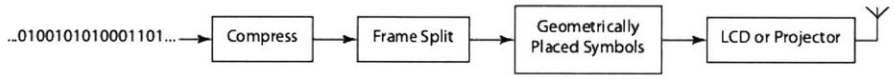

Figure 3.15: Basic transmitter chain using ordinary LCD screen showing colors

the bitrates received at each one of the cameras, with respect to angle and distance as related to users in a physical space.

Transmitter The basic transmitter block is shown in Figure 3.15. We consider a scenario where incoming bits are first compressed. These bits may be split into several frames to guarantee independence. Next, forward error correcting is implemented to protect the bits through the lossy channel. An additional block in this chain creates a feedback system to determine how these protection bits are placed in the physical space. The block diagram suggests a system with multiple receive cameras, and a single transmitting display.

Receiver The basic receive block diagram is shown in Figure 3.16. It is split into two sections where many of the preprocessing elements are borrowed from image processing. The postprocessing elements are more reminiscent of the decoding mechanisms invented for RF design blocks. Here, we explain the role of each component.

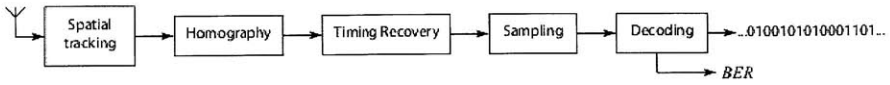

Figure 3.16: Basic receiver chain using mechanically focused camera on the LCD display

Spatial Tracking: The purpose of spatial tracking is to determine where in the scene the data is located. We search for the corners using a modified corner algorithm. As there might be several corners in the scene, each of the corner candidates using the fast corner detector is compared to a large quadrilateral generated from the second frame. This is done in the system using a function *find_corners* which takes as input the incoming frame and the sampled frames.

The quadrilateral from the second frame is obtained by repeatedly blurring and low-passing the second sampling frame and then converting to a high contrast black and white image. The square is found by discovering all white regions in the scene and calculating a "square score" given by:

$$\left(\frac{\min\left(\sqrt{\text{area}}, \frac{\text{perimeter}}{4}\right)}{\max\left(\sqrt{\text{area}}, \frac{\text{perimeter}}{4}\right)} \right)^2 \tag{3.8}$$

All distances are calculated from the corner candidates in the first frame to the centroid of the square found in the next frame. The four corners with distances closest to each other are considered the corner calibration points for this block. The brightest pixel point associated with these corner calibration points are considered the corners for this round.

Homography Once the scene is found, the perspective scene is restored using an image matrix transformation. This is done in the system using a function *frame_recover*. A transfer function is formed using the corners from *find_corners* and a homogeneous inversion may be performed from the formed matrix C. All the following frames are cropped using corners found from *find_corners* and the transform coordinates found from *frame_recover*. These are passed through the rest of the receive chain until another dark calibration frame is found.

Timing recovery is done by detecting all dots from *sampoints_frame*. *sampoints_frame* recovers the image by adaptively equalizing the entire image and then passing it through a high-contrast filter with contrast limits of 0.2. This is then converted to a high-contrast black and white image. The

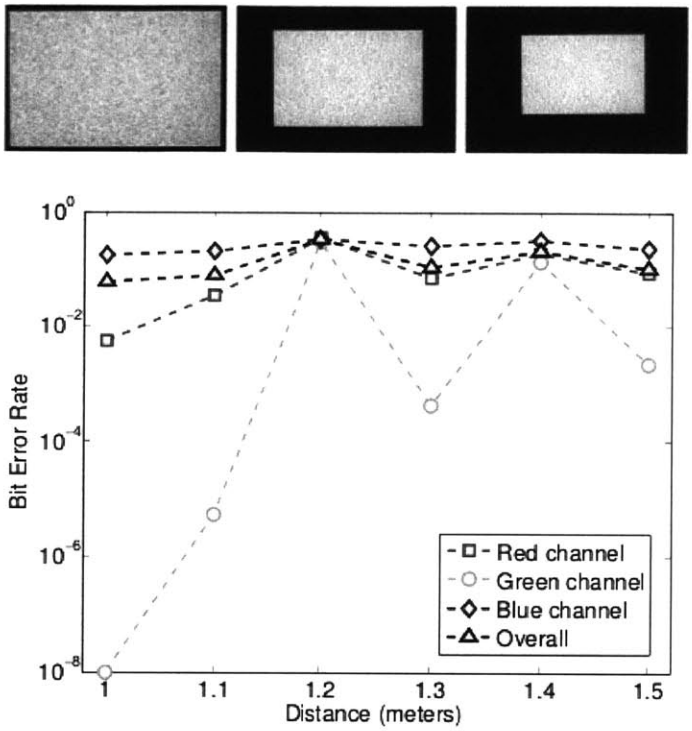

Figure 3.17: Bit error rate with respect to distance

centroids of each one of these points are found using standard logic functions.

This points are assumed to be contained on a near perfect grid. The points of this mesh are reordered as such. This mesh is interpolated by an upsampling factor of 4 to allow for localized timing offset. Here, a match filter from the grid length is used as a kernel for the image. The convolved image is used to find the optimal sampling points.

Sampling is done from the result of the timing recovery frame where all maximum values corresponding to a region within a found grid point is sampled. *Sampling* takes as input an incoming color image and slices the image into three slices and considers each slice with grayscale levels. As a result of the interference discussed, each slice is adaptively equalized followed by high contrast adjustment. The result is sampled with sampling points found during timing recovery.

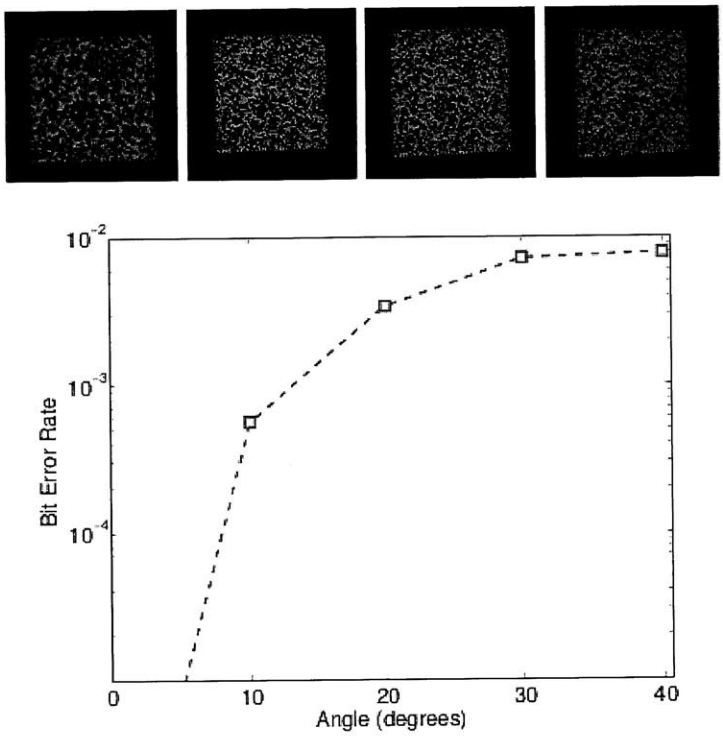

Figure 3.18: Bit error rate with respect to angle

Figure 3.17 reports BER using this system with respect to distance for encoding in all color channels for a 4.96Mbits/frame/screen in n x n array of cameras and screens where $n = 3$. Here, the total transmit size of each frame is 1599x1035 pixels. In this experiment, the array setup is such that each camera is focused on each display.

Angle evaluation is done using a square black and white checkerboard carrying a throughput of 7.86Mbits. Figure 3.18 shows BER as a function of different viewing angles and a fixed distance of about 2m. We can clearly see from these results that the BER increases to 10^{-2} as the angle increases.

Figure 3.19 shows that although the channel model for these results are analytical, the resulting bit errors still have randomness. One of the key observations is that so long as the noise model for a bit can be represented using a simple Gaussian noise model, the improvement schemes such as the

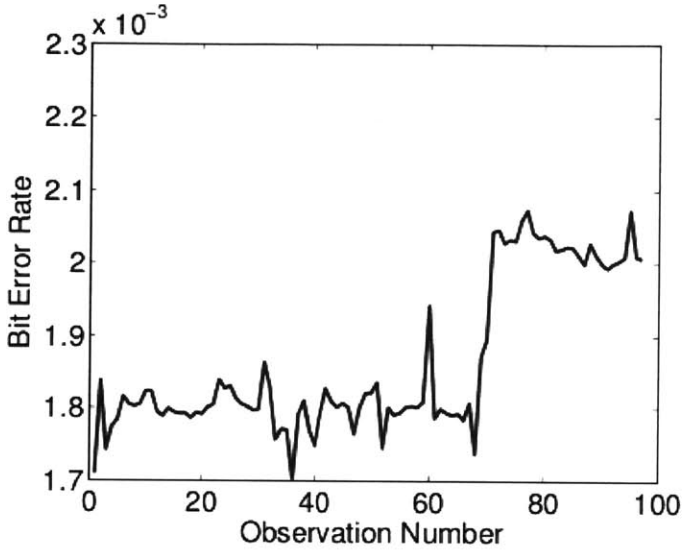

Figure 3.19: Randomness in Bit error rate

ones used for VRcodes can continuously be improved.

These experiments present a screen to camera optical system built with commodity hardware which may be used for many interactive scenarios. Depending on the application and needs of this network, there are many modifications which can be made which will ultimately result in significantly higher bit rates. These techniques can continuously be improved analogous to how the RF channel can continuously be improved.

State-of-the-art techniques embedding information in our environment which utilize all physical dimensions of space, time and wavelengths are either intrusive to the eye or require significant new hardware for deployment. The VRCode design presented in this paper encodes information in a manner that is unobtrusive and necessarily decodeable by a well-designed software decoder. VRCodes provide maximum flexibility for a designer looking to incorporate more interaction through the visible light domain. We believe that the use of unobtrusive visual markers which carry information opens up new avenues in the areas of ubiquitous computing and human interaction. Additionally, we believe the VRCode design may benefit other applications in augmented reality including camera calibration, hidden debug screens and

novel screen-to-camera interfaces.

Chapter 4

Hardware Device Trends Based on a Parameterized Model

Digital cameras, from professional SLRs to cellphone cameras, are ubiquitous and are used not only for photography but also for accessing information by recognizing and understanding scene elements. As devices change in the future, we should gain insight for techniques such as VRCodes as well as how they can be adopted for new devices. Meanwhile, by now, we would like to show that the picture in Figure 4.1 is a good one to keep in mind.

This diagram shows that by now, we can think of devices as frontends which can either emit light or receive light (or both). We think of the general interface between this frontend and backend and allow for the signal processing in the backend to serve as a direct interface with the application on the mobile phone.

Even though the devices themselves may change in the future, the idea is that the software can continue to be developed in parallel. Some of the potential devices that can be built are shown back in Chapter 1 and explained

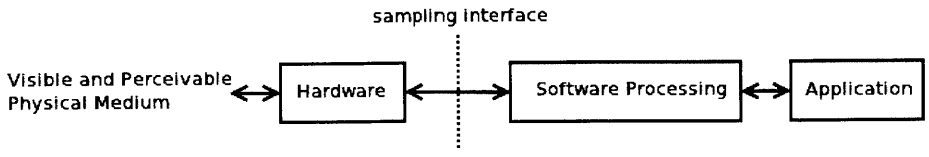

Figure 4.1: Devices can be used both to emit and sense data in visible light. Can the hardware and software continue to be developed separately?

in terms of public infrastructure and private devices. John Underkoffler showed in 1999 that the I/O bulb could be used as a device which could both transmit and receive light (Underkoffler et al., 1999). Natan Linder showed in 2011 that public devices might look like AR lamps which can turn any surface into an interactive space (Linder and Maes, 2010). Susanne Seitinger showed in her thesis (Seitinger, 2010) that nearly any public space can be turned into a light-emitting synthentic visual medium.

