CIDER: Enabling Robustness-Power Tradeoffs on a Computational Eyeglass

Addison Mayberry Yamin Tun Pan Hu Duncan Smith-Freedman
Deepak Ganesan Benjamin M. Marlin Christopher Salthouse
University of Massachusetts, Amherst
Amherst, MA 01003
{amayberr, ytun, panhu, dganesan, marlin}@cs.umass.edu
salthouse@ecs.umass.edu

ABSTRACT

The human eye offers a fascinating window into an individual's health, cognitive attention, and decision making, but we lack the ability to continually measure these parameters in the natural environment. The challenges lie in: a) handling the complexity of continuous high-rate sensing from a camera and processing the image stream to estimate eye parameters, and b) dealing with the wide variability in illumination conditions in the natural environment. This paper explores the power–robustness tradeoffs inherent in the design of a wearable eye tracker, and proposes a novel staged architecture that enables graceful adaptation across the spectrum of real-world illumination. We propose CIDER, a system that operates in a highly optimized low-power mode under indoor settings by using a fast Search-Refine controller to track the eye, but detects when the environment switches to more challenging outdoor sunlight and switches models to operate robustly under this condition. Our design is holistic and tackles a) power consumption in digitizing pixels, estimating pupillary parameters, and illuminating the eye via near-infrared, b) error in estimating pupil center and pupil dilation, and c) model training procedures that involve zero effort from a user. We demonstrate that CIDER can estimate pupil center with error less than two pixels (0.6°), and pupil diameter with error of one pixel (0.22mm). Our end-to-end results show that we can operate at power levels of roughly 7mW at a 4Hz eye tracking rate, or roughly 32mW at rates upwards of 250Hz.

Categories and Subject Descriptors

C.3 [**Special-Purpose and Application-Based Systems**]: Real-time and embedded systems; J.3 [**Life and Medical Sciences**]: Health

Keywords

eye tracking; pupil; neural network; power robustness tradeoff; near-infrared; wearable

MobiCom'15, September 7–11, 2015, Paris, France.
© 2015 ACM. ISBN 978-1-4503-3619-2/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2789168.2790096.

1. INTRODUCTION

The human eye offers a fascinating window into an individual's personality traits, medical problems, brain abnormalities, behavioral conditions, cognitive attention, and decision making. These characteristics have made it the subject of decades of research by experts in cognition, ophthalmology, neuroscience, epidemiology, behavior, and psychiatry, who have not only enhanced our understanding of how the eye works, but also revealed new ways of diagnosing health concerns. For example, nearly every health condition that affects the brain causes substantial variations in eye movement patterns including ADHD [15], Autism [29], Williams syndrome [30], Schizophrenia [6], Parkinsons [1, 5, 7, 31], Alzheimers disease [26], Depression [8], and others. The eye also reveals a great deal about our current cognitive state, thereby providing surprising benefits for even healthy individuals. In the landmark book "Thinking Fast and Slow" [20], Nobel laureate Daniel Kahneman eloquently describes how an individual's System 2, which is our slow, deliberate, analytical and consciously effortful mode of reasoning, tires after too much cognitive effort, resulting in greater reliance on the unreliable but less effortful System 1, leading to poor decision making (also known as ego depletion). The effects are wide-ranging: judges are more likely to deny parole at the end of the day [12], clinicians have been found to prescribe unnecessary antibiotics [23], soldiers make poor decisions in operational environments [17], people buy more junk food [3], consume more alcohol and cigarettes [4, 10], and so on. How would we detect such cognitive "fatigue?" By looking at the eye and measuring pupil dilation.

Despite the enormous potential for advancing detection of health states and understanding of human decision making by measuring the eye, progress has been stymied by the lack of wearable eye trackers that are integrated into a regular pair of eyeglasses. But the design of a low-power wearable eye tracker is remarkably challenging from the computation, sensing, communication, and aesthetic design perspectives. A real-time eye tracker involves an eye-facing imager sampling at frame rates of tens of Hz (up to 100Hz to detect fine-grained eye movements or saccades) thereby generating megabits of data per second and making communication to a phone extremely power-hungry. As a reference point, the Google Glass lasts only a few hours when streaming from its outward facing camera, while running too hot for comfort [22]. Real-time computation on the eyeglass is also remarkably challenging, particularly given the volume of data and complexity of the image processing techniques. While our focus in this paper is on the computation and power aspects, aesthetic design presents an equally significant challenge since the sensors need to be embedded in an unobtrusive manner within an eyeglass frame.

Several initial efforts have been made to design low-power wearable eye trackers (e.g. iShadow [25], iGaze [39]), but many challenges remain. We tackle two in this work — power and robustness. Power consumption is a major avenue for improvement in eye trackers. The iGaze eye tracker consumes 1.5W, and a more optimized eye tracker, iShadow [25] has a power budget at around 70mW. These numbers are still much higher than typical wearables which only consume a few milliwatts of power, so there is a significant gap that we need to bridge to enable long-term operation of eye trackers on small wearable batteries.

Another major avenue for improvement is robustness. Eye trackers simply do not work outdoors given the variability in outdoor lighting conditions. More generally, achieving robust operation in environments with different illumination conditions is extraordinarily challenging and hasn't been achieved so far by either research prototypes or bulkier commercial products. Some eye trackers such as iShadow and iGaze rely on visible light, but clearly this fails under poorly illuminated conditions. Many commercial eye trackers use near-infrared illumination of the eye, but these do not operate outdoors since they are overwhelmed by ambient infrared light.

Our fundamental contribution is the design of a staged architecture for computational eyeglasses that can trade off between power and robustness to illumination conditions. The principle underlying our architecture is well-known to systems researchers — we optimize heavily for the common case but provide more power-hungry features to deal with the more difficult but uncommon scenarios that occur.

The common case is that a) we spend a substantial fraction of time indoors (homes, shopping malls, etc), and b) we spend 80% of the time fixating on points, during which time the eye moves only a small amount (referred to as microsaccades, which are typically less than 0.4°). We optimize for this regime by using a small amount of near-infrared illumination, a few tens of pixels sampled per estimate, and a few dozen instructions executed per pixel to estimate eye gaze and pupil dilation parameters. The power consumption for the common case is, therefore, only about 7mW — in contrast, iGaze [39] consumes 1.5W (three orders of magnitude difference), and iShadow [25] consumes 70mW (order of magnitude difference).

The question, however, is how to deal with much more antagonistic environments involving outdoor sunlight, shadows, and specular reflection of an onboard illumination source from off the cornea. To tackle this regime, we propose several adaptation methods, where we sacrifice energy-efficiency and switch between the sensing and computational blocks that we activate depending on how much noise and variability we observe. This includes sampling more pixels to get a better estimate of the pupil, performing more computation on the pixels to deal with noise, and using more complex models to estimate eye parameters. These stages of the pipeline consume 20mW, which is more than an order of magnitude higher than the typical power consumption, but these are triggered less than 10% of the time, therefore our overall efficiency does not increase significantly.

One of the interesting auxiliary benefits of our staged processing pipeline is that it can operate at very high frame rates during typical operation. Our optimized pipeline can operate at rates exceeding 100 fps, which is at the the high end of tethered remote eye trackers. This capability is particularly useful for detecting small fine-grained saccadic movements which happen while reading or when fatigued, providing further window into an individual's neural activities. Our algorithm, CIDER (CIrcle Detection of Edges with Reinforcement), is the first wearable eye tracker to achieve such high frame rates.

Our experiments show that

- CIDER can track pupil center with accuracy of roughly 1 pixel (0.3°) and pupil dilation with accuracy of approximately 1 pixel (0.22mm) in indoor lighting conditions.
- CIDER adjusts to indoor and outdoor illumination using an indoor-outdoor NIR detector in conjunction with different models and hardware settings. We show that the pupil center error increases only by a modest amount in outdoor settings (4 pixels or 1.2°).
- We operate end-to-end at a total power budget of 7.5mW when running at 4Hz, which is 10× less than previous state-of-art in this area [25]. Alternatively, we can achieve eye tracking rates of upwards of 250 frames/second by scaling power consumption up to 32mW.

2. DESIGN TRADEOFFS

Robust estimation of eye measures on a computational eyeglass presents a number of technical issues that cut across many aspects of design. In this section, we separate each component of the system and identify the key robustness–power tradeoffs that they present.

