PupilNet, Measuring Task Evoked Pupillary Response using Commodity RGB Tablet Cameras: Comparison to Mobile, Infrared Gaze Trackers for Inferring Cognitive Load

CHATCHAI WANGWIWATTANA*, XINYI DING, and ERIC C. LARSON, Southern Methodist University, USA

Pupillary diameter monitoring has been proven successful at objectively measuring cognitive load that might otherwise be unobservable. This paper compares three different algorithms for measuring cognitive load using commodity cameras. We compare the performance of modified starburst algorithm (from previous work) and propose two new algorithms: 2 Level Snakuscules and a convolutional neural network which we call PupilNet. In a user study with eleven participants, our comparisons show PupilNet outperforms other algorithms in measuring pupil dilation, is robust to various lighting conditions, and robust to different eye colors. We show that the difference between PupilNet and a gold standard head-mounted gaze tracker varies only from -2.6% to 2.8%. Finally, we also show that PupilNet gives similar conclusions about cognitive load during a longer duration typing task.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → **Neural networks**; • **Computer systems organization** → *Sensors and actuators*;

Additional Key Words and Phrases: Cognitive Load, Pupillary Response, Machine Learning

ACM Reference Format:
Chatchai Wangwiwattana, Xinyi Ding, and Eric C. Larson. 2017. PupilNet, Measuring Task Evoked Pupillary Response using Commodity RGB Tablet Cameras: Comparison to Mobile, Infrared Gaze Trackers for Inferring Cognitive Load . *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 171 (December 2017), 26 pages. https://doi.org/10.1145/3161164

1 INTRODUCTION

Pupillary monitoring is a well-established tool used in psychology for measuring stress, emotional arousal, cognitive load, and a number of different factors objectively. Pupillary monitoring is used in varied scenarios because pupillary diameter is controlled by the automatic nervous system. The automatic nervous system is involved in a number of behavioral operations of the body. Thus, when behavioral stimulus exhibits in the autonomic nervous system, dilations and constrictions of the pupil can also result. For example, stimuli that exhibit sexual arousal [44] or stress [24] can dilate the pupil independently of the amount of environmental lighting. In his seminal work, Kahneman demonstrates pupillary monitoring can objectively measure cognitive load—pupillary diameter increases during high cognitive load and decreases when those tasks are finished processing. This phenomenon is known as task-evoked pupillary response [7, 9]. Monitoring pupil dilations, then, can help machines to understand when an individual is experiencing high cognitive load. To be sure,

*The corresponding author

Authors' address: Chatchai Wangwiwattana, cwangwiwatta@smu.edu; Xinyi Ding, xding@smu.edu; Eric C. Larson, Southern Methodist University, School of Engineering, Dallas, TX, 75205, USA.

2474-9567/2017/12-ART171 $15.00
https://doi.org/10.1145/3161164

pupil monitoring is an extremely coarse measure of the thought process for an individual, but can serve as a physiological indicator that a variety of behavioral processes might be occurring. In this way, pupillary monitoring can be seen as another tool for cognitive, affective, and contextual computing.

In our previous work, we piloted experiments showing that sub-millimeter pupil dilations can be tracked from commodity web-cameras and smartphone cameras [42, 43]. The ultimate goal was to characterize whether pupil monitoring from commodity cameras was feasible in day-to-day ubiquitous applications, such as while commonly interacting with a smartphone. However, many research questions remained including evaluations in different lighting conditions and with different levels of common occlusion expected from using a forward facing mobile phone camera. Furthermore, our pilot experiments only ascertained pupil dilations for one specific task (difficulty of memorization) and did not perform well on individuals with darkly colored eyes. It was unclear how to apply these measures of pupil dilation in longer duration, more cognitively complex tasks, such as when an individual is learning. This paper dramatically extends the analysis of our original work to include more systematic evaluation of pupil monitoring in different lighting, occlusions, and more realistic situations. Furthermore, we present methods for analyzing pupil dilation when individuals practice typing, which is a longer duration and cognitively more complex task. Because of the complexity of our evaluation, we completely redesigned the algorithms in our original work, incorporating convolutional neural networks that extract pupil response from low-resolution RGB images of the eyes. Because the new system is markedly different from our original design (called PupilWare), we change its name to **PupilNet**. We outline our contributions for PupilNet as follows:

- The PupilNet algorithms, which use convolution neural networks to extract pupil size from the forward facing camera of a mobile tablet.
- An evaluation of PupilNet in a human subject study approved by the SMU institutional review board. Our evaluation investigates lighting conditions and occlusion while individuals conduct a memorization task [23]
- An evaluation of PupilNet as to whether the system requires personalized calibration or if a general model of pupil dilation can be created.
- Algorithms for monitoring pupil dilation to infer typing ability. To the best of our knowledge, this is one of the first investigations of monitoring pupil dilation in longer duration, more cognitively complex task.

Our results show that PupilNet performs similarly to a head worn eye tracker and is robust to changes in lighting variation and eye color. We show that the difference between PupilNet and a gold standard head-mounted gaze tracker varies only from -2.6% to 2.8%.

2 PUPILLARY RESPONSE AND COGNITIVE LOAD

William Shakespeare once wrote that "the eyes are the windows to the soul." This famous quote was based mostly on intuition. Most anyone in most any culture can attest that the eyes of a person intuit emotion. However, the exact phenomenon is hard to explain. Modern psychology might rephrase Shakespeare's words to be "the eyes are a complicated mechanism that connects to many functional areas of the nervous system, revealing behavioral cues through subtle and overt changes which engineers might describe as noise." In this way, pupillary monitoring is the ubiquitous computing equivalent of infrastructure mediated sensing [39]. The infrastructure that we want more information about is the nervous system (*i.e.*, the user) and the point of mediated sensing is the pupil, which has a complex relationship with the underlying processes of the nervous system. As we concentrate, focus, relax, become aroused, and generally think, noise manifests on the pupil. While we cannot explain every artifact, we can map many pupillary responses to different processes of the brain.

Pupillary diameter has a strong correlation with tonic activation of the Locus-Coeruleus-Norepinephrine(LC-NE) system [4]. Essentially, this means that the pupil fluctuates in size with the constant excitation of low-level brain function. The LC-NE is responsible for the modulating of signals that enter the neocortex, which is why the

LC-NE is hypothesized to control reward-seeking behavior, cognitive processes, and arousal [3, 10, 22]. Although a direct link between these phenomena and the responses of the LC-NE is unclear, a growing body of evidence supports the notion that a relationship exists. For example, the LC-NE is hypothesized to regulate arousal and help to focus attention through supplying norepinephrine to the pre-frontal lobe which neuroscientists believe is the center of executive function. This is exciting because it means such high level "thought" functions might partially manifest through the LC-NE and generate changes in the pupil. Therefore, we argue that measuring the pupil reliably in day-to-day scenarios could be used to augment a wide variety of affective sensing and context-aware applications. While this scenario might seem a distant use case, a number of pupillary response characteristics in response to external stimuli are well understood by the neuroscience and psychology communities [49]. To orient the reader to this community, we shift our focus to the tools employed in these experiments.

2.1 Tools for Pupillary Monitoring

The simplest measurement of pupil size can be captured with a pupillary diameter chart—a cylinder with printed circles in sizes ranging from 2 mm to 9 mm. A user subjectively compares the size of a subject's pupil to each circle. While this method is cheap and simple, it cannot readily capture changes in pupil size in response to stimulus from physiological experiments (about 0.25-0.5 mm changes).

The most accurate measurement of pupillary response is through a medical device known as a "pupillometer." A pupillometer can refer to two different devices. In our context, we refer to the monocular device that measures pupil dilation. This device covers the eye with a silicon cylinder and bathes the eye in infrared light. This infrared light uniquely illuminates the iris versus the pupil. The pupillometer provides the average and standard deviation of pupil diameter over a short period (2-4 seconds). This device is commonly used by ophthalmologists for eye exams and other diagnostic purposes such as detecting neurosyphilis. Nonetheless, the measurement time is typically too short for measuring cognitive load and these devices typically cost upwards to $5000 USD.

Infrared-based eye trackers have become the device of choice for most modern pupillary monitoring experiments. The work of Klingner *et al.* has been instrumental in establishing the idea that IR gaze trackers are reliable measures of pupillary response for cognitive load monitoring [30]. Gaze trackers commonly come in two types: remote and head mounted. Both use infrared light sources to increase contrast between pupil and iris. Gaze trackers typically do not inhibit the user's movements and therefore can be used to track pupil size in a number of different scenarios. Moreover, with the advent of mobile, glasses worn eye trackers, a number of experiments can be carried out in ambulatory or driving scenarios. While gaze trackers can be expensive (for example, mobile eye trackers are typically upwards of $10,000 USD) and are not suitable for measuring pupil size in day-to-day applications, they are excellent choices for laboratory studies in pupil dilation. In our experiments, we employ a TobiiPro glasses 2 which retails for about $12,000 USD and can be used in ambulatory settings as well as streamed to a mobile device. We use this functionality in our testing to collect ground truth pupil size concurrently with user-facing video in a synchronized manner from our tablet application.

Various experiments exist relating pupillary response with different stimulus and different behaviors. These experiments include traditional cognitive load (task-evoked) experiments, such as digit span tasks and target tests, as well as more recent experiments such as monitoring driver fatigue [2] and interruptibility in human-computer interaction [20, 33]. These experiments are typically carried out using mobile, head-worn gaze trackers. While a survey of the experiments in the field is out of the scope of this paper, the interested reader can find summaries and examples in [9, 14, 25, 32, 37, 43, 49].

Pupillary response is sensitive to both internal and external stimulus [49]. The most prominent external factors are lighting and eye fixation distance. In fact, lighting and eye fixation can change pupil diameter by as much as 9 mm independently [9]. Because of these external factors, most experiments use time-locked, short duration analysis to keep changes in pupil dilation more easily interpretable. In the future, we anticipate analysis

of longer duration pupillary response will be used to interpret more complex behaviors. To be competitive, then, PupilNet must also have the ability to measure pupillary response in these longer duration scenarios. To understand if PupilNet can measure these longer duration responses properly, we carry out a non-time locked typing experiment, comparing the pupil response inferred by PupilNet and the ground truth eye tracker.