We present parameterization of some of the existing hardware devices and also present some new setups. The hope is that these devices can be explained in terms of the overall bitrate. We outline some of the relationships between the amount of information that can actually be transmitted from a screen to a camera to be dependent on the parameters of the screen as well as the parameters of the camera.

4.1 Screens and Cameras

First, we can revisit the encoding scheme created in Figure 4.2. Now, we see that the range of allowable encoding schemes allows for choosing appropriate pairs within the range of displayable television colors. The region outlined shows the range of metamers that can create the target color for the human eye is determined by the following parameters:

- critical fusion frequency threshold of the person viewing the screen
- desired target color
- level of sensitivity that the eye will be susceptible to movement
- modulation chosen for showing target color
- number of metamers to iterate over

Correspondingly, once the region surrounding the target color is chosen corresponding to the perception of the eye, the following hardware parameters will determine the information capacity performance of the emitting device as well as the range of available target colors:

- temporal transmission rate of the active surface
- range of color that can be emitted by the active surface
- overall pixel count of the active surface

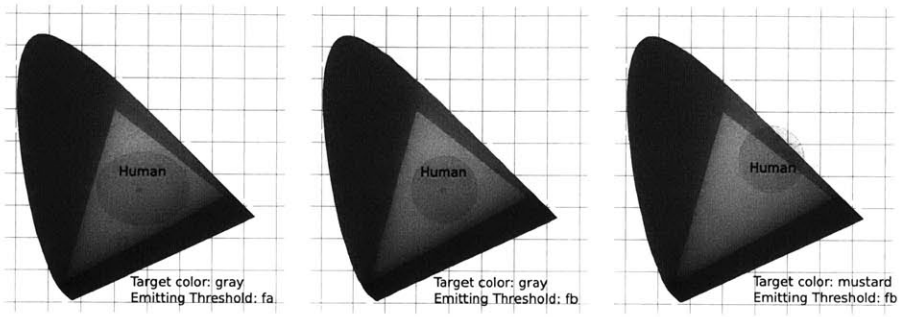

Figure 4.2: Range of emittable colors embedded on a CIE graph with human boundaries drawn in for a particular target color

A potential new screen which implements an additional yellow pixel to the existing RGB subpixels for the screen would allow for a very different design. Recently, a screen with an extra yellow (Y) pixel has been introduced to create a RGBY pixel. Such an addition would introduce a new set of techniques available for the encoder to choose from and change the region of colors that can be represented.

Consider an ideal camera with infinite pixel and timing resolution as well as a perfect communications channel. Then, we could think only about the screen parameters. In the case where the screen has a h x w resolution where each pixel has an n-bit pixel depth and a refresh rate of t_r per second, then at the very least there could be $R_{\text{ideal}} = h$ x w x n x t_r number of bits per second. Considering the number of pixels that an ordinary LCD screen has (1280x1040), this could turn out to be a considerably high bitrate!

Determining the characteristics of the camera which maximize bitrate is simpler if we think of the camera separated between the optics and the sampling itself. As we will see below, cameras can be modelled as the optics followed by the light that gets collected on the sensor board. The optics themselves may be parameterized to see the exact effect that each element has on the resulting bitrate as well as any mechanical aligning required. We leave this open. The criteria most likely are not different from the optics that make cameras enjoyable for us.

The sensor qualities on the sensor board which determine the most effective transmission schemes, however, are the following:

- ability to distinguish amongst hues
- level of sensitivity to amplitude changes

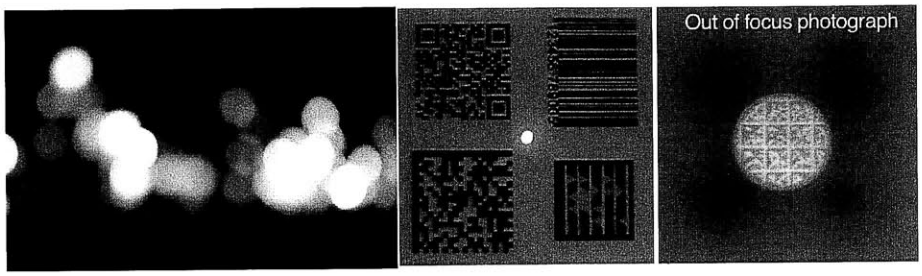

Figure 4.3: Information can be encoded in what we often call the "bokeh" effect. This optical setup, which may be a setup for the future, can still benefit from VRCodes

- "cross-talk" between neighboring pixels
- total number of sensors

Together with the optics, a camera with these favorable qualities will allow for an encoding scheme that packs more data into the "allowable" region without disturbing the human's perception.

4.2 A Novel Optical Hardware Setup

Based on this discussion of important qualities which allow for better datarates, we show an alternative screen-to-camera based interaction with a new optical setup that allows for an uncalibrated and defocused camera to resolve an embedded digital pattern. We include this work here to show that even an alternative optical device setup in the future may benefit from the design and implementation of VRCodes.

In particular, this optical setup allows for the human to see a tiny little dot with the eye as shown in Figure 4.3. The technique here also takes advantage of how our eyes see differently from our cameras by observing that as humans we don't notice the beautiful bokeh effects in the evening! This is due to the fact that our eyes have tiny apertures whereas consumer cameras have larger apertures that can change their focus and blur. If we could introduce a new optical mechanism which would allow us to embed data in that visual bokeh, then we would be able to transfer information in a different way through visible light to our cameras without bothering our eyes.

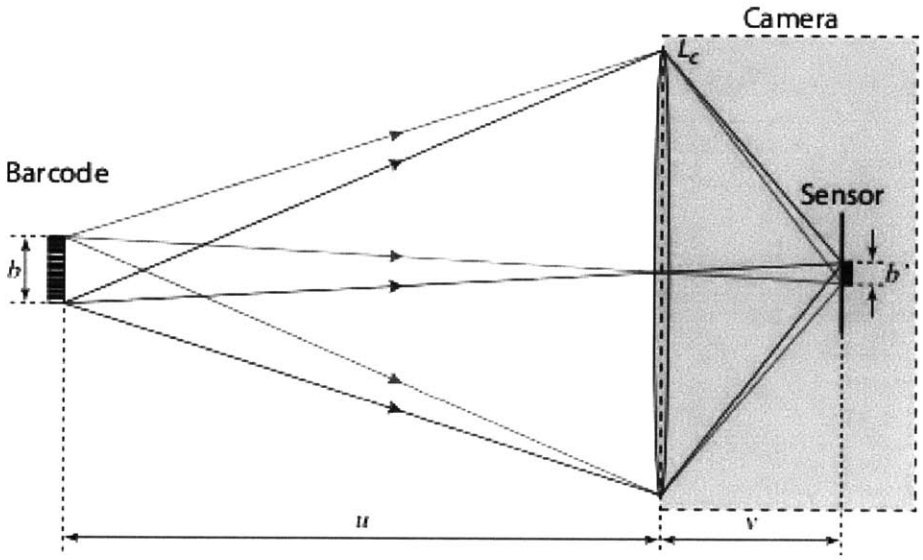

Figure 4.4: A camera looking at a visual barcode pattern in a similar way our eye looks at a visual pattern

As a result, we introduce an alternative hardware setup which allows us to create machine data that is embedded in what appears like a tiny little dot to our eyes. Figure 4.3 (Mohan et al., 2009) shows the novel optical setup that improves some of the challenging aspects of decoding traditional visual codes at a distance. The mentality behind this optical setup is to modify the hardware so that the variety of optical factors that make detection of a traditional barcode challenging can be eliminated.

Figure 4.4 shows a small passive pattern on the left and a camera on the right. Note that this is the same setup as a classical pinhole camera focusing on a barcode image.

The idea is to create a new transmitter device which allows information to be encoded in the bokeh. To begin with, the pin-hole based code provides a good intuition for bokeh based capture of directionally varying rays (Figure 4.5). This simple design is useful to understand the relationship among viewable barcode size b, aperture size a, barcode to pinhole distance d and pinhole to camera distance u.

From Figure 4.5, the size of the visible barcode pattern is given by:

$$b = \frac{ad}{u} \tag{4.1}$$

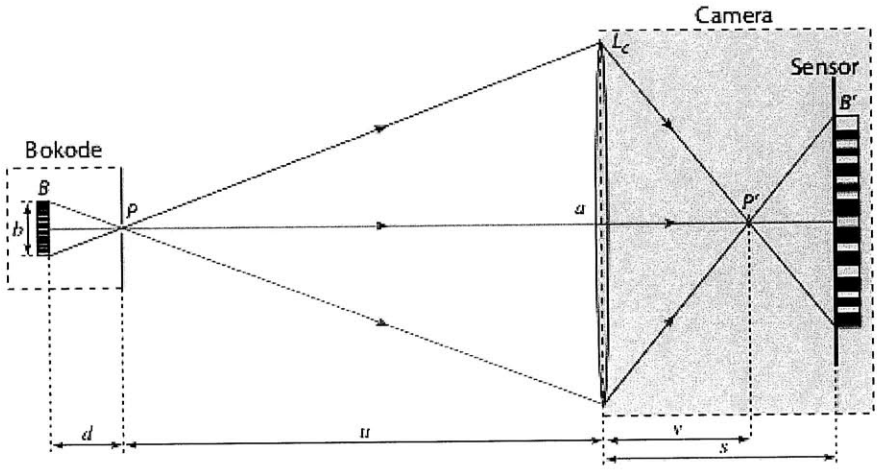

Figure 4.5: A pinhole based setup where the flat barcode pattern has been replaced with a barcode behind a pinhole

For a typical camera with an aperture size of 25mm, the distance between the barcode and the pinhole of 5mm, and the distance of the camera lens from the pinhole of 5m, the resulting viewable pattern size is approximately $25\mu m$. Clearly, this barcode size is much smaller than that of traditional printed barcodes.

Next, consider the magnification achieved at the image sensor to observe this tiny code. The lense images the pinhole P to a point P' at a distance v from the lens. According to the thing lens equation, we have:

$$\frac{1}{f_c} = \frac{1}{u} + \frac{1}{v} \tag{4.2}$$

where f_c is the focal length of the camera lens. Assuming that the pinhole P is infinitely small, the size of the image at a distance v from the lens is also infinitely small. However, as we place the sensor out of focus at a distance s from the lens, we get a highly magnified image of the barcode image. The size of the barcode image is given by:

$$b' = (v - s)a/v \tag{4.3}$$

In the pinhole model, the magnification properties are easy to understand

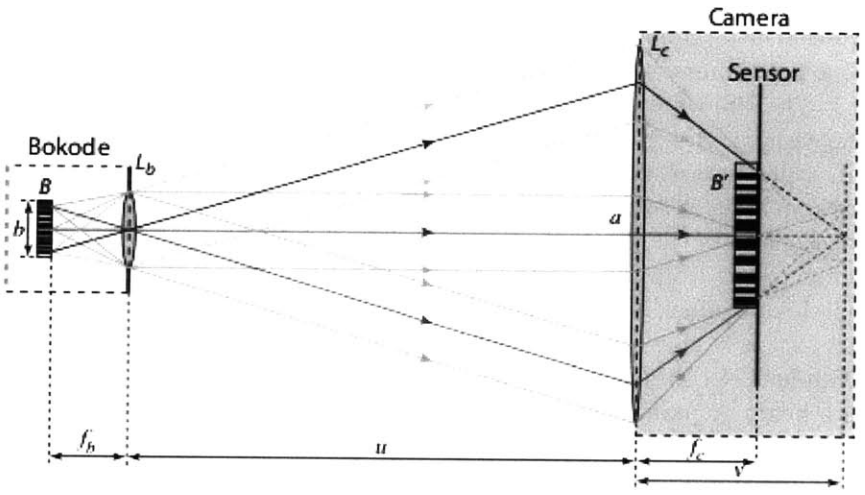

Figure 4.6: A camera looking at a visual barcode pattern with a small lenslet placed carefully at focal length in front of the pattern thus collimating the rays

from the ray diagram, even without Equation 4.3. As seen in Figure 4.4, the barcode image can be made arbitrarily large by simply moving the lens more out of focus and increasing s. Additionally, a larger lens aperture also gives larger magnfication.