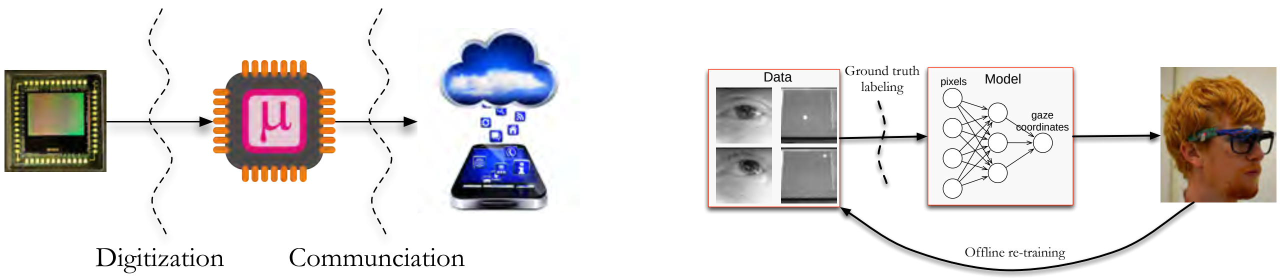
Sensing: Continuous operation of a camera is power-hungry. The energy cost of sensing primarily arises from digitization of pixels — while the analog electronics of a CMOS camera has very low power consumption (few milliwatts), digitizing hundreds of pixels per image at high frame rate (upwards of 30fps) consumes orders of magnitude more power, resulting in a power consumption of several tens to hundreds of milliwatts for typical cameras.

To reduce power consumption of such a camera, we would need to throttle the rate at which pixels are digitized. This can be done in one of several ways, including sub-sampling the image, reducing the resolution of the image, or acquiring fewer frames. However, reducing power consumption in this manner can be detrimental to dealing with variations — for example, in the presence of variable illumination or shadows, acquiring more pixels and more frames is better since it provides more contextual information and facilitates more robust de-noising methods.

Computation: Once pixels are acquired from the imagers, we need to process them to estimate eye parameters. The computational demands of continuous high-rate image processing (filtering, feature extraction, and detection) are significant, and require CPUs that have more resources and are more power-hungry than low-power MCU-class processors. One way of addressing this issue is simply to use an MCU with a higher clock rate that can meet the processing requirements. However, as with using a more advanced image sensor, such capability would come with a significant increase in power needs. This tradeoff makes higher-end processors generally infeasible for use with wearables.

If using a more powerful MCU is not an option, the most natural alternative is to trim the computational requirements by using a specialized model. For example, iShadow [25] uses a neural network rather than a typical image processing pipeline, which greatly reduces the amount of computation. However, this often comes at the cost of robustness since such a one-size-fits-all model is not always able to deal with variations in illumination conditions and still achieve high accuracy.

Communication: Can we solve some of the computation issues by offloading it to the phone and cloud? The challenge is dealing



(a) Architecture Challenge

(b) Modeling Challenge

Figure 1: Design challenges. (left) Two major challenges in system design are the power consumed for digitizing pixels from the camera and the power consumed for high-rate communication to a mobile phone. (right) Another major challenge is calibrating the system during online operation without requiring explicit interaction with the user.

with the power-efficiency of radios at high data rates — while radios like BLE, Zigbee, and even WiFi Direct are all relatively low power when used intermittently in a duty-cycled manner, continuous transfer at 30fps from a camera requires always-on radios and increases the power consumption to several tens or hundreds of milliwatts. However, offload has the benefit of being able to leverage vast computational resources to re-train models, adjust parameters, and deal with noise in sensor data, so when judiciously used, it can be effective.

Illumination: One of the key challenges in CIDER is how to deal with the peculiarities of indoor and outdoor lighting to enable robust estimation of eye parameters in both environments. Existing techniques rely either on natural illumination [25, 39] or artificial illumination using a near infrared light source (e.g. [34]). Both techniques have significant advantages and drawbacks. Natural illumination has power savings since typical near-infrared LEDs consume many tens or hundreds of milliwatts during continuous operation. However, this technique is highly sensitive to ambient lighting, and has poor performance when operating in low lighting conditions such as while driving a car at night. In contrast, illuminating the eye via near-infrared (NIR) can provide higher signal to noise ratio, but NIR illumination does not work outdoors since sunlight has significant infrared content, which overwhelms the illumination from the NIR LEDs. Thus, the conditions observed by the IR camera vary significantly, and require more robust methods to process images and extract eye parameters.

Calibrating to new users: An under-explored challenge in designing eye trackers is how to deal with calibration to a new user. The training process is cumbersome and typically performed each time a user wears the device. This was also the case in our prior work [25], where the neural network model is learnt via a calibration procedure requiring the user to look at dots on a computer monitor before wearing the glasses. One question that we ask is whether we can train these systems for a new user with zero-effort i.e. no user involvement whatsoever. Such capability can enable us to train the system without asking the user to alter their normal behavior. This also opens up the possibility to automatically retrain the system when needed, thereby allowing greater flexibility in dealing with robustness issues.

In the following section, we outline a design that provides power efficiency while also creating robustness to variability.

3. CIDER OVERVIEW

At a high level, CIDER uses two different approaches to trade off between robustness and power — the first is a two-stage rapid eye tracking controller, and the second is indoor-outdoor model switching to deal with different illumination conditions and noise. Under typical indoor illumination, CIDER relies on a "Search–Refine" two-stage controller and a small amount of NIR illumination of the eye to estimate eye parameters in a fast, efficient, and accurate manner.

The search stage operates with no prior knowledge of pupil location, and uses a neural network to obtain an estimate of pupil center and size from a sub-sampling of pixels. The refine stage takes an estimate from the search stage, and uses a very fast and accurate procedure to locate and track the eye. When the refine stage loses track of the pupil due to specular reflections or other unpredictable variations, it reverts to the search stage to get another estimate of the eye location. The two stages differ in terms of the amount of sensing required (i.e. number of pixels acquired per frame from the imager) as well as the amount of computation performed to extract eye parameters from the pixels.

In outdoor settings, CIDER turns off NIR illumination (since there is too much ambient infrared), switches to different camera hardware parameters to deal with outdoor lighting, and switches the neural network model to one that is trained for outdoor conditions. We detect indoor vs outdoor conditions using an NIR photodiode that tracks the level of ambient infrared light. In the outdoor mode, CIDER does not use the refine stage and only relies on the neural network to track the eye. We generally find that there is significant variability, which makes it difficult for a more optimized model to operate in a reliable manner.

Overall our pipeline achieves a graceful tradeoff between robustness and power — under typical indoor illumination, CIDER spends most of its time in the fastest and most efficient stage while occasionally using the neural network to provide estimates. In outdoor illumination, CIDER spends all of its time in the slower but more robust neural network stage.

CIDER also addresses robustness issues by designing a model training pipeline that operates with no input from the user. Whenever the eyeglass is fully charged or has good connectivity, a block of images can be communicated to the phone, and a new model trained offline. The enabler is an offline image processing pipeline that generates accurate labels of pupil center and dilation from noisy image data (collected either indoors or outdoors), which makes it possible to learn the neural network models with zero effort from the user.

Finally, CIDER also improves on prior work in that we can simultaneously estimate pupil center and pupil dilation, thereby providing two key measures of the eye in real-time. In principle, CIDER's estimate of pupil center can be used to determine the

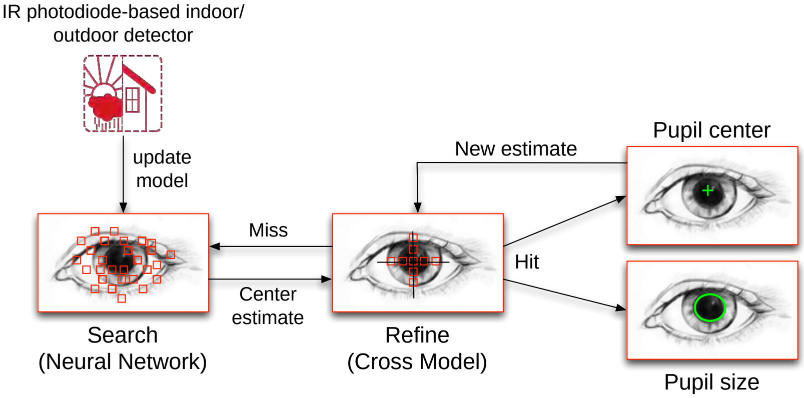


Figure 2: The CIDER pipeline: a) search stage using a neural network to get an initial estimate of pupil location, b) refine stage to zone in on exact pupil center and perform rapid tracking unless the pupil is missed, and c) NIR-photodiode-based detection of indoor/outdoor mode to update neural network model.

gaze direction of the user by leveraging geometric mapping methods from the inward-facing image plane to outward-facing image plane as done in iGaze [39].

4. CIDER DESIGN

In this section, we discuss the details of how CIDER works. We first discuss operation in the indoor case, for which CIDER is highly optimized, and then discuss how we handle the more variable outdoor case.

4.1 Search–Refine Controller

CIDER achieves high speed, high accuracy, and low power by using a rapid switching loop between the search and the refine stage.

Search stage – neural network model: The search stage is an artificial neural network (ANN) prediction model that operates over a subsampled set of pixels, based on the design outlined in [25]. We provide a high-level overview for completeness, and refer the interested reader to [25] for a more thorough overview. The process involves setting up the prediction problem as a neural network where the inputs are the pixel values obtained from the imager, and the output is a predicted (x,y) coordinate pair. The problem is set up as a bi-objective optimization, where one objective is to minimize the set of pixels that need to be sampled to reduce power consumption, and the second objective is to minimize loss in pupil center prediction accuracy. This is achieved using a neural network learning algorithm together with a regularizer that penalizes models that select more pixels. The optimization problem has two terms: a) an error term that captures how well the algorithm predicts gaze coordinates, and b) a penalty term that increases with the number of pixels selected. To promote sparsity i.e. to select a small active pixel set to sample, the algorithm uses a sparsity-inducing ℓ_1 regularization function, which minimizes the number of pixels sampled. The optimization function is solved offline using labeled training data, and the parameters are hard-coded into the eyeglass platform for real-time prediction of the pupil center.

