PupilScreen: Using Smartphones to Assess Traumatic Brain Injury

ALEX MARIAKAKIS, JACOB BAUDIN, ERIC WHITMIRE, VARDHMAN MEHTA, MEGAN A. BANKS, ANTHONY LAW, LYNN MCGRATH, and SHWETAK N. PATEL, University of Washington

Before a person suffering from a traumatic brain injury reaches a medical facility, measuring their pupillary light reflex (PLR) is one of the few quantitative measures a clinician can use to predict their outcome. We propose PupilScreen, a smartphone app and accompanying 3D-printed box that combines the repeatability, accuracy, and precision of a clinical device with the ubiquity and convenience of the penlight test that clinicians regularly use in emergency situations. The PupilScreen app stimulates the patient's eyes using the smartphone's flash and records the response using the camera. The PupilScreen box, akin to a head-mounted virtual reality display, controls the eyes' exposure to light. The recorded video is processed using convolutional neural networks that track the pupil diameter over time, allowing for the derivation of clinically relevant measures. We tested two different network architectures and found that a fully convolutional neural network was able to track pupil diameter with a median error of 0.30 mm. We also conducted a pilot clinical evaluation with six patients who had suffered a TBI and found that clinicians were almost perfect when separating unhealthy pupillary light reflexes from healthy ones using PupilScreen alone.

CCS Concepts: •**Applied computing** →**Consumer health;** •**Human-centered computing** →*Smartphones;*

Additional Key Words and Phrases: Health sensing, smartphones, pupillometer, pupillary light reflex, pupil dilation, convolutional neural network

ACM Reference format:
Alex Mariakakis, Jacob Baudin, Eric Whitmire, Vardhman Mehta, Megan A. Banks, Anthony Law, Lynn McGrath, and Shwetak N. Patel. 2017. PupilScreen: Using Smartphones to Assess Traumatic Brain Injury. *PACM Interact. Mob. Wearable Ubiquitous Technol.* 0, 0, Article 0 (2017), 27 pages.
DOI: 0000001.0000001

1 INTRODUCTION

Traumatic brain injury (TBI) accounts for 30% of all injury-related deaths in the United States [9]. TBI can occur in a variety of situations, including car accidents, falls, and blunt force trauma. A concussion is a specific form of TBI caused by a swift blow to the head; these injuries tend not to be life-threatening, but can have serious and long-term effects on a person's memory, motor abilities, and overall cognition [37]. One area in which concussions have garnered national attention is sports, particularly contact sports such as boxing, hockey, and American football. The CDC estimates that there are roughly 3.8 million concussions per year in the US, and about half of them will go undiagnosed [19]. Patients suffering a concussion have a 600% increased risk of a future head injury and 15% increased risk of permanent cognitive deficits [19]. This is particularly more problematic

This work is supported by the National Science Foundation Graduate Research Fellowship and the ARCS Foundation.

Author's addresses: A. Mariakakis and E. Whitmire and Vardhman Mehta and M. A. Banks and S. N. Patel, School of Computer Science and Engineering, University of Washington; J. Baudin, Department of Physiology and Biophysics, University of Washington; A. Law, Department of Otolaryngology - Head and Neck Surgery, University of Washington. L. McGrath - Neurological Surgery, University of Washington.

DOI: 0000001.0000001

for younger athletes who are not as well-educated on concussion prevention measures such as proper tackling technique. Roughly 250,000 young Americans (<20 years old) were treated for sports-related concussions in 2009 [7]. High school football players are 3 times more likely to suffer a catastrophic head injury than college football players [4]. Athletic departments with major funding can afford to have a team doctor with years of experience on-hand to diagnose concussions. For teams that are not as well-funded (e.g., pee-wee, middle school, high school), a school nurse, volunteer, or parent must put themselves in the same position as those doctors, but without the same tools or knowledge at their disposal. Identifying concussions immediately is essential because allowing a concussed athlete to return to play can lead to further significant injury [34]. There exists a need for accessible concussion screening that anyone can use at any moment. Our proposed system, PupilScreen, is meant to address this need by using a technology that most people have within arm's reach: a smartphone.

The methods that team doctors currently use to assess the probability of a concussion on the sidelines fall in one of two categories. The first category is task-based methods, which grade the performance of an athlete at a particular task using quantitative measures. For example, the King-Devick test [15] requires an athlete to read single digit numbers from left-to-right in different configurations. The second category includes survey-based methods, such as the Sport Concussion Assessment Tool (SCAT) [1]. Although a great deal of research supports the efficacy of these methods [14, 16], they capture indirect effects of concussions, require the athlete to be responsive, and take minutes to complete. These methods also require baseline measurements taken at the beginning of the season, which Broglio et al. [5] found were not repeatable for 118 healthy student volunteers. Furthermore, there is anecdotal evidence that athletes sometimes intentionally fail the baseline assessment so that there is little difference following an injury and they can remain in play [33].

A more quantitative method to assess a TBI is to check a person's pupillary light reflex (PLR), or the manner in which their pupils react to a light stimulus. The PLR of those who have suffered a TBI is typically either slower or not as pronounced [6]. The clinical gold standard for measuring the PLR uses a device called a pupillometer. Pupillometers are expensive (~$4,500 USD) and are therefore mainly used in hospital intensive care units. Another method for assessing the PLR is through a penlight exam, in which a clinician directs a penlight towards each of the patient's eyes and observes the pupils' responses. This procedure is simple to perform, but has many drawbacks, including a lack of standardization, a need for deliberate training, and poor inter-observer reliability [38]. Those who provide first aid in emergency situations (e.g., EMTs and battlefield medics) will often conduct penlight exams despite these limitations because rapid assessment is prioritized over precision.

PupilScreen combines the repeatability, accuracy, and precision of a pupillometer with the ubiquity and convenience of the penlight test for quantifying a person's PLR. The PupilScreen system consists of two ubiquitous components: a smartphone app and a box (Fig. 1). Most people own a smartphone, and the box can be easily created since it does not require any wiring or expensive components. This means that PupilScreen can be available to almost anyone just hours before a sports event. The PupilScreen app records an 8-second video of a person's eyes as the pupils constrict in response to the smartphone's flash. The video is analyzed by convolutional neural networks (CNNs) in order to estimate the diameter of the pupils in each frame. We explored two different architectures. The first architecture uses two CNNs in sequence, where the first estimates the locations of the pupils and the second estimates their diameters given images cropped around their locations. The second architecture uses a fully convolutional network to perform pixelwise segmentation. By examining how the pupil diameter changes over time, we extract metrics used by clinicians for diagnosis (e.g., constriction velocity, magnitude of diameter change). To standardize the results of the PupilScreen app, the smartphone is placed in a 3D-printed box. The box simultaneously eliminates ambient lighting conditions and controls the distance between the person's face and the flash.

[1] http://www.sportconcussions.com/html/SCAT3.pdf

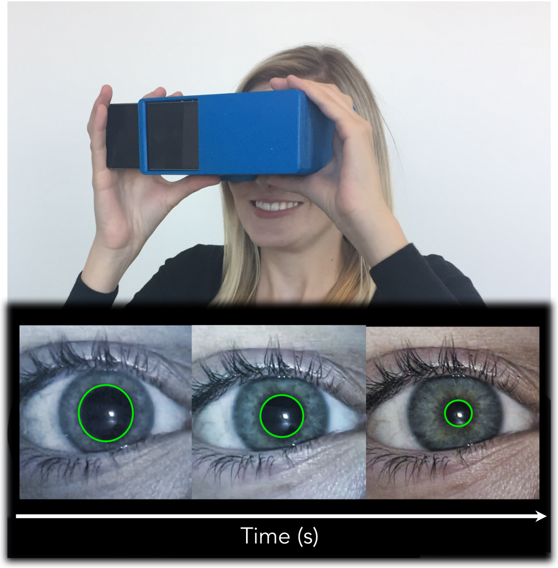

Fig. 1. PupilScreen is a system that measures the pupillary light reflex to determine the severity of a traumatic brain injury. A smartphone app records a video of the patient's eyes as the camera's flash illuminates them. The VR headset-like box controls the position of the phone and the lighting that reaches the eyes.

Training CNNs requires a large quantity of diverse data, which is difficult to collect from patients with TBI. Therefore, we evaluated PupilScreen's ability to track the PLR on a dataset from 42 healthy adults. The range of pupil sizes encountered in non-reactive pupils is a subset of that encountered in reactive pupils; because the networks are trained on video frames in isolation, training PupilScreen on data from healthy individuals allows it to measure pupil diameter in individual video frames regardless of pupil reactivity. We found through our analysis that the PupilScreen was able to track pupil diameter with a median error of 0.30 mm with the fully convolutional network, the more accurate of the two approaches. Meeker et al. [35] found that manual pupil examination has a median error of 0.5 mm, and a clinical pupillometer has a median error of 0.23 mm, which places the accuracy of PupilScreen between the two. PupilScreen was also able to track the pupil center with a median error of 0.20 mm. Using information about the pupil diameter over time, PupilScreen extracts three clinically relevant measurements: constriction amplitude, percentage, and velocity. We found that PupilScreen estimates constriction amplitude with a mean absolute error of 0.62 mm for a range of measured amplitudes that spanned 0.32-6.02 mm, constriction percentage with a mean absolute error of 6.43% for a range that spanned 6.21-62.00%, and max constriction velocity with a mean absolute error of 1.78 mm/s for a range that spanned 1.37-8.99 mm/s. To support PupilScreen's efficacy as a diagnostic tool, we conducted a pilot clinical evaluation with six patients who had suffered a TBI. We found that clinicians were able to distinguish between normal and abnormal PLR curves produced by PupilScreen with almost perfect accuracy.

In designing a smartphone-based pupillometry system, our main challenges are:

(1) Designing a controlled setup that is portable and inexpensive, and
(2) Accurately identifying the pupils in video using only visible light.

Our contribution comes in four parts:

(1) The design and implementation of the PupilScreen system, which allows a smartphone to perform repeatable PLR tests at a fraction of the cost of a clinical device,
(2) Two different CNN-based approaches for estimating the pupil diameter in videos,
(3) An evaluation of PupilScreen's accuracy on 42 healthy participants, and
(4) An evaluation of PupilScreen's ability to assist with diagnosis on 6 individuals who have suffered a TBI.

2 PUPILLARY LIGHT REFLEX BACKGROUND

Papers by Martnez-Ricarte et al. [30], Larson and Behrends [22], and Zafar and Suarez [55] provide thorough discussions on the mechanics of the pupil, the pathophysiology of the PLR, and the diagnostic power of the PLR. We summarize their content here for a broader audience, but refer the reader to their papers for a more detailed discussion of the PLR.

2.1 The Characteristics of the PLR

A normal PLR is defined as symmetric constriction or dilation of both pupils in response to a light stimulus or its absence, respectively. The pupil size must change by a non-trivial amount within a specified time frame and should change in both eyes, regardless of which eye is stimulated. For example, when a person covers one eye while the other is exposed to bright light, the pupils of both the covered and exposed eyes should constrict, producing a phenomenon known as the consensual response.

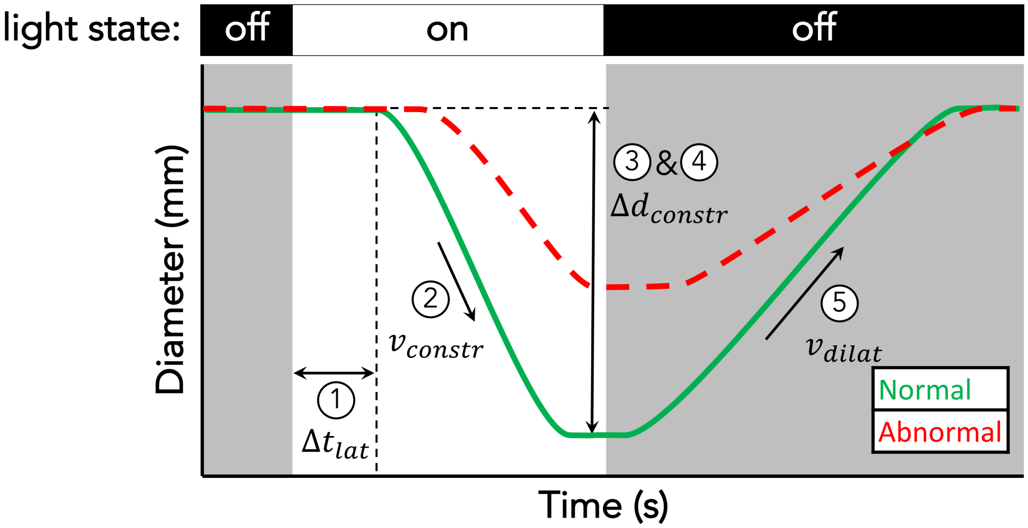

Fig. 2. A PLR curve annotated with the five common descriptive measures: (1) latency, (2) constriction velocity, (3) constriction amplitude, (4) constriction percentage, and (5) dilation velocity. An abnormal PLR curve with increased latency, slower velocities, and diminished amplitude is also included for comparison.