When comparing the usability of PupilNet and a gaze tracker for pupillary monitoring, the major benefit of using a web camera is that it is much more pervasive and much lower cost. Most mobile devices and laptop computers have built-in cameras. Our previous work has shown that a commodity web camera shows similar performance to a more expensive gaze tracker in extracting pupil size and predicting levels of cognitive load in controlled environments [43]. Nonetheless, RGB images introduce additional challenges when compared to images captured from an infrared based gaze tracker. Images from a web camera have lower resolution, lower signal to noise ratio, and are more sensitive to environmental lighting conditions. Moreover, darker iris colors lower contrast between the pupil and iris. PupilNet uniquely addresses many of these issues through machine learning and signal processing.

3 RELATED WORK

Measuring cognitive load can be divided into two categories: (1) subjective measurement using self-report and (2) objective measurement using indicators like invested time, learning outcomes, and physiological signals like pupil dilation, heart rate, and galvanic skin reaction. An overview of these different measurements is beyond the scope of this paper, but the interested readers can find an excellent survey by Plass *et al.* [41]. Iqbal *et al.* [20] proposed that pupil dilation is highly correlated with mental workload while a user interacts with computing devices. When a user is interacting with a UI, one can predict the mental effort by measuring the pupil dilation. We find inspiration from the work of Iqbal and are motivated to create a robust PupilNet system that can be deployed in the use cases investigated by her [20]. Similarly, Pfeuffer *et al.* [40] proposed a proof of concept model to classify mental workload from the pupillogram in three different lighting conditions. However, the main aim of this experiment was to understand if cognitive load measures can be more consistently measured across lighting (not for detecting pupil change from an RGB image). His experiments were performed with a high-resolution remote gaze tracker.

There is also evidence that cognitive load measurement is important in some types of learning applications. For instance, Beste *et al.* [54] developed a piano learning tool called BACh that could automatically adjust musical score sheet difficulty levels based on the cognitive load of a user. Cognitive load was measured by a brain-computer interface employing near-infrared spectroscopy. When the measured cognitive load was below a user-specific threshold, the system increased the difficulty of the sheet music. They showed that users learned significantly more efficiently using BACh.

There are also a number of related works describing algorithms for detecting that a pupil exists in an infrared image and roughly where the center of the pupil lies. These works are typically used as the initial steps in recognizing the gaze of a user. There are two well-known algorithms commonly used for pupil detection. The first one is Hough transform [51]. Hough transform is used to detect shapes in an image through a voting or peak detection in the Hough parameter space. The algorithm starts by roughly segmenting a pupil with a binary thresholding, then applying a Canny edge detector to find edges of the pupil. Then, the image is transformed to a parameter space (also known as the Hough space). In doing so, the algorithm scans all edge pixels to find a shape that satisfies its parametric form. Local maxima in this parameter space are shapes within an image. Since the Hough transform relies on the quality of edge detection, it performs well in high contrast images such as in infrared images. However, most pupil and iris boundaries in RGB images are significantly lower contrast than in infrared images, and, therefore, the Hough transform tends to generate high false-positive rates in such conditions.

The second well-known pupil detection algorithm is known as the integro-differential operator proposed by Daugman [12]. The algorithm searches over a three dimensional space (x, y, and size) looking for the parameters that maximize pixel intensity differences. This technique is arguably more straightforward than Hough transform, but it is computationally expensive. Many combinations of these two techniques with other techniques have also been proposed. For example, [18, 31, 45] estimates rough boundaries of the iris with the Hough transform then places an active contour to find a pupil boundary. It reduces computational cost because an active contour only searches within a region of interest: the iris and pupil area. Even so, this technique heavily relies on the performance of the Hough transform and, thus, suffers from many of the same downsides when applied to RGB images rather than infrared images of the eye.

Finally, another well-known algorithm for pupil detection and pupil center location is the Starburst algorithm. Li *et al.* [36] proposed the hybrid Starburst algorithm that combines feature-based and model-based algorithms. The algorithm starts by removing the corneal reflection from the image and estimates the iris boundaries with the derivatives along 18 rays from the center of the eye. After detecting the edges, a RANSAC algorithm is applied to eliminate outliers. Later, an extension of the algorithm was proposed by Ryan *et al.* [46]. Recently, Świrski *et al.* [47] used Haar-like feature detector to roughly detect the pupil location and further refine the pupil center using k-means segmentation. Javadi *et al.* [21] proposed SET method which decomposes contours in binary images into sinusoidal components. Fuhl *et al.* [15] proposed the ExCuse algorithm, which is based on edge filtering and oriented histogram. The method they use to refine the coarse pupil center is similar to that is used in the original Starburst work [36]. Despite the success of these algorithms for infrared image pupil location tracking, none are evaluated on RGB images with relatively low resolution of the eye. In contrast, our algorithms are evaluated using commodity user-facing cameras available on tablets. Thus, the resolution of our eye images is significantly lower than most studies in the literature. In our previous work, we [43] applied a modified starburst algorithm to commodity cameras and achieved fair performance, though a number of challenges remained.

More in line with our current work, some low-resolution RGB image-based algorithm exists for tracking gaze, but they are not directly designed for pupil segmentation or pupil size measurement. For example, OpenFace [6] is an open source video-based tool for tracking head pose, face tracking, facial landmark detection, and gaze tracking. OpenFace uses a Constrained Local Neural Fields (CLNF) [5] to detect both the eye and center of the eye. Similarly, a number of different works have investigated RGB images for tracking gaze [19, 52]. However, these works mainly focus on detecting the center of the pupil, not the size of the pupil.

4 ALGORITHMS

In this paper, we present three different algorithms for tracking the size of the pupil from relatively low resolution RGB images. In this section, we describe the three algorithms: (1) Modified Starburst (MSB), (2) Two-Levels Snakuscules (2LSN), and (3) PupilNet (PN). In previous work, we proposed the modified starburst algorithm, MSB,[43]. The MSB algorithm searches within a set of potential pixels that are located roughly around the pupil and iris boundary and estimates an ellipse from these points. In contrast, 2LSN tries to fit edge locations to existing predefined circular ring structures of the pupil and the iris. Both MSB and 2LSN use traditional model-based segmentation techniques where features are hand-crafted by engineers. On the other hand, PupilNet is a convolutional neural network (CNN) that automatically learns and extracts features from the raw images.

Preprocessing. We preprocessed the images captured by the iPad camera to find regions of interest (ROI) centered on the eyes. The area around an eye was extracted by using a facial landmark detector trained with iBUG 300-W facial landmark dataset and an ensemble of regression trees [27]. Once an eye ROI was detected, we automatically cropped and scaled all the eyes to the same size (120 by 120 pixels). After resizing, we extracted tighter regions of interest for both the left and right eyes using the landmarks around the eyelids and eye corners.

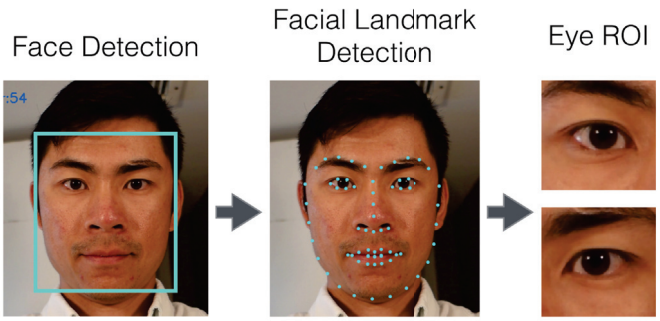

Fig. 1. Example of extracting the eyes with our pre-processing algorithm.

This final cropping resulted in 32 by 32 pixel frames with 3 RGB channels. These eye images were further processed with our proposed algorithms. Figure 1 illustrates the image preprocessing steps.

4.1 Modified Starburst Algorithm

The modified starburst algorithm starts by identifying the rough pupil area using binary thresholding to extract the dark area of the pupil. Noisy pixels are eliminated with a morphological closing operation. We calculate the center of mass of the largest extracted binary area and use it as a seed point. From this seed, the algorithm marches in different directions, looking for strong edges (*i.e.*, the edge of the pupil). The algorithm only marches from 45 degree to 135 degree and each direction is separated by 20 degrees. The motivation for restricting the march is to avoid the shadow from eyelashes and be more robust to squinting. The pixel intensity difference between the seed point and the current point is calculated. If the pixel intensity difference is greater than a threshold t, we record it as an edge point. In our previous work, we used a fixed scalar threshold obtained from a calibration process. However, in this paper, the threshold we selected is identified by a percentile of the cumulative histogram of the 32 by 32 pixel eye ROI. From our experiments, 1.5% of the cumulative histogram performs reasonably well at detecting the edges of a pupil. In later analysis, we use grid searching to validate this percentage.

Once the algorithm has marched all directions and stored all the edge points, it uses the average of the x and y points to calculate a new seed for the next iteration. We randomly add ±2 pixels to the new seed point to prevent the algorithm from stopping in a local optimum (similar to simulated annealing). These edge points from all iterations are sent to an ellipse-fitting algorithm which returns a best-fit ellipse. We run this algorithm separately on each eye.

4.2 2-Levels Snakuscules

Snakuscules [48] is an adaptation of traditional active contours tracing algorithm [26]. Traditional active contours use planar parametric curves and their parameters are optimized with existing data, monitoring shape, and prior knowledge. Thevenaz *et al.* proposed a simplified version of traditional active contours optimization for detecting circular structures in an image [48]. Later, Garg *et al.* manipulated this algorithm to locate eye center in an image [17]. In this work, we propose the use of 2-Levels Snakuscules, which attempts to exploit the concentric, circular structure of pupil and iris. Instead of solely relying on detecting the edge of the pupil, the algorithm uses the pixel intensity of the entire iris and pupil area to estimate the pupil location and size. Figure 2 (right) shows an overview of the proposed 2-Levels Snakuscules algorithm.