With this optical setup, the information of the barcode is embedded in the angular and not in the saptial dimension. By throwing the camera out of focus, we capture this angular information in the defocus blur formed on the sensor. The pinhole is blurred, but the information encoded in the bokeh is sharp.

Adding a Lenslet Unfortunately, the pinhole setup in Figure 4.5 is impractical due to limited light efficiency and diffraction. We replace the pinhole with a small lenslet carefully positioned at a distance equal to its focal length (f_b) away from the barcode pattern as shown in Figure 4.6).

Essentially, a barcode of size b is placed at a distance u from a small lenslet. The size of the barcode image is $b' = (v/u)b$. The effective magnifcation scaling is $M_1 = (b'/b) = (v/u)$. For a typical case where a 50mm focal length lens takes a photo from about $u \approx 5m$, the focused image is at $v \approx 50mm$ and the magnification is $M_1 \approx 0.01$. The magnification reduces as the distance

of the camera from the barcode increases.

The solution is based on defocus blur also known as the bokeh effect. It provides depth-independent magnification of features. The defocus blur of a point light source on the image sensor is called the point spread function (PSF). The PSF of a camera depends on depth u of the point with respect to the camera, the camera's plane of focus. A camera focuses by changing the distance between the lens and the image snsor. When the camera lens focuses on a point source, the image is very close to a point (ignoring lens aberrations). As the plane of focus moves away from the point source, the image expands from a point to fill a circular disc. The disc is also called the circle of confusion. This optical setup encodes useful information in the bokeh that is visible only when the camer is out of focus.

The lenslet collimates the rays coming from a point on the barcode to form a beam or parallel ray bundle. Parallel rays for each point means the virtual image of the barcode is at infinity. The camera focuses at infinity by positioning the camera lens at a distance equal to its focal length f_c from the sensor, and forms an image of the barcode on the sensor.

The viewable part of the barcode produces an image of size:

$$b' = (v - f_c)a/v = f_c a/u \tag{4.4}$$

Finally, substituting $d = f_b$ and using Equations 4.1 and 4.4, we get

$$M_b = f_c/f_b \tag{4.5}$$

The resulting optics of this setup are very similar to that of an infinity corrected microscope. The lenslet acts like the microscope's objective and the camera lens is similar to the eyepiece. Unlike a traditional microscope, however, there is no tube connecting the two lenses, and we have multiple microscopes sharing the same eye piece in a scene with more than one pattern.

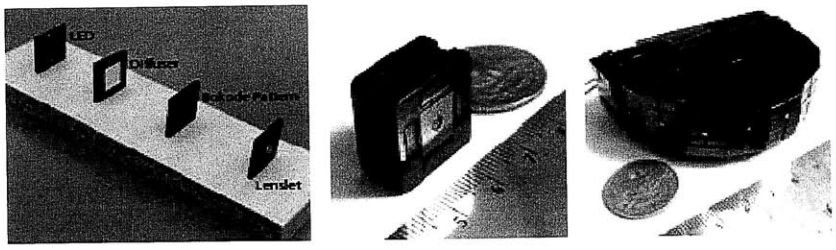

Figure 4.7: An exploded view of the optical setup for the little emitter

For a typical setup, we use a lenslet with focal length of 5mm, and a camera lens with focal length of 50mm. The effective magnification factor around $M_b = 100$. Compare this to the $M_1 \approx 0.01$ obtained in the case of a traditional barcode, the image is 1000 time larger. Futhermore, for a typical viewable lenslet-setup with a region size of $b = 25\mu m$ (obtained above), we have $b' = 250\mu m$ which implies a coverage of around 50-150 pixels on the image sensor (depending on the sensor pixel size). This also means, for the same sized barcode we can potentially back 1000 times more pixels in both dimensions and effectively one million times more data. Equation 4.1, reminds us that if we reduce lenslet-lens distance u to zero, i.e. by taking photo touching the lenslet, we can recover all this information.

Benefits of Optical Setup In addition to the small physical size of this alternative emitter, the optical setup has several useful properties that make the it well suited for a range of at-distance tagging, and angle estimation problems compatible with commmodity cameras.

Camera focus: In order to image the pattern, the camera focuses at infinity. This is independent of the lenslet-camera distance. This is a significant advantage over the traditional barcode where cameras require refocusing when the depth changes. When the scene is in sharp focus, the pattern (3mm diameter in our pro- totype), occupies very few pixels in the image and does not intrude into the rest of the scene.

Camera lens aperture: The size of viewable barcode pattern is proportional to the camera aperture size (Equation 4.1). A relatively large lens aperture is required to see a reasonable part of the pattern. This explains why the pattern is effectively 'in- visible' to the human eye which has a relatively small pupil size of 2mm to 6mm. In Section 5.2 we discuss ways capturing the pattern even with a small camera aperture.

Lenslet to camera distance: As shown in Equation 5, the magnification of the optical system is independent of the distance between the lenslet and the camera. This is different from a traditional barcode and makes decoding the pattern much easier. The size of viewable barcode pattern is inversely proportional to the camera-lenslet distance (Equation 1). Hence, a human eye or small aperture camera, can view the larger pattern by holding it up close. The distance-dependent viewing region may be used to encode hierarchical information into barcodes so that cameras recover more viewable bits as they get closer.

Orientation: The camera views a different part of the pattern depending on its position relative to the lenslet. The viewable region of the pattern

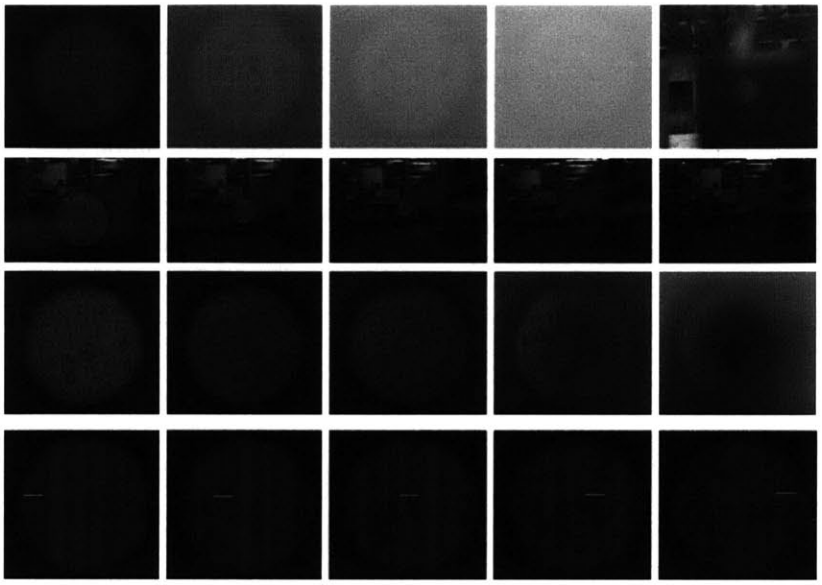

Figure 4.8: Four tests showing this novel optical hardware setup in context of ambient lighting, distance, angle and translation

is a function of the angle formed between the camera and the lenslet's optical axis. Unlike a traditional barcode, with appropriate pattern design, the pattern gives completely different pieces of information to cameras in different directions. We use this property for estimating the angle in AR applications.

Prototypes and Implementation We explore several techniques to create and capture the patterns. Each design has advantages and limitations, and we believe that the best technique will depend on the exact application.

Figure 4.7 shows an exploded view of an active pprototype. We use a plano-convex lens with a 3mm diameter and 8mm focal length as the lenslet, and a battery powered LED for backlight. We print the pattern with a pixel size of $15\mu m$ on a transparency using a Heidelberg Herkules printer. The acrylic housing ensures that the distance between the lenslet and the pattern is exactly equal to the lenslet's focal length (8mm). The whole assemblage is prototyped for easy modification and a smaller housing is possible for manufacturing.

This optical setup can also be completely passive by replacing the LED with a *retroreflector*. We use the camera flash to illuminate the pattern behind the lenslet, and the retroreflector reflects the light back towards the camera lens. We place a polarizer in front of the camera lens, and another in front of the flash such that their polarization direction is perpendicular to one another. This elminates specular reflection from the lenslet and enhances the pattern. We also experimented with a beamsplitter to position the flash at the center of projection of the camera.

To capture images using this optical setup, we use readily available consumer cameras to demonstrate the point. Canon Digital Rebel XSi and the Canon 5D IL, paired with reasonably large aperture lenses, EF *85 mm f/1.8* and the EF *50mm f/1.8* for majority of the demonstrations. The lenses were used at their largest aperture setting, and manually focused to infinity.

For AR applications, we capture two photos (one focused at the scene, and another at infinity) to simultaneously capture the scene and the embedded information. A beamsplitter is used with two synchronized cameras that share the same center of projection. In the future, this can be achieved with a camera changing aperture from narrowest to widest, with an auto-focus like mechanism for changing focus from the lenslet distance to infinity in successive frames, or with an extra out of focus sensor for cameras with multiple CCDs.

A camera with a much smaller aperture (such as a mobile phone camera) will see a limited part of the embedded pattern unless it is at a close range. However, even from a large depth, we can make a larger part of the code viewable. We translate the camera within one exposure time with the lens focused at infinity. Different parts of the pattern are then imaged on different parts of the sensor.

Optical Properties Properties of the prototype can have performance evaluated under several conditions. This analysis is specific to our current prototype that uses hand-assembled components and a pattern with a feature size of $p = 15\mu m$. A higher resolution pattern would greatly improve many of these characteristics.

Resolution limit: The printing resolution currently sets a limit on the information content of the pattern design. Additionally, diffraction due to the finite size of the lenslet sets a hard limit on the maximum resolving power of the lenslet-setup. The angular resolution of the system is given by $\sin(\theta) = 1.22(\lambda/\alpha_s)$, where λ is the wavelength of light, and α_s is the diameter of the lenslet. For our prototype with a lenslet aperture of $3mm$

and a focal length of $8mm$, the resolution limit is approximately $1.8\mu m$ for visible light.

Ambient light: The contrast of the image depends on the ambient light and texture around the lens. The first figure in Figure 4.8 shows the image of the pattern captured under varying illumination conditions, including outdoors. High frequency texture surrounding the code also reduces contrast.

Working distance and angle range: The second image in Figure 4.8 shows photos of the back pattern captured at different lenslet-to-camera distances. Consistent to the described theoretical setup, the maximum distance at which one can see the co mplete Data Matrix is around $2m$. Reducing the pattern feature size from $15\mu m$ to $5\mu m$ increases this distance to $6m$ while still remaining within the diffraction limit. We obtain photos with good contrast and detail from over $4m$ away from the lenslet. The third figure in Figure 4.8 shows photos captured at different camera angles to the optical axis. We obtain robustly decodable codes for a cone of approximately 20 degrees around the optical axis. The angular range is limited by lens aberrations and vignetting due to the directional nature of the LED. A higher quality lens and more diffuse LED may increase this range, but may result in an increased cost and reduced brightness.