When given pupil diameter as a function of time, clinicians focus on five simpler quantitative measures (Fig. 2):

- **Latency (ms):** the time between the beginning of the light stimulus and the start of pupil constriction
- **Constriction velocity (mm/s):** the speed at which pupil constricts; reported as mean or max

- **Constriction amplitude (mm):** the difference between the maximum pupil diameter before light stimulation and minimum pupil diameter after light stimulation
- **Constriction percentage (%):** the constriction amplitude expressed as a percentage of the initial size
- **Dilation velocity (mm/s):** the speed at which the pupil dilates; reported as mean or max

2.2 Diagnostic Significance of the PLR

Because the neural pathways underlying the PLR include multiple brain regions and traverse many others, it is sensitive to a variety of injuries [52]. Our motivating use case is traumatic brain injury. When the brain shifts inside the skull, it has the potential to injure both the cranial nerves carrying signals necessary for the production of the PLR or the brain regions that process these signals. A survey by Zafar et al. [55] in 2014 notes that the literature relating PLR to concussions is limited because it often includes a small number of patients (≤10 patients with TBI) or individual case studies; however, researchers such as Ciuffreda et al. have recently published the results of studies with larger datasets. In 2015, Thiagarajan et al. [46] quantitatively evaluated the PLRs of individuals with non-blast-induced, chronic, mild TBI (mTBI). That study included 15 healthy individuals and 17 patients with mTBI. Thiagarajan et al. [46] found statistically significant differences between the two populations for most of the PLR metrics listed in Section 2.1. In a study published a year later, Truong et al. [48] carried out a larger study with 40 healthy individuals and 32 patients with mTBI. Beyond the larger study population, Truong et al. also studied how different light stimuli (e.g., pulses, step changes, different colors) could be used to better discriminate certain PLR metrics. Populations of the same size were later examined to determine how pupillary asymmetry [50], photosensitivity [49], and refractive errors [51] affected the PLR. With more accessible pupillometry, such as that provided by PupilScreen, we believe that larger scale studies will be easier than ever before, particularly for examining the immediate effects on the PLR following a crisis.

Changes in the PLR are much better described by the literature in the context of severe TBI since those patients are often hospitalized and the changes are more obvious as a result of the severe cerebral dysfunction. Taylor et al. [45], for example, found that elevated intracranial pressure (ICP) for >15 minutes in patients with midline shift was associated with a decrease in pupillary constriction velocity. The PLR has also been examined as an indicator of the outcomes for patients following cardiac arrest. In a case study with 30 patients, Behrends et al. [3] found that the presence of a reactive pupil during the first five minutes of CPR was associated with increased survival and good neurologic outcome.

2.3 Techniques for Measuring the PLR

There are two methods used by clinicians to measure the PLR. The clinical gold standard method uses a device called a pupillometer. Infrared-based pupillometry takes advantage of the fact that there is a better demarcated boundary between the pupil and the iris when infrared imaging is used. While pupil diameter is tracked using infrared light, a ring of white LEDs stimulates the eye, causing the pupillary constriction. The components needed to make a pupillometer can be inexpensive, but the total product costs ~$4,500 USD because, among other reasons, it is a self-contained system with proprietary algorithms and strict hardware requirements. Nevertheless, pupillometers provide two main benefits: precision and consistency. A study conducted by Meeker et al. [35] revealed that, for a modest participant pool, a pupillometer can track the pupil diameter with a median error of 0.23 mm. Couret et al. [8] asked multiple clinicians to perform PLR measurements on 200 healthy volunteers in a variety of ambient lighting conditions. They found high intra-class correlation for maximum resting pupil size (0.95) and minimum pupil size after light stimulation (0.87) regardless of ambient lighting or device operator.

A low-cost alternative for measuring the PLR involves using a penlight - a pen-sized flashlight (Fig. 3). A penlight test is performed by directing the penlight toward and away from the patient's eye. Because the PLR is manually observed by a clinician, penlight-based pupil measurements are more likely to be inaccurate and imprecise. Meeker et al. [35] found that manual measurement of pupil diameter resulted in a median error of

Fig. 3. A penlight test being performed by a clinician.

0.5 mm, more than twice that of a pupillometer. Couret et al. [8] found a poor Spearman's rank correlation coefficient (0.75) between manual pupil size measurements and pupillometer readings. Only 64% of the cases when volunteers had pupils smaller than 2 mm were properly identified, and only half of the cases of anisocoria (i.e., unequal pupil sizes) were caught. Larson et al. [23] note the inability of clinicians to detect small, but clinically significant responses. Characteristics such as constriction velocity and amplitude also cannot be measured in absolute terms when using a penlight; instead of reporting a constriction velocity as 3.8 mm/s, observers can only describe the PLR as "normal", "sluggish", or "fixed". Penlight exams lack standardization as well. Clinicians purchase penlights from different companies, each with their own brightness specifications. Even if two health care providers use the same penlight, the patient may not experience the same light stimulus because of how the clinicians hold their penlights (i.e., distance and angle) or due to differences in ambient lighting conditions. Prior work has also discussed how penlight tests can lead to poor inter-observer reliability in PLR characteristics. Olson et al. [38] performed a single-blinded observational study where two practitioners were asked provide subjective scores for pupil reactivity. Across 2,329 paired assessments, Cohen's kappa coefficient was only moderate for pupil size (κ = 0.54), shape (κ = 0.62), and reactivity (κ = 0.40). In fact, only 33.3% of the pupils that were judged to be non-reactive by the practitioners were scored as non-reactive by pupillometry.

Our prototype of PupilScreen is the first step towards combining the advantages of a pupillometer (repeatability, accuracy, precision) with the advantages of a penlight test (ubiquity, convenience). Before discussing how PupilScreen works, we will first provide an overview of pertinent related work.

3 RELATED WORK

The ubiquity of smartphones has enabled them to become a platform for the effective deployment of health applications. Mobile health is growing at an exponential rate, but we focus our review of related work on applications pertaining to the eye. We also summarize previous work surrounding gaze tracking and pupil measurement.

3.1 Ocular Diagnostic Applications

A series of ocular diagnostic applications have been proposed by the Camera Culture Lab at MIT. CATRA [40] detects cataracts in the eye's lens by scanning the eye with a beam of collimated light and asking for feedback from the user regarding whether the beam appears clear or blurry. NETRA [39] asks the user to align patterns projected through a microlens display and pinhole plane to identify refractive errors in the eye. Finally, EyeMITRA [24] is a wearable camera for performing mobile retinal imaging. By asking the user to focus onto points shown in the other eye, indirect diffuse illumination allows the camera to view the retina in the back of the eye. These projects combine optical components with user input on simple tasks to perform otherwise complicated diagnostic procedures. Our work differs from theirs in two ways: (1) once started, the PupilScreen application is completely automated, and (2) PupilScreen utilizes hardware that is either already ubiquitous (the smartphone) or simple to create with loose tolerance (the box).

There have been a number of other projects that utilize smartphones for diagnosing ocular conditions. D-Eye[2] is a smartphone adapter for performing fundoscopy. Abdolvahabi et al. [1] describe how to catch the early onset of rare eye cancers in newborns based on the color of their pupils in digital photos; many are familiar with the common "red-eye" effect in pictures, but tumors in the back of the eye can make the pupil appear white in photos. Bastawrous et al. [2] and Giardini et al. [17] have deployed a suite of tools for diagnosing ocular conditions, including visual acuity and glaucoma.

3.2 Concussion Diagnostic Applications

Regarding concussions, metrics other than the PLR have been examined for diagnosis. Maruta et al. [31, 32] measured visual tracking performance in terms of gaze positional error relative to a target and found that the performance variability increased for those with a TBI. Joiiv Lindsay [27] is one of the many researchers who have noted that involuntary eye movements are more prevalent in those with a TBI. Such work has been conducted in a clinical setting with dedicated devices; along with measuring the PLR, we look forward to investigating these metrics with PupilScreen in the future.

Lee et al. [25] provide a thorough survey of publicly available smartphone and tablet apps that are intended for assessing sports-related concussions. We refer the reader to their survey for a complete list of the smartphone apps that were examined, which includes both apps that are intended for non-medical personnel (e.g., coaches or parents) and medical personnel (e.g., team doctors). Lee et al. compared the purpose of each app to the SCAT2 and found that all of them exhibited partial or imperfect compliance to it. Furthermore, they found that the apps serve as a means of presenting, managing, and documenting various aspects of the SCAT2 rather than automating them.

3.3 Gaze Tracking

Our work proposes a novel method for measuring pupil diameter. Although gaze tracking is a different problem - one that cares about the position of the pupil relative to the eye - the techniques used in both problems share many similarities.

The easiest way to track gaze involves the use of infrared light to emphasize the pupils. Infrared light is invisible to the naked eye and reflects off of the cornea, a fact which is leveraged in one of two ways. In bright pupil tracking, the light source is aligned with the camera so that the reflection can be tracked; in dark pupil tracking, the light source is off-angle so the pupil remains darker than the rest of the eye. There are a variety of commercial products by companies such as Tobii and LC Technologies that leverage this phenomenon for pupil detection. These products are primarily intended for controlled, desktop situations, but researchers have proposed form factors meant for on-site and outdoor scenarios. Fischer and van den Heever [10], Świrski et al. [44], and Kassner et al. [20] are just three examples of techniques that take advantage of custom-designed headsets with an infrared camera pointed directly at the eyes for gaze tracking. All three of those systems are intended for gaze tracking and are evaluated as such, but their algorithms calculate both the pupil center and diameter as a means to that end. It should also be noted that Fisher and van den Heever's device uses gaze tracking alongside visual

[2] https://www.d-eyecare.com/

tasks like the King-Devick test with the intent of diagnosing sports-related concussions on the sideline, although there is no formal study on how that data improves the power of those tests.

There is a variety of methods for tracking the pupil without the help of infrared light. Qualcomm's SnapDragon SDK[3] provides facial features like gaze direction using a smartphone's front-facing camera, but their algorithm is proprietary. Timm and Barth [47] propose a mean of gradients approach for identifying the pupil center; essentially, the center is found using an optimization technique that identifies the pixel where a vector field of image intensity gradients is most likely to converge. For smartphones and tablets, EyeTab [54] relies on the observation that the pupil and the iris are normally concentric, so the center of the ellipse that best fits the edge between the iris and the sclera also corresponds to the center of the pupil.

Fuhl et al. have proposed a number of methods for detecting the pupil center. ExCuSe [11] utilizes two different techniques depending on whether the image contains a reflection or not. If there is a reflection, curved edges are found using dynamic thresholding and morphological operations; if there is no reflection, the coarse center is estimated using histograms oriented at various angles and then refined using an iterative ellipse-fitting technique [26]. ElSe [13] defines the pupil as the location where an image of the eye responds to two pre-determined convolutional filters: a circular mean filter and a surface difference filter. Finally, PupilNet [12] uses two CNNs for gaze tracking; the first CNN returns a coarse pupil center estimate, which is used to select a region of interest that is fed into a second CNN to refine the prediction.