The only major change to the ANN model in [25] for this work is the output target of the model. The goal of the original iShadow work was to predict the gaze location of the subject based on the orientation of their eye in the image. For this work, we instead want to identify the pupil within the eye image and report relevant parameters - center xy-coordinate, and radius of the shape. Since the original ANN model has several desirable properties, including input feature reduction and good accuracy, we chose to use the same model to estimate these parameters. The input to the neural network is still the subsampled pixels, however it is now trained to minimize error over three target values- center x, center y, and radius.

Refine stage – cross-search model: The refine stage is a cross-search model (shown in Figure 3) that leverages the estimate from the neural network to track the center of the pupil and pupil size with minimal sampling overhead. The first step of the cross model is to sample one row and one column of pixels at the estimated location provided by the neural network. There is substantial fixed pattern noise from our camera, so we need to first remove this noise (described in more detail in §5.2). Once the fixed pattern noise has been removed, the pixel values are median filtered and segmented into several regions — Sclera, Iris, Pupil, and Sclera. The segmentation process convolves the pixels with a box filter to detect edges. Since we run the same operations on a column and a row of pixels, we have two chords corresponding to the pupil along the vertical and horizontal axes. We assume that the pupil is a circle for simplicity of computation, and then it is straightforward to compute the center of the circle from the mid-point of the two chords.

Rapid switching between stages: Switching between the two stages works as follows. The ANN executes once, and then hands control over to the cross model to see if it can handle further refinements without requiring the ANN. The cross model is extremely fast, and takes a fraction of the time of the ANN, so it can execute quickly and check if further tracking of the eye can be handled entirely by using the cross model. To determine when to switch back, the cross model performs an internal validity check to see if the results it has obtained are consistent. Specifically, the cross model checks if the two chords (horizontal and vertical) result in a consistent solution. If there is too much error, it falls back to the ANN model. Since the cross model is fast, any misses are quickly handed by the ANN within a short time window, so the time window during which we do not have an estimate of the eye parameters is tiny.

The speed at which the cross model operates means that it is not only refining the estimate from the ANN, but is also tracking the eye. The cross model can operate at frame rates of several hundreds of Hz, which is much faster than the speed at which larger saccades occur. As a result, even if the eyeball is moving, the cross model makes small adjustments each frame, thereby tracking the eye. The

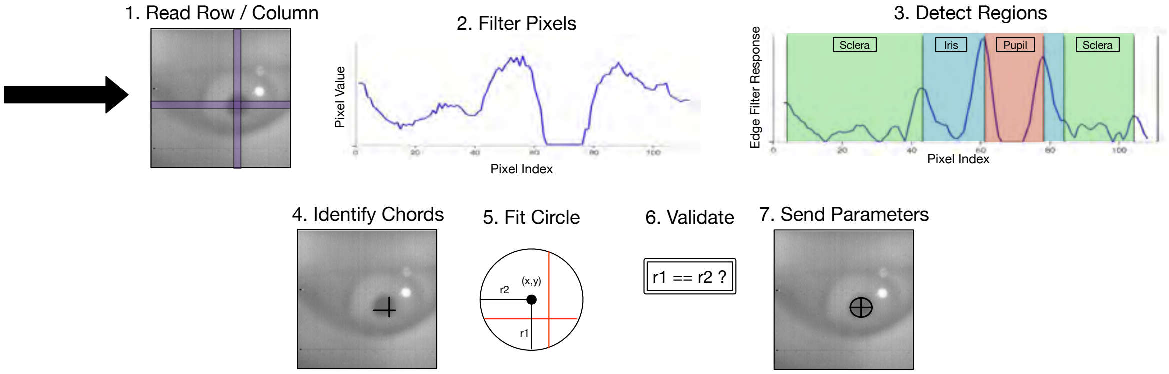


Figure 3: The CIDER cross-search model: 1) a row and column of pixels near the pupil center are sampled, 2) values are median filtered, 3) regions of the eye are detected using edge detection, 4) chords within the pupil are detected, 5) circle is fitted to the chords, 6) consistency check is performed to detect hit or miss, 7) the pupil center and size are estimated.

only occasions when the cross model fails is when there are blinks, specular reflections, shadows, or other artifacts, in which case it switches to the neural network.

Optimizing NIR power consumption: One of the key enablers of the rapid switching controller described above is NIR-based illumination of the eye. Even though indoor lighting can vary significantly, there is virtually zero infrared content in the light emitted by lightbulbs (FL, CFL, LED, etc). This gives us an opportunity to use a small NIR light source to illuminate the eye, and use an NIR-pass filter on the camera to make sure that only the NIR illuminated content is captured. This gives us very controlled lighting conditions despite potential changes in the indoor lighting level.

One issue that we face is that typical NIR LEDs have high power consumption (the one we use consumes 180mW at peak power) — this is small compared to the overall power budget of typical eye trackers that consume watts of power, but it is exorbitant when we are attempting to operate at a few milliwatts of power. Thus, one question that we faced is how to reduce this power consumption.

There are two ways to reduce NIR power consumption — one is to duty-cycle the NIR photodiode, and the other is to reduce the operating voltage of the LED. NIR duty-cycling can be done between frames, therefore the reduction in number of pixels acquired using the cross-search model plays a significant role in the duty-cycling benefits. Reducing the operating voltage of the LED is effective as well — we found that NIR LEDs operate down to about 1.15V, and while reducing the voltage results in increased noise, there is sufficient signal for the neural network to learn a robust mapping. A small downside of this approach is that we lose some efficiency since NIR LEDs are typically most efficient at the high-end of their voltage range, however, this is balanced by the substantial power benefits that can be obtained. The combination of duty-cycling and low voltage operation reduces the NIR power budget by roughly two orders of magnitude, from 180mW to less than a milliwatt.

We note that our use of NIR is very different from methods used by commercial eye trackers. Typical eye trackers use multiple narrow NIR beams, and process the image data to locate these NIR beams, before combining this information with an eye model. However, this process requires several NIR LEDs, more complex geometric methods for estimating eye parameters, and does not generalize to outdoor settings. We operate with just two NIR LEDs, very simple models, and our computational methods continue to work in outdoor settings albeit at higher cost (as we describe below).

4.2 Indoor-Outdoor Model Switching

A second switching mechanism in CIDER is between indoor and outdoor modes of operation. Indoor and outdoor operation are very different for two reasons: a) NIR illumination is useful in indoor settings since it provides a controlled environment for eye tracking, but not for outdoor settings where there is too much ambient IR, and b) camera gain parameters need to be adjusted for outdoor settings and this requires modification of the neural network parameters.

Our idea is to track ambient IR conditions using a separate infrared photodiode that is built into our eyeglass (facing outward rather than inward). We use the IR levels to switch between different camera parameters (gain settings), as well as different neural networks trained for different conditions. This mechanism can be viewed as a camera gain control mechanism that is tightly integrated with the eye parameter estimation pipeline. Typical cameras use automated gain control (AGC) to deal with lighting variations, but a downside is that the pixel values are continually changing depending on the gain parameters. This makes it difficult to run a specialized computational function such as neural network, particularly when subsampling pixels to operate with as few pixels as possible. Rather than continuous adjustments, our method can be viewed as a discrete approach, where we have two models corresponding to specific ambient IR settings, and we switch both the hardware parameters of the camera and the model based on the observed settings.

The switching process itself is extremely simple from the perspective of the firmware, requiring only a few MCU instructions to sample the photodiode at regular intervals. Since lighting conditions can be reasonably expected not to change with high frequency (i.e. more than once every few seconds), this sampling can be done as infrequently as once a second or less. If the MCU detects a significant change in lighting conditions, altering the camera gain parameters also only requires a small handful of instruction cycles. Thus, the overall power and time cost of the switching process is negligible.