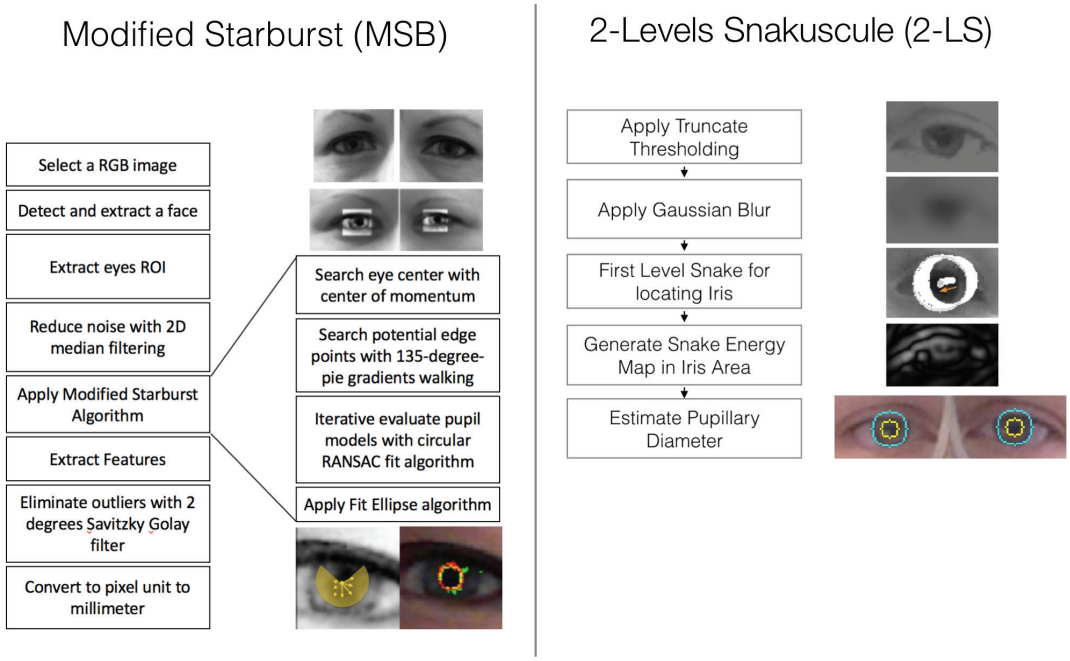

Fig. 2. Original Pupilware Algorithm (left), 2-Levels Snakuscules Algorithm (right)

The algorithm starts by applying a masking truncation threshold to reduce specular reflection and bright objects such as bright spots on glasses lenses. This essentially clips the bright areas of an image to a maximum threshold and it reduces the chance of creating multiple local optimums. The value of thresholding is 3% of cumulative histogram. This parameter can be adjusted depending on image resolution. After clipping, a Gaussian filter is applied to the eye images. This step is essential to increase the stability of a snake moving smoothly to the center of eye. The size of a kernel depends on image resolution (A larger image requires a larger kernel size). We used 11×11 kernel size for locating the iris. Next a circle "snake" is placed at the center of the image. Snakuscules exploits the circular structure of the iris against the lighter (or nearly white) cornea. A snake that has maximum energy is the one that has the greatest difference between average pixel intensity of the outer ring (the iris and cornea boundary) and the inner ring (the iris and pupil boundary). The circles are constrained to have the same center. Equation 1 shows how to calculate the energy of a snake:

$$E = \frac{\iint_{\rho R<\|x-x_0\|<R} f(x,x_0)dx_1dx_2}{\iint_{\rho R<\|x-x_0\|<R} dx_1dx_2} - \frac{\iint_{\|x-x_0\|<\rho R} f(x,x_0)dx_1dx_2}{\iint_{\|x-x_0\|<\rho R} dx_1dx_2} \quad (1)$$

where E is the energy of the snake with outer radius R and inner radius ρR. This equation ensures that we can adjust the snake body with fine-grained precision (inner circle). A simplified equation is similarly used by Garg *et al.* [17] in finding eye position. Figure 3 shows the energy map after applying an iris snake (middle) and a pupil snake (right) on a gray eye image. The brightest area is the iris/pupil location.

A snake iteratively searches for more energy in four directions (up, down, left, right). If a nearby location has more energy, it greedily moves along that direction, searching for the highest energy until converged. In this

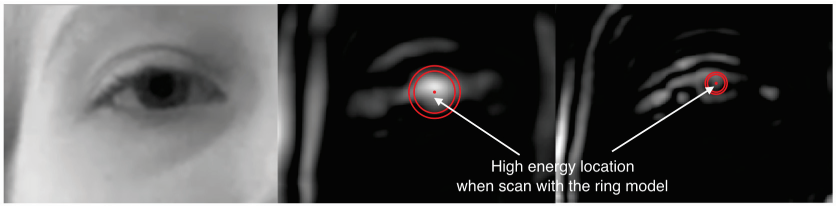

Fig. 3. An energy map generated by the 2LS algorithm. (left) the original image, (mid) an iris energy map, (right) a pupil energy map

work, we set a maximum of 15 iterations to increase the speed of the algorithm. Once the iris and eye center are located, a smaller second snake (a pupil snake) is applied only in the iris area. Note that a pupil snake is placed in an original eye image, not the blurred image so that contrast is preserved. A pupil snake can only grow to be a maximum of 90% of the iris snake to avoid detecting edges of the iris and cornea. All energy points are memorized in an energy vector (E). This energy vector is used to estimate pupil boundary with sub-pixel accuracy. Once a pupil snake is converged (about 5-15 iterations), the final pupil size is determined by a linear combination of radius and normalized energy weights. Intuitively, a higher energy snake boundary is more likely to be a pupil boundary. If there is more than one high energy step, then the pupil boundary might be somewhere in between those pixels as shown in equation 2 and 3.

$$E_{norm} = \frac{E}{\sum_0^n E_i} \tag{2}$$

$$r_{final} = \sum_0^n E_{norm,i} \cdot r_i \tag{3}$$

where E_{norm} is a normalized vector of energy. E is an energy vector. The algorithm is performed on both eyes separately.

The hyper-parameters of MSB (primer, a percentage of cumulative histogram cutoff, and degree offsets) and hyper-parameters of 2LSN (Gaussian blur kernel size, and kernel standard deviation, and maximum growth percentage) reported in the earlier section were set as the optimal values. These optimized hyper-parameters were chosen using a randomized grid search with the mean square error between the ground truth and the inferred pupil size as the objective function value.

4.3 PupilNet: Convolution Neural Network Structure

We propose a convolution neural network, called PupilNet, for measuring pupil size. Figure 4 shows the network architecture of PupilNet. We noted that other architectures were investigated, but did not significantly increase or decrease the reliability of the network. The input image data used by PupilNet is tightly segmented around the iris area as a result of our preprocessing. As such, there is not much variation in the overall structure of the input images. Because of this, we found that two convolutional layers are sufficient to extract features related to pupil size. Using only two layers also helps to prevent over-fitting. Using more than two layers only marginally increased the network performance (results were not statistically significant) but training required much more computation power. With more data, a larger model may better extract pupillary features; however, with the current amount of training data, we did not find that more convolutional layers increased performance. We also

experimented with larger eye area segmentations (i.e., 64×64 input image size that encompassed more of the eyes including surrounding skin). To attain the same performance as the 32×32 image inputs, the architecture required 3-4 convolutional layers. The overall result, however, was not significantly better than the 32×32 image size. Segmenting only the iris area allows us to train a simpler network with fewer data. This supports a conclusion that a simpler network using tighter segmentation has about the same expressivity as a larger segmentation with a more complex network. With the amount of training data in our current dataset, that additional complexity is not warranted.

Before training our model, we preprocessed the eye images by applying histogram equalization. This helps to enhance the contrast between pupil and iris, and helps reduce the impact of different lighting conditions in the images. Even so, for dimly lit images, it also greatly increases noise artifacts. We applied the filters of size 3×3×3 to capture variation in each color channel of the input images. In addition to using the RGB representation, we also investigated using only one color channel and using the hue-saturation-value representation. We found that training with all three RGB channels performed better and the network converged faster. After the convolutional layers, we flattened the architecture and used two fully connected layers. We used Rectified Linear Unit (ReLU) activation functions and 50% dropout rate in the hidden layer before the output layer. The loss function employed was the mean squared error between calculated pupil diameter and ground truth pupil diameter. For additional regularization, we employed data expansion techniques including a random rotation (up to 10 degrees) to the images as well as vertical and horizontal perturbations (up to 10% of the image size). Finally, we used Adaptive Moment Estimation (Adam) as out adaptive gradient optimization procedure [28].

We implemented the network using Tensorflow r0.12 implemented in python 3.5. We used a batch size of 128 frames per batch and ran all models for 400 epochs. The loss function on the training data began to approach a stable value after 400 epochs. We also employ L2 regularization of each convolutional and dense layer, with scaling factor of C=0.01 for all layers. Training of the network parameters was completed on a machine with an Intel Xeon@2.6GHz CPU (24 cores) and 320GB of RAM. The specifics of constructing and back propagating through a convolutional network as is beyond the scope of the paper. We refer interested readers to a number of classical works: [16, 34, 35, 50].

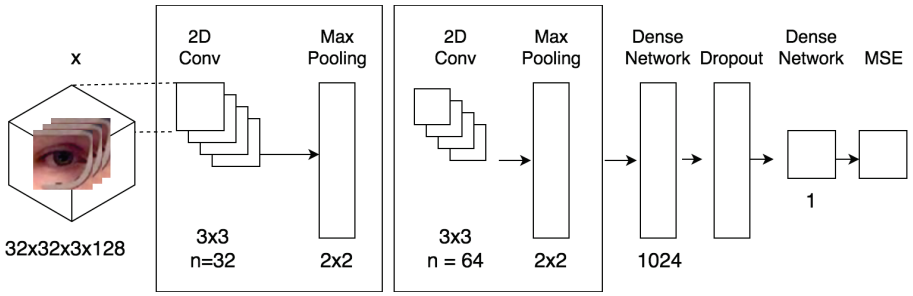

Fig. 4. Convolutional neural network architecture used in PupilNet.

The output of PupilNet is an estimate of the pupil dilation for each frame of the video sequence. Once we received these frame-level estimates from PupilNet, we post-process the signal as follows: We eliminate blinks from the time series in the same manner as performed for the gaze tracker (*i.e.*, via median filter comparison). Left eye images and right eye images were treated as separate examples while training the PupilNet architecture. During post-processing, we averaged the result of the left eye and right eye pupil estimates. We then applied a median filter of length approximately 2 seconds to the estimated pupil size signal over time. This filtered signal was then compared to the output of the gaze tracker pupil size. Each participant performed 36 total trials for the

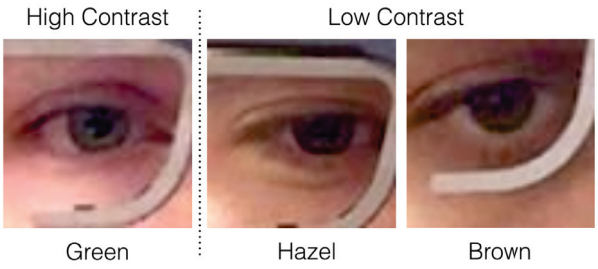

Fig. 5. High contrast eye image(left) and low contrast eye image(right) . Here, eye color is used a proxy for eye contrast because eye color is typically the best indicator of contrast between pupil and iris.

digit span task, each of which has about 800 frames. The total number of eye images collected was about 633,600 (800 frames×36 trials×2 eyes×11 participants).