Angle and depth estimate: We use a 1D DeBruijn sequence to test the angular resolution of our prototype. For our current prototype, we have an angular resolution of approximately $\alpha = \arctan(p/f_b) \approx 0.1$.

We also compare angle estimation robustness of the prototype to ARToolKit for small changes in angle, with the camera positioned along the optical axis, and exactly overhaed for the planar tag. ARToolKit relies on changes in the shape of the black rectangle to estimate the angle. The changes in shape are significant when the tag is viewed from oblique angles, and provide reliable angle estimates. However, these changes are subtle when the camera is exactly overhead, result in jitter noise. Angle estimation is more robust in this case because it primarily relies on the digital information contained in the visible Data Matrix codes. We estimate the depth from the size of the circle of confusion it produces on the sensor. Like other depth from defocus systems (Pentland, 1987), the depth resolution falls of inversely with distance.

The small physical size and the unobtrusive nature of this lenslet-setup makes them suitable for estimating camera pose and distance for some of augmented reality and motion capture applications we outline. Unfortunately, the angular range of the current prototype is limited to 20 degrees, and this may limit its direct applicability in classic AR applications. Combining this optical setup with existing planar fiducial based AR tags may provide

reliable angular estimates for a wide range of angles. Unlike many other AR techniques, the point source itself occupies very few pixels in the focused photo, and is relatively unobtrusive to the scene. We require the use of a second camera focused at infinity to capture the angular information. For motion capture, we get identification in addition to position and orientation, and the system does not have to deal with marker swapping and marker reacquisition, even when the markers go oustide the scene and come back.

While the specific optical setup suffers from limitations in providing a motioncapture system, it may be used in context of many of the applications outlined in Chapter 5. Higher resolution patterns allow for the use of smaller aperture cameras, and initial experiments with electron beam lithography suggest that we can easily get close to a micron feature size. The physical thickness of the lenslet is greater than the traditional barcode or optical marker. It may even be possible to reduce the depth to that of an origami lens, a Fresnel lenslet or a reflective transmission mode holographic optical element as a Fourier hologram or a single hogel (holographic pixel). While the prototype requires a relatively dark area with a low frequency texture surrounding it, experiments reveal that it is reasonably robust to ambient light. Auto-exposure and motion blur on some cameras may result in poor image quality, but a special mode on the camera may help. We can improve the current passive design by using better retroreflectors, or fluorescent reflectors coupled with a UV camera.

Figure 4.9: A third person shot of a SLR camera taking an out of focus image of the small lenslet setup

The clever optical setup once put together according to Figure 4.9 can be used in conjuction with the software frameowrk described in this thesis. By replacing the passive transparent pattern in the back with a tiny active screen, *VRCodes* may be used together with the hardware setup as well as many other setups to further improve throughput. Further, as asserted throughout this thesis, the software components may be improved independently and concurrently with these evolving hardware setups.

4.3 A Practical Method for Helping to Achieve Screen-to-Camera Capacity

Just like how the basic setup in Figure 4.1 allows for parallel development of new hardware form factors and software form factors, it also allows many of the techniques that we are already very familiar with in the radio frequency design to be adapted for direct application to these seemingly different devices.

Based on this hypothesis, we show that many of the schemes used for radio frequency platforms can be adapted for improving the performance of VRCodes due to the simple noise models that our decoder chain show. Figure 4.10 shows typical distribution of soft values exposed in a radio testbed experiment conducted in (Woo et al., 2007). The simple noise model suggests that many of the techniques which are used for radio frequency systems can be adapted here to work with systems through visible light.

We present one method for improving error probabilities within this framework as adapted from SOFT (Woo et al., 2007), a method for improving detection probability in a wireless RF channel. The premise of SOFT is to utilize the inherent spatial diversity of a screen to camera channel. The numerous observations can be exploited to recover from errors in the final bit estimation.

Consider an extreme scenario where the bit error rate is about 10e-3 and the packet size is 1500B (i.e., 12000 bits). In this case, the probability that an the first observation correctly decodes a transmitted packet is 0.99912000 $\approx$ 10e-5. Say we have two frames with two different observations, and the bit errors are independent. If one can combine the correct bits across multiple observations of the scene to produce a clean packet, the delivery probability becomes 0.99.2 The problem, however, is that when multiple observations differ on the value of a bit, it is unclear which observation is correct. Prior cooperation proposal attempts to resolve conflicts between observations by trying all possible combinations and accepting the combination that satisfies

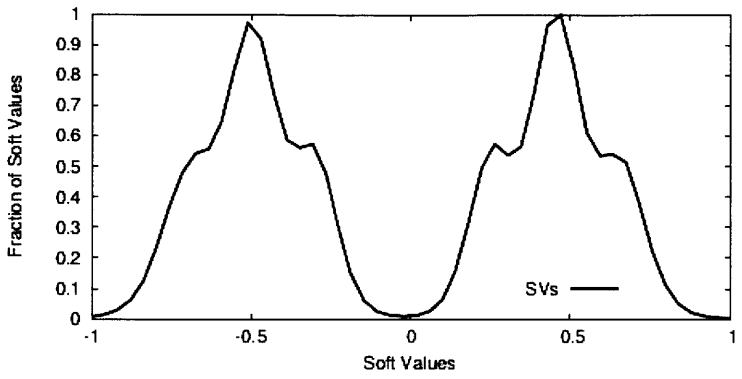

Figure 4.10: A typical distribution of the soft values seen in the Universal Software Radio Peripheral testbed. The two modes corresponding to "0" and "1" bits

the packet checksum (Miu et al., 2005). Such an approach, however, has an exponential cost, limiting its applicability to packets with only a handful of corrupted bits.

Instead, SOFT, a typically cross-layer architecture in RF may be adapted for the visible light medium and recover faulty packets from a single observation. Currently, software-based visible light decoders such as the QR-code decoder (18004:2006, 2011) computes a confidence measure on their 0-1 decision for each bit. However, due to the limitation of the interface between the physical and data link layer, this confidence information is thrown away during the thresholding step. SOFT may be used to export this confidence measure from the physical to the data link layer. It shows that this new interface can combat null-reads and dramatically increase the overal packet delivery rate.

In essence, SOFT works by:

- combining confidence values across multiple faulty receptions to recover a clean packet. Corrupted packets may be combined in software by taking into account the confidence estimates for every decoded packet.
- faulty packets combine the received packet with another to reduce the total amount of time for decoding.

One engineering challenge in the design of SOFT is to identify the strategy that maximizes the likelihood of correctly reconstructing a packet

from multiple faulty receptions. Say that we have multiple receptions of the same packet annotated with physical-layer confidence values, but they do not agree on the value of the i th bit in the packet. How should one resolve the conflict? One could take a majority vote. Alternatively, one could assign to the bit the value associated with the highest confidence. Both of these strategies are suboptimal, and may be even destructive. For example, some of the repeated observations may be affected by a collision or an interfering light signal, and thus end up polluting the information rather than enhancing it. By estimating the noise variance at each receiver and taking it into account, we do better than the above strategies. We demonstrate analytically and show experimentally the superiority of this approach.

In essence, SOFT shows significantly higher delivery rates than prior cooperation proposals that do not use physical layer information. It can provide up to 9-fold reduction in comparison to PPR and MRD (Miu et al., 2005) methods.

The use of spatial and temporal diversity, i.e. combining information from multiple observations to boos throughput is a highly active area. Some work is purely theoretical with no system implementation or experimental results. The main difficulty in implementing these theoretical proposals is their reliance on tight synchroniza- tion across nodes. Thus, practical systems that exploit spatial diversity usually operate with multiple antennas on the same card (i.e., MIMO systems). This analogy is quite strong as the screen-to-camera system can effectively be seen as a multiple-input-to-multiple-output system.

Work on diversity also appears in the area of soft-handoff in CDMA cellular networks. While a mobile phone is moving from one cell to another, the two base stations transmit the same information to the mobile phone and listen to the phone's signal. The stronger signal is utilized for each frame in the call or both signals are combined to produce a clearer copy of the signal. Combining the signals is termed softer handoff and is possible when the cells involved in the handoff have a single cell site (i.e., they are the same node). Further, a few theoretical works compute the capacity improvements from combining 802.11 signals in a way similar to CDMA networks.

SOFT is inspired by signal combining technologies used in RAKE receivers and softer handoff in CDMA cellular networks. SOFT, however, does not combine signals; it combines confidence values associated with individual bits. SOFT defines a simple interface between the physical and datalink layers; each bit is assigned a normalized confidence value that can be expressed with a few bits (3 bits in our results). This interface limits the communication overhead and simplifies the combining algorithm.

A close piece of work is by Miu (Miu et al., 2005) et al, who propose combining multiple receptions. Their work does not use physical layer information but rather dvides the packet into multiple blocks. When the access points receive conflicting block, this proposal attempts to resolve the conflict by trying all block combinations and checking whether any of the resulting combinations produce a packet that passes the checksum. The overhead of their approach is exponential in the number of erroneous blocks and thus too costly when there is a signficant loss.

Soft Decoding The use of confidence values in decoding is usually referred to as *soft decoding*. Prior work in information theory has discussed soft decoding and applied it to various codes including Reed Solomon and LDPC (Gallager, 2007). The Viterbi decoding algorithm has a soft extension (SOVA) that includes physical information as a priori probabilities of the input symbols. SOFT presents a systems architecture that is inspired by soft decoding. It pushes soft decoding up to higher layers and incorporates it withing a cooperative design where multiple observations may be cached and combined to recover fault packets. Futhermore, it focuses on practical network issues, such as compatibility with current applications and the framework presented in this thesis.

In parallel, other groups have used soft information for a different purpose. They use soft information to find incorrect chunks in packets and retransmit them. In contrast, SOFT reconstructs correct packets by combining soft information across multiple erroneous receptions.

SOFT sits below the MAC, between the physical layer and the data link layer. It modifies the interface between the two layers. Instead of directly using 0's and 1's, the physical layer returns for each bit a number in [-1, 1], which we call a *soft* value (SV). The SV gives us an estimate of how confident the physical layer is about the decoded bit. The decoded bit is zero when the corresponding SV is negative, and is one otherwise. The smaller the absolute value of the soft value, the less confident the physical layer is in whether the sent bit is a 0 or 1.

SOFT associates soft values with each bit, we refer to this replacement of bits with their corresponding soft values as *soft packet* or *s-packet*. This data link layer in SOFT decode such *s-packets* to obtain the transmitted packets.

Although the details of the physical layers differ across devices and within th esame technology, physical layer design can easily expose a SV for each decoded bit/symbol. Generally speaking, the function of a wireless receiver is to map a received signal to the symbol that was transmitted. To do so, the

receiver first computes a soft value from the received signal for each symbol, which is a real number that gives an estimate of the transmitted symbol. The receiver then maps the soft value to the symbol which is closest to it. For example, when the symbols correspond to a single bit, the soft value is decoded to a "1" bit if it is positive and "0" otherwise. The soft value can be easily generalized to systems where each symbol represents more than one bit.

Combining Algorithm How should one combine the information from multiple corrupted receptions in order to maximize the chances of recovering a clean packet? None of these s-packets is sufficient on its own to recover the original transmission. Yet, because these are s-packets, they carry SVs for each bith, which give the combining algorithm a hint about the noise level in that bit.