This indoor–outdoor model switching can be viewed as a form of power vs robustness adaptation. In indoor settings, we consume more power due to NIR illumination of the eye, but save much more power by reducing the number of pixels sampled and associated computation. In outdoor settings, we shut off the NIR LED and opportunistically leverage ambient IR to save power. We rely on

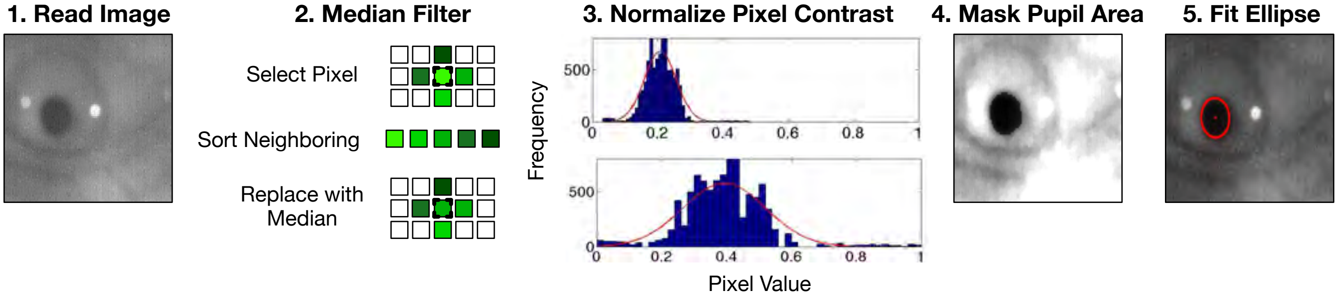


Figure 4: Labeling pipeline with select stages shown

a more complex neural network model, which implies more pixels and more computation, but gain robustness in the process.

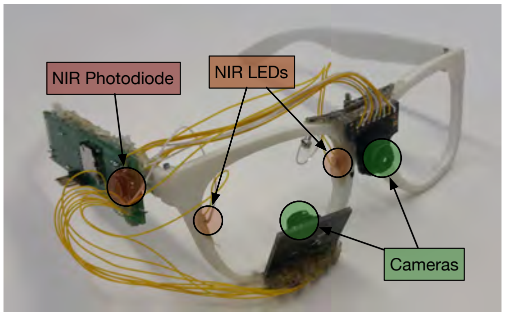


Figure 5: Eyeglass platform

4.3 Zero-Effort Model Training

One remaining question in our system is how to train the models for each user. Ideally, we would want such training to be completely automated to minimize burden on the user. This problem often goes unaddressed in existing approaches, most of which are designed to require a period of explicit user participation in order to generate model training data (e.g. [25]). Zero-effort training could greatly increase the likelihood of broader applicability of our system.

The core question in training is how to develop a robust offline method for generating labels from noisy images collected by the camera. The offline procedure that we use for training the neural network is shown in Figure 4. The raw image is processed through a median filtering stage, from which the region corresponding to the eye is extracted. This region is further contrast-adjusted and filtered, and segmented to extract dark regions in the image. In good conditions, only the pupil shows up as a dark region, but we faced two additional challenges.

First, we see specular reflection of the NIR LED from the eye, and when the specular reflection overlaps with the pupil, the dark region can look like a disk, or like a disk with a bite on the side. To address this, we fill holes that we might observe in the segmented shape using standard image-fill techniques that identify distinctive regions of color within a larger area (the pupil) and adjust them using the surrounding pixels. Since the specular reflection is small relative to the size of the pupil, these simple techniques work extremely well in practice. Second, in outdoor conditions, we often see shadows caused by the sun's position relative to the eyeglass frame, and these shadow regions are also picked up by the segmentation block. To isolate the pupil, we look for the roundest segment to detect the pupil.

Given the target pupil location from the above image processing pipeline, we then divide the data into train and test sets and learn the neural network parameters. The new model is then uploaded to the glasses.

5. CIDER SYSTEM

In this section, we describe the main components of our eyeglass system and our implementation of CIDER, and the improvements over other prototypes that have been designed in past work.

5.1 CIDER Platform

Our eyeglass platform has a low-power camera that is mounted in the lower part of the frame facing the eye, as well as an NIR illuminator, as shown in Figure 5. We also have another outward-facing camera, as well as other sensors, but we do not discuss them in detail since they are not pertinent to the methods in this paper. We use the standard optics on the image sensors, which give a 36° field of view. The eye-facing camera has an NIR filter to capture the illuminated eye.

Our platform is similar to iShadow [25], with three key differences. First, we mount the eye-facing camera at the bottom of the frame rather than the top as in iShadow. The difference in how we mount the camera has implications on the tracking accuracy, as well as robustness to different conditions. One major consideration in this decision is that people naturally tend to look down with their eyes much more frequently than they look up. If the camera is mounted in the lower position we observed that, when looking down, the user's pupil is pointed nearly directly at the camera, making detection easier. In addition, when a person looks down for any reason, the upper eyelid naturally lowers a little. This can obscure the eye features when viewed from a higher vantage point (we have observed this in practice). Note that the lower eyelid does not noticeably raise when looking up, so the eye does not become obscured even when viewed from a lower angle. Thus, we concluded that mounting the camera on the lower portion of the frame was a strict improvement over the upper portion.

The second major change from iShadow is that we illuminate the eye with a pair of NIR LEDs shown in Figure 5. The placement point for the LEDs was chosen after careful characterization of what location would provide best illumination while minimizing issues due to specular reflections. Similarly, the choice of NIR LED was made after a rigorous measurement study involving more than a dozen types of NIR LEDs to understand their power-illumination profiles. The third and final difference is that we have an NIR photodiode that detects the level of ambient NIR and allows us to detect indoor conditions vs outdoor conditions.

Our platform is very different from other prototypes such as iGaze [39]. The iGaze device consumes more than a watt, and uses a Raspberry Pi attached to a glasses frame with cameras and sensors. The difference in power between our prototype and iGaze is between two and three orders of magnitude, and virtually every component in our system from algorithm to hardware components is optimized to achieve the power reduction.

Microcontroller: The iShadow platform uses an MCU with an ARM Cortex M3 core [11]. Our implementation of the platform uses an STM32L151 microcontroller, which is manufactured by STMicro Corporation [32] and is an implementation of the Cortex M3 standard. The STM32L1 family emphasizes low power consumption and includes a wide variety of processor sleep modes that are useful for reducing power draw by inserting timed sleep cycles where possible. It also includes several built-in peripherals for handling common communication protocols such as USB, reducing the firmware development burden significantly.

Image sensors: Our hardware framework is built around the Stonyman Vision Chip produced by Centeye, Inc.[9]

The Stonyman camera has a resolution of 112x112 pixels, each of which is characterized by a logarithmic voltage response to lighting conditions. These pixels have a high dynamic range, and more importantly, allow a random-access interface which the Stonyman provides via a register-based control scheme. Besides the extremely low power consumption compared to off-the-shelf cameras (3mW), the main advantage of the imager is that it allows for random access to individual pixel values. This feature allows us to sub-select specific pixels that we need for CIDER, and results in significant reduction in the digitization cost.

Another important characteristic of the Stonyman imager that is the fact that the camera gain parameters are controlled programmatically rather than automatically (i.e. there is no automatic gain control like in other cameras). While this could be viewed as disadvantage, we find the ability to control gain to be beneficial for us in that we can adjust gain parameters and the model parameters in tandem when triggered by the NIR photodiode.

Finally, the Stonyman camera also provides features such as a sleep mode, during which the pixel acquisition circuitry can be powered down. The low-power state has power consumption less than half a microwatt since only a few control registers are running to maintain camera state.

5.2 Handling Camera Noise

We faced several implementation challenges, particularly in how we deal with the camera noise and identify camera gain parameters, which we briefly list in this section

Fixed pattern noise: One of the biggest challenges that we face in designing CIDER is dealing with low-level noise and the way in which the noise is intertwined with the hardware circuitry of the Stonyman camera. For example, one issue was that the fixed pattern noise of the pixels looked different when we were reading pixels along a horizontal line vs along a vertical line for the cross model. We identified that this issue was related to the way in which the pixel readout circuitry is designed on the Stonyman camera. The pixels along each row are daisy-chained to a single readout circuit, therefore, once we started reading out pixels from the beginning of the row, all the pixels along that row were activated. This resulted in varying noise along the row since the pixels at the end of the row had higher magnitude signal and noise. This issue does not occur when sampling pixels along a column since each only one pixel is read from each row. To address this problem, we learned a different fixed pattern noise mask per column and per row through offline calibration, and we subtracted the mask from the measured values to obtain the actual signal.

Gain settings: Another issue that we faced is that the Stonyman camera provides four gain parameters, each of which has roughly 20 settings. This results in a huge search space (20^4) to determine which is the best parameter setting for indoor and outdoor conditions. The search process is largely mechanical but time-consuming since an image has to be captured for each setting, and the values checked to see if it is appropriate. The setting is important, however, since indoor gain values simply do not work outdoors and result in the pixel values saturating. While the settings we identified works well under different outdoor conditions, it may be possible to perform further fine-tuning to specific conditions and get better results than those which we have reported.