In this work, we explored two different network architectures. The first architecture is similar to PupilNet in that it involves two CNNs in sequence. However, instead of using the second network to provide a more precise estimate of the pupil center, we use the second network to estimate the pupil diameter. Although the first network only provides a coarse estimate of the pupil center, we demonstrate that it is sufficiently accurate for our purposes. The second architecture is an implementation of FCN-8, a fully-convolutional neural network proposed by Long et al. [28] for achieving pixelwise segmentation.

3.4 Pupil Measurement

Researchers have extended existing techniques for identifying the pupil center to measure the contour of the pupil. Starburst [26] initializes an estimate of the pupil center using the mean of gradients approach. The algorithm then increments a marker in different directions from that seed until the first strong edge (defined by the gradient along this path crossing some threshold, which is expected to occur between the iris and pupil) is reached. An ellipse is fit to those edge points and its center is used as the seed for subsequent iterations of the same procedure until convergence.

A subset of the work in this area is particularly motivated by the use of pupil dilation as a proxy for assessing cognitive load. PupilWare [42] proposes improvements on the Starburst technique for use with a desktop web camera. These improvements include avoiding directions that could contain eyelash shadows and adding randomness to seed selection. Klingner, Kuman, and Hanrahan [21] do not discuss their pupil measurement algorithm in great detail, but provide a deeper analysis on task-evoked pupillary responses.

Many of the non-infrared-based techniques anecdotally cite issues for people with dark irises, even going as far as removing users with extremely dark irises from their studies. They primarily rely on the presence of an edge between the iris and the pupil. PupilScreen uses a completely model-based approach that can learn features beyond edges (e.g., gradients and contiguous black pixels) for tracking the pupil.

4 DATA COLLECTION

We collected video recordings using the PupilScreen app and box to train its CNNs and evaluate its ability to track pupil diameter. Since our approach to segmenting pupils relies on CNNs, we require a large number of

[3] https://developer.qualcomm.com/software/snapdragon-sdk-android

Table 1. Participant demographics (N = 42)

SEX - N (%)	
Male	16 (38.1%)
Female	26 (61.9%)
IRIS COLOR - N (%)	
Blue	17 (40.5%)
Brown	20 (47.6%)
Mixed	5 (11.9%)

training examples from individuals with various pupil sizes and iris colors. This is difficult to attain through a patient population with TBI. Cases of TBI are limited, and the pupils of those with TBI usually stay a fixed size. Because of this, our networks are trained on data from healthy volunteers at the University of Washington and Harborview Medical Center. Below, we elaborate on the diversity of the participant pool. We then describe our data collection procedure, including the design of the PupilScreen box and our methods for gathering ground truth measurements. In Section 6.4, we present a preliminary evaluation conducted on six individuals with TBI to examine PupilScreen's clinical efficacy. All facets of our study were approved by the University of Washington's Institutional Review Board.

4.1 Enrollment

Our training dataset comes from 42 volunteers: 16 males and 26 females. Typical non-infrared computer vision-based systems are reliant on determining the border between the iris and the pupil, which is more obvious for those with light blue eyes than those with dark brown eyes. For this reason, it was important to recruit participants with various iris colors. Our study includes a balanced mix of iris colors: 17 blue, 20 brown, and 5 with a noticeable gradient between different colors. In most cases, the irises that were classified as mixed were light brown near the pupil but primarily blue.

Ideally, ethnicity should have no effect on PupilScreen's ability to measure the pupil diameter since the two are uncorrelated. We crop the images beforehand to reduce the number of skin-related pixels that are utilized by the CNNs; however, since our model-based approach for tracking the pupil is agnostic to the eye's structure, we can make no guarantee that the CNNs will not learn to estimate the pupil center or diameter from skin tone features. Although we did not specifically ask participants for ethnicity information, we note that one-sixth of the participants had a darker skin complexion.

4.2 Data Collection Application

All of the data was collected by the researchers using a custom app on an iPhone SE. The phone was placed into a slot in the back of the PupilScreen box (Fig. 4). The design of the box is the same as the one used in BiliScreen [29], a project by a subset of this work's authors that aims to estimate the color of a person's sclera to detect cases of jaundice. The box-phone combination serves three purposes: (1) the box controls the position of the phone relative to the person's face, including the distance to and alignment with the face, (2) the box eliminates the effects of ambient lighting conditions, and (3) the phone provides its own lighting using the flash. The dimensions of the box are loosely modeled after the Google Cardboard. Besides the fact that the camera is centered for the PupilScreen box, rather than the screen as in the Google Cardboard, the main difference between the two is the fact that the PupilScreen box is deeper. Having the camera close to the participant's face increases the effective resolution of their eyes, which allows PupilScreen to detect smaller changes in pupil diameter and measure the

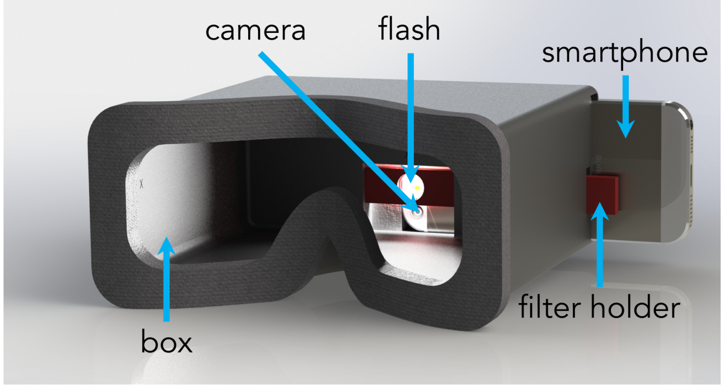

Fig. 4. A 3D rendering of the PupilScreen box. The smartphone's flash lies in the horizontal center of the box. The box has a hole on the side so that a neutral density filter and a diffuser can be aligned with the flash using a sliding stick.

PLR with increased precision. On the other hand, moving the phone further away allows the camera to see both eyes at once and reduces the discomfort caused by the intense flash.

Although the box used in this study was 3D-printed for durability, we believe that it could be made with an even cheaper material like cardboard (provided that it is sturdy enough to support the weight of the phone). Also note that there is no electronic connection between the phone and the box, simplifying its manufacturing requirements. Apple iOS 9 does not provide complete dimming control over the brightness of the flash LED. At close distances, participants from a pilot study found the intensity of the light to be uncomfortable. To make the light more manageable, a neutral density filter and diffuser were placed directly in front of the flash using a sliding stick. These components were chosen because they had precise specifications available online, but they could be replaced with a cheaper alternative like a sheet of white computer paper in the future.

Prior to putting the box up to their face, participants were asked to take off glasses if they wore them. Once the phone was placed in the box and the participant held it up to their face, the flash was turned on briefly and autofocus was enabled. The resulting camera focus was fixed for the remainder of the study to avoid blurriness as the lighting in the box changed. The flash was then turned off and after a brief pause to allow the pupils to recover, data collection commenced. The video was recorded at 30 fps with 1920×1080 resolution. After an audible 3-second countdown from the phone's speakers, the flash illuminated the participant's eyes. The stark change in lighting maximized the degree to which the pupil constricted, akin to the difference experienced when using a pupillometer. The recording stayed on for another five seconds, resulting in an 8-second long recording. The five second period after the introduction of the light stimulus was far longer than what was needed to capture the PLR, but provided extra video frames for evaluation. For each study participant, the PLR was recorded three times. Between recordings, a one-minute break was added to allow the participant to rest their eyes.

4.3 Ground Truth Measurements

Videos were manually annotated to generate ground truth labels. Using custom software, two researchers labeled frames by selecting points along the edges of the pupils and letting OpenCV's ellipse fitting algorithm generate a corresponding outline (Fig. 5). The researchers could see and adjust the outlines to better fit the images. If the pupil was difficult to distinguish from the iris, the researchers could adjust the contrast to make it more visible. If the pupil was still too difficult to see after that, either because of poor focus or lighting, the frame was skipped; this only happened for 1.8% of the total frames encountered. The points were fit to an ellipse because not all

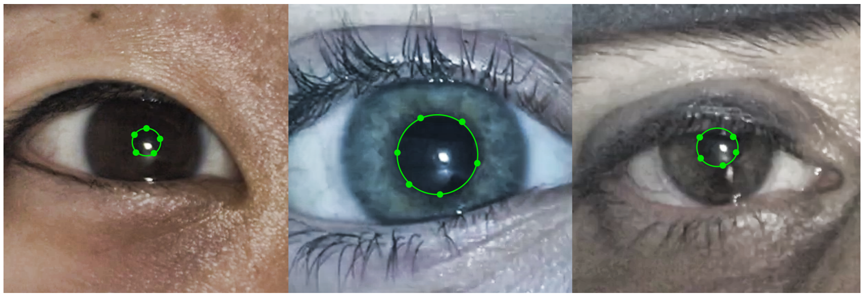

Fig. 5. A selection of manually annotated images of pupils zoomed in on the region of interest. Note that although the pupil may seem indistinct from the iris in some of the images above, the labeling was performed on much larger monitors with better contrast than what appears in print.

pupils are circular. Since pupillometry is only concerned with a single pupil diameter, the ellipses were converted to circles by averaging their axes. With this method, the pupil diameters were labeled in pixels. The researchers labeled every fifth frame in the three videos from each user. Each video was 8 seconds long, but the first 3 seconds occur before the flash was turned on, resulting in 5 seconds × 30 frames/second × (1/5 frames) × 3 videos = 90 labeled frames per person. Frames were labeled independently of one another to avoid biases between frames; however, this led to greater variation between consecutive frames that can be primarily attributed to human error. A 3$^{\text{rd}}$-order Savitzky-Golay filter was applied to temporally smooth the pupil center and diameter labels. To quantify the agreement of the labels across the researchers, both labeled a common set of 5 users (15 videos, 450 frames). The average difference between the smoothed pupil center labels was 3.46 px, which translates to 0.27 mm. The average difference between the smoothed pupil diameter labels was 2.00 px, which translates to 0.16 mm. Note that these variations are not independent; if a researcher underestimated the extent of an edge, the labeled center would move away from that edge and the labeled diameter would be lower than the actual value. The degree of inter-researcher agreement can also be quantified using the intersection-over-union (IoU) measure, a standard metric for segmentation agreement between two regions. The mean IoU for the researchers' labels was 83.0%. Note that the IoU measure is calculated relative to the total area of the two labeled pupils. If the pupil center labels for a 3 mm pupil were only off by a single pixel, that difference alone would lead to an IoU score of 93.8%.

Although a clinical-grade pupillometer could have provided an alternative method for quantifying the PLR, its results would not have been directly comparable to PupilScreen. The two setups have light stimuli with different intensities, which would result in different magnitudes of pupil constriction. Furthermore, PupilScreen eliminates the effect of ambient lighting because the box completely encloses the patient's eyes, whereas pupillometers do not since they are used in hospitals with roughly standard lighting conditions. Infrared imaging could have been used to provide a comparative ground truth measurement of pupil diameter; however, an algorithm still would have been needed to turn those frames into pupil diameters, and that algorithm would have needed its own validation.

5 PUPILSCREEN ALGORITHM

In this section, we will describe how the video data was pre-processed before being input to the CNNs. We then follow by describing the architecture of the CNNs used to estimate the pupil center and the pupil diameter, the post-processing of the CNN outputs, and the specifics of the CNN training.

5.1 Pre-processing

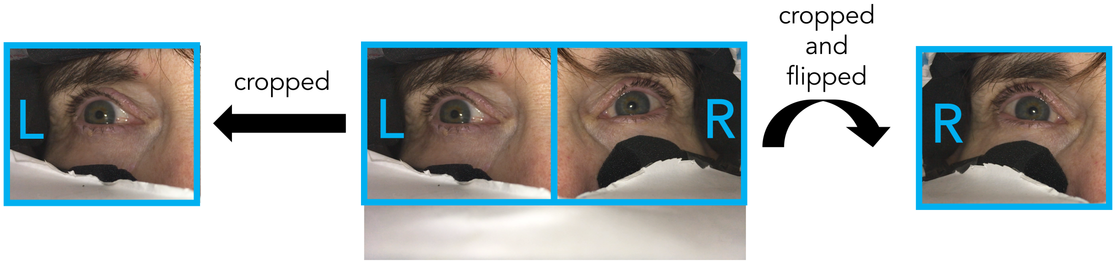

Fig. 6. Each frame was cropped to create two input images for the CNNs: one for the left eye and one for the right eye. The image of the right eye and its label were flipped to make the two images comparable.