5 EXPERIMENTAL SETUP

All human subjects experiments in the evaluation of PupilNet were approved by the university human subjects board. Participants were recruited by mouth, email, and flyers and were compensated with a $10 gift card for their time. In all, we recruited 11 participants. Table 1 gives an overview of the participant demographics. All participants reported normal or corrected to normal vision. One of the participants wore eye glasses and one wore colored contact lenses. We group the participants by their iris color. Because eye color is the largest contributor to contrast between the pupil and iris, we further subgroup the participants as follows: **low contrast** consisted of brown and hazel eyes (6 participants) and **high contrast** consisted of blue and green colored eyes (5 participants). We chose to combine hazel and brown eyes because the relative image contrast from the camera was qualitatively similar, as shown in Figure 5. That is, participants in the experiments with hazel eyes tended to have more brown color than green color in their eyes, subjectively. In our results, we use this high- and low-contrast grouping to understand how resilient the different algorithms are to contrast changes resulting from eye color differences.

Participant Demographics	
Age	Average age = 22.25, sd = 6.181
Gender	6 males, 5 females
Eye color	Blue 4, Green 1, Hazel 2, Brown 4

Table 1. Participant Demographics

	Average lux	**std lux**
Dark	5.5	2.5
Dim	7.4	2.5
Bright	105.0	14.5

Table 2. Lighting conditions

5.1 Environment setup

Experiments were conducted in a closed room with controlled lighting conditions. Three lighting conditions were used to simulate real-world situations. We did not constrain participants' heads during the test; they were free to move their heads and look around as they saw fit. In each lighting condition, we measured the light level with an ambient light meter next to the eyes of the participant with the meter pointing toward the iPad's display. The lighting measurements recorded are summarized in Table 2. The room used for data collection was windowless and the only light sources used were carefully controlled in terms of their position relative to the participant. **Bright:** Firstly, we use a bright, well-lit room to conduct our experiments. Lighting was controlled via two overhead lights that provided ample light. The lights used are fluorescent bulbs which are known to produce

flicker. We did not observe any noticeable flicker in the room, but we did not explicitly control for bulb flickering during our tests. **Dim:** Next we turn off all overhead lights and use a small office lamp to illuminate the room. This lamp is situated behind the participant as to light up what they are seeing, not their face. Participants described this lighting as dim and not an environment in which they would typically work. We chose this configuration to produce visible grain artifacts in the images of the eye. The iPad tablet we employ uses a number of "low-light" techniques to boost contrast and, as such, lighting had to be decreased considerably to collect a lower quality image. **Dark:** in this configuration we turned all lighting off in the room. The only lighting to illuminate the face came from the screen of the iPad tablet. Example images and the visible artifacts are shown in Figure 6. Histogram equalization is used to enhance the contrast of these images. Note that in addition to the increased grain noise, the baseline pupil size is noticeably wider in the darker environments. Finally, we did not instruct participants to hold their head constant or to maintain any other pose. As a result, we observed a number of artifacts in the dataset such as blurring from head movement, occlusion, and participants sitting too close or far from the screen. Figure 7 gives some samples of issues observed in the dataset. We mitigate some of the artifacts with head pose estimation and face detection. When the face is not detected in the image, we do not estimate the pupil size.

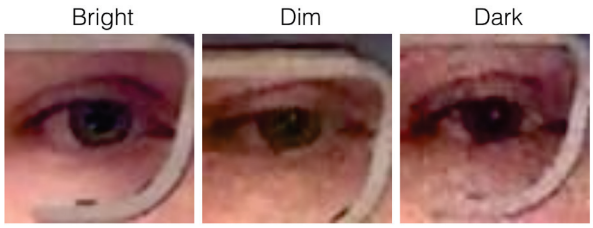

Fig. 6. Eye images in different lighting conditions. Histogram equalization has been applied to increase contrast.

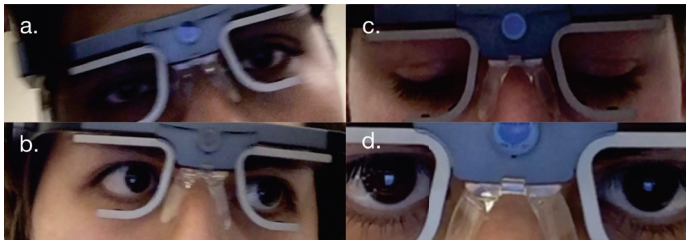

Fig. 7. (a) blurred image caused by rapid movement. (b) rapid eye movement while recalling memory. (c) occlusion when participants looked down at the keyboard. (d) large screen reflection occluding an entire pupil when participant move too close the screen.

5.2 Equipment and Gaze Signal Conditioning

Blinks cause distinct artifacts in the raw signals of the pupil diameter measurement. We remove blinks using a custom filtering process. First, we smooth the signal with a median filter of size 101 points, about 2 seconds. We then compare the filtered and raw gaze tracker signals, discarding data that differs more than three standard deviations. This is the same procedure of blink removal we use from our previous work [42, 43]. After removing blinks, we save the raw gaze tracker signal as our ground truth output for each eye.

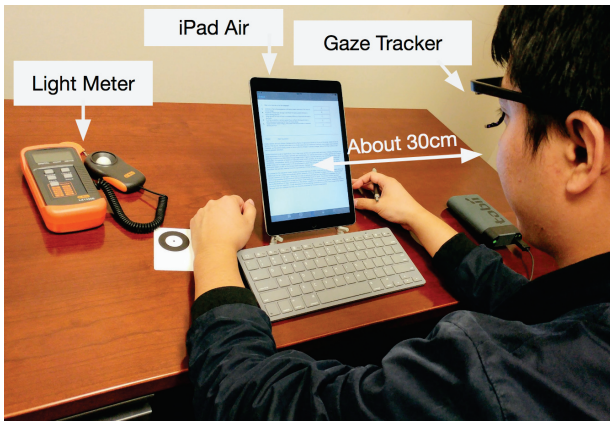

Fig. 8. A participant getting ready to perform the experiment.

We collect video data of the participants from the forward facing camera of an iPad Air 2. All camera settings are left as default, 50% brightness, and auto-brightness turned off. We note that the iPad provides some basic image processing and contrast enhancement for low light images as part of its default configurations. We collect videos having 720p HD (1028x720) at 30 FPS, in all lighting conditions. The sensor compensates the lack of light by increasing sensitivity (ISO). This does not affect the device frame rate in low lighting condition, but it affects the quality of the images as shown in Figure 6. Because the Gaze tracker sample rate is 50 FPS and the video frame rate is 30 FPS, we downsample the gaze tracker frames (with interpolation) to match the video frame rate.

5.3 Tasks

Participants were asked to perform two different tasks: the digit span task (memorization) and a typing task. These tasks are described in more detail in later sections. In each task, participants sit at a comfortable distance from the tablet screen and perform typing or memorization (Figure 8). We use these tasks to evoke cognitive load that can be measured by the Tobii Pro Glasses 2 gaze tracker. Pupil dilation is sensitive to the lighting condition, so we record lighting condition from a light meter next to participant's eye pointing toward the iPad before the beginning of each task. During experiments, we did not observe visible light flickering, shimmer, or ripple. We change the lighting condition in a controlled manner before the experiment begins. All trials for that lighting condition are then completed. That is, the ambient lighting condition does not change within a specific trial of an experiment, it only changes between trials. Moreover, we applied counterbalancing for the lighting condition in our experimental design. Thus, lighting level ordering was chosen randomly for each participant.

5.4 Ground truth selection

In the early 2000's, Klingner demonstrated that high-resolution, remote eye-trackers could provide sufficient precision for measuring cognitive load for task evoked pupillary response (TEPR). TEPR is measured by monitoring percent changes in pupil size from the baseline pupil size (*i.e.*, the pupil size after external stimuli is presented) [29]. This is described mathematically by equation 4:

$$d_{change} = \frac{d_i - mean(d_{baseline})}{mean(d_{baseline})} \tag{4}$$

where d_{change} denotes the percentage change in pupillary diameter from baseline (pre-stimulated stage). $d_{baseline}$ denotes pupillary diameter during a pre-stimulated stage which is typically a few seconds before stimulus is applied. TEPR researchers have found the percentage pupillary change advantageous because the unit can be standardized regardless of a unit (mm or pixels) and original baseline size. Even so, there is a changing relationship between pupil dilation in response to TEPR and the lighting in the room. A number of researchers have shown that percentage pupil dilation is not consistent for the same increases or decreases in cognitive load across different lighting [53]. Even so, a linear correction factor can often be applied to correct these changes across lighting [40].

We use a Tobii Pro glasses 2 [1] for ground truth comparison in this study. The Tobii Pro glasses 2 has several advantages over remote gaze trackers. It is head mounted and therefore does not suffer from problems caused by head pose or distance. It provides fixation estimation, which can be useful for discounting the accommodation reflex (pupil dilation when fixating to far objects). The device also provides video of the participant's point of view and where she or he is looking. This kind of information is useful for further cognitive load analysis. The pupil size is recorded at a resolution of 50 FPS. Our previous work showed that a more economical remote gaze tracker also works, as long as it provides high-resolution IR images with a hough circle algorithm in a controlled environment (*i.e.* office or classroom)[43].