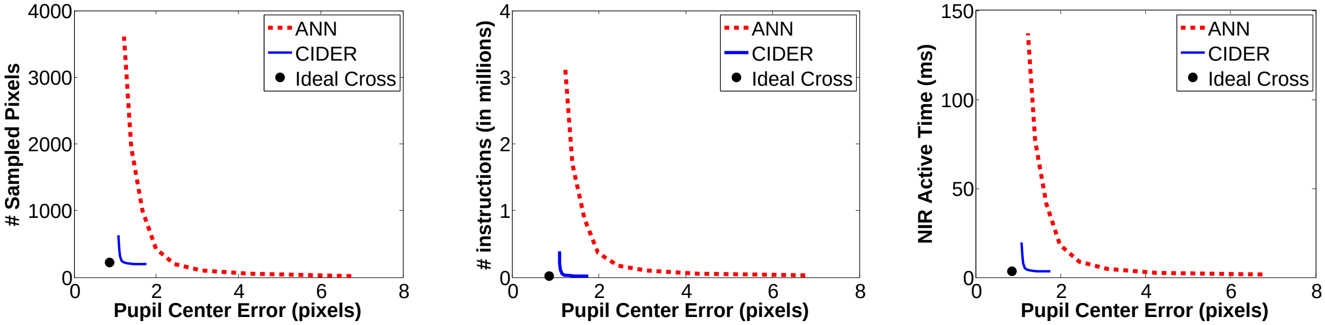
6. EVALUATION

We first describe the datasets that we have collected and the evaluation metrics we use and then describe our experimental evaluation.

6.1 Datasets and Ground Truth Labeling

We evaluate CIDER with four datasets that correspond to different environments and dynamics. All data collection experiments involving human subjects received approval from an institutional review board.

- **Indoor-Stable data (fixed pupil, fixed illumination)** We collected data from 16 users, 12 male and 4 female. Each subject performed a video calibration routine where they looked at a high contrast dot moving on a computer monitor for several minutes. This gives us good coverage of eye positions, and allows us to train a good model as well as determine robustness to position of the eye. The illumination was held constant during this period, and subjects' pupils were roughly 5–7 pixels wide in this illumination. We generated approximately 2500 eye images for each user. The subjects involved in the data collection represent a range of ethnic groups with different eye shapes and iris colorations. We refer to this dataset as `indoor-stable`. All subjects in the other datasets were also in the `indoor-stable` dataset.
- **Indoor-Variable data (variable pupil, variable illumination)** We collected this data for 14 users, 10 male and 4 female. We varied the lighting conditions in five discrete levels using a combination of different sets of ceiling lights as well as target spotlights. The subjects pupils dilated between 5–15 pixels during this period, which gives us a fairly large range of pupil dilations that is representative of what one would observe in real-world settings. The screen brightness was kept low enough to not impact dilation much. The above computer-based calibration routine was executed for each setting to obtain data. We refer to this dataset as `indoor-variable`.
- **Outdoor data (uncontrolled illumination)** We collected this data for three users, all male, under outdoor settings. The conditions were generally bright. We obtained several minutes of data from each participant generally gazing at the outside scene under different orientations.
- **Indoor-Outdoor switching data** Our indoor-outdoor data was collected for one user, who walked between indoor and outdoor conditions repeatedly for four iterations, while spending roughly a minute in each environment. This dynamic setting helps us evaluate whether the NIR photodiode-based



(a) Sensing pixels vs center accuracy

(b) Computation cycles vs center accuracy

(c) NIR time vs center accuracy

Figure 6: Cost vs Accuracy

model switching algorithm works effectively with real scene changes.

Ground truth labeling: All data collected above was labeled for pupil center and pupil size using the process described in §4.3. Once labeled, we trained the neural network to identify the pupil center and radius of the best-fit circle approximating the pupil shape using a standard five-fold cross-validation scheme. We averaged the test set error over the five folds to get an average score. Pupil center error is computed as the L2 (Euclidean) distance between the estimate and the label, pupil size error as the difference between the estimated radius and the label. The errors were averaged over all subjects per model size to get a final set of error estimation accuracies over a range of neural network model sizes.

6.2 Evaluation Metrics

We use several performance metrics to evaluate our system. Since our power numbers are specific to our platform, we provide both a more general metric that could generalize to any platform as well as a more specific metric for our platform given the hardware components that we choose.

Cost metrics: We use three performance metrics to evaluate our system.

- **Sensing cost** We measure sensing cost in two ways: a) the number of pixels sub-sampled from the imager, and b) the power consumed for sampling the pixels for the Stonyman camera. The former measure generalizes to any camera that can be sub-sampled, while the latter measure provides a real measurement that includes constant overheads of switching the camera from sleep to active mode, sampling the pixels, and switching back to sleep mode.
- **Computation cost** We measure computation cost in two ways as well: a) the number of instructions that need to be executed for each model, and b) the power consumed for executing instructions.
- **NIR cost** Similar to the above two metrics, we measure the NIR cost in terms of active time (i.e. time for which the NIR is turned on), as well as the power consumed for our NIR LED with duty-cycling and voltage optimizations described in §2.

Accuracy metrics: We measure accuracy of estimating eye parameters using two metrics

- **Pupil center** The accuracy in measuring pupil center is measured in pixels in the image captured by the eye-facing imager. This measure gives us an idea of how far we are from the best-case performance given the sampling granularity of the imager. From an application perspective, the key metric of interest is the degree error in estimating gaze. We estimate that each pixel corresponds to roughly 0.3°, so this gives us a mapping from pupil center error measured in pixels to gaze error. Commercial (tethered) gaze trackers achieve errors of roughly 0.5° [35], which is slightly less than two pixel error on our system.
- **Pupil radius** Similar to pupil center, we measure pupil dilation in pixels. Each pixel in in the camera's visual field corresponds to roughly 0.22mm when measured on the pupil, which is similar to the resolution of high-end gaze trackers.

6.3 CIDER Performance

As we have outlined, the search and refine stages of CIDER are intended to maximize estimation accuracy and power efficiency over a range of environmental parameters. Our first set of results evaluates the performance of CIDER by comparing it against the two stages (search and refine) independently. We use the `indoor-stable` data in this evaluation, which gives us an understanding of best case performance under limited dynamics. We compare the following schemes in this evaluation:

1. **Neural network** The neural network model is learnt as described in §4.1 — we vary λ (regularization parameter) to learn various models that have different tradeoffs between accuracy and pixels (which translates to power). This gives us a pareto optimal set of solutions i.e. a set of solutions that shows the tradeoff between the two objectives.

2. **Idealized cross** The idealized cross method is initialized by the pupil center estimated by our offline algorithm. The cross model then estimates the pupil center and pupil size, and we compare the accuracy against ground truth. Clearly, this is an idealized scenario where the cross model should perform very well, but it still helps us understand how well the edge detection and parameter estimation methods work in the best case.

3. **CIDER** The CIDER method is the fast switching technique. Since the CIDER pipeline involves switching between the ANN and cross model, we expect performance to be in-between the above models.

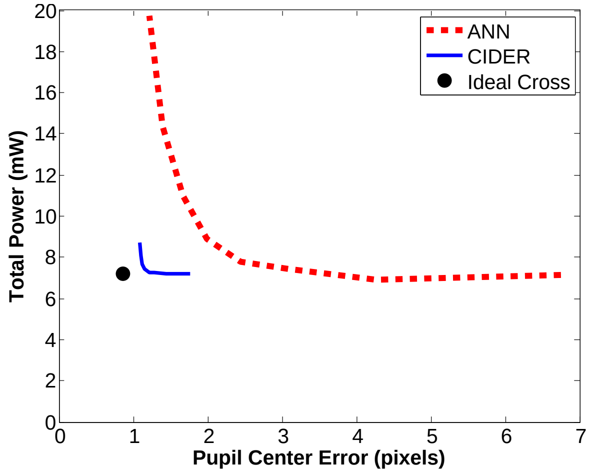


Figure 7: Aggregate power vs accuracy

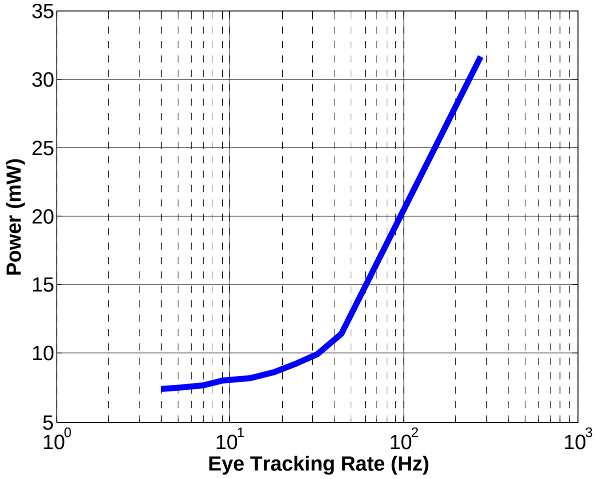


Figure 8: Aggregate power vs eye tracking rate (log scale)

Sensing, computation and NIR cost: Figure 6a shows the curve of sensing cost (in number of pixels sampled) against pupil center estimation error (in number of pixels). The curve is obtained by tuning the neural network regularization parameter, which allows for the generation of a number of network sizes with varying power needs and corresponding accuracy. The result clearly shows that there is a significant gap between any pareto optimal solution that can be obtained for the neural network vs the solution provided by the idealized cross model. CIDER operates between the two but closer to the idealized cross model. This can be explained by the fact that the neural network is triggered only about 10-15% of the time whereas the cross model operates the remaining 85-90% of the time.