Videos were recorded at 30 fps with 1920×1080 resolution. Treating each pixel as an individual input feature produces a very large input layer with a significant amount of unnecessary information; pixels around the eye socket provide no information about the pupil, and pixels on the left and right sides of the image should be considered independently in order to catch cases in which the pupils behave differently. We attempted to crop around the eyes using off-the-shelf eye detection algorithms, but found that they failed in many cases. This may have been because the detection algorithms rely on the presence of other facial features (e.g., nose) that are obscured by the PupilScreen box. Instead, the conservative cropping bounds in Fig. 6 are used. The bottom third is cropped off because it only contains the box. The remainder of the video frame is split into two halves - left and right - to produce one image per pupil. To make the images comparable and allow a single CNN to handle each task, the image of the right eye and the coordinates of its pupil center label are flipped horizontally. To emphasize the pupil, the image is converted to the HSL color space and contrast-limited adaptive histogram equalization (CLAHE) [41] is applied to the lightness (L) channel. In short, CLAHE avoids the pitfalls of global histogram equalization by dividing an image into small tiles (88 px in our case) and then equalizing only within those individual tiles.

5.2 CNN Architectures

Two different architectures were tested for measuring the size of the pupil. We describe their inspiration and implementation details below.

5.2.1 First Architecture: Sequential CNNs.

The first architecture was similar to that of PupilNet by Fuhl et al. [12], which uses two networks in sequence to arrive at a precise estimate of the pupil center. The intuition behind their approach was that the first network reduces the search space for the pupil by roughly localizing the pupil center, allowing for the second network to ignore irrelevant pixels and examine a specific region in more detail. Inspired by that intuition, we also explored the use of two networks for pupil measurement. The first network serves the same purpose, but the two applications differ in the second network. Rather than learning a finer pupil center measurement, we train the second network to learn the pupil diameter. We demonstrate that even if the pupil is not exactly centered using the output of the first network, the second network can be robust enough to handle those issues.

Fig. 7 illustrates the details of the first architecture. The first network (Fig. 7, top) is trained to accept an image from the pre-processing step as input and return the location of the pupil center. Before being input to the network, the image is downsampled by a factor of 4. The network has 5 convolutional layers, each with a rectified

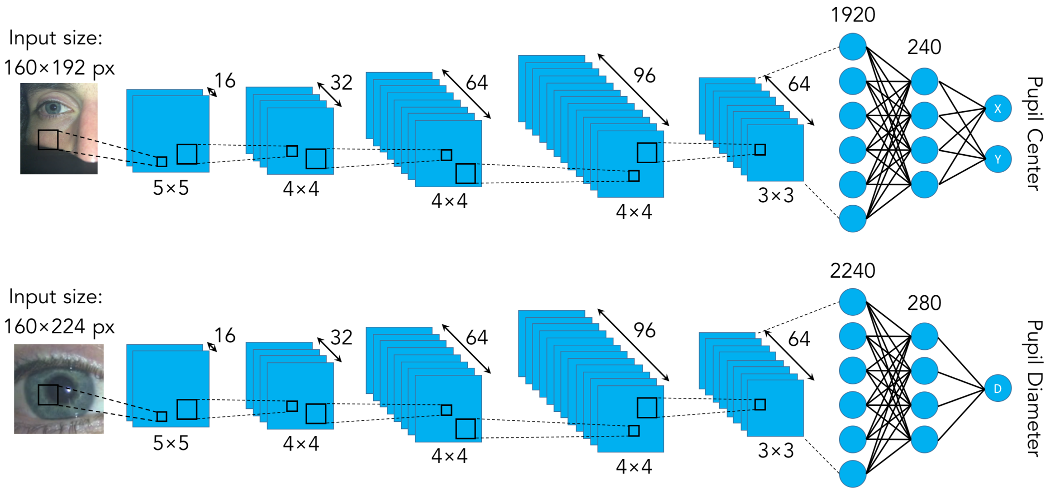

Fig. 7. The first architecture that was explored for PupilScreen. The top numbers indicate the number of filters in the convolutional layers or neurons in the fully-connected layers. The bottom numbers specify filter dimensions. For example, the first convolutional layer in both networks applies 16 5×5 px filters. There are 2×2 px mean-pooling layers after each convolutional layer, but they are omitted for space. **(top)** The first CNN takes the original image as an input and returns an estimate of the pupil's location. **(bottom)** Given the location of the pupil center, a region of interest is cropped from the original image and provided to the second CNN to estimate pupil diameter.

linear (ReLU) activation function followed by 2×2 px mean-pooling layers. Mean-pooling was chosen over max-pooling because max-pooling results in translation-independent behavior that would have been undesirable for capturing location information. The final layer of the first network is fully-connected to compress information across all filters and sub-regions to an x- and y-coordinate estimate. The output labels were normalized according to the mean and standard deviation of the pupil location across the entire dataset. This was done to ensure that the same error in either direction would equally affect the network's weights during backpropagation.

Using the output of the first network, a region of interest that is roughly 1/9$^{\text{th}}$ of the original image's size is cropped and centered about the estimated pupil. That region is provided to the second network (Fig. 7, bottom), which is trained to estimate the pupil diameter. The network has the same architecture as the first one except for the fact that it produces a single output: the pupil diameter.

The number of layers was determined empirically to balance the tradeoff between network size and accuracy. Smaller networks are desirable so that they can fit on the smartphone, but we found that using fewer layers did not yield satisfactory results. The other specifics of the networks (e.g., more smaller filters as the network gets deeper, pooling after each set of convolutional filters) were based on suggestions from literature [18], but are certainly an area for future investigation.

5.2.2 Second Architecture: Fully Convolutional. The first network architecture learns the pixel indices of the pupil center and the diameter of the pupil, but treats them just like any other continuous outputs rather than explicit location and size information. The second network architecture takes a different approach, viewing the problem as one of explicit segmentation. The goal of segmentation is to produce a label for every single pixel that specifies the object to which it belongs; as illustrated in Fig. 8, there are two classes for the purposes

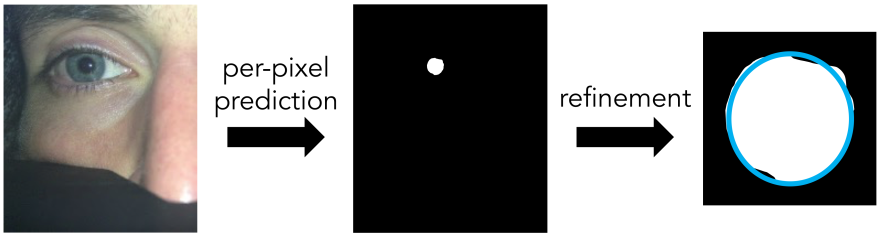

Fig. 8. The second architecture assigns each pixel to one of two classes: “pupil” (white) or “non-pupil” (black). The largest contiguous cluster of “pupil” pixels is assumed to be the pupil, and its border is smoothed so that it can be fit to an ellipse.

of PupilScreen: “pupil” and “non-pupil”. We implemented FCN-8, a fully convolutional architecture proposed by Long et al. [28]. In short, fully convolutional networks are normally based on a pre-trained convolutional network for image classification (e.g., VGG16 [43]). The final classifier layer is removed and replaced by layers that deconvolve, or upsample, the downsampled predictions to their original resolution. For the sake of network size, we downsample images by a factor of 2 before inputting them to the network.

Once pixelwise predictions are produced, there is still the matter of measuring a pupil diameter. The largest contiguous cluster of pixels with the “pupil” label is treated as the pupil. The border of that cluster is smoothed using median blurring and then fit to an ellipse. The mean of the ellipse’s two axes is treated as the pupil diameter for that frame.

5.3 Training

Both architectures were trained with backpropagation using batches composed of 10 images randomly sampled from the training set. To ensure that there was no overlap between training and testing data, the evaluation was conducted using 5-fold cross-validation across users; in other words, if there are N users, N/5 users are held out each time for testing and the remaining 4×N/5 users are used for training. Recall that three videos were recorded for each user. All networks were trained for 10 epochs per fold; this number was determined empirically based on the convergence of the smoothed loss function outputs across the training data. On average, training the first network architecture took 14 mins per fold, resulting in a total training time of 14 mins × 5 folds × 2 networks = 2 hours 20 mins. Training the second network architecture took 1 hours 59 mins per fold, resulting in a total training time of 119 mins × 5 folds = 9 hours 55 mins. Computation was carried out by a single NVidia GeForce Titan X GPU. Testing an individual frame through either network architecture took approximately 2 ms, which means that it would take the system roughly 2 ms × 30 frame/second × 5 seconds = 300 ms to test an entire video. The networks in the sequential CNN architecture were trained using batch gradient descent in order to minimize the L_2 loss. The fully convolutional network was trained in the same way to minimize the per-pixel multinomial logistic loss.

To ensure that the dataset was not significantly biased towards images of fully constricted pupils, only frames within the first 3 seconds of the light stimulus were used for training. To both generate more training samples and further promote training data diversity, training images and their associated labels were randomly jittered together (i.e., translated by a small amount). That amount was at most 10% of the input image dimensions for the first network, which was determined based on the variation of the pupil center observed in the videos. The jitter amount was at most 15% of the input image dimensions for the second network in order to sufficiently cover the spread of pupil center predictions from the first network. In this latter case, jittering the input images allows the second network to be trained to tolerate such errors.

5.4 Extracting PLR Metrics

In the end, the consecutive CNNs in PupilScreen take an individual image as input and return the pupil's diameter as output. A PLR curve shows a patient's pupil diameter as a function of time following a light stimulus. To construct this, videos are passed through the networks frame-by-frame. From that point, there are three post-processing steps to make the resulting curve more comparable to the curves provided by pupillometers: (1) Extreme prediction outliers are removed using heuristics based on human physiology: pupils should not be smaller than 1 mm or larger than 10 mm, and the pupil diameter should not change by more than 10 mm/s [6]. (2) Like the ground truth labels, the predictions are smoothed using a 3^{rd}-order Savitzky-Golay filter. This removes undesirable fluctuations between frames that occur because the pupil diameter is estimated from each frame individually. (3) Predictions are scaled from pixels to millimeters using a constant factor that was estimated through a device calibration procedure. A fiducial of known dimensions was placed in front of the camera at roughly the same distance as the user's eyes; its dimensions were measured in pixels and the calculated ratio was applied to all videos. This approach is not perfect since different people have different eye socket depths. Nevertheless, the ground truth labels used for analyses are all in pixels, so the conversion is primarily used to transform the results into more relevant units.

Relevant clinical measures (Section 2.1) can be extracted from the smoothed and scaled PLR curve. Calculations for the constriction amplitude and the constriction percentage require the minimum and maximum pupil diameter. The maximum pupil diameter always occurs at the beginning of the video since the pupil is most dilated before the light stimulus. After the pupil constricts, its diameter can fluctuate as it reaches its final equilibrium size. Because of this, the minimum diameter is identified by taking the average diameter in the last second. The maximum constriction velocity is calculated by computing the maximum of the centered derivatives across the entire curve. Although PupilScreen is designed to measure the latency between the time of the light stimulus and when the pupil begins to constrict, we found that the frame rate limits the granularity of the calculation ($(30 \text{ fps})^{-1} = 0.03$ s/frame) and the usefulness of that measure, so we ignore it for this study.

6 RESULTS

Since PupilScreen is a data-driven algorithm, the diversity of the data used to train the algorithm is important. Section 4.1 details the diversity of the participants, but in Section 6.1, we describe the quantitative diversity of the pupil center and diameter. We then present the accuracy of PupilScreen's ability to localize and measure the pupil with the two different architectures that were explored, followed by an examination of how the errors manifest in the PLR curves and affect the PLR metrics. We conclude with a brief evaluation of PupilScreen's clinical efficacy, including how accurately clinicians can make diagnostic decisions based on PupilScreen's estimated PLR curves and their comments on PupilScreen's design.