Previous research has shown that remote gaze trackers are appropriate for measuring cognitive load [29, 30]; however, the Tobii Pro Glasses is not marketed or commercially tested for its ability to discern pupillary changes. To test the performance of Tobii Pro Glasses 2 in measuring pupil size, we compared it with a medical grade pupillometer. In particular, we use a Neuro-optics VIP hand-held pupillometer. This device allows static measurement of pupil size and is a gold standard device for measuring pupil diameter in millimeters. To compare the results of the pupillometer and the Tobii Glasses Pro, we measured pupil size of seven volunteers in three lighting conditions each: bright, dim, and dark, corresponding to the same lighting levels used in our experiments. For each participant, we conducted ten iterations in each lighting condition for a total of 7×10×3=210 measurements. The pupil size was captured at the same time from both devices for three seconds each (the medical pupillometer can only record three seconds of data at a time). The pupillometer contains a rubber encapsulating shield completely around the eye. To measure the pupil size from both devices at the same time, we placed the rubber shield inside the lens of the Tobii Glasses. The results, averaged across each iteration and each participant, are shown in Figure 9. Error bars correspond to the standard error in the mean. From the graph, we can observe that the Tobii Pro Glasses 2 consistently captures pupil size slightly lower than the medical grade pupillometer by 0.1 mm on average (sd=0.78). This difference increases slightly (to about 0.4 mm) when the environment lighting is very dark. Because our interest is in tracking changes in pupil size (not absolute size), we calculated the percent change from the "baseline" pupil size (in the bright condition) as shown in Figure 10. In this calculation, we divided all of a participant's pupillometer measurements and all of a participant's gaze tracker measurements by the average pupil size for that participant during the bright condition. In this way, the number reflects the percentage change in the pupil across lighting, rather than the raw millimeter difference. The results show there is not a statistical difference (based on a two-tailed t-test, $p>0.5$) between the pupillometer percentage measures and the gaze tracker percentage measures for pupil size. This evidence suggests that the Tobii Glasses Pro is a suitable ground truth measure of percentage pupillary change.

To further examine if the Tobii glasses are an appropriate measure of cognitive load, we conducted analyses of the percentage pupil dilation for individuals when memorizing sequences of different lengths. This experiments is commonly referred to as the digit span task and is described in more detail later on (we use this task to evaluate PupilNet). We averaged the pupillary response over time (commonly called the pupillogram) from the Tobii Glasses grouped by 6-digit, 7-digit, and 8-digit across different lighting conditions and participants as shown in Figure 11. In this experiment, our pre-stimulus period is 10 seconds. The result is consistent with the results from other TEPR literature [8, 9, 29] where pupil dilations are increased, on average, when participants memorize

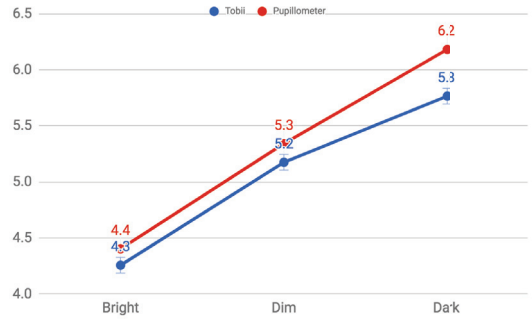

Fig. 9. Averaged absolute pupil size affected by pupillary light reflex across each iteration and each participant in three lighting conditions

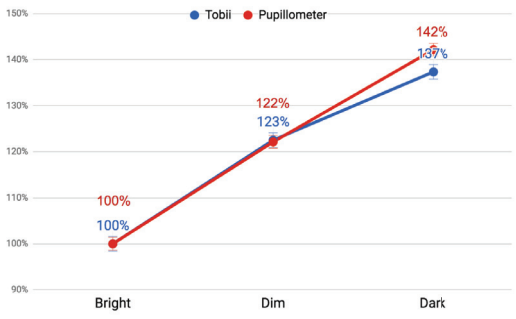

Fig. 10. Averaged percent change in pupil size affected by pupillary light reflex across each iteration and each participant in three lighting conditions

longer sequences. Pupil size for 6-digit sequences is clearly separated from 7- and 8-length sequences. However, 7- and 8-digit sequences are not as clearly separated. This result is also consistent with previous research, and one theory is that human short-term memory capacity peaks at either 7 or 8 digits for most individuals [11, 38]. In summary, Tobii Pro Glasses 2 estimates pupil size consistently smaller than a medical pupillometer. However, because we are interested in measuring percentage changes in pupil size (and the Tobii Glasses exhibit similar behavior to previous research), we consider the Tobii Pro Glasses 2 a suitable ground truth measurement device.

6 EXPERIMENT ONE: DIGIT SPAN TASK

6.1 Digit Span Task

The digit span task is used by many seminal studies in cognitive psychology [7, 9, 30]. During the task, a participant is asked to memorize a sequence of digits, normally between five to ten digits [30], and, after being presented with the entire sequence, the participant repeats those digits back. This artificially induces the cognitive load of an individual which, in turn, increases the pupil diameter. In other words, the pupil diameter gradually increases when a participant memorizes each digit in the sequence. Once the digits are spoken, the pupil size gradually decreases back to the baseline. The amount of dilation in the pupil is typically small, sometimes less the one millimeter. In our study, each participant is asked to memorize sequences of length 6, 7, and 8 digits. Each sequence length if repeated for four iterations. These iterations are repeated under three different lighting conditions (bright, dim, dark), resulting in a total of (3 sequences)x(3 lighting levels)x(4 iterations) = 36 trials

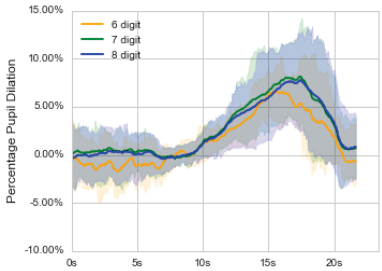

Fig. 11. lighting condition 1 (bright)

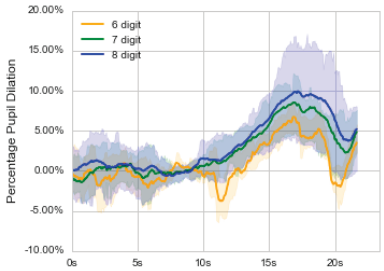

Fig. 12. lighting condition 2 (dim)

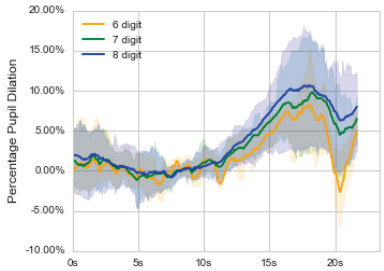

Fig. 13. lighting condition 3 (dark)

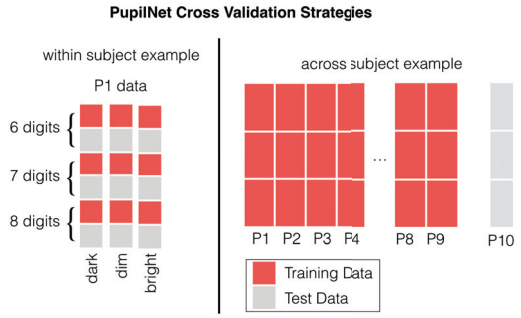

Fig. 14. Cross Validation Strategy

per participant. When the task begins, a sequence of digits appears, one-by-one, on the screen of the iPad every second. Only one digit is displayed on the screen at a time. Participants memorize the sequence of digits and, after being presented with all the digits, orally recites the entire sequence. This results in gradually increasing cognitive load as the participant memorizes the sequence. Longer sequences induce more cognitive load and, thus, larger dilations in the pupil. The experimenter ensures that the sequence is repeated without errors to ensure that the participant is sufficiently focused on the task. The displayed digits for each sequence are chosen randomly using a random number generator on the iPad. We use counter balancing at all stages of the experiments. That is, the ordering of lighting condition is chosen randomly. Once a lighting condition is selected, the participant will randomly be given a 6, 7, or 8 length sequence. The participant does not know the length of the sequence beforehand. We continue presenting a random 6, 7, or 8 length sequence until four iterations of each sequence have been displayed to the participant. We then randomly choose another lighting condition and repeat the process.

6.2 Cross Validation

To train our machine learning algorithm, we employ two different cross-validation strategies: leave-one-subject-out (LOSO) cross-validation and within-subject cross-validation. We perform LOSO cross-validation on 11 participants (*i.e.*, 11 models are created using data from 10 participants and tested on the left out user). This mirrors a use case for an out-of-the-box solution that requires no specific user calibration. Ideally, this is the preferred scenario for the PupilNet system, where no user calibration is needed.

However, we also wanted to investigate the benefit of personalizing the model with a calibration procedure. The within-subject cross-validation mirrors a use case where calibration is available for a specific participant. That is, we assume that there will be a calibration procedure using a remote eye tracker to retrain the PupilNet architecture. We choose half of the data for training and half of the data for testing, stratified across sequence length and lighting levels. So for one participant, 18 trials would be used for training and 18 trials for testing. In this way, the data are separated in contiguous time spans which is imperative for a realistic evaluation of time series data. Such a situation (requiring calibration data) limits the utility of the PupilNet system. However, in some scenarios, we anticipate that the calibration step could be warranted.

6.3 Results: Characterizing Pupil Dilation in the Digit Span Task

During the digit span task, we expect the pupil to dilate while the participant memorizes the digit sequences and for the pupil diameter to constrict quickly after speaking the digits. As the length of the memorized sequence increases, the pupil should dilate more. However, every individual has a slightly different "baseline" pupil size

which shifts in response to fatigue, sleepiness, lighting, and focal length. Therefore, we wish to characterize the percentage increase in the pupil size above baseline. We utilize three different evaluation criteria for ascertaining this difference. The first evaluation metric we choose is a custom metric based upon the difference between peak pupil diameter and baseline diameter (discussed next). We also employ the correlation coefficient for one iteration of the digit span task between the estimate of percentage pupil diameter and actual percentage pupil diameter. Finally, we take the raw point-by-point difference between the estimated and actual percentage pupil dilation. We abbreviate each algorithm according to: PupilNet (PN), Modified Starburst (MSB), and 2-Level Snakuscules (2LSN).