The performance difference in terms of computation cost is substantial as well, in fact, even more than in the case of sensing (Figure 6b). The neural network computation is much more involved than the cross model, so there are significantly more operations per pixel. In addition, since the cross model requires fewer pixels, the number of times the computation needs to be performed is also much lower. Thus, the number of instructions that need to be computed for the cross model is orders of magnitude lower than for the neural network.

Finally, the time spent with the NIR LED on is also substantially lower for the idealized cross and CIDER models (Figure 6c). Since the cross model needs very little time to sense, the NIR LED needs to be turned on for a minuscule amount of time for each frame.

Energy savings: We now look at how the benefits in sensing, computation and NIR translate into energy savings on our platform. We measure the average power over a 10 second window of operation using a DAQ running at a 10 kHz sampling rate. To measure power consumption for all three models, we fix the pixel capture + predict rate of the system to 4 Hz by inserting MCU sleep periods as needed. The 4Hz rate is chosen to enable us to measure a sufficiently large range of neural network model sizes to plot the pareto optimal graph.

Figure 7 shows the aggregate power consumption of CIDER and compares against the two other baselines. We see similar trends as we saw earlier in that CIDER operates in between the idealized cross and ANN model with roughly a 3× reduction (compared to neural network models that have low error). The overall power budget for CIDER is roughly 7mW, which is a huge improvement over state-of-art (order of magnitude less power consumption than [25]), and a substantial achievement considering that the system is operating a camera, estimation algorithm, and NIR LED.

One curious feature of the graph is that the baseline for all schemes is shifted by about 6mW. The baseline shift corresponds to constant overheads incurred by our platform, and for configuring various parameters for the camera upon wakeup and shutdown. We suspect that there are various sources of power leakage that contribute significantly to the baseline, but we have not yet been able to fully debug these issues. Looking forward, we expect that some of this constant overhead can be eliminated with a more optimized computational block such as an FPGA rather than a general-purpose MCU.

Power vs tracking rate: Another benefit of CIDER is that it can achieve high tracking rates. We plot the power vs pupil tracking rate in Figure 8, which shows the total system power consumed as the tracking rate is varied. To generate this graph, we used the same model as was used for the measurements in Table 1, and inserted sleep periods of variable length between each single execution of the CIDER pipeline. The measurements were again taken using a DAQ sampling at 10kHz.

Component	**Power (4 Hz)**	**Power (278 Hz)**
Camera	7.30 μW	30.8 μW
MCU (digitization)	2.67 mW	11.3 mW
MCU (computation)	4.79 mW	20.2 mW
NIR	8.24 μW	34.8 μW
Overall	7.48 mW	31.6 mW

Table 1: CIDER power breakdown

Table 1 shows a finer-grained breakdown of the power vs tracking rate for each component of CIDER (with a moderately large neural network chosen to use 10% of the pixels). We give two power measurements - one taken at the maximum eye tracking rate possible for this model size, namely, 278 Hz, and one taken at the 4Hz rate used for the rest of the evaluation results. There are several useful observations that can be made from this result. Interestingly, the camera and NIR consume virtually no power compared to other components since they are turned on for a very tiny amount of time. The acquisition consumes a significant amount of power — this is because digitization of the analog signal output from the camera is expensive. One of the major improvements that CIDER provides is reduction of the digitization overhead. The MCU computation is also expensive, however some of this cost could be reduced by using a more optimized computation block such as an FPGA.

Estimation accuracy: The above discussions emphasize power consumption, but it is instructive to look at the absolute accuracies that can be achieved by CIDER. The above results show that CIDER achieves pupil center estimation accuracy within 1.2 pixels. The neural network method cannot achieve such accurate estimation even when consuming considerably more power and resources.

This result may seem surprising at first, since it is natural to expect a more power-hungry technique to have a corresponding increase in performance. The main reason is that the NIR-illuminated eye (indoors) presents very strong edges that are easier to accurately identify using edge detection techniques (the cross model) than using a neural network. So, the accuracies tend to be higher for CIDER even though the power consumption is much lower. This is not the case in the outdoor environment, however, hence the need for the indoor-outdoor switching model. Thus, not only are we able to achieve substantially reduced power consumption, we also do so while simultaneously improving accuracy to within a small amount of the lower bound of what is achievable with our camera.

Model	**Pupil Size Error (pixels)**
Neural Network	0.50
CIDER	0.85

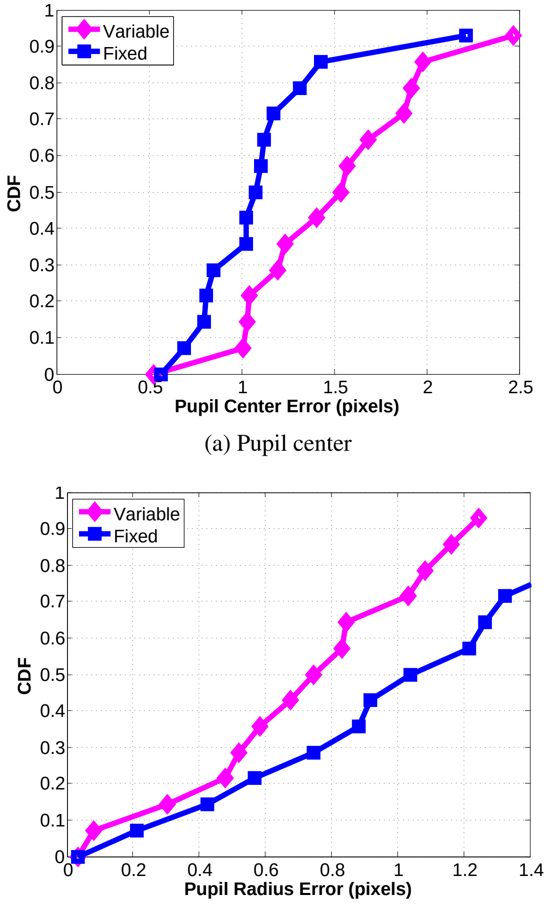
Table 2: Pupil radius estimation accuracy of CIDER

Table 2 shows the results for pupil size estimation when using only the neural network and when using CIDER. We do not show the entire power-accuracy profile for pupil size since we find that even the smaller ANN models perform well in estimating the pupil size, and there is not much difference in using a larger model. So, we present only the mean performance across all model sizes. We see that the pupil size estimation error is typically less than one pixel, which suggests that both stages can do an excellent job in estimating pupil size. Indeed, we find that the error for CIDER may be over-estimated since we often see that the cross model's estimates are closer to the real value than even ground truth labels.

6.4 CIDER Under Variable Conditions

Having evaluated CIDER under relatively stable conditions, we turn to situations that have more variability. Specifically, we look at three cases: a) variability in the pupil dilation of the user, b) an outdoor setting with variable illumination, and c) the user moving from an indoor to an outdoor setting.

Variable pupil dilation: The results in §6.3 were taken under fixed pupil dilation, so one question is whether the results are robust to varying pupil sizes. Figure 9 compares the pupil center and pupil size estimation errors of CIDER for the 14 users in `indoor-variable`, all of whom are also in the `indoor-stable` dataset. Figure 9a compares the pupil center prediction results for fixed and variable illumination conditions, each as an error CDF, and Figure 9b gives the same comparison for size prediction. The center prediction accuracy under varying pixel sizes is marginally worse than the accuracy under fixed pixel sizes, but the difference is not significant. For the size prediction task, CIDER actually generated slightly better estimates on the variable pupil size dataset. This seems counter-intuitive at first, however, in the `indoor-variable` dataset, the pupil size is generally larger than in `indoor-stable`, as the lighting conditions were darker for most of the experiment.



(a) Pupil center

(b) Pupil size

Figure 9: Performance comparison - fixed and variable pupil size

This makes accurate detection of the size slightly easier for both the ANN and the cross model. Overall, we see that performance of CIDER is robust to variation in pupil size.

Outdoor dataset: The outdoor scenario represents another high variability situation for CIDER. The cross model does not work in this situation, so the system relies primarily on the neural network that is trained for outdoor settings. We find that the accuracy with CIDER under outdoor settings is roughly 4 pixels (for moderately sized ANNs). The results are worse than accuracy in indoor settings, but not far off. In fact, the accuracy that we obtain in outdoor settings is better than the results that were obtained in [25] under indoor settings. One of the main reasons for the performance difference is the vastly improved labeling pipeline that we have developed, which allows us to label noisy data quite well.

We get about 1 pixel pupil dilation error, but we find that that this is an over-estimate of the real error for reasons described above. There is about a 1 pixel offset between the radius estimated by the offline labeling algorithm (which performs filtering), and by the cross model. For transparency, we have reported the error as observed, but we think the error is about one pixel smaller than that reported.