6.1 Data Distribution

The left side of Fig. 9 shows the distribution of the pupil center location across all users after the video frames were cropped, flipped, and scaled to millimeters. The distribution is centered at the mean pupil center for reference. The distribution has a standard deviation of 3.22 mm in the x-direction. This spread can be attributed to variation in interpupillary distance and the fact that participants did not perfectly align their face within the PupilScreen box. The distribution has a standard deviation of 4.18 mm in the y-direction, which can also be attributed to different face shapes and the placement of the PupilScreen box relative to the participant's face.

The right half of Fig. 9 shows the distribution of the pupil diameter scaled to millimeters. The distribution has a mean of 4.39 mm and a standard deviation of 1.38 mm. However, the distribution is non-normal because the

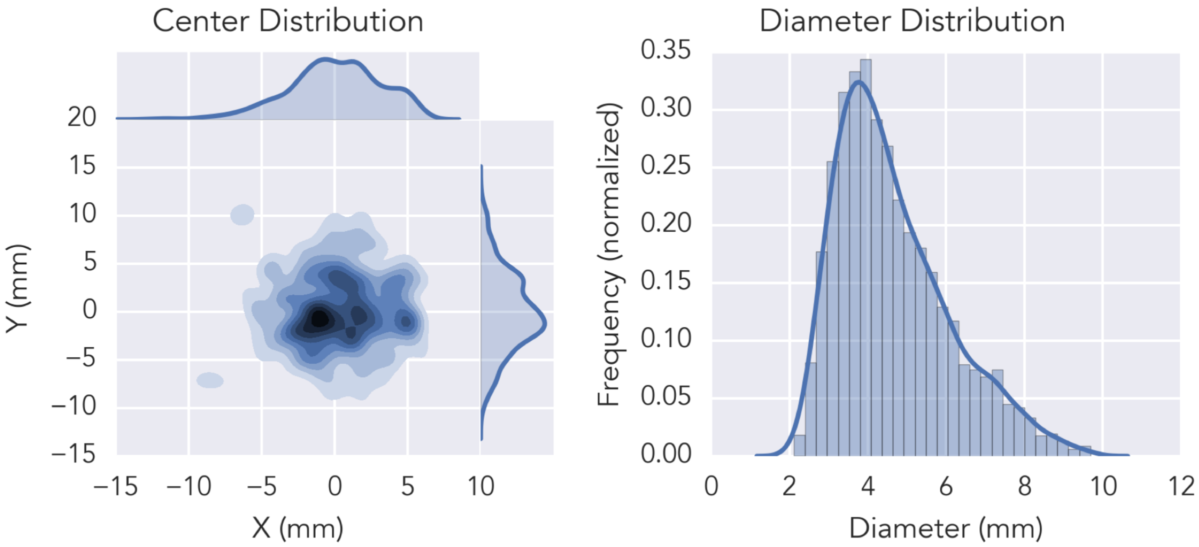

Fig. 9. **(left)** The distribution of the pupil centers across all users. **(right)** The distribution of the pupil diameters across all users.

pupil constricts in a logarithmic fashion, which means that the pupil only spends a small amount of time in its fully dilated state.

6.2 CNN Results

The cumulative distribution functions (CDFs) at the top of Fig. 10 show the distribution of the absolute errors for the sequential network architecture. The thick dashed line in both plots compares the results to a baseline that assumes the mean predictions for all users; this is not meant to serve as a comparable algorithm, but rather ground the results relative to some other estimator. Improvement over the baseline demonstrates that the networks are learning more than just the mean value.

The top-left of Fig. 10 shows the CDF for the errors of the first network, which estimates the pupil center for a cropped input video frame. Across all users, the distribution of Euclidean errors has a median of 0.43 mm and a 90th percentile of 0.87 mm. The error distributions across the different iris colors are nearly identical. The magnitude of the error can partly be attributed to the pre-processing of the video frame. Input images are downsampled by a factor of 4, which reduces the resolution of the pupil center estimation to 0.31 mm. Despite the loss of resolution, the errors are well within the diameter of the iris (10-12 mm). In fact, most are within the smallest observed pupil diameters (~2 mm). Although it is ideal for the pupil to be centered in the image that is input to the second network, the most important result is that the eye always remains in the region of interest that is cropped around the center prediction. By jittering the training data, the second network is trained to handle shifted images.

The top-right of Fig. 10 shows a similar CDF plot for the errors of the second network, which estimates the pupil diameter given an image cropped using the pupil center output by the first network. Across all users, the distribution of absolute errors has a median of 0.36 mm and a 90th percentile of 1.09 mm. According to Meeker et al. [35], the error of PupilScreen's diameter estimation is better than that of manual examination (0.5 mm), but worse than that of a clinical pupillometer (0.23 mm). To determine if the error of the first network leads to greater errors in the second network, we examined the accuracy of the second network given input images cropped around the ground truth pupil center. We found that there was little difference between using the predicted pupil centers and the ground truth pupil centers (50th: 0.36 mm, 90th: 1.19 mm vs. 50th: 0.36 mm, 90th: 1.15 mm). The fact that using the ground truth centers did not improve the accuracy of the pupil diameter estimation may be a byproduct of the fact that the training data was jittered, leading the network to be invariant to exact pupil location.

The Bland-Altmann plots in the bottom half of Fig. 10 show a different representation of the diameter prediction errors split across the different iris colors. In all cases, the sequential network architecture tends to overestimate the pupil diameter. If the CNN relies upon convolutional filters that look for edges, overestimation could be happening because those filters are more likely to respond to regions outside of the pupil's actual boundary. The mean pupil diameter errors are +0.24 mm, +0.27 mm, and +0.07 mm for blue, brown, and mixed eyes, respectively. We find that the most extreme outliers belong to a small subset of participants who had particularly dark irises. We believe that this error can be reduced with more training data from participants with similarly dark irises.

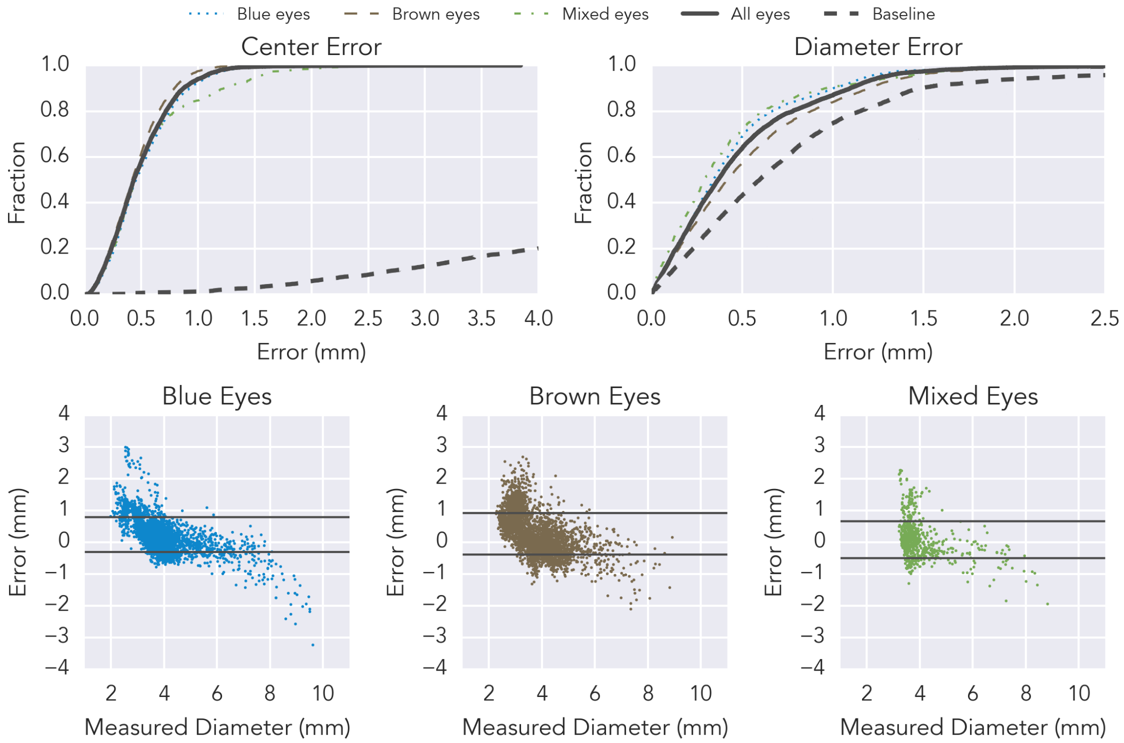

Fig. 10. The accuracy results for the sequential network architecture. **(top-left)** The CDF of the pupil center prediction error. **(top-right)** The CDF of the pupil diameter prediction error. **(bottom)** Bland-Altmann plots showing the residuals of the pupil diameter predictions split across the different iris colors: blue, brown, and mixed from left to right. The black lines indicate one standard deviation from the mean.

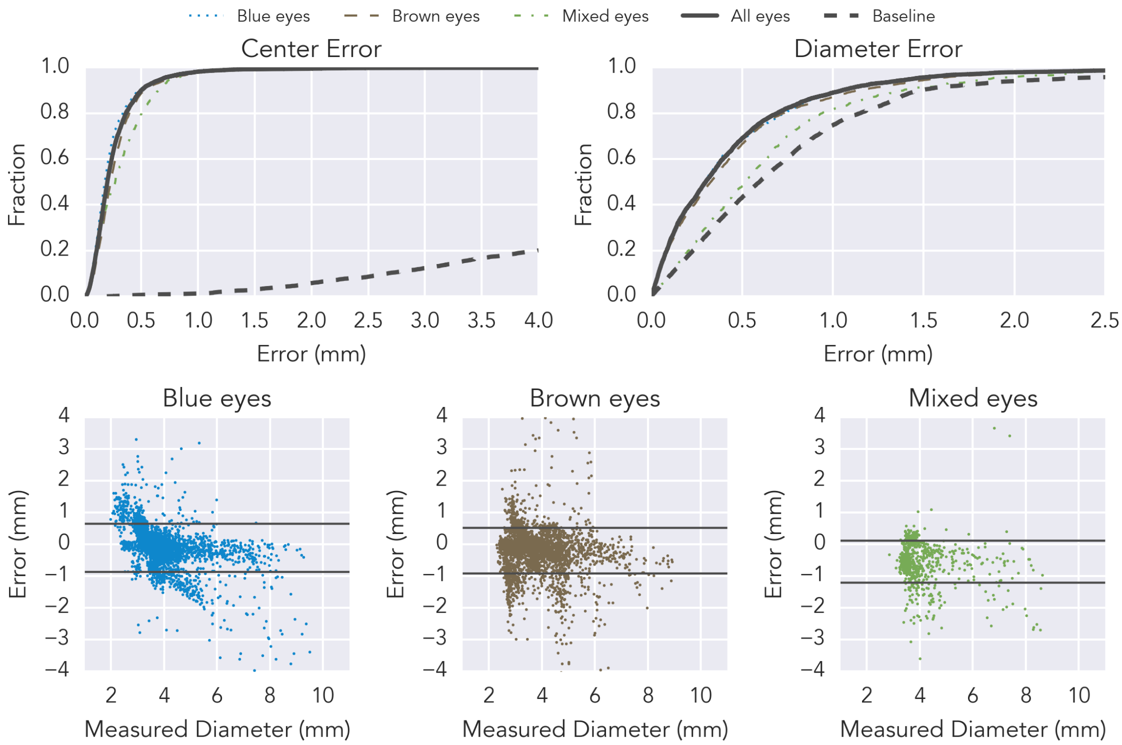

Fig. 11. The accuracy results for the fully convolutional architecture. **(top-left)** The CDF of the pupil center prediction error. **(top-right)** The CDF of the pupil diameter prediction error. **(bottom)** Bland-Altmann plots showing the residuals of the pupil diameter predictions split across the different iris colors: blue, brown, and mixed from left to right. The black lines indicate one standard deviation from the mean.