We define an evaluation metric that is the difference between the percentage pupil dilation in the gaze tracker and the algorithm. To calculate this custom evaluation metric, we first take the average size in millimeters during the 5 seconds before the digits appear on the screen. For the gaze tracker we denote this as GT_{start} and for PupilNet we denote this as PN_{start}. The peak pupil dilation is calculated by the maximum dilation value while the participant speaks back the memorized digits, denoted as GT_{peak} and PN_{peak} for the gaze tracker and PupilNet, respectively. Figure 15 shows two example digit span sequences for one participant with each measurement called out. Our evaluation metric is difference in these values. To compare the gaze tracker and an algorithm X, for example, we calculate the dilation difference as follows:

$$\Delta_X = \frac{GT_{peak}}{GT_{start}} - \frac{X_{peak}}{X_{start}} - \mu_{\Delta_X} \tag{5}$$

We denote this difference for different algorithms by the subscript X. For PupilNet, the dilation difference would be written as Δ_{PN}. We note that we are taking the raw subtractive difference between two percentages. This dilation Therefore if $\Delta_{PN} = -1\%$ it would mean that PupilNet calculated a 1% greater increase in pupil diameter than the gaze tracker. The term μ_{Δ_X} is a constant, scalar correction factor for algorithm X that eliminates any constant bias in the results. We calculate this correction factor by the average difference between $\frac{GT_{peak}}{GT_{start}}$ and $\frac{X_{peak}}{X_{start}}$ for each algorithm X. We apply a constant correction factor because we are most interested in the variation of the residuals and are less concerned with a systematic bias in the entire output. For each algorithm, we apply correction factors of $\mu_{\Delta_{MSB}} = -5.0\%$, $\mu_{\Delta_{2LSN}} = -8.0\%$, and $\mu_{\Delta_{PN}} = 6.2\%$. In practice, this correction factor would

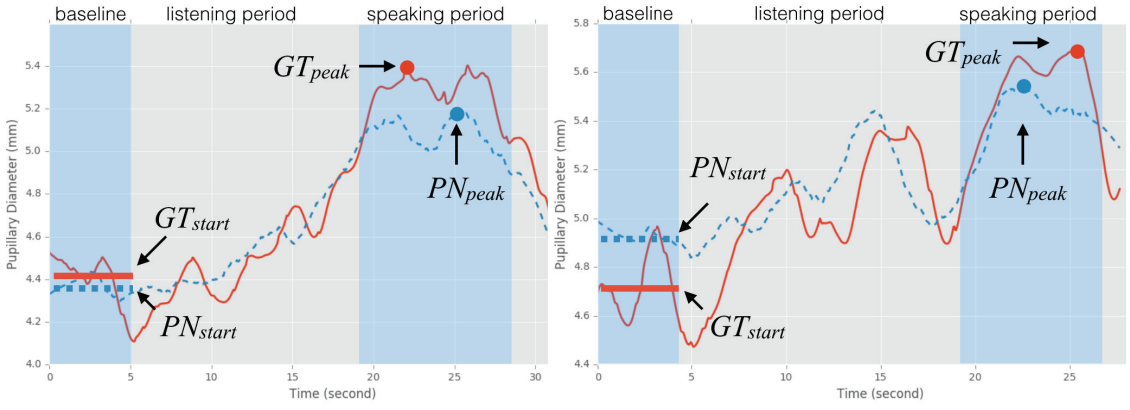

Fig. 15. Comparison of PupilNet and gaze tracker for two different participants. Also shown are the measures used in evaluating pupil dilation inference.

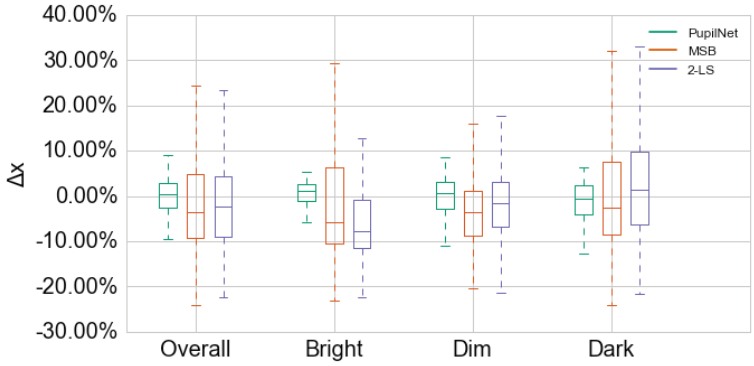

Fig. 16. Summary of Δ_{PN}, Δ_{MSB}, and Δ_{2LSN}. Results are separated by the average lighting level in the room, bright (≈100 lux), dim (≈7.5 lux), and dark (≈5 lux). Please note that the plots vary from -30% up to 50%.

need to be added to the output of the pupil dilation estimate, $\frac{X_{peak}}{X_{start}}$, when the system was deployed. Only one correction factor is applied to each algorithm. That is, we do not apply correction factors based upon subgroups of the data, only the overall average. We also note that reporting mean differences with the correction factor would be inappropriate. Instead, we report a variation of the residual differences. Smaller residual magnitudes centered around the correction factor are desired.

This metric ignores whether both PupilNet and the Gaze tracker recover the same exact millimeter measurement, as long as the percentage change from baseline to peak dilation is similar. Using Δ_X, we can understand if the dilation during an iteration of the digit span task is estimated closely by an algorithm. To help understand the expected values for the percentage dilation for the gaze tracker, $\frac{GT_{peak}}{GT_{start}}$, we look at the variations seen by the gaze tracker in Figures 11, 12, and 13. These figures show, on average, what the expected percentage difference from baseline should be for each digit sequence, and thus, the expected percentage differences for different cognitive loads. From the literature [23], the most marked cognitive differences occur between sequences of length 6 and length 8. From Figures 11, 12, and 13 we can observe that the difference between these sequence lengths is anywhere from 3%–6%, on average. Therefore, a good algorithm would need to consistently have $\Delta_X < \pm 3\%$. We note that this is only an estimate of the needed Δ_X, which could still discern useful differences even if Δ_X is greater than ±3%, especially if the value constantly under- or over-estimates the value. Moreover, the average pupil dilation across participants contains considerable overlap, so many pupil diameter differences could be greater than 10% fluctuation, depending on the effort required for a particular individual to memorize a sequence.

6.3.1 Within-subject Δ_{PN}, Δ_{MSB}, and Δ_{2LSN} Across Lighting. We break up our results into several sections based upon the research question addressed by the experiment or the analysis. We begin with analysis from the digit span task sequences at different lighting levels. In our first question, we ask: what algorithm performed most accurately at detecting pupil dilation during the digit span task? Figure 16 shows boxplots of the results across algorithm for each lighting level using within-subjects training. Each boxplot represents the distribution of differences, Δ_{PN}, Δ_{MSB}, and Δ_{2LSN}, for all 11 participants combined. They are grouped by the lighting in the room as bright, dim, and dark. Please recall that the lighting is chosen to produce noticeable, prominent artifacts in the images of the eyes representing worst-case lighting scenarios. Across all lighting conditions, PupilNet performs superior to all other algorithms. Overall, the interquartile range for Δ_{PN} is -2.6%–2.8%, compared to -9.2%–4.8% for Δ_{MSB} and -8.9%–4.3% for Δ_{2LSN}. When looking across lighting conditions, all algorithms perform worse in low lighting conditions. However, the decrease in performance for PupilNet is markedly less than the

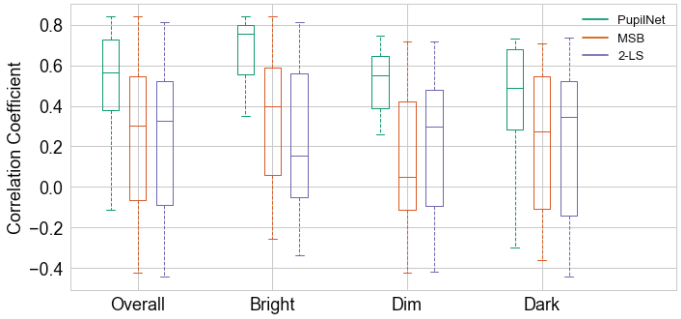

Fig. 17. Results of correlation coefficient

other algorithms. In dark lighting (≈5 lux), PupilNet has an interquartile range of -4.0%−2.4%, which is only marginally increased from the bright condition. However, the overall maximum and minimum variation (as given by the whiskers of the boxplot) is increased. While these extrema are concerning, the vast majority of trials have Δ_{PN} within 5% of the actual value. Given that we expect the cognitive load to induce dilations in the range of 3%−6%, we conclude that PupilNet follows nearly the same conclusion as the gaze tracker in brightly lit environments and is robust to most lighting changes. However, one should expect inaccurate dilation measures in dim and dark lighting conditions.

To further investigate the performance of each algorithm across lighting, we present results with a different evaluation metric, linear correlation. Correlation is agnostic to systematic biases in the outputs. For our purposes, correlation measures how closely two time-series increase and decrease together. Figure 17 is a boxplot of correlation coefficients between the gaze tracker and the algorithm under test: PN, MSB, or 2LSN. The correlation coefficient is calculated between the gaze tracker and the algorithm separately for each iteration of the digit span task. In this way, we seek to understand if the algorithm and gaze tracker increase and decrease linearly during each iteration. All of the correlations for all participants are grouped for the boxplot. The conclusions of correlation are similar to that of Δ_{PN}. PupilNet and the gaze tracker produce similar trends for each iteration of the experiment, with overall correlation typically above 0.5. PupilNet by far outperforms MSB and 2LSN, which have significantly smaller median correlations based upon a Wilcoxon rank-sum test (p<0.01). Across lighting conditions, we also see that PupilNet has some degradation in performance for dim and dark environments. The degradation in the median correlation is significantly smaller than when in a bright environment based on a Wilcoxon rank-sum test (p<0.01). This leads to a similar conclusion as previously: dim and dark environments will cause some inaccurate dilation measures, but overall, PupilNet is far more robust than other algorithms across lighting. Because of the clear performance advantage of PupilNet compared to the other algorithms, we focus primarily on the results of PupilNet for the remainder of the paper.

While our previous evaluation criteria show trends in the overall conclusion from each digit span task iteration, they do not elucidate the point-by-point accuracy of PupilNet. That is, how well does PupilNet match gaze tracker time series output. We use the percentage pupillary response as measured from baseline, d_{change}, that is customarily used when reporting pupillograms in studies of cognitive load (as discussed in equation 4). We report the results of this point-by-point difference using a modified Bland-Altman plot. The plot is shown in Figure 18. The vertical axis is differences between ground truth and PupilNet.