Say that the combiner is presented with 3 s-packets. Recall that a positive SV means "1", a negative SV means "0", and that the magnitude of the SV refers to the confidence that the physical layer has in the decision. Let the SV corresponding to the i^{th} bit is "0" or "1"?

The most straightforward approach would take the maximum corresponding SV since it is the reception with the most confidence. Thus, in the above example, one owuld say that bit i is a "1" because the reception with the highest confidence (i.e., 0.3) has mapped it to "1". But is this the best answer? Clearly, there are other approaches that look equally good. For example, one may resort to a majority vote. In this case, two receptions map this bit to "0" (-0.15 and -0.2) whereas one reception maps it to "1" (one positive value of 0.3). Thus, a majority vote would decide that the bit should be "0", which is the opposite of the previous answer.

There are other approaches too, such as comparing the sum of the positive values to the sum of the negative values, or comparing the sum of the squares of the positive values to the sum of the squares of the negative values, etc. We would like to pick the combining strategy that maximizes the recovery probability.

We analyze the packet recovery probability and adopt a combining strategy that maximizes the likelihood of recov- ering a packet. Our strategy uses the sum of SVs weighted by the noise variance of the capture. It is based on the following lemma, which gives the optimal decision assuming additive Gaussian white noise (AGWN).

Let $y_1, ..., y_k$ be SVs that correspond to multiple receptions of the same bit over different AWGN channels to maximize the recovery probability, one

should map the bit to "0" or "1" according to the following rule:
if $\Sigma_i \frac{y_i}{\sigma_i^2} \geq 0$, then the bit is "1", otherwise it is a "0",
where σ^2, is the noise variance in the i^th AWGN channel.

The proof is as follows:

proof Let $\vec{y} = (y_1, y_2, ...y_k)$ be soft values associated with a particular bit. Given a transmitted bit value x, the y's are conditionally independent, and correspond to multiple receptions of the same bit over different independent AWGN channels n, with $\mu = 0$ and σ^2. I.e. $y_x = hx + n_i$, where h is constant and $x = -1$ for a "0" bit or $x = 1$ for a "1". We model h as a constant independent of the channel because our experimental results show that with the automatic contrast control, h becomes irrelevant.

Let H_0 be the hypothesis that $x = -1$ and let H_1 be the hypothesis that $x = 1$. Note that the two hypothesis are equally likely a priori, i.e., before a reception the likelihood of a "0" is the same as the likelihood of a "1".

Here we consider the likelihood functions:

$$\begin{aligned} H_0 : P_{\vec{y}|H}(\vec{y}|H_0) &= \amalg \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(y_i-(-h))^2}{2\sigma_i^2}} \\ H_1 : P_{\vec{y}|H}(\vec{y}|H_1) &= \amalg \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(y_i-(-h))^2}{2\sigma_i^2}} \end{aligned} \tag{4.6}$$

The optimum decision is to choose:

$$
\begin{aligned}
P(H_0|\vec{y}) &\gtrless P(H_1|\vec{y}) \\
\frac{P(H_0,\vec{y})}{P(vecy)} &\gtrless \frac{P(H_1,\vec{y})}{P(vecy)} \\
P(\vec{y}|H_0)\frac{P(H_0)}{P(vecy)} &\gtrless P(\vec{y}|H_1)\frac{P(H_1)}{P(vecy)} \\
P_{\vec{y}|H}(\vec{y}|H_0) &\gtrless P_{\vec{y}|H}(\vec{y}|H_1) \\
ln\left(\amalg\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{\frac{-(y_i-(-h))^2}{2\sigma_i^2}}\right) &\gtrless ln\left(\amalg\frac{1}{\sqrt{2\pi\sigma_i^2}}e^{\frac{-(y_i-(-h))^2}{2\sigma_i^2}}\right) \\
\sum_i \frac{-(y_i-(-h))^2}{2\sigma_i^2} &\gtrless \sum_i \frac{-(y_i-h)^2}{2\sigma_i^2} \\
\sum_i -\frac{y_i}{\sigma_i^2} &\gtrless \sum_i \frac{y_i}{\sigma_i^2}
\end{aligned}
\tag{4.7}
$$

The above result is equivalent to making a "1" or "0" decision.

We can note the following points based on this simple proof. Straightforward approaches for combining multiple receptions include taking the SV with the highest confidence or doing a majority vote are suboptimal.

Treating all receptions or observations equally is suboptimal. This is particularly important if subsequent observations are separated by a small time gap. It's not necessarily true that more observations directly correspond to a better result. Making an accurate prediction for σ^2 is important for improving reliability. This is similar to maximal ratio combining but since the SVs are normalized by the physical layer, there is no need to multiply by the mean strength of the signal. Another difference is that with SOFT, a combined packet must still pass higher level checksums.

The decision rule above requires the noise variance, σ_i^2, on each channel. Each value is computed by looking at the variance in the received signal. The result assumes an AWGN model. The accuracy of this assumption depends on the precise modulation scheme.

When combined with other methods such as a direct sequence spread spectrum, SOFT can be modified to share a confidence value for a sequence. Each corrected chipc may be mapped to codewords corresponding to "0" and "1" bits in the packet. Although this approach works, a soft value is sent for each received chip. Instead, to improve speed in decodability, a single SV may be sent per codeword.

Figure 4.11: Block diagram for each of the different components

Estimating the Noise Variance In order to compare the soft values observed at different receivers, we require the corresponding σ_i^2 of the channels. In SOFT, the physical layer estimates a channel's variance experimentally using the empirical distribution of the SVs before quantization. The variance is passed to the higher layer in each s-packet.

Thus, for the i^{th} AWGN channel, the SV value of a bit can be expressed as:

$$y_i = hx + n \tag{4.8}$$

where n_i is a white Gaussian noise, h is the channel attenuation, and $x = -1$ for a "0" bit and $x = +1$ for a "1" bit. Thus, conditioned on the bit value, y_i, is also a Gaussian variable with the same variance as n_i. Thus, we can estimate the variance in n_i using the variance in the y's.

Note that the soft values have the same distribution as the n_i's only when we fix the value of the transmitted bit (i.e., a fixed x in the equation above). In general, we cannot tell for sure which SV corresponds to a "0" and which corresponds to a "1". But because the vast majority of soft values already correspond to correctly decodable bits, the statistical variance of the absolute value of the soft values is a good estimate for σ_i^2. We compute this for every packet using all the soft values from it.

Figure 4.11 shows where the SOFT decoding is inserted. Despite the big difference in throughput achieved when using a SOFT decoder, it is important to keep the overhead bounded. Thus, if a SV is expressed as a 16-bit number, then each packet will be amplified 16 times. To reduce the overhead, we need to quantize the SVs. Optimal quantization depends on the σ^2 of the channel. That is, for a particular distribution of SVs, one needs to integrate the total area and divide into 2^k bins where k is the number of bits used to express each SV. In practice, however, we found that a uniform quantization works equally well. Thus the quantization works as follows. It picks a constant cutoff for all SVs across all experiments. The cutoff value depends on the modulation scheme and thus should be calibrated for each

specific rate. Values above the cutoff are reduced to the cutoff value. Every SV is expressed using 3 bits, one for the sign and 2 bits for the magnitude.

Implementation Approach The specific modulation scheme used to test the SOFT approach is Gaussian Minimum Shift Keying (GMSK). However, the ideas developed are independent of this scheme. The main reason for using GMSK to test the GNU radio approach is that the GNU radio project has a mature GMSK implementation. At the time of these results, GNU implementations of other modulation techniques were either non-existent or buggy. We later implemented SOFT with Differential Binary Phase Shift Keying (DBPSK) and the results are qualitatively similar to those obtained with GMSK.

To give an idea of how the SOFT approach performs for the USRP platform, three nodes are used as access points (APs). Each randomly chosen sender transmits 500 packets, where each packet size is 1500B. Two copies of the received packets are used to compare the SOFT approach to the other state of the art approaches.

The approach is compared to alternative combining algorithms (all of which may be also tested in this VRCode framework):

- SOFT: This approach adopts the maximum likelihood technique
- Max Confidence: This is similar to SOFT but with a max-confidence combining strategy. Specifically, when combining multiple SVs that correspond to the same bit, the bit is mapped to a "0" or "1" according to the SV with the maximum absolute value.
- Majority Vote: This is also similar to SOFT but it uses a majority vote to combine multiple receptions. Specifically, each bit is assigned the value agreed on by the majority of the receptions (i.e. the majority of the APs)
- Current Apprpoach: This approach is similar to the current WLAN scenario which does not allow the APs to cooperate, and does not combine bits across different receptions.
- MRD: This is a prior for packet recovery. MRD does not use physical layer information. It works as follows. It divides each packet into blocks for each block, it assumes that at least one of the APs has correctly received the bit values for that block. It attempts to recover a faulty packet by trying every version received for each block, and

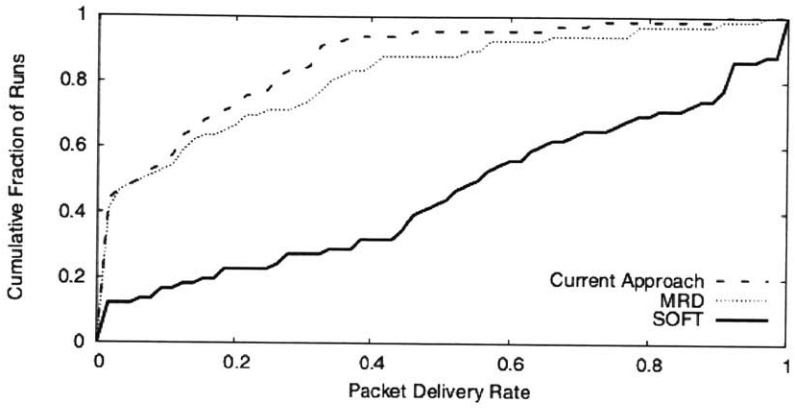

Figure 4.12: CDF of Packet Delivery Rates comparing SOFT against the current approach and the state-of-the-art MRD approach

checking whether such combined packet passes the 802.11 checksum. MRD computational cost is exceptional in the number of blocks. Thus, in order to be computationally feasible. MRD recommends using 6 blocks per packet, which is what we use in our experiments.

We summarize the packet delivery rates with a few results characterized for the RF medium. For these results, the delivery rate is computed as the fraction of uncorrupted packets received at the best access point and the one with the lowest loss rate for the particular sender. For the combining approaches, the delivery rate is the fraction of packets that are successfully recovered either immediately at the radio receiver, or after running the combining algorithm. The delivery rate is computed for each approach without retransmission since this is how it would be tested for a visible light channel.

Figure 4.12 shows Soft significantly improves the packet delivery rate in lossy and low-coverage networks. While the mean delivery rate in the current approach and MRD is less than 7%, the mean SOFT delivery rate is as high as 62%. Retransmissions are turned off in these experiments, thus the improvement in delivery rate is mainly due to the use of soft values and the effect of spatial diversity.

Figure 4.13 compares the delivery rate under three combining strategies: SOFT, Max-Confidence, and Majority Vote. This shows that SOFT's

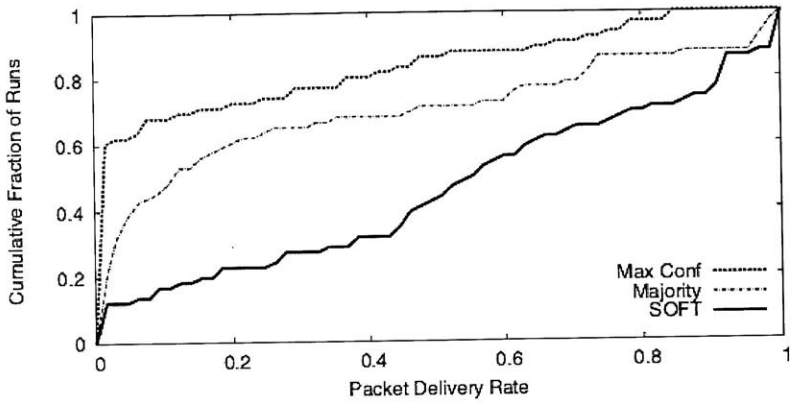

Figure 4.13: CDF of Delivery Rates for different combining strategies that could be alternatives to SOFT's strategy

weighted-sum strategy performs significantly better than the others, which is consistent with analytical results.