Fig. 11 shows the same performance measures for the fully convolutional architecture. The CDFs at the top of the figure show that the fully convolutional network was generally more accurate than using sequential networks. Across all users, the distribution of Euclidean errors for the pupil center has a median of 0.20 mm and a 90th percentile of 0.50 mm. The distribution of absolute errors for the pupil diameter has a median of 0.30 mm, which is closer to the observed accuracy of a clinical pupillometer than the 0.36 mm median error of the sequential network architecture. Examining the Bland-Altmann plots in Fig. 11, we find that the fully convolutional architecture tends to underestimate the pupil diameter. The mean pupil diameter errors are -0.11 mm, -0.20 mm, and -0.55 mm for blue, brown, and mixed eyes, respectively. Beyond the inherent differences between the two architectures from a deep learning standpoint, one reason for the improved results could be the fact that explicit morphological operations could be performed on the pixel labels; rather than hoping that the network could learn some attribute in regards to smooth edges, it is easier exercise domain-knowledge and enforce such rules afterwards. The post-processing could also explain why this architecture underestimated diameters; although smoothing can remove protrusions from a jagged pupil boundary estimate, it can also shrink an otherwise correct, smooth pupil boundary estimate.

There is a noticeable difference between the results for different iris colors. For both architectures, images of brown eyes led to the worst results. The sequential network architecture had a median error of 0.41 mm and a 90th percentile error of 1.19 mm, and the fully convolutional architecture had a median error of 0.33 mm and a 90th percentile error of 1.14 mm. This may be because the boundary between the pupil and the iris is less noticeable for people with darker irises, so the convolutional filters in the networks are less likely to respond to the appropriate regions of the eye. We also hypothesize that this is the reason for why the measured diameter error for brown eyes does not correlate with the pupil size as it does with the lighter iris colors, a phenomenon noted by Meeker et al. when pupils were manually examined.

6.3 Metric Evaluation

The outputs of PupilScreen's networks are irrelevant unless they are combined sequentially in PLR curves. For the sake of brevity, the results from here on out come from the fully convolutional architecture since it was slightly more accurate. To quantify how well the predicted PLR curves track the human-labeled PLR curves, their normalized cross-correlation was calculated. The average normalized cross-correlation across all videos is 0.91. Fig. 12 compares several examples of PLR curves produced by PupilScreen with ground truth PLR curves from manual annotations.

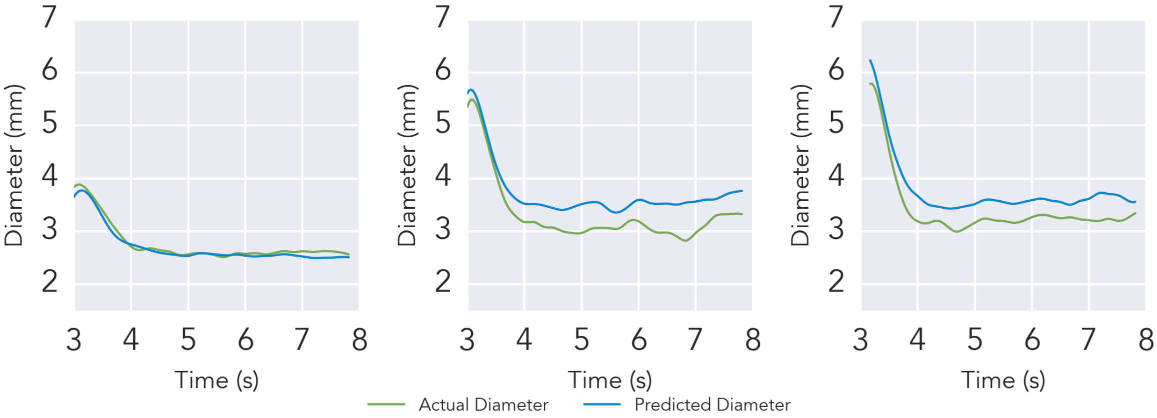

Fig. 12. Examples of predicted and ground truth PLR curves. **(left)** An example where PupilScreen accurately estimates all PLR metrics. **(center)** An example where PupilScreen accurately estimates the max constriction velocity, but underestimates the constriction amplitude and percentage. **(right)** An example where PupilScreen accurately estimates the constriction amplitude and max constriction velocity, but underestimates the constriction percentage.

Table 2 describes how well PupilScreen is able to predict PLR metrics relative to those measured from the manually labeled dataset. Table 2 also shows the range of those metrics across all participants as a point of comparison for the error magnitude. PupilScreen can track constriction amplitude with a mean error of 0.62 mm, constriction percentage within a mean error of 6.43%, and max constriction velocity with a mean error of 1.78 mm/s. As a point of comparison from the literature, an evaluation of PupilWare by Rafiqi et al. [42] demonstrated that their system tracked constriction and dilation percentages with an accuracy such that 90% of their predictions fell within 10% of the ground truth. However, there are many differences between PupilWare and PupilScreen that make these results difficult to compare. PupilScreen was evaluated on many more participants than PupilWare (42 vs. 9), but the evaluation of PupilWare aggregated a time series of percent change values rather than the single

Table 2. PLR metric evaluation

CONSTRICTION AMPLITUDE - mm	
Ground truth range	0.32-6.02
Mean absolute error	0.62
Standard deviation of absolute error	0.72
CONSTRICTION PERCENTAGE - %	
Ground truth range	6.21-62.00
Mean absolute error	6.43
Standard deviation of absolute error	6.74
MAX CONSTRICTION VELOCITY - mm/s	
Ground truth range	1.37-8.99
Mean absolute error	1.78
Standard deviation of absolute error	0.67

summary statistic like PupilScreen. The two systems are also intended for different applications. PupilWare is designed to track changes in pupil size attributed to varying cognitive load, which tend to be smaller in amplitude than the changes induced in PupilScreen.

Examining the predicted PLR curves further provides insight into the nature of these errors. The center and right plots in Fig. 12 show cases where a repeated error across frames led to the the inaccurate estimation of some PLR metrics, but not others. In the center, PupilScreen correctly tracks the pupil diameter during constriction, but then overestimates the final diameter of the pupil after constriction. The max constriction velocity is correctly estimated in these situations, but the constriction amplitude and percentage are not. On the right, PupilScreen follows the ground truth PLR curve with a roughly constant offset. This means that although the absolute estimate of the pupil diameter may be off, the change between the minimum and maximum pupil remains unchanged. This behavior only affects the constriction percentage since it relies on an absolute baseline; the constriction velocity and amplitude remain unaffected. Although not shown in Fig. 12, errors in all three metrics can also be attributed to pupil diameter predictions that deviated from nearby frames in a manner that failed PupilScreen's outlier criteria but were significant enough to create a deflection in the filtered PLR curve.

6.4 Preliminary Clinical Evaluation

To gauge PupilScreen's diagnostic efficacy, we supplemented our dataset with videos from six patients at Harborview Medical Center's trauma ward and neuro-intensive care unit (neuro-ICU). These individuals had sustained significant head trauma, but were stable enough at the time to be recruited for the study. Their doctors and nurses knew beforehand that they had non-reactive pupils. Non-reactive pupils are frequently observed in patients whose condition is unstable, making it difficult to use our research prototype without interfering with the clinician's workflow. As before, three videos were recorded for each patient; however, there were complications in collecting these videos, including the inability of the patients to keep their eyes open and the inability of the clinician to maintain the position of the box while recording the videos. Because of these issues, only 24 of the 36 possible PLR curves (3 videos per patient × 2 eyes per patient × 6 patients) were suitable for analysis.

To evaluate PupilScreen's accuracy on non-reactive pupils, we randomly selected one of the folds created during our initial training and analysis. The patient videos were processed using the CNNs that were trained on that fold's training data to produce pathologic PLR curves. An equal number of healthy PLR curves were generated using randomly selected videos from that fold's test set. Using the same network for both sets of videos guaranteed that the PLR curves were generated from networks that were trained on the same data. Fig. 13 shows examples of both responsive and non-responsive pupils that were collected with PupilScreen. The PLR curves from healthy individuals have a noticeable exponential decay, whereas the PLR curves from the patients do not.

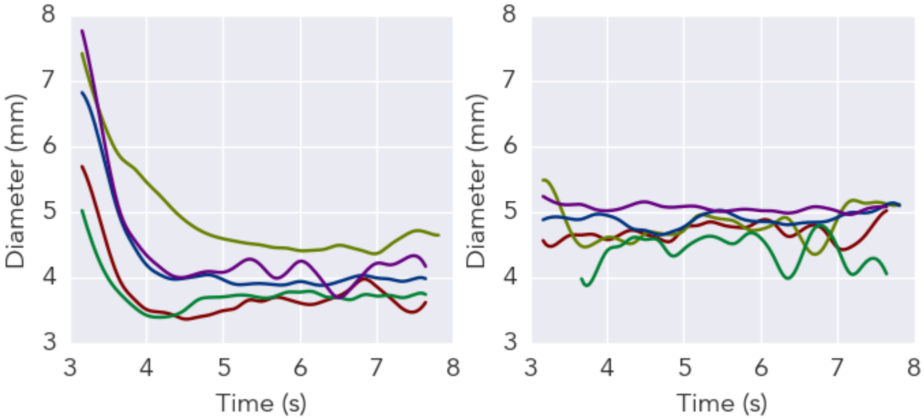

Fig. 13. A subset of **(left)** responsive and **(right)** non-responsive PLR curves that were shown to clinicians for our preliminary clinical evaluation.

The PLR curves were anonymized, shuffled, and then sent to two clinicians familiar with pupillometry. The clinicians were asked to classify the PLRs as either "responsive" or "non-responsive". They were not told how many curves would be in each category, nor were they shown the video recordings themselves. The first clinician was able to correctly classify every curve in our dataset. The second clinician misclassified one non-responsive PLR curve as responsive. In that particular case, PupilScreen estimated that the person's pupil constricted in a slow and almost linear manner, but by a significant amplitude. The second clinician also misclassified one responsive PLR curve as non-responsive, again, due to the borderline pupil constriction amplitude.

6.5 Clinician Feedback

Throughout our design process, we asked clinicians about their personal experiences with pupillometry and for feedback on PupilScreen's design. These clinicians included surgeons, nurses, and other personnel at the Harborview Medical Center's neuro-ICU. Although PupilScreen is proposed as a tool to be used by team doctors and parents, clinicians who work with TBI are far more familiar with existing pupillometry methods and their tradeoffs and could provide far more insight beyond novelty.

One of the surprising findings early on was that although the clinicians were familiar with the purpose of a pupillometer and its advantages over a penlight test, the pupillometer was hardly used in the clinical setting. The pupillometer was mainly used to track changes in PLR over a long period of time to identify worsening injuries as quickly as possible in otherwise unresponsive patients. For diagnosis or triage, penlights are strongly preferred for their simplicity and ease of access, despite the limited precision and lack of consistency they afford. As one clinician stated, "If whatever you ask an EMT to do adds twenty seconds or so, it's not worth it". In fact, we found that some clinicians use their smartphone's flash instead of a penlight, validating aspects of our idea.

When we asked the clinicians about the prospect of PupilScreen's convenience, they were excited by the idea of a smartphone app that would be in their pockets at all times. Unsurprisingly, clinicians pointed out that the PupilScreen box was still a bulky object that needed to be carried to conduct the test, although some reasoned that it would be far cheaper to place multiple boxes in the neuro-ICU than multiple pupillometers. One clinician recommended a foldable box that would be easier to transport. Another clinician suggested a monocular design that would record one eye at a time; such a system would still require a separate component from the phone, but it would be roughly half the size of the PupilScreen box. The most popular suggestion was a system where no box was required at all. Eliminating the box would make PupilScreen even more convenient than a penlight, but removing the box eliminates control over lighting, which is crucial for ensuring that the pupil is visible and that the light stimulus provided to the eyes is standardized. Nonetheless, this is a goal that we hope to strive for in the future.