The horizontal axis is the mean of differences between ground truth and PupilNet. This type of plot helps to show any bias and variance in the estimates across the range of the ground truth measurement. We do not apply any constant correction factor to the plots, as was done with the Δ_X measure. The majority of d_{change}

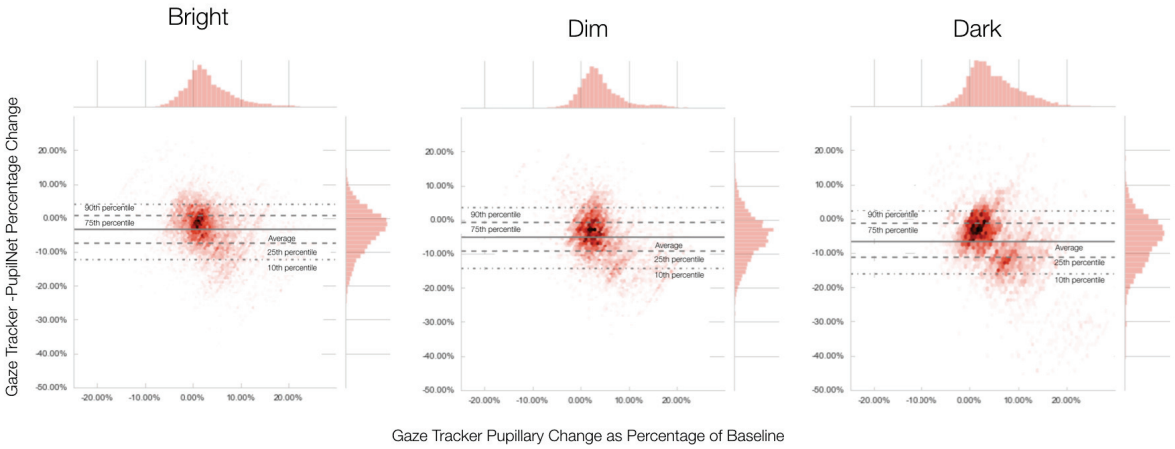

Fig. 18. Bland-Altman plot between With-in subject PupilNet and Gaze Tracker

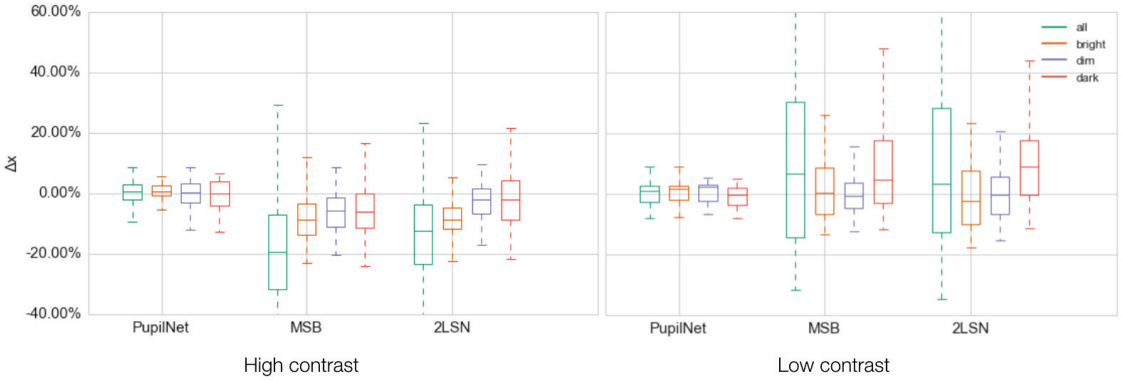

Fig. 19. Summary of dilation percentage differences for individual with high contrast (5 subjects) and low contrast eyes (6 subjects). The distributions for each boxplot are calculated only in the bright lighting configuration.

measurements are zero during the tests. This is because there is a relatively long duration of silence when the participant waits for the digits in the sequence to appear on the tablet, compared to the relatively quick memorization and speaking of the digits. We can see that PupilNet has a constant underreporting bias. However, close inspection of the plot reveals that there is also a small linear bias in the residual values that becomes more apparent in darker lighting conditions. This supports previous conclusions that darker lighting conditions degrade the performance of PupilNet. However, in bright conditions, there is almost no linear trend in the Bland-Altman plot, and the bounds of variation from percentile measures are mostly within 5% of the actual value. This supports a conclusion that PupilNet estimates follow the gaze tracker output closely.

6.3.2 Within-subject Δ_{PN}, Δ_{MSB}, and Δ_{2LSN} Across Different Eye Contrasts. In our second analysis, we wish to understand how PupilNet performs with individuals having a different contrast between iris and pupil. As discussed, gaze trackers use infrared lighting source and therefore are quite robust to different pupil/iris contrasts

because the infrared light uniquely illuminates either the pupil or the iris. However, the contrast between the iris and pupil is critical for RGB-based algorithms like PupilNet.

Because the conclusions between evaluation criteria are similar, we choose to report results only for evaluation metric Δ_X for the remainder of analyses. Figure 19 shows boxplots of Δ_{PN}, Δ_{MSB}, and Δ_{2LSN} using within-subjects training in a bright environment. We group the results by the contrast of the participant's eyes. From Figure 19 we can see that for participants with high contrast eyes, the algorithms have more similar performance, but PupilNet still has markedly smaller residuals. In the low contrast group, however, PupilNet is the only algorithm that maintains performance. In particular, 2LSN dramatically drops in performance for low contrast eyes but is a relatively good performer for high contrast eyes. This indicates that the performance gap in MSB and 2LSN might be explained by a combination of eye contrast and lighting condition. Figure 19 also shows multiple subgroups for each algorithm, lighting condition, and eye contrast group. For PupilNet, the degradation in performance due to lighting differences is about the same regardless of whether the participant was in the high- or low-contrast group. However, MSB and 2LSN had dramatically different performance across lighting, depending on the contrast group. We therefore conclude that PupilNet works reliably across different eye contrasts and the degradation in performance across lighting conditions is similar for different eye contrasts. MSB and 2LSN do not exhibit this reliability across eye contrast or lighting.

6.3.3 Δ_{PN} for Within-subject and Across subject Training. In this analysis, we want to understand if PupilNet generalizes across subjects. That is, all results thus far presented assume that a calibration stage has been used to collect gaze tracker data and RGB eye data. In this analysis, we compare within-subjects cross-validation to leave-one-subject-out (LOSO) cross-validation. Ten participants' data are used to train a PupilNet model and then tested on the remaining participant. The results are summarized in Figure 20. In the figure, p01 means the model is trained using participants p02 to p11 and tested on participant p01. This mirrors a use case when no gaze tracker calibration would be available. All lighting conditions and all eye contrast groups are combined. We show results grouped by each participant in the study and for all users combined. From Figure 20 we can see that most of the models perform slightly worse than the within-subjects models. However, the degradation in performance is not consistent. Some users show dramatic degradations in performance. This is especially true for p09, who is degraded in performance much more than other participants, with an interquartile range that includes 20% dilation difference. We reviewed the video for this participant and discovered that this participant had trouble wearing the Tobii Glasses frame because her head diameter was below average. This resulted in the glasses slipping down from the bridge of her nose and occluding part of the eye. The personalized, within-subject training was able to account for this occlusion, but the LOSO cross validation had no training examples with this type of occlusion. However, when we disregard p09, there are still a number of users with significantly degraded performance. In Figure 20, we denote users with significantly different results between within subject and LOSO with an asterisk (*). All but three users have statistically worse residuals.

It is unclear if gathering more data from more diverse participants would boost the performance of the LOSO algorithm. It is possible that collecting more data and creating a deeper, more complex network could improve results. This can only be determined empirically. Even so, from this result, we can conclude that PupilNet can be used as an out-of-the-box system for some individuals but not all users. From our previous results, the results can be readily and significantly improved through a calibration process.

7 EXPERIMENT TWO: TOUCH TYPING VERSUS NON-TOUCH TYPING

While we have presented results of PupilNet for capturing pupil size in the digit span task, it is still unknown if PupilNet can be used outside of this specific cognitive task. The motivation for PupilNet was to understanding individual's cognitive load in general, but the current results only display an ability to measure task evoked cognitive load. In general, using pupil diameter to understand general cognitive load is still an open research

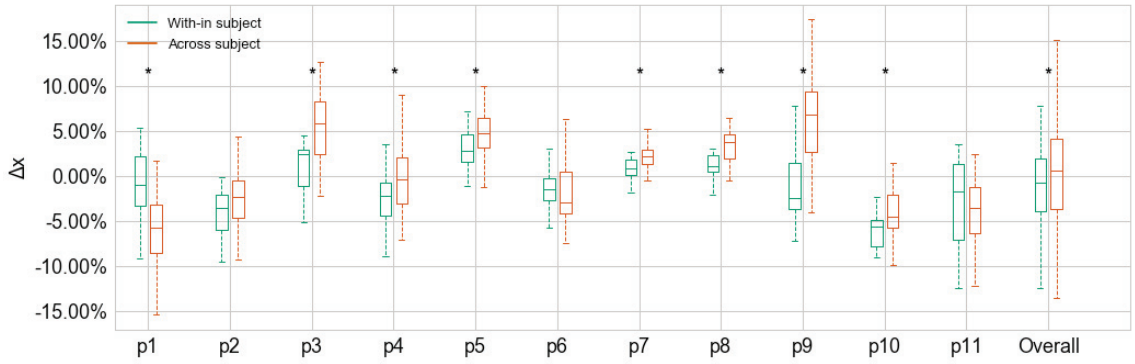

Fig. 20. Results of across-subject leave one out cross validation for each model. An asterisk (*) indicated that the distributions are statistically different based on a two-tailed T-test (p<0.01)

problem. However, it is important to understand if PupilNet can achieve the same conclusions as a gaze tracker for tasks that requires more than simple digit memorization. To investigate this, we chose to evaluate PupilNet while subjects perform a typing exercise. We consider the results of this experiment to be exploratory because the relationship between pupil diameter and cognitive load outside of simple tasks is not a well-understood phenomenon. In this way, our aim in the analysis is to discover if PupilNet and the gaze tracker deliver similar conclusions.

The digit span task is designed to evoke cognitive load and is therefore straightforward to evaluate. However, the exact degree of cognitive load for a typing task is not well defined. We hypothesized that there would be a difference in the cognitive load exerted by individuals that can and cannot perform *touch typing*. Touch typing is defined as a skill in which individuals use all of one's fingers to type, without looking at the keys. Touch type requires a mix of memorization and muscle memory. We assume the memorization will result in different pupil dilation behaviors between participants that can perform touch typing and individuals that cannot perform touch typing. Moreover, because typing involves muscle memory, we anticipate cognitive load differences will manifest for participants that focus less on the movement of their hands. We know that cognitive load is affected by tasks involving muscle memory from previous research by Yuksel *et al.* [54], who used cognitive load differences (measured via brain interface) in a system to train individuals to play piano. Therefore, we will explore if it is possible to observe similar shifts in cognitive load using PupilNet and the gaze tracker between touch typing and non-touch typing behaviors. We note that throughout this experiment we use within-subject trained models to infer pupil size because it was shown to have more consistent accuracy across a wider range of subjects.