Finally, we show that anticipating the calibration noise rate for using the SOFT approach can significantly improve the delivery performance as shown in Figure 4.14

To reduce the overhead of utilizing and computing the SOFT metric, the SV values may be quantized and expressed using only 3 bits. Although (Woo et al., 2007) shows the results on a radio testbed for how much can be saved, these results can not be extrapolated to a different set of hardware. However, the results that the 3 bits is indeed enough to improve performance for a single Gaussian noise profile can be applied.

Lossy links and poor connectivity are two main problems in wireless networks that have direct impact on application perforomance and user satisfaction. Even in the wireless communication field, these techniques continue to be improved as shown with this specific example of (Woo et al., 2007). Given this characterization of the encoding and decoding schemes, however, these new techniques can directly be applied to VRCodes to improve performance. Wireless media has long tried to exploit the inherent spatial and temporal diversity models by reversing complex channel models. The characterization of VRCodes as a solution for a software defined interface allows these techniques to be applied directly.

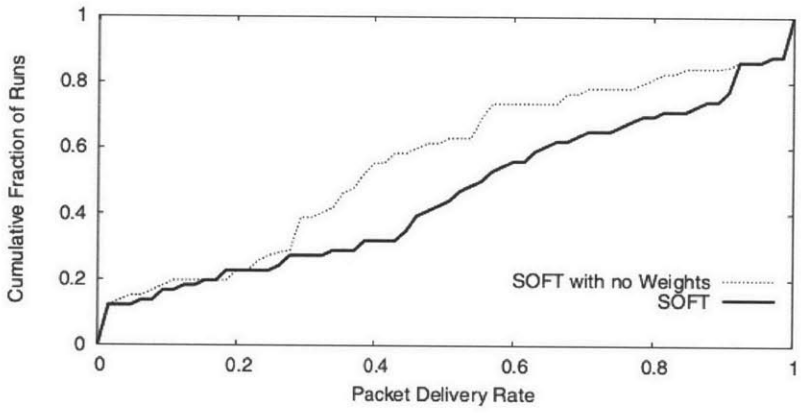

Figure 4.14: CDF of Delivery Rates with and without normalization by the noise variance

Chapter 5

Demonstrating Novel Interactions as a Result of Proposed Framework

Creating visual environments that are immersive has attracted a wide variety of participants who wish to experience new interactions in the context of a physical space. We first discuss much of the prior work in this space and then point out specific techniques which may bring some of those experiences into reality with commodity hardware.

5.1 Related

Specific techniques that involve spatio-temporal data modulation on an active screen are known. But as far as we know, ours is the first system designed to be compatible with widely-deployed active screens and rolling-shutter cameras while at the same time satisfying the primary goal of producing an unobtrusive visual environment. Perhaps the most enabling and successful data transferring methods through the visible light medium are barcodes and Visible Light Communication (VLC) (Komine and Nakagawa, 2002; Little et al., 2008). The problem of creating a tag that can be used with existing visible light architecture has given rise to many passive as well as active solutions which use projectors, LEDs and television screens.

Visual Barcodes Passive barcodes in print mediums are usually decoded using a flying spot scanning laser and a single photodetector which picks up the absence or presence of light reflected from the 0-1 stripes on a 1-D barcode (Gallo and Manduchi, 2011). This is done to avoid the physical

focusing and depth of focus issues common with consumer-grade mobile cameras. Newer codes exploit 2D imaging of advanced camera scanners and pack more information (Swartz and Wang, 1990). They include Data Matrix and both use Reed Solomon (MacKay, 2002) error correction. Shotcode (López de Ipiña et al., 2002) is a circular barcode designed specially for consumer-grade regular cameras thus making it robust toward blur. (Belani, 2011) compared various barcode standards for use with cellphone cameras as the reader.

Active visual codes (Langlotz and Bimber, 2007) use multiple color channels and temporally changing codes displayed on either a LCD screen or projector to maximize the data throughput and the robustness of the barcode recognition with respect to a parametrized camera. The small color liquid-crystal display of a mobile-terminal may also be used as a data-transmitting device (Hranilovic and Kschischang, 2006). Our method can use any of these schemes as the higher-layer transmission protocol. In order for visual codes on active displays to gain widespread use, designers should have maximum flexibility in designing the system content worry-free of the visual obtrusiveness of the code itself. Active visual codes contains much more information and can be read by a well-designed decoder on a camera. Embedding more data for the camera on the active screen implies codes that appear as flashing screens to the human viewer. On the other hand, VRCodes limit the number of perceived colors as a function of the display parameters available; however, media prints and posters are a few examples of custom-designed displays. To increase dynamic range, we use a higher-end 120-Hz screen (consumer 3D-ready display in the sub 1000-dollar range at this time of writing).

Steganography Steganography hides information in the visual medium and is generally considered a form of cryptography. MSU StegoVideo (Qingzhong Liu, 2008) is a standalone software that hides a message in a video stream by re-encoding the primary stream with error-protection. Infrared Data Association (IRDA) is a non-profit standards committee setup to regulate line-of-sight communications in the infrared medium and is regulated as part of the IEEE 802.11 specifications (7, 2011). Ultra-Violet invisible passive ink is a passive method of achieving "secret messaging" (Berson and Auslander, 1994) where the reflective substrate itself is not visible to the human eye until a UV-light is used to reveal the ink. Bokode (Mohan et al., 2009) is an optical technique of embedding data in the bokeh using a special lens setup that can be captured using an out-of-focus camera.

Radio Frequency Identification (RFID) tags (Li, 2007) are used to determine presence of an object with range. In the case of RFID, many phones do come with a near-field-communications (NFC) reader but can not extract relative positioning and orientation. In order to leverage IRDA, UV-based tags and Bokodes, additional specialized readers must be deployed on each consumer mobile device.

Augmented reality and relative positioning Relative orientation and positioning is useful for interactions which need to know relative camera-pose. Normally, this is done by placing the identifier on the peripheral. Here, a camera-enabled peripheral can be used to infer its own positioning. Most techniques require finding at least four point correspondences and thus the accuracy and precision improves with the physical size of the tag. ARTag (Fiala, 2005) and ARToolKit (Billinghurst and Kato, 1999) use four points on the periphery of the pattern to infer position. Tateno et al. (Tateno et al., 2006) proposed nesting smaller markers within other markers to improve AR performance from a large range of distances. Claus and Fitzgibbon (Claus and Fitzgibbon, 2005) perform a survey of several fiducial marker systems, assessing the processing time, identification, and image position accuracy with respect to viewing angle and distance. VRcodes provide identification, distance, and angle from each frame. We describe how to diminish the need for a second camera using post-processing.

Sampling-Based Setup Our specific designs are based on how the camera works differently from the eye. The concept behind modeling the behavior of a rolling shutter camera comes from the motivation to remove metameric effects in displays (Hong et al., 2001). Nayar et. al (Gu et al., 2010) study the rolling shutter camera. Directly using the moire pattern seen only by the camera and not by the human is somewhat opportunistic. This approach is similar to using a microscopic system which reads pixels that the human eye cannot resolve (Traub, 1974) or a light-modulating light bulb which is flickering beyond a perceivable rate. The techniques presented here assume a rolling-shutter camera which does require synchronization with the active screen. However, these techniques may be extended so long as we are using a receiver (with a shutter) that behaves more like a sampling device than like our eyes.

5.2 Novel Interactions

We explore a few scenarios which may benefit from the use of a software-defined pixel space and a personal peripheral capable of interacting with this environment. We point out ahead of time that many of the applications described here are a result of the limitations introduced by long-range RF communications. Although the use of visible light as a data medium for bidirectional surfaces will ultimately require specialized hardware such as LuminAR (Linder and Maes, 2010) or BiDi Screen (Hirsch et al., 2009), a wide range of interfacing devices including screens and cameras already exist for human-compatible applications. The following descriptions outline a few use-cases which may benefit from an assumed ecosystem of televisions and mobile-cameras. We compare the resulting interactions amongst each other and point out differentiating properties.

5.2.1 Audio Line of Sight

In 2010, we demonstrated a pair of headphones with a small camera mounted on top (Woo et al., 2010). By creating this personal device with a small camera, this allowed the pair of headphones to become aware of the context that it's wearer was looking in. As captured in Figure 5.1, the setup was two large 47" screens placed side-by-side to one another. Depending on which direction the wearer looks in, she always hears the audio stream associated with the moving pictures.

Using visible light to transfer this information allows the whole system to become scalable. Since it does not suffer from the same sort of interference problems as open sound or RF, many users can stand in front of the installation at the same time and experience independent personal streams. The directional nature of visible light also realizes the natural expectation for us to hear from the direction we're looking in.

5.2.2 NewsFlash

In 2011, a surge of news-related technologies evolved experimenting with new displays of information. Most of these installations involved new interactions which involved screens and displays. Papereye, a project in collaboration with the Boston GlobeLab (Becker, 2012), involved snapping pictures of paper headlines and linking them with a shareable link. In order to add more interaction, new ways of embedding information needed to be incorporated.

Figure 5.1: Audio Line of Sight: A personal device which allows one to hear only from the direction that one is looking in

In many cases, active screen real estate is limited and there must be more information to understand what is going on.

Newsflash (see Figure 5.3) is the result of this bridge from print to medium. The installation is a collaborative way to experience news events from around the world. The public display shows only the well-designed frontpages of newspapers. The article text is not resolvable. In order to get more information about each frontpage, *VRCodes* are used to embed data in a manner that is not obtrusive to the human eye but visible and decodable by a camera. Although each news page appears to be "static", there are actually video reels that embed data underneath.

Newsflash was built with the use of VRCodes embedded on a display which was capable of emitting 60Hz. It used lighter contrast modulation to create a code that could be decoded using a mobile phone's consumer camera. The directionality can also be combined with this decoding of data so that different users can obtain different streams of data.

5.2.3 Additional New Interactions

There are many other applications that already exist in some form today that can actually benefit from the alternative approach through visible light

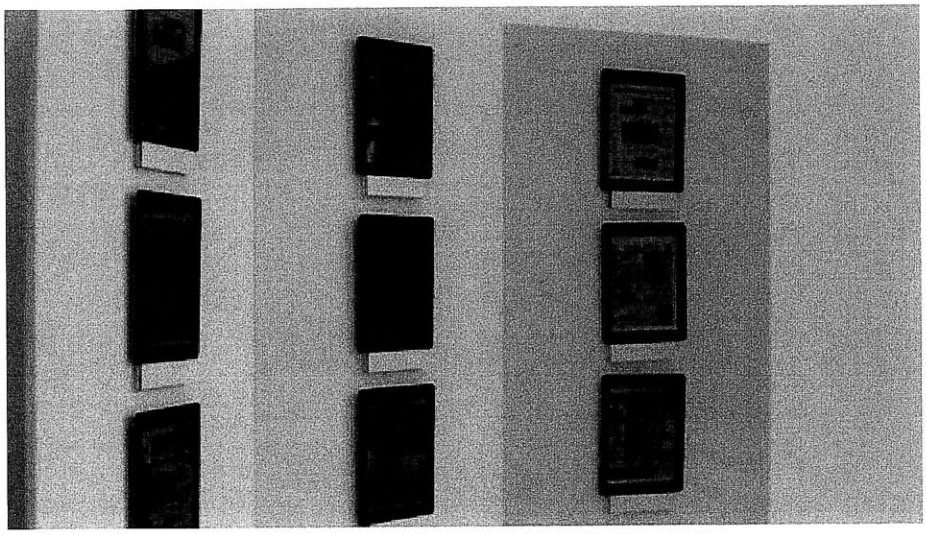

Figure 5.2: Newsflash: a set of displays that each display a different newspaper from around the world.

presented in this thesis. We outline many of these here and draw comparisons by using a series of tables.