Another issue raised about PupilScreen's design is the difficulty of using PupilScreen on patients who are unconscious. In the sports-related concussion scenario, the cases that most warrant the use of pupillometry are when the patients are conscious and can comply with most verbal instructions. In the neuro-ICU, penlights and pupillometers are often used on unconscious patients, and clinicians must hold those patients' eyelids open with one hand in order to expose the pupil(s). This is a manageable, but clumsy maneuver to conduct with the PupilScreen box. Manipulating the patient's face in this manner can also allow extra light to seep in from the top of the box, which reduces the control over the lighting within it.

From our interviews, we believe that PupilScreen's design will be suitable for use by team doctors and parents, but requires further improvement for use by EMTs and other hospital clinicians.

7 DISCUSSION

Our goal was to develop a system that could quantitatively assess the severity of TBIs by measuring a person's pupillary light reflex. Furthermore, we imposed the requirements that the system should be automated and easy to deploy. We believe that PupilScreen is the first step toward these goals. The PupilScreen box allows anyone to use their phone as an inexpensive pupillometer. It does so by blocking out ambient lighting while allowing the smartphone to provide its own light stimulus from the flash. Using two sequential CNNs, PupilScreen measures the pupil center with a median error of 0.43 mm and the pupil diameter with a median error of 0.36 mm. Using a fully convolutional network, PupilScreen achieves median errors of 0.20 mm and 0.30 mm for those same two measures, respectively. Once we found that PupilScreen could track the PLR with reasonable accuracy, we conducted a preliminary clinical trial with six patients who had suffered a TBI. When clinicians were given PLR curves from both healthy and injured individuals, they were almost always able to reach the correct diagnosis.

7.1 Hardware

The low-fidelity nature of the PupilScreen box has advantages and disadvantages. The only requirements on the box were that it needed to block out light from the environment and that it allowed the smartphone's flash to illuminate the patient's eyes. A variety of materials for the box could have satisfied these requirements. We 3D-printed the box using PLA plastic for durability over the course of the study. The PupilScreen box could easily be mass-produced using injection molding for similar results. Since the box does not require any embedded electronics outside of the user's smartphone, people can even construct their own PupilScreen box using stiff cardboard. This last idea is particularly enticing because it could allow for the generalization of our system throughout the diverse smartphone ecosystem. The PupilScreen box used in the study was made specifically with iPhones in mind since they have a more unified design. Later models (iPhone 4 or after) have both the camera and flash on the top-left corner at the back of the phone, which lent itself to the design shown in Fig. 4. Android phones come in all sorts of different configurations and shapes, which would require a dedicated box design for each model or a configurable box to cover all of them.

Beyond the design of the PupilScreen box, the diverse smartphone ecosystem could influence the diagnostic efficacy of PupilScreen, although we believe these effects would be minimal. Different smartphone models may have different flash LEDs, but most are bright enough to cause a similarly significant PLR. PupilScreen could tune its thresholds for various PLR metrics based on information about the flash LED that can be stored in a lookup table. There is larger variation in smartphone cameras across specifications, including sensitivity and resolution. Cameras can respond to various wavelengths of light in different ways. We believe this should have minimal impact on PupilScreen's CNN-based approach since the convolutional filters should still respond in a similar manner if preprocessing or calibration can be employed to standardize input frames. With regards to camera resolution, a higher resolution translates to a higher pixel-per-mm ratio given a fixed camera placement and focus. A higher pixel-per-mm ratio allows PupilScreen to detect smaller changes in pupil diameter and measure the PLR with increased precision. In cases when the resolution is too low, PupilScreen could incorporate a zooming procedure that maximizes the pixel-per-mm ratio without sacrificing focus. However, too much variability in resolution could lead to issues since the filters in PupilScreen's CNNs have fixed pixel sizes and may be trained to only recognize contours within certain scales.

By relinquishing lighting control to the smartphone, the current PupilScreen design is limited in what kind of responses it can capture. In our evaluation, we only examined pupil constriction, not dilation. This is because there is no intermediate lighting state between the on and off stages of the smartphone's flash, and when the flash is off, the camera cannot see the patient's eyes. Some smartphone models are beginning to provide multiple flash LEDs (e.g., iPhone 6), but we found there was not enough of a difference between them to induce significant pupil diameter changes. Early in our design phase, we briefly experimented with using the smartphone's screen as the lighting source. We decided against this design because most smartphone screens are not sufficiently bright to make the eyes visible within the PupilScreen box. Furthermore, since most front-facing cameras are on the corner of the smartphone, the screen illuminates the patient's face at an angle when the camera is centered between their eyes. This can form a light gradient across the patient's face, or worse, a shadow on an eye, creating undesirable noise in the data.

7.2 Software

One might argue that we did not collect enough data to sufficiently train the networks' thousands of parameters. We attempted to mitigate some of these issues by jittering our data during training and starting with a pre-trained network in the case of the fully convolutional architecture; however, we recognize that more data is always better. Beyond collecting more data in the same manner as we have in the past, we plan to incorporate synthetic datasets, such as SynthesEyes by Wood et al. [53], to develop a more diverse dataset. We may also explore ways of scraping the web for images to further bolster our dataset.

There is more exploration left to be done concerning the optimal CNN architecture for identifying the pupil. One drawback from using CNNs on individual video frames in general is that consecutive frames are treated independently until predictions are combined for the PLR curve. This approach does not account for the fact that the pupil changes size continuously and, therefore, nearby frames should have correlated pupil diameters. PupilScreen uses low-pass filtering to reduce unnecessary variation between nearby frames. Another way to account for frame continuity would have been to use an algorithm that trains on entire sequences, such as a continuous-time recurrent neural network. We chose not to do this because it requires a significant number of examples for both reactive and non-reactive pupils, which would only be feasible with a larger deployment. There is also the possibility that such an approach could bias towards learning the typical PLR, leading to diagnostic false negatives. Although using two sequential CNNs led to slightly worse results, the full range of possible structures for those networks was not explored. As pointed out by Chellapa [53], factors related to network size (e.g., memory footprint, number of parameters, training time) are still an open challenge in the deep learning community. Staying up to date with advancements in that field while focusing on the our specific task will be important for eventually moving PupilScreen to a configuration that does not require a server.

Most of our participants complied with PupilScreen's procedure, meaning that they blinked as little as possible and kept their gaze toward the camera. These constraints are also imposed by pupillometers; if the patient does not comply, the pupillometer rejects the trial and requests a retest. Both pupillometers and PupilScreen currently handle blinking in different ways that lead to similar results. Pupillometers explicitly localize the pupil using infrared light. If they cannot find the pupil, the PLR curve for those frames has a null value. As long are there are not too many null values in the PLR curve, the pupillometer interpolates the pupil diameter for those frames. PupilScreen does not include an explicit blinking detection step, so all frames are tested through the CNNs regardless of the whether the pupil is visible in them or not. That being said, the CNNs are only trained on images where the pupil is visible, so cases when the pupil is not visible lead to outlier results that are handled through the post-processing described in Section 5.4. We found that cases of blinking were not a significant source of error in PupilScreen's results, but a blink detector [36] could be incorporated at the beginning of PupilScreen's pipeline so that irrelevant frames are accounted for sooner.

Handling different gaze directions is a simpler matter for both PupilScreen and pupillometers. Pupillometers fit an ellipse, not a circle, to the pupil. If the ellipse's eccentricity is too low (e.g., its axes are uneven), the frame is rejected just as a frame with a blink. The data for PupilScreen was also originally labeled as ellipses. The elliptical labels were converted to a circular representation where the diameter was defined as the average of the ellipse's axes, so the CNNs are trained to interpret the ellipses in that manner. The maximum of the ellipse's axes could have been a better summary of the pupil since the dimension parallel to the direction of the rotation decreases in size; however, we chose to use the mean as a compromise between this phenomenon and the fact that some pupils have small protrusions along their perimeter that artificially extend their clinically significant boundary.

7.3 Future Applications

PupilScreen is primarily targeted toward individuals interested in assessing the severity of head trauma, whether it be a high school coach checking for concussions or an EMT checking the extent of a more general TBI. Zafar and Suarez [55] note that most of the studies involving the diagnostic significance of pupillometry are limited due to small sample sizes. The clinical study we conducted has the same issue since it only included six individuals who had suffered significant head trauma. We were limited to individuals who were in a stable condition because clinicians were hesitant of introducing yet another instrument into their workflow during time-critical situations. Following their suggestions, we plan to explore the possibility of removing the PupilScreen box. Rather than imagining PupilScreen as an inexpensive pupillometer, removing the box would turn PupilScreen into a more quantitative penlight exam, sacrificing consistency and standardization in favor of convenience. Ensuring that the penlight exam is conducted in a reasonable manner would become the responsibility of the user interface. Visual guides could show an inexperienced user how close the phone should be from the patient's face, and feedback could be provided if the pupils were not sufficiently stimulated by the light.

We believe that by making pupillometry more accessible in this manner, we can enable researchers to reassess previous studies with greater sample sizes. In fact, we plan to conduct a follow-up study looking at the correlation between PupilScreen, a clinical-grade pupillometer, and the tools currently used by American football teams for assessing concussions (e.g., the King-Devick test and the SCAT). We also plan on examining how our technique can be used to check for other eye-related conditions that may indicate a TBI, such as involuntary eye movement [27] and poor visual tracking performance [31, 32].

8 CONCLUSION

Traumatic brain injuries require rapid assessment to ensure that the proper measures are put into place to maximize a patient's chances of a recovery. Measuring the pupillary light reflex (PLR) is a quantitative method that screens for injury to multiple brain regions. Although methods exist for performing such a test, they are either expensive or inexact. To this end, we have presented PupilScreen, a smartphone app and accompanying 3D-printed box that enables anyone to automatically measure a person's PLR. PupilScreen relies on convolutional neural networks to estimate pupil diameter within a video as the smartphone's flash causes the pupil to constrict.

We evaluated two different architectures on data collected from 42 healthy individuals and found that using a fully convolutional network could achieve a median error of 0.30 mm when measuring the pupil diameter. We also conducted a pilot clinical evaluation with six patients who had suffered a TBI, in which clinicians were able to correctly differentiate between patients and healthy individuals with almost perfect accuracy using PupilScreen's output alone. It is our hope that PupilScreen will enable researchers to conduct additional future studies evaluating the diagnostic significance of the PLR.

9 ACKNOWLEDGMENTS

We thank the National Science Foundation and the ARCS Foundation for their funding for the graduate students on the project. We also thank Ruth Ravichandran, Edward Wang, Elliot Saba, and others who provided their feedback on the work.

REFERENCES

[1] Alireza Abdolvahabi, Brandon W. Taylor, Rebecca L. Holden, Elizabeth V Shaw, Alex Kentsis, Carlos Rodriguez-Galindo, Shizuo Mukai, and Bryan F. Shaw. Colorimetric and longitudinal analysis of leukocoria in recreational photographs of children with retinoblastoma. *PloS one*, 8(10):e76677, oct 2013.

[2] Andrew Bastawrous, Hillary K. Rono, Iain A. T. Livingstone, Helen A. Weiss, Stewart Jordan, Hannah Kuper, and Matthew J. Burton. Development and Validation of a Smartphone-Based Visual Acuity Test (Peek Acuity) for Clinical Practice and Community-Based Fieldwork. *JAMA Ophthalmology*, 133(8):930, aug 2015.

[3] Matthias Behrends, Claus U Niemann, and Merlin D Larson. Infrared pupillometry to detect the light reflex during cardiopulmonary resuscitation: A case series. *Resuscitation*, 83(10):1223–1228, 2012.

[4] Barry P Boden, Robin L Tacchetti, Robert C Cantu, Sara B Knowles, and Frederick O Mueller. Catastrophic Head Injuries in High School and College Football Players. *The American Journal of Sports Medicine*, 35(7):1075–1081, mar 2007.

[5] Steven P Broglio, Michael S Ferrara, Stephen N Macciocchi, Ted A Baumgartner, and Ronald Elliott. Test-retest reliability of computerized concussion assessment programs. *Journal of Athletic Training*, 42(4):509–514, 2007.