Indoor-Outdoor switching: We now look at a situation where a user is moving between an indoor and outdoor environment, and show how well our IR photodiode-based model switching performs. Figure 10 shows the error distribution during the indoor segments

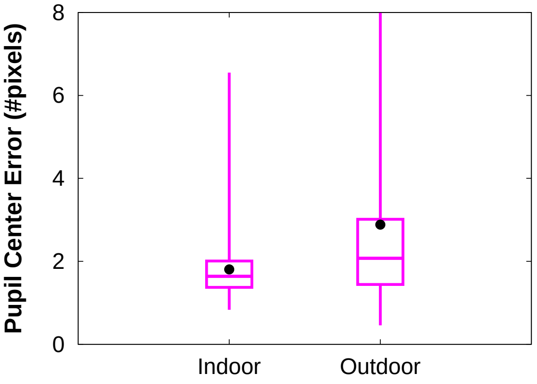


Figure 10: Indoor-Outdoor switching

vs outdoor segments. This is shown as a box plot, where the three lines in the box corresponds to the quartiles (25 percentile, median, and 75 percentile), the whiskers correspond to the max and min, and the dot is the mean. We truncate the max error whisker for the outdoor case since there are some cases where CIDER returns more than 10 pixel error.

We also verified from the traces that the NIR-based switching works effectively, and switches models between the indoor and outdoor modes whenever the user changes environments. As observed in §4.2, the instruction cycle and power cost of the detection and switching process itself is negligible. The error distribution of the predictions is higher for the outdoor case, but it is still relatively low with a mean of less than three pixels. The error when indoors is lower with a mean of less than two pixels.

6.5 CIDER High-Speed Eye Tracking

One of the major benefits of CIDER is the eye tracking speeds that it can achieve. High-speed eye tracking is useful for understanding fast saccadic movements of the eye, which is one of the neural mechanisms for maintaining visibility. For example, one of the interesting use-cases for measuring micro saccades is as a diagnostic test for ADHD [16], and there are other applications of such measurements [24].

However, high speed eye tracking is also very challenging on a wearable device. Commercial high-speed eye trackers achieve several hundred hertz tracking rates (e.g. the Eyelink high-speed eye tracker samples at 500Hz, and the ASL H7-HS can sample at rates up to 360Hz). Of course, these eye trackers are also bulky, tethered for power and connected to a computer for data storage and processing. One interesting question is, how fast CIDER can operate if it is not duty-cycled and is allowed to perform pupil estimation as fast as possible?

To evaluate the maximum speed achievable by CIDER, we run it continuously on our eyeglass without duty-cycling. We measure the rate at which it generates pupil center measures, and find that CIDER achieves frame rates of 250–350 Hz (depending on whether a medium-sized or small ANN is used). These speeds are comparable to the rates achieved by high-speed eye trackers. One caveat is that CIDER is not uniformly sampling since it occasionally uses the ANN. However, the irregularity during the use of ANN can be mitigated by using a smaller ANN model. The power consumption at this frame rate is several tens of milliwatts since the system is operating in always-ON mode. Therefore, many of the optimizations that we used earlier no longer work. However, we don't anticipate that the high-speed mode will be used continuously; rather, this mode may be triggered when appropriate. Overall, we think that the ability to sample at high speed has substantial implications for a wide range of health and cognition applications.

6.6 Accuracy of Labeling

To evaluate the accuracy of the labeling scheme described in §4.3, we hand-labeled 100 eye images from one subject's data. For each image, we selected an elliptical region that visually seemed to best fit the pupil area. We then compared the pupil center and size estimate with those provided by the automatic labeling system for the same frames. The results are given in Table 3. Note that for both measures, the hand-labeling and automatic labeling techniques yield very similar results. The pupil size is slightly higher, but this is most likely due to the fact that the low-resolution images do no provide as sharp of an edge as would be expected with a higher-resolution camera. Thus, the pupil edge appears spread over a one- to two-pixel area, and distinguishing the exact pupil boundary within that region is difficult for a human to do visually.

Feature	**Mean Difference (pixels)**
Pupil Center	0.853
Pupil Size	1.52

Table 3: Automatic labeling vs hand labeling of pupil

7. RELATED WORK

At a high level, the idea of trading off energy consumption against robustness is present in a plethora of computing systems, but our novelty lies in enabling such tradeoffs by optimizing cost of digitizing pixels while robustly tracking eye parameters. We briefly highlight related efforts that have not already been mentioned in this paper.

Dynamic energy-aware adaptation: At a high level, our work can also be viewed as an instance of runtime energy-aware adaptation. Many techniques can be used to achieve such adaptation, including techniques such as varying application fidelity to achieve desired battery lifetime (e.g. changing the fidelity of a speech recognizer or video playback on a mobile device) [14] and adaptive sampling and communication that leverages spatial and temporal structure in sensor signals [36, 13, 21], among others. Our work is a very specialized instance of such adaptation in the context of eye trackers, and proposes the use of two models that are optimized to extract eye parameters at different costs (cross model and neural network), and techniques for switching between the models based on observed dynamics.

Eye tracking overview: Eye and gaze tracking has been a field of study for several decades [18, 28, 38]. Until recently, gaze tracking had mostly been applied to laboratory settings, where a subject sits facing one or more cameras that record video of their eyes and estimates eye parameters, often with their heads in a chin rest so as to remain motionless.

Algorithms to compute gaze positions fall into three categories: shape-based, appearance-based and hybrid algorithms [18]. The shape-based approach uses features of the eye image to fit an ellipse to the boundary between the pupil and the iris [18]. This approach typically works best with near-infrared (NIR) illumination sources, which make it easier to detect the pupil-iris boundary as previously discussed [28]. The cross-search model used in CIDER falls into this category of algorithm.

Appearance-based gaze tracking algorithms attempt to predict the gaze location directly from the pixels of the eye image without an intermediate geometric representation of the pupil. This approach essentially treats the gaze inference problem as a regression problem where the inputs are the pixels of the eye image and the outputs are the vertical and horizontal components of the point of gaze in the outward facing image plane. Due to the generality of the gaze inference problem when formulated in this way, predictions can be based on essentially any multivariate regression approach. Two prominent approaches used in the gaze tracking literature are multi-layer neural networks [2] and manifold-based regression [33]. The neural-network model used in the iShadow work [25] and adapted for CIDER falls into this category.

Mobile eye tracking devices: Recently, there has been tremendous interest in mobile eye trackers that can track eye parameters "in the wild." Companies such as Tobii and SMI have produced devices which have shown great promise for opening up new avenues of research [27, 37]. However, existing industrial-grade mobile eye tracking devices are predominantly a condensed version of a standard remote tracking system, including carefully calibrated on-board illumination and multiple high-definition cameras. While the engineering required to condense such a complex system to a wearable form-factor (usually in the shape of a standard pair of eyeglasses) is impressive, the power requirements of such systems are inordinately large by wearables standards - the user is required to carry a large battery pack, and even then their average run time is less than four hours. In addition, the devices only perform video recording and storage. Since processing the eye data often means running computations over thousands or hundreds of thousands of high resolution simultaneous image streams, even with a full desktop machine the processing can be a time-intensive task. And finally, the cost of these devices is often in the tens of thousands of dollars, making any kind of large-scale study infeasible to all but the most well-funded organizations.

Another notable device of relevance to our work is iGaze. The goal of iGaze is to detect gaze fixations and determine whether the user is looking at a networking-capable device, and if so, the user can initiate a wireless connection to facilitate some useful exchange of data between the iGaze platform and the device in question. The platform itself is comprised of a head-mounted camera monitoring the eye, very similar to this work, and a Raspberry Pi device carried by the user to do image processing and gaze computation. iGaze's predictive error and power consumption are relatively high compared to our work - they report average gaze error of 5° and average power consumption in excess of 1 W, whereas our error is below 1° and power consumption is 7mW. The high power consumption is presumably due to the use of traditional computer vision algorithms for pupil identification, which incur high cost as they require relatively high-resolution images and do not allow for subsampling.

8. FUTURE WORK

To close, we discuss in this section some avenues of future work that we have not addressed in this paper.

Adaptation to dynamics: While our zero-effort pupil-labeling approach tackles the initial calibration problem, one question that we have not addressed is how to deal with dynamics in eyeglass positioning or the environment. For example, shifts in the position of the glasses relative to the user's eye would increase error in ANN predictions. Similarly, prediction error would also increase if the ambient infrared in outdoor settings is significantly different from the data used in training the ANN model. To address such dynamics, we need approaches to detect in real-time that the existing model is performing less accurately than expected, and dynamically adjust the model, perhaps by leveraging our zero-effort training procedures described in this paper. We are exploring techniques for dynamic re-calibration in ongoing work.

Form-factor: The form-factor of the eyeglass is a particularly important problem to tackle for more widespread use of such devices. The biggest challenge in the design is finding an unobtrusive placement of the cameras while simultaneously achieving good coverage of the eye to enable robust estimation of eye parameters. Even with our current prototype, there were instances where we could not obtain complete coverage of the eye due to differences in face shape and the limited field-of-view of the camera. This problem is exacerbated when the camera needs to be embedded in the eyeglass frame since the positioning of the camera is closer to the eye and more sensitive to changes in placement. One of the questions that we are currently exploring is how to embed the cameras in the eyeglass frame so as to reduce form-factor without sacrificing accuracy.