Payments, proximal public and private key exchange in Table 5.3 fundamentally relies on a physically secure bidirectional communications channel. The public space requires a camera to interpret a machine-visible transmitting code on a human-perceivable surface. The participant carries a private key on a personal peripheral which is communicated from screen to camera in the public environment. After key verification, the system returns with a response which requires the user to authenticate using a secondary method. The components of security come from the perceivable and directional nature of the channel as well as additional authentication methods which use the same hardware for camera-based verification of natural features including the face.

Design-conscious display public display installations which are an evolution from print require more embedded data to enable proximal interaction. The same codes used for print appear obtrusive and unwieldy for a designer to incorporate. VRCodes as shown in Table 5.2 allow for beautiful designs which can embed data as well as position and orientation. Together with a cloud backend, multiple users can be served at once with differing

Figure 5.3: When a phone is held up to each of the newspapers, it reveals the VRCode which is used to identify each of the individual papers.

data streams which may also be location dependent.

Directional and audio surround-sound from visual gaze allows a participant in the public space to receive audio only from a specific line of sight from multiple directions. Specifically, when used in conjunction with camera-enabled glasses, one can "hear" audio only from the direction he is looking in according to the dimensional sound that we hear in the real world. Many participants may be packed into a small space and each can hear individual digital streams of audio data. This allows automatic synchronization between audio and visual perception. Further, it can allow for true surround sound simulation from afar.

Gaming peripheral creates an image positioning system that allows relative orientation of images given a specific marker. In a similar domain as photosynth, multiple people can take images from different angles. Each picture can be stretched and morphed according to the precise position coordinates obtained from an active marker. The resulting captured images can be displayed across a screens. This same concept can be used on a microscale where an optical-based pen can be used on a surface which embeds digital position data.

	Deployment	**Public/Private Key Exchange**	**Reconfiguration**	**Ease of Use**
VRCodes	bi-directional displays and personal peripheral in participant's hand	spatial transceivers result in higher capacity and security	new codes easily redisplayed on mobile phones	requires line of sight similar to the credit card
RFID	specialized RFID readers as well as transmitters in public space	low capacity and easier to snoop	reprogrammable RFID transmitter requires full software-defined radio	tap-and-go response
Infrared	Phototransistors integrated into existing devices	Without integration of receiver results in slow datarate limited by framerate of existing cameras	new codes easily redeployed	requires line of sight similar to credit card

Table 5.1: Comparison for existing technologies using Public/Private Key Exchange

	Deployment	**Design Aesthetic**	**Scalability**	**Distance**	Positioning
VRCodes	screens and mobile cameras are ubiquitous	unobtrusive	due to large data rates and accurate positioning, can support many users	usable from afar with minor trade-offs	digital positioning
2D Barcodes	screens and mobile cameras are ubiquitous	must be printed large to be effective, fully visible	many users with same data stream	major trade offs in spatial real-estate vs. effectiveness	none
Water marks	screens and mobile cameras are ubiquitous	unobtrusive	can not support many users	usable from afar with minor trade offs	none
AR Tags	screens and mobile cameras are ubiquitous	must be printed large to be effective, fully visible	no data stream, but can support many users for position	usable from afar with minor trade-offs	analog positioning
Active codes	screens and mobile cameras are ubiquitous	flashing and distracting	carry a lot of data but have no relative positioning	usable from afar with minor trade-offs	none

Table 5.2: Comparison for existing technologies using barcode and digital information designs

	Deployment	Audio Quality	Directionality	Tethering
VRCodes	Displayed on ordinary screens and read with camera	Digital transmission allows for digital audio	placement and alignment of audio	untethered
Audio Headphones	Earphone buds and audio jacks	analog transmission limited by clarity of line	none	tethered
Wireless Audio	Radio transceivers on device and audio peripheral	Analog transmission resulting in analog-level clarity	reprogrammable RFID transmitter requires full software-defined radio	untethered
Infrared	Phototransistors	Analog or digital transmission depending on deployment	scattered transmission (not highly directional)	untethered

Table 5.3: Comparison for existing technologies using directional and audio surround-sound from visual gaze

Chapter 6

Conclusion

The goal of this thesis is to present the opportunity for every light in the environment around us to be used as a data transmitting source as well as for position/orientation in a manner that is not obtrusive to humans. By defining a software-only platform, we are able to design new techniques such as VRCodes. We demonstrate a few interactions that are enabled using existing hardware such as mobile cameras and television screens.

6.1 A Systems Argument for Proximal Light-Based Interactions

As a final point, we present an alternative systems-level argument for light-based proximal interactions. As networked devices such as cellphones, i-pads, netbooks, e-books, mp3-players, cameras and laptops fill up schools and enterprises, the demand for small wireless devices in dense networks increases. Cisco now recommends a 10:1 wireless device to access point ratio resulting in on average 1 access point every $9m^2$ in an enterprise setting (Cisco, 2004). Motivated by user demand, industry concerns gradually shift from addressing coverage to mitigating interference.

As a result, abstractions developed to simplify network design are desperately broken apart to deal with the inevitable saturation of the shared (regulated) wireless medium. Previously confined to the physical layer, low level communications concepts increasingly show presence in recent state-of-the-art link-layer and network-layer protocols. Concurrent efforts pursue increasingly complex modeling methods in order to optimize performance of current off-the-shelf devices (Liu et al., 2009; Shrivastava et al., 2009).

Abstractions in wireless networks are seriously hindered by the shortage of reliable models.

Looking ahead, there is a trend toward the use of higher frequencies from the EM spectrum. As engineers develop devices employing high frequencies with corresponding wavelengths less than a few millimeters, they encounter line of sight obstacles in their deployment. Up and coming 60GHz communication devices exhibit up to 36dB attenuation as a result of architectural and mobile occlusions (Smulders, 2002). Likewise, optical links offer multi-gigabit (IRDA specifications, 2008) and even terabit (Hirata et al., 2003) bitrates which demand highly geometrically-aware placement. Propagation at these millimeter-wave frequencies effectively approaches strict LOS requirements as a result of the typical materials present in urban and indoor environments.

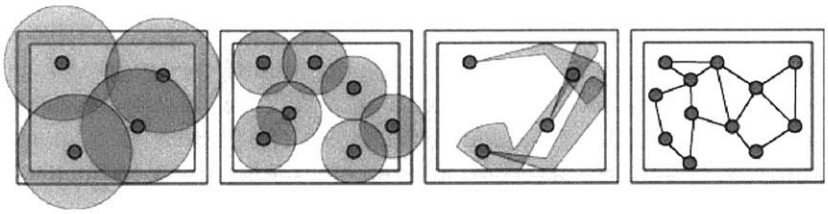

Figure 6.1: 802.11, 802.11n, cell phones towers must manage power and frequency dimensions, (b) bluetooth, RFID are designed for low power and increased spatial reuse, (c) free-space optics and directional antennas use beamforming but must still address multipath, (d) new low-power directional devices may be safely reduced to a directed graph.

Despite the common sentiment that the LOS requirement is a serious "obstacle", we propose that the inherent directionality of unregulated high frequency EM spectrum should be embraced as shown in Figure 6.1.

Rather than view LOS as a hindrance, directionality dramatically reduces interference thus making the problem of modeling and abstraction tractable. Therefore, to build very dense indoor wireless networks, we look toward devices operating in the micrometer wavelengths of visible and infrared light which offer orders of magnitude increase in raw bandwidth as well as the benefit of very high directionality.

High directionality and interference-free operation of visible light devices is enabled by our ability to focus high frequency radiation with simple physical means. Consider a single LED captured by a camera where SNR is perceived only as a function of ambient light. Given a completely dark environment, an automatic focusing mechanism can easily resolve a source which is miles

away. These relatively simple and fully-realizable techniques are already common in environments such as extra-terrestrial satellite communications, developing-world networks and urban last-mile outdoor links(Hemmati, 2006; Kedar and Arnon, 2004). In contrast, directing RF radiation requires "phase engineering" via complex antenna design or MIMO algorithms(Gesbert et al., 2003).

Using our own framework, we realize experimental physical layer devices which allow us to analyze a simple network-layer protocol. We can then demonstrate in practice that interference-free devices using high-frequencies actually have generous "antenna-packing" limits [1] and are realizable in practice.

In other words, the human-compatible methodologies presented in this work can be seen to have the following contributions: (1) a prototype system supporting our assertion that low-power directional devices are ideal for scalability in an enterprise wireless network (2) a demonstration that theoretical models which achieve dramatic capacity gains can be realized in practice (3) an abstraction argument which extends to system deployability and applications of wireless proximal networks.

Prior Scalability Arguments Gupta and Kumar's (Gupta and Kumar, 1999) analysis of wireless network scalability ultimately relied on a non-spatially-directional model which caused collision if not appropriately managed. However, as shown in our demonstration with LED emitter and sensor in Figure 2, there exist several physical device form-factors which do not require medium control. This gives rise to entirely new scalability laws which Gupta and Kumar mention themselves as future work.

Likewise, when Shepard(Shepard, 1996) first highlighted the use of dense multi-radio wireless networks, it was envisioned to be ubiquitous for an *outdoor* environment. Dense indoor networks have brashly adopted all challenges which come with multi radio devices including interference, collision, unclear abstractions and poor-scalability. Worst of all, the lack of a directional and proximal communicating system has resulted in a lack of new interactions.

The use of directional physical components to build networks has been explored in many communities. Komine et. al presented prototypes and a fundamental framework for transmission of white LED light in (Komine and Nakagawa, 2002). They find the use of white LED light transmission to be

[1] This is in contrast to RF MIMO solutions which require at least half a wavelength of separation between antennas (a few cm) thus making scalability on a small device unrealistic.

an efficient and low-overhead method of transmitting bits in an on-off keying (OOK) time modulating manner. Little et. al (Little et al., 2008) build an indoor wireless lighting system also employing OOK time modulation. Their efforts to build practical transceivers brings out many of the challenges reminiscent of classic RF communications including crosstalk and contention of light-waves. They also demonstrate an end-to-end system all the way to the user level.

Using directional antennas which steer line-of-sight beams has also drawn attention with recent work considering the 60GHz bandwidth. Building networks from 60GHz devices has drawn attention due to different LOS and reflection characteristics (Liu et al., 2009; Shrivastava et al., 2009). Many measurement-based works show that alternative network-layer approaches must be used with physical devices that are fundamentally different. In (Liu et al., 2009), the authors offer several insights toward scalability, angle-dependence and resulting protocol-design.

Despite prior work and theory for indoor optical systems, directional devices which provide bitrates on the order of gigabits are just now becoming widely available. Gigabit-infrared links are already on the market. Advanced semiconductor research turn their attention toward small LED form factor multi-gigabit (even terrabit) emitters and detectors (IRDA specifications, 2008). Resonant Cavity Light Emitting Diodes (RCLEDs) find their way to deliver multi-gigabit free-space optical speeds. The low-power, and low-cost of these hardware developments will ultimately motivate a complete system implementation and opportunity to rethink current megabit architectures with wireless RF without regard toward regulations and interference limits.