[6] Jose E Capó-Aponté, Thomas G Urosevich, David V Walsh, Leonard A Temme, and Aaron K Tarbett. Pupillary Light Reflex as an Objective Biomarker for Early Identification of Blast-Induced mTBI. *Journal of Spine*, 2013.

[7] Centers for Disease Control. TBI: Get the Facts, 2016.

[8] David Couret, Delphine Boumaza, Coline Grisotto, Thibaut Triglia, Lionel Pellegrini, Philippe Ocquidant, Nicolas J Bruder, and Lionel J Velly. Reliability of standard pupillometry practice in neurocritical care: an observational, double-blinded study. *Critical Care*, 20(1):99, dec 2016.

[9] Mark Faul, Likang Xu, MM Wald, and VG Coronado. Traumatic Brain Injury in the United States. Technical report, 2010.

[10] JD Fischer and DJ van den Heever. Portable video-oculography device for implementation in side-line concussion assessments: a prototype. In *Proc. EMBC '16*, 2016.

[11] Wolfgang Fuhl, Thomas Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. ExCuSe: Robust Pupil Detection in Real-World Scenarios. In *International Conference on Computer Analysis of Images and Patterns*, pages 39–51. Springer International Publishing, 2015.

[12] Wolfgang Fuhl, Thiago Santini, Gjergji Kasneci, and Enkelejda Kasneci. PupilNet: Convolutional Neural Networks for Robust Pupil Detection. (1-10), 2016.

[13] Wolfgang Fuhl, Thiago C Santini, Thomas Kübler, and Enkelejda Kasneci. ElSe : Ellipse Selection for Robust Pupil Detection in Real-World Environments. In *Eye Tracking Research & Applications*, pages 123–130, 2016.

[14] Kristin M Galetta, Lauren E Brandes, Karl Maki, Mark S Dziemianowicz, Eric Laudano, Megan Allen, Kathy Lawler, Brian Sennett, Douglas Wiebe, Steve Devick, Leonard V Messner, Steven L Galetta, and Laura J Balcer. Journal of the Neurological Sciences The King-Devick test and sports-related concussion : Study of a rapid visual screening tool in a collegiate cohort. *Journal of the neurological sciences*, 309(1-2):34–39, 2011.

[15] Kristin M Galetta, Mengling Liu, Danielle F Leong, Rachel E Ventura, Steven L Galetta, and Laura J Balcer. The King-Devick test of rapid number naming for concussion detection: meta-analysis and systematic review of the literature. *Concussion*, 1(2):cnc.15.8, 2015.

[16] Matthew S. Galetta, Kristin M. Galetta, Jim McCrossin, James A Wilson, Stephen Moster, Steven L Galetta, Laura J Balcer, Gary W Dorshimer, and Christina L Master. Saccades and memory: Baseline associations of the King-Devick and SCAT2 SAC tests in professional ice hockey players. *Journal of the Neurological Sciences*, 328(1-2):28–31, 2013.

[17] Mario E Giardini, Iain A T Livingstone, Stewart Jordan, Nigel M Bolster, Tunde Peto, Matthew Burton, and Andrew Bastawrous. A smartphone based ophthalmoscope. *Proc. EMBC '14*, 2014:2177–2180, 2014.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[19] Kimberly G Harmon, Jonathan A Drezner, Matthew Gammons, Kevin M Guskiewicz, Mark Halstead, Stanley A Herring, Jeffrey S Kutcher, Andrea Pana, Margot Putukian, and William O Roberts. American Medical Society for Sports Medicine position statement: concussion in sport. *British Journal of Sports Medicine*, 47(1):15–26, jan 2013.

[20] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct*, pages 1151–1160, New York, New York, USA, 2014. ACM Press.

[21] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. Measuring the task-evoked pupillary response with a remote eye tracker. *Proceedings of the 2008 symposium on Eye tracking research & applications - ETRA '08*, 1(212):69, 2008.

[22] Merlin D Larson and Matthias Behrends. Portable Infrared Pupillometry. *Anesthesia & Analgesia*, 120(6):1242–1253, 2015.

[23] Merlin D Larson and Isobel Muhiudeen. Pupillometric analysis of the 'absent light reflex'. *Archives of neurology*, 52(4):369–72, 1995.

[24] Everett Lawson, Jason Boggess, Siddharth Khullar, Alex Olwal, Gordon Wetzstein, and Ramesh Raskar. Computational retinal imaging via binocular coupling and indirect illumination. In *Proc. SIGGRAPH '12*, page 51, 2012.

[25] Hopin Lee, S John Sullivan, Anthony G Schneiders, Osman Hassan Ahmed, Arun Prasad Balasundaram, David Williams, Willem H Meeuwisse, and Paul McCrory. Smartphone and tablet apps for concussion road warriors (team clinicians): a systematic review for practical users. *British journal of sports medicine*, 49(8):1–2, apr 2014.

[26] Dongheng Li, David Winfield, and Derrick J Parkhurst. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, volume 3, pages 79–79. IEEE, 2005.

[27] Joiiiv R Lindsay. The significance of a positional nystagmus in otoneurological diagnosis. *The Laryngoscope*, 55(10):527–551, 1945.

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 3431–3440, mar 2015.

[29] Alex Mariakakis, Megan A. Banks, Lauren Phillipi, Lei Yu, James Taylor, and Shwetak N. Patel. Biliscreen: smartphone-based scleral jaundice monitoring for liver and pancreatic disorders. In *Proceedings of the 2017 ACM Interactive, Mobile, Wearable, Ubiquitous Technologies*. ACM, 2017.

[30] F Martínez-Ricarte, A Castro, MA Poca, J Sahuquillo, L Expósito, M Arribas, and J Aparicio. Infrared pupillometry. Basic principles and their application in the non-invasive monitoring of neurocritical patients. *Neurología (Barcelona, Spain)*, 28(1):41–51, 2013.

[31] Jun Maruta, Stephanie W. Lee, Emily F. Jacobs, and Jamshid Ghajar. A unified science of concussion. *Annals of the New York Academy of Sciences*, 1208(1):58–66, oct 2010.

[32] Jun Maruta, J L Tong, Stephanie W Lee, Zarah Iqbal, Alison Schonberger, and Jamshid Ghajar. EYE-TRAC: monitoring attention and utility for mTBI. *Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring Ii: And Biometric Technology for Human Identification Ix*, 8371:11, 2012.

[33] Alex Marvez. Players could try to beat concussion tests — FOX Sports, 2011.

[34] Paul McCrory, Willem H Meeuwisse, Mark Aubry, Bob Cantu, Jiri Dvorak, Ruben Echemendia, Lars Engebretsen, Karen Johnson, Jeffery S Kutcher, Martin Raftery, Allen Sills, Brian W Benson, Gavin A Davis, Richard G Ellenbogen, Kevin Guskiewicz, Stanley A Herring, Grant L Iverson, Barry D Jordan, James Kissick, Michael McCrea, Andrew S McIntosh, David Maddocks, Michael Makdissi, Laura Purcell, Margot Putukain, Kathryn Schneider, Charles H Tator, and Michael Turner. Consensus Statement on Concussion in Sport: The 4th International Conference on Concussion in Sport Held in Zurick, November 2012. *British Journal of Sports Medicine*, 47(2):250–258, apr 2013.

[35] Michele Meeker, Rose Du, Peter Bacchetti, Claudio M Privitera, Merlin D Larson, Martin C Holland, and Geoffrey Manley. Pupil examination: validity and clinical utility of an automated pupillometer. *The Journal of neuroscience nursing : journal of the American Association of Neuroscience Nurses*, 37(1):34–40, 2005.

[36] Tim Morris, Paul Blenkhorn, and Farhan Zaidi. Blink detection for real-time eye tracking. *Journal of Network and Computer Applications*, 25(2):129–143, 2002.

[37] Rosemarie Scolaro Moser, Grant L Iverson, Ruben J Echemendia, Mark R Lovell, Philip Schatz, Frank M Webbe, Ronald M Ruff, Jeffrey T Barth, Donna K. NAN Policy and Planning Committee, Shane S. Donna K. Broshek, Shane S. Bush, Sandra P. Koffler, Cecil R. Reynolds, Cheryl H. Silver, Sandra P. Koffler, Cecil R. Reynolds, and Cheryl H. Silver. Neuropsychological evaluation in the diagnosis and management of sports-related concussion. *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists*, 22(8):909–16, nov 2007.

[38] Dai-Wai M Olson, Sonja Stutzman, Ciji Saju, Margaret Wilson, Weidan Zhao, and Venkatesh Aiyagari. Interrater Reliability of Pupillary Assessments. *Neurocritical Care*, 24(2):251–257, apr 2016.

[39] Vitor F Pamplona, Ankit Mohan, Manuel M Oliveira, and Ramesh Raskar. NETRA: interactive display for estimating refractive errors and focal range. *ACM transactions on graphics (TOG)*, 29(4):77, 2010.

[40] Vitor F Pamplona, Erick B Passos, Jan Zizka, Manuel M Oliveira, Everett Lawson, Esteban Clua, and Ramesh Raskar. Catra: cataract probe with a lightfield display and a snap-on eyepiece for mobile phones. In *Proc. SIGGRAPH '11*, pages 7–11, 2011.

[41] Stephen M Pizer, E. Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.

[42] Sohail Rafiqi, Chatchai Wangwiwattana, Jasmine Kim, Ephrem Fernandez, Suku Nair, and Eric C. Larson. PupilWare: towards pervasive cognitive load measurement using commodity devices. In *Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments - PETRA '15*, pages 1–8, New York, New York, USA, 2015. ACM Press.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] Lech Swirski, Andreas Bulling, and Neil Dodgson. Robust real-time pupil tracking in highly off-axis images. *Etra*, pages 1–4, 2012.

[45] William R Taylor, Jeff W Chen, Hal Meltzer, Thomas A Gennarelli, Cynthia Kelbch, Sharen Knowlton, Jenny Richardson, Matthew J Lutch, Azadeh Farin, Kathryn N Hults, and Lawrence F Marshall. Quantitative pupillometry, a new technology: normative data and preliminary observations in patients with acute head injury. Technical note. *Journal of neurosurgery*, 98(1):205–213, 2003.

[46] Preethi Thiagarajan and Kenneth J Ciuffreda. Pupillary responses to light in chronic non-blast-induced mTBI. *Brain Injury*, 29(12):1420–1425, oct 2015.

[47] Fabian Timm and Erhardt Barth. Accurate Eye Centre Localisation by Means of Gradients. In *VISAPP*, pages 125–130, 2011.

[48] James Q Truong and Kenneth J Ciuffreda. Comparison of pupillary dynamics to light in the mild traumatic brain injury (mTBI) and normal populations. *Brain Injury*, 30(11):1378–1389, sep 2016.

[49] James Q Truong and Kenneth J Ciuffreda. Objective Pupillary Correlates of Photosensitivity in the Normal and Mild Traumatic Brain Injury Populations. *Military Medicine*, 181(10):1382–1390, oct 2016.

[50] James Q Truong and Kenneth J Ciuffreda. Quantifying pupillary asymmetry through objective binocular pupillometry in the normal and mild traumatic brain injury (mTBI) populations. *Brain Injury*, 30(11):1372–1377, sep 2016.

[51] James Q Truong, Nabin R Joshi, and Kenneth J Ciuffreda. Influence of refractive error on pupillary dynamics in the normal and mild traumatic brain injury (mTBI) populations. *Journal of Optometry*, mar 2017.

[52] H Kenneth Walker, W Dallas Hall, and J Willis Hurst. *Clinical Methods*. Butterworths, 1990.

[53] E Wood, T Baltruaitis, X Zhang, Y Sugano, P Robinson, and A Bulling. Rendering of Eyes for Eye-Shape Registration and Gaze Estimation, 2015.

[54] Erroll Wood and Andreas Bulling. Eyetab: Model-based gaze estimation on unmodified tablet computers. *Etra*, pages 3–6, 2014.

[55] Sahar F Zafar and Jose I Suarez. Automated pupillometer for monitoring the critically ill patient: A critical appraisal, 2014.

Received November 2017; accepted June 2017