9. CONCLUSIONS

In summary, this paper describes a new method, CIDER, for estimating eye parameters on a computational eyeglass using a staged architecture that trades off power for robustness. Our architecture uses an optimized detector for the "common case" involving a user being indoors and in limited-noise settings, and tremendously reduces the overall power consumption down to numbers that are within the range of typical wearable devices. CIDER deals with more noise and variable illumination settings by using more computational and sensing heft to filter out noise and deal with variability. Finally, we achieve very high frame rates, which gives us the ability to sense fine-grained eye parameters. Most surprisingly, we enable all of this functionally while operating on a small ARM Cortex M3 micro controller. To see a video demonstration of CIDER, visit [19].

Acknowledgments

The authors thank REU student Jose LaSalle for his contributions in the early stages of the glasses design process and for his assistance in exploring various avenues of hardware improvement. We would like to thank Geoffrey Barrows, of Centeye, Inc. for his continued support with the development and use of the Stonyman cameras. His assistance has relieved significant development burden and made much of this work possible. We also thank anonymous reviewers for their suggestions to improve this work. This research was partially supported by NSF grants #1217606, #1218586, and #1239341, as well as NIH grant #1U54EB020404.

10. REFERENCES

[1] M. B. Aerts, R. A. J. Esselink, W. F. Abdo, F. J. A. Meijer, G. Drost, N. Norgren, M. J. R. Janssen, G. F. Borm, B. R. Bloem, and M. M. Verbeek. Ancillary investigations to diagnose parkinsonism: a prospective clinical study. *J. Neurol.*, Nov 2014.

[2] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. Technical report, Pittsburgh, PA, USA, 1994.

[3] R. Baumeister and J. L. Alquist. Self-regulation as a limited resource: Strength model of control and depletion. *Psychology of self-regulation: Cognitive, affective, and motivational processes*, 11:21–33, 2009.

[4] R. F. Baumeister and T. F. Heatherton. Self-regulation failure: An overview. *Psychological inquiry*, 7(1):1–15, 1996.

[5] H. Bekkering, S. F. Neggers, R. Walker, B. Gleissner, W. H. Dittrich, and C. Kennard. The preparation and execution of saccadic eye and

goal-directed hand movements in patients with Parkinson's disease. *Neuropsychologia*, 39(2):173–83, 2001.

[6] M. S. Bolding, A. C. Lahti, D. White, C. Moore, D. Gurler, T. J. Gawne, and P. D. Gamlin. Vergence eye movements in patients with schizophrenia. *Vision Res.*, 102:64–70, Sep 2014.

[7] C. Bonnet, J. Rusz, M. Megrelishvili, T. Sieger, O. Matoušková, M. Okujava, H. Brožová, T. Nikolai, J. Hanuška, M. Kapianidze, N. Mikeladze, N. Botchorishvili, I. Khatiashvili, M. Janelidze, T. Serranová, O. Fiala, J. Roth, J. Bergquist, R. Jech, S. Rivaud-Péchoux, B. Gaymard, and E. Růžička. Eye movements in ephedrone-induced parkinsonism. *PLoS ONE*, 9(8):e104784, 2014.

[8] N. Carvalho, N. Noiret, P. Vandel, J. Monnin, G. Chopard, and E. Laurent. Saccadic eye movements in depressed elderly patients. *PLoS ONE*, 9(8):e105355, 2014.

[9] `http://www.centeye.com/%20products/current-centeye-vision-chips/`. "Current Centeye Vision Chips", Accessed: 2015-06-24.

[10] P. Christiansen, J. C. Cole, and M. Field. Ego depletion increases ad-lib alcohol consumption: Investigating cognitive mediators and moderators. *Experimental and clinical psychopharmacology*, 20(2):118, 2012.

[11] `http://www.arm.com/products/processors/cortex-m/cortex-m3.php`. "Cortex-M3 Processor - ARM", Accessed: 2015-06-24.

[12] S. Danziger, J. Levav, and L. Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

[13] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 588–599. VLDB Endowment, 2004.

[14] J. Flinn and M. Satyanarayanan. Energy-aware adaptation for mobile applications. In *ACM SIGOPS Operating Systems Review*, volume 33, pages 186–201. ACM, 1999.

[15] M. Fried, E. Tsitsiashvili, Y. S. Bonneh, A. Sterkin, T. Wygnanski-Jaffe, T. Epstein, and U. Polat. ADHD subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. *Vision Res.*, 101:62–72, Aug 2014.

[16] M. Fried, E. Tsitsiashvili, Y. S. Bonneh, A. Sterkin, T. Wygnanski-Jaffe, T. Epstein, and U. Polat. ADHD subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. *Vision Res.*, 101:62–72, Aug 2014.

[17] K. E. Friedl, S. J. Grate, S. P. Proctor, J. W. Ness, B. J. Lukey, and R. L. Kane. Army research needs for automated neuropsychological tests: monitoring soldier health and performance status. *Archives of Clinical Neuropsychology*, 22:7–14, 2007.

[18] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500, 2010.

[19] `http://sensors.cs.umass.edu/projects/eyeglass/`. "iShadow: The Computational Eyeglass", Accessed: 2015-06-24.

[20] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[21] M. Li, D. Ganesan, and P. Shenoy. Presto: feedback-driven data management in sensor networks. *IEEE/ACM Transactions on Networking (TON)*, 17(4):1256–1269, 2009.

[22] R. LiKamWa, Z. Wang, A. Carroll, F. X. Lin, and L. Zhong. Draining our glass: An energy and heat characterization of google glass. In *Proceedings of 5th Asia-Pacific Workshop on Systems*, APSys '14, 2014.

[23] J. A. Linder, J. N. Doctor, M. W. Friedberg, H. Reyes Nieva, C. Birks, D. Meeker, and C. R. Fox. Time of Day and the Decision to Prescribe Antibiotics. *JAMA Intern Med*, Oct 2014.

[24] S. Martinez-Conde, S. L. Macknik, X. G. Troncoso, and D. H. Hubel. Microsaccades: a neurophysiological analysis. *Trends Neurosci.*, 32(9):463–475, Sep 2009.

[25] A. Mayberry, P. Hu, B. Marlin, C. Salthouse, and D. Ganesan. ishadow: design of a wearable, real-time mobile gaze tracker. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, pages 82–94. ACM, 2014.

[26] R. J. Molitor, P. C. Ko, and B. A. Ally. Eye Movements in Alzheimer's Disease. *J. Alzheimers Dis.*, Sep 2014.

[27] J. D. Morgante, R. Zolfaghari, and S. P. Johnson. A critical test of temporal and spatial accuracy of the tobii t60xl eye tracker. *Infancy*, 17(1):9–32, 2012.

[28] C. Morimoto and M. Mimica. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1):4–24, 2005.

[29] L. M. Schmitt, E. H. Cook, J. A. Sweeney, and M. W. Mosconi. Saccadic eye movement abnormalities in autism spectrum disorder indicate dysfunctions in cerebellum and brainstem. *Mol Autism*, 5(1):47, 2014.

[30] W. Seiple, R. B. Rosen, and P. M. T. Garcia. Abnormal fixation in individuals with age-related macular degeneration when viewing an image of a face. *Optom Vis Sci*, 90(1):45–56, Jan 2013.

[31] A. Srivastava, R. Sharma, S. K. Sood, G. Shukla, V. Goyal, and M. Behari. Saccadic eye movements in Parkinson's disease. *Indian J Ophthalmol*, 62(5):538–44, May 2014.

[32] `http://www.st.com/stm32`. "STM32 32-bit ARM Cortex MCUs - STMicroelectronics", Accessed: 2015-06-24.

[33] K.-H. Tan, D. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision, 2002.*, pages 191–195, 2002.

[34] `http://www.tobiiglasses.com/`. "Tobii Glasses: Mobile Eye Tracker for real world research," Accessed: 2015-06-24.

[35] `http://www.tobii.com/Global/Analysis/Marketing/Brochures/ProductBrochures/Tobii_X1_Light_Eye_Tracker_Technical_Specifcation_Leaflet.pdf?epslanguage=en`. "Tobii X1: Gaze Precision and Gaze Accuracy", Accessed: 2015-06-24.

[36] R. Willett, A. Martin, and R. Nowak. Backcasting: adaptive sampling for sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 124–133. ACM, 2004.

[37] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg. Detecting eye contact using wearable eye-tracking glasses. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 699–704. ACM, 2012.

[38] L. Young and D. Sheena. Survey of eye movement recording methods. *Behavior Research Methods*, 7(5):397–429, 1975.

[39] L. Zhang, X.-Y. Li, W. Huang, K. Liu, S. Zong, X. Jian, P. Feng, T. Jung, and Y. Liu. It starts with igaze: Visual attention driven networking with smart glasses. In *Proceedings of the 20th annual international conference on Mobile computing and networking*, pages 91–102. ACM, 2014.