With clear abstractions in place, we demonstrate the potentials of network-layer development using proof-of-concept physical layer devices. We currently pursue more sophisticated physical layer devices for a realistic testbed. The suggestions in this work frame how to address other scenarios such as mobility and handoff. Based on this work, the intuition is that network-layer development offers far greater gains in overall throughput and spatial reuse than abstract-breaking techniques. The directionality of the visible light spectrum may be adapted for mobility scenarios.

With this in mind, we present a few systems-level descriptions of this scenario:

Plug-n-Play Shared Medium Despite the effort to reduce interference and collision, users seem to have accept the notion of "dead spots". Bluetooth devices which connect to handsets in a 4ft x 4ft area, experience interference and suffer from lack of data rate despite appearing as having maintained connection. Access points perform active load balancing and drop packets

intermittently as temporarily relief. A *Plug-n-Play* Ethernet-like deployment does not suffer from the same drawbacks in practice as these densely packed wireless devices.

Unregulated Privacy Networks In addition to high frequencies in the spectrum being unregulated, home indoor networks do not fall into complete jurisdiction of FCC laws. A directional indoor network need not go through 3rd parties even in initial phases for applications such as telephony and VOIP

High Throughput Meshing RM mesh networks are a valuable solution to enterprise and outdoor networks due to the benefits they offer such as ease of deployability and high-throughput. In comparison, the directional LOS networks which offer scalability and higher throughput can be deployed incrementally. They are also backward compatible and able to co-exist with wired Ethernet bridges. Mesh networks formed from these physical devices offer at least 10x throughput over existing wireless hardware. Devices in the visible light spectrum will offer further wireless throughput.

Distributed Storage With an infrastructure already in place, loading and off-loading data to an existing architecture seems a small step. Optical devices, as they are priced today, are low-power devices which may offer substantial benefits in an everyday environment. Augmenting storage is a natural application.

Summary Although confronting interference, congestion and lack of directionality below the network-layer seems a good idea, new opportunities emerge where the capacity gains of directional links may outweigh the computation necessary to combat interference and collision. End-to-end principles and layered abstraction are solid design concepts which may be adhered to in indoor systems without sacrificing capacity nor complexity. In fact, with (a) small devices in a dense network on the rise and (b) a trend toward the use of higher frequencies in the RF spectrum, directional and line-of-sight links are the key towards wireless networks with increase capacity, clear abstraction, and ease of deployment. Looking ahead toward the use of up-and-coming directional physical-layer devices in addition to the ones presented in this thesis, this offers great opportunities for improving practical network throughput. Most importantly, pursuing these software methods for embedding information in visible light allows us to use the devices we already have as a general platform. Ultimately, it allows us to reconnect with the physical world by giving us flexibility to design our own visual environment while at the same time not being obtrusive to the human eye.

Bibliography

18004:2006, I. (2011). Information technology. automatic identification and data capture techniques. bar code symbology. qr code.

7, I. . W. T. G. (2011). Visible light communication. IEEE Standards Association.

Becker, D. (2012). Globelab: Breathing new life into journalism.

Belani, H. (2011). Barcodes for mobile devices by hiroko kato, keng t. tan and douglas chai. *ACM SIGSOFT Software Engineering Notes* **36**, 32–33.

Berson, W. and Auslander, J. D. (1994). Bar code scanner for reading a visible link and a luminescent invisible ink. United State Patent.

Billinghurst, M. and Kato, H. (1999). Collaborative mixed reality. In *Proceedings of International Symposium on Mixed Reality (ISMR '99). Mixed Reality–Merging Real and Virtual Worlds.*

Bimler, D. (2012). Direction in the color plane as a factor in chromatic flicker and chromatic motion. *J. Opt. Soc. Am. A* **29**, A74–A81.

Bose, V. G., Guttag, J. V., Tennenhouse, D. L. and Smith, A. C. (1999). Design and implementation of software radios using a general purpose processor.

Cisco (2004). How cisco deployed wireless access points worldwide. In *http://www.cisco.com/web/about/ciscoitatwork/downloads/ciscoitatwork/pdf/Cisco_IT_Case_Study_Wireless_LAN_2004.pdf*.

Claus, D. and Fitzgibbon, A. W. (2005). Reliable automatic calibration of a marker-based position tracking system.

Fiala, M. (2005). Artag, a fiducial marker system using digital techniques. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05. Washington, DC, USA: IEEE Computer Society.

Galbi, D. A. (2003). Revolutionary ideas for radio regulation. Law and economics, Federal Communications Commission.

Gallager, R. G. (2007). 6.450: Principles of digital communication.

Gallo, O. and Manduchi, R. (2011). Reading 1-d barcodes with mobile phones using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Gesbert, D., Shafi, M., shan Shiu, D., Smith, P. and Naguib, A. (2003). From theory to practice: an overview of mimo space-time coded wireless systems. *Selected Areas in Communications, IEEE Journal on* **21**, 281–302. ISSN 0733-8716.

Gregory, R. (1979). *Eye and Brain, the psychology of seeing*. New York, Toronto: McGraw-Hill Paperbacks. ISBN xxx.

Grey, P. (2012). How your camera works.

Gu, J., Hitomi, Y., Mitsunaga, T. and Nayar, S. K. (2010). Coded Rolling Shutter Photography: Flexible Space-Time Sampling. In *IEEE International Conference on Computational Photography (ICCP)*.

Gupta, P. and Kumar, P. (1999). Capacity of wireless networks.

Hemmati, H. (2006). *Deep Space Optical Communications*. Wiley amp.

Hirata, A., Harada, M. and Nagatsuma, T. (2003). 120-ghz wireless link using photonic techniques for generation, modulation, and emission of millimeter-wave signals. *J. Lightwave Technol.* **21**, 2145.

Hirsch, M., Lanman, D., Holtzman, H. and Raskar, R. (2009). BiDi screen: A thin, depth-sensing LCD for 3D interaction using lights fields. *ACM Trans. Graph.* **28**.

Hong, G., Luo, R. and Rhodes, P. (2001). A study of digital camera colorimetric characterisation based on polynomial modelling. *John Wiley and Sons* .

Hranilovic, S. and Kschischang, F. (2006). A pixelated mimo wireless optical communication system. *Selected Topics in Quantum Electronics, IEEE Journal of* **12**, 859–874.

IRDA specifications (2008). gigaIR. In *http://www.irda.org*.

Kedar, D. and Arnon, S. (2004). Urban optical wireless communication networks: the main challenges and possible solutions. *Communications Magazine, IEEE* **42**, S2–S7. ISSN 0163-6804.

Komine, T. and Nakagawa, M. (2002). Integrated system of white led visible-light communication and power-line communication. In *13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2002.*

Langlotz, T. and Bimber, O. (2007). Unsynchronized 4d barcodes: coding and decoding time-multiplexed 2d colorcodes. In *Proceedings of the 3rd international conference on Advances in visual computing - Volume Part I.*

Li, T. Wang, G. (2007). Security analysis of two ultra-lightweight rfid authentication protocols , 109–120.

Linder, N. and Maes, P. (2010). Luminar: portable robotic augmented reality interface design and prototype. *Proceedings of UIST2010* .

Little, T. D. C., Dib, P., Shah, K., Barraford, N. and Gallagher, B. (2008). Using led lighting for ubiquitous indoor wireless networking. In *Proceedings of the 2008 IEEE International Conference on Wireless & Mobile Computing, Networking & Communication*, WIMOB '08. Washington, DC, USA: IEEE Computer Society. ISBN 978-0-7695-3393-3.

Liu, X., Sheth, A., Kaminsky, M., Papagiannaki, K., Seshan, S. and Steenkiste, P. (2009). Dirc: increasing indoor wireless capacity using directional antennas. *SIGCOMM Comput. Commun. Rev.* **39**, 171–182. ISSN 0146-4833.

López de Ipiña, D., Mendonça, P. R. S. and Hopper, A. (2002). Trip: A low-cost vision-based location system for ubiquitous computing. *Personal Ubiquitous Comput.* **6**, 206–219. ISSN 1617-4909.

MacKay, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. New York, NY, USA: Cambridge University Press. ISBN 0521642981.

Minsky, M. (1986). *The society of mind*. New York, NY, USA: Simon & Schuster, Inc. ISBN 0-671-60740-5.

Miu, A. K., Balakrishnan, H. and Koksal, C. E. (2005). Improving Loss Resilience with Multi-Radio Diversity in Wireless Networks. In *11th ACM MOBICOM Conference*. Cologne, Germany.

Mohan, A., Woo, G., Hiura, S., Smithwick, Q. and Raskar, R. (2009). Bokode: imperceptible visual tags for camera based interaction from a distance. *ACM Trans. Graph.* **28**.

Pentland, A. P. (1987). A new sense for depth of field. *IEEE Trans. Pattern Anal. Mach. Intell.* **9**, 523–531. ISSN 0162-8828.

Qingzhong Liu, Andrew H. Sung, M. Q. (2008). Video steganalysis based on the expanded markov and joint distribution on the transform domains detecting msu stegovideo , 671–674.

Raskar, R., Beardsley, P., van Baar, J., Wang, Y., Dietz, P., Lee, J., Leigh, D. and Willwacher, T. (2004). Rfig lamps: interacting with a self-describing world via photosensing wireless tags and projectors. *ACM Trans. Graph.* **23**, 406–415. ISSN 0730-0301.

Seitinger, S. (2010). Liberated pixels : alternative narratives for lighting future cities. *PhD Thesis* .

Shepard, T. J. (1996). A channel access scheme for large dense packet radio networks. In *In Proc. ACM SIGCOMM*.

Shrivastava, V., Ahmed, N., Rayanchu, S., Banerjee, S., Keshav, S., Papagiannaki, K. and Mishra, A. (2009). Centaur: realizing the full potential of centralized wlans through a hybrid data path. In *Proceedings of the 15th annual international conference on Mobile computing and networking*. New York, NY, USA: ACM. ISBN 978-1-60558-702-8.

Smulders, P. (2002). Exploiting the 60 ghz band for local wireless multimedia access: prospects and future directions. *Communications Magazine, IEEE* **40**, 140–147. ISSN 0163-6804.

Swartz, J. and Wang, Y. P. (1990). Fundamentals of bar code information theory. *Computer* **23**, 74–86. ISSN 0018-9162.

Tateno, K., Kitahara, I. and Ohta, Y. (2006). A nested marker for augmented reality. In *ACM SIGGRAPH 2006 Sketches*, SIGGRAPH '06. New York, NY, USA: ACM.

Traub, A. (1974). Light pen reading. United State Patent.

Underkoffler, J., Ullmer, B. and Ishii, H. (1999). Emancipated pixels: real-world graphics in the luminous room. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '99. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co. ISBN 0-201-48560-5.

Woo, G., Borovoy, R. and Lippman, A. (2010). Audio headlight: Directional light-based sound.

Woo, G., Lippman, A. and Raskar, R. (2012). Vrcodes: Unobtrusive and active visual codes for interaction by exploiting rolling shutter. *International Symposium for Mixed and Augmented Reality* .

Woo, G., Mohan, A., Raskar, R. and Katabi, D. (2011). Simple lcd transmitter camera receiver data link. *CSAIL Technical Report* .

Woo, G. R., Kheradpour, P., Shen, D. and Katabi, D. (2007). Beyond the bits: cooperative packet recovery using physical layer information. In *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, MobiCom '07. New York, NY, USA: ACM. ISBN 978-1-59593-681-3.