7.1 Procedure and Recruitment

We designed an experiment in which users carried out a typing task using a custom interface (Figure 21). Participants were asked to type a given paragraph of 163 words with no time limit. The interface shows 18-20 words at a time to minimize variation in lighting intensity and help mitigate up-down eye movement. A cursor moved over the characters as the user typed. The cursor would stop moving if the typed character was not the correct character. Typing experiments were carried out only in the brightly lit room and only using within-subject trained models. We recruited participants that had already completed the digit span task experiment. In this way, we could use the already trained within-subject PupilNet model for that participant to generate the PupilNet prediction of their pupillary response during typing. In this way, we used the digit span task to calibrate the PupilNet model for each participant.

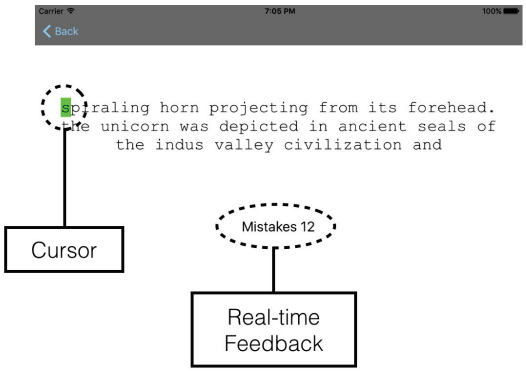

Fig. 21. Typing Task User Interface

Five users agreed to participate in the typing test. Three self-identified as being able to perform touch typing and two identified as not being able to perform touch typing. Because the sample size for this experiment was smaller, we present only qualitative analysis of the results. The content of the paragraph was taken from a Wikipedia article about unicorns. This was chosen because the subject matter is familiar but the exact words used are more diverse. Users are likely not familiar with typing many of the words in the article such as proper Greek names like "Ctesias" and "Strabo." Therefore, we can expect some cognitive load to be induced by memorizing the names and some cognitive load to be reduced by muscle memory. On average, subjects took about 5 to 10 minutes to complete the typing exercise depending on their skill level.

Participants wore the gaze tracker while performing the typing task and were instructed to complete the text quickly but take care to not make many mistakes. We observed that individuals moved their heads and eyes in this experiment considerably more than during the digit span task. As such, there were increased motion artifacts. The gaze tracker, because it was head worn, was not very sensitive to the movement of the head during this task. However, PupilNet uses a remote, embedded tablet camera. In this way, we rely more on the landmark detection and facial head pose estimation to correct angles and motion in this task than in the digit span task.

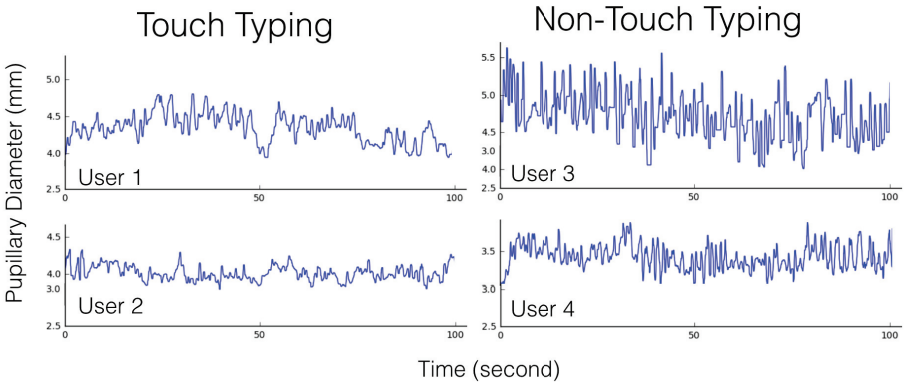

Fig. 22. Gaze tracker data for each of the fours users during a subset of the typing task.

7.2 Results: Within-subject Modeling for Typing Behavior

Figure 22 shows the gaze tracker data for four participants for a two-minute interval during typing. There are marked differences in the pupil dilation for this task depending on the whether the user reported as being able to perform touch typing or not. The touch typing subjects (left) had considerably fewer peaks in their pupillary response than the non-touch typing participants. It is important to note that not all of peaks are due to cognitive load increases. When users gaze down at the keyboard, the pupil size will change in response to the different focal length. Also, the illumination from the tablet interface may dilate and constrict the pupil in response to head motions. Because of the exploratory nature of this experiment, we are less concerned with the mechanism for why the pupil dilates or constricts; we want to understand if the same behavior is apparent in PupilNet estimate.

We predict the pupil size for each participant using PupilNet. The training examples for each participant come from the same models used to predict their digit span task from experiment one (within subject). In this way, our method mirrors a use case where the participant has provided some calibration data at an earlier date. To understand if PupilNet can discern the differences in peaks, we extract the peaks in each time series using the continuous wavelet transform as proposed by Du *et al.* [13]. This algorithm convolves the signal with an array of wavelets with different frequency and time bandwidths. Peaks are uniquely apparent by setting a global threshold for the response of the wavelets. We use this algorithm to count the number of peaks in each 30-second interval for each participant during the typing task. Figure 23 shows a box and swarm plot of the number of peaks for each 30-second interval grouped by touch and non-touch typing. It is apparent from the gaze tracker data that a clear difference exists in the number of peaks extracted for non-touch typing and touch typing. Non-touch typing individuals have considerably more peaks as extracted by the algorithm. This trend is also well established in the PupilNet box and swarm plots. However, the difference between touch typing and non-touch typing is not quite as pronounced. We can observe this because there is a slight overlap in the interquartile range for each group using PupilNet, but not the gaze tracker. We also investigated if there are any meaningful statistical differences in the distributions and we find that the results follow an intuitive pattern. A two-tailed t-test between each group reveals that: (1) no significant difference ($p = 0.15$) exists between groups *Non-touch GT* and *Non-touch PN*, (2) no significant difference ($p = 0.24$) exists between groups *Touch GT* and *Touch PN*, and (3) a significant difference does exist ($p < 0.01$) between the touch typing group and non-touch typing group for both the gaze tracker and PupilNet.

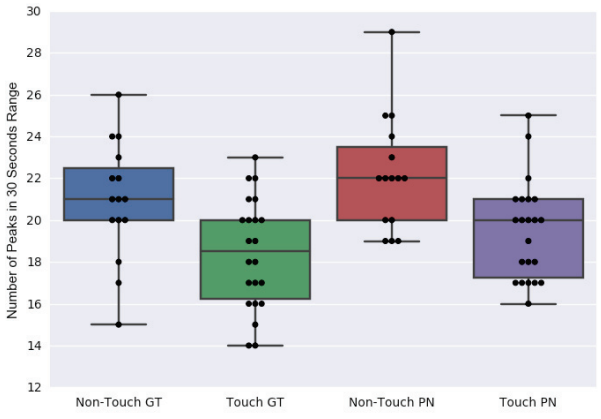

Fig. 23. This figure shows a box and swarm plot of the number of peaks for each 30-second interval grouped by touch and non-touch typing.

While the gaze tracker and PupilNet do not give identical measurements for the pupillary dilation during the typing task, the results are largely similar and the conclusions are identical. We, therefore, conclude that PupilNet shows great promise in inferring cognitive load in cognitive tasks beyond the digit span task. We have only shown this difference for a typing task, but we hypothesize that the results extend to a number of different tasks where a user interacts with a tablet. As we begin to understand how cognitive load changes for different tasks, PupilNet can be evaluated in other cognitively complex tasks. We hypothesize that PupilNet can effectively capture aspects of many pupillary response tasks but leave the evaluation of other tasks to future work. In future work, we also wish to understand if PupilNet and a gaze tracker can be used to judge the quality of a number of tasks like reading, writing, and test-taking. Understanding the cognitive load of subjects during these tasks could be key to understanding confusion, attention, and engagement. We also want to explore the combination of different sensing modalities with PupilNet in these scenarios, such as facial expressions, gaze, and heart rate. Combining these different sensing protocols could be key in the creation of a robust context-aware computing environment.

8 CONCLUSION

In conclusion, we presented three algorithms for estimating pupil size using off-the-shelf commodity cameras. The most promising algorithm was shown to be PupilNet, a convolutional neural network architecture. The results show the PupilNet model is robust to lighting condition and participants with different eye color which are major challenges for traditional image processing methods. Moreover, we have shown that conclusions are similar in a cognitively more complex typing task whether using PupilNet or a head worn eye tracker.

REFERENCES

[1] 2015. Tobii Pro Glasses 2 wearable eye tracker. (Jun 2015). https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

[2] Alexandra Branzan Albu, Ben Widsten, Tiange Wang, Julie Lan, and Jordana Mah. 2008. A computer vision-based system for real-time detection of sleep onset in fatigued drivers. In *2008 IEEE Intelligent Vehicles Symposium*. IEEE, 25–30. https://doi.org/10.1109/IVS.2008.4621133

[3] Gary Aston-Jones and Jonathan D Cohen. 2005. An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance. *Annual review of neuroscience* 28 (2005), 403–50. https://doi.org/10.1146/

[4] Gary Aston-Jones, Janusz Rajkowski, Piotr Kubiak, and Tatiana Alexinsky. 1994. Locus coeruleus neurons in monkey are selectively activated by attended cues in vigilance tasks. *Journal of Neuroscience* 14 (1994), 4467–4480.

[5] Tadas Baltrušaitis, Peter Robinson, and Louis Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. *Proceedings of the IEEE International Conference on Computer Vision* (2013), 354–361. https://doi.org/10.1109/ICCVW.2013.54

[6] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*.

[7] Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin* 91, 2 (1982), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

[8] Jackson Beatty. 1982. Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. (1982), 276–292 pages.

[9] Jackson Beatty and Brennis Lucero-Wagoner. 2000. *The pupillary system*. 142–162 pages. http://prx.library.gatech.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2000-03927-005&site=ehost-live

[10] Craig W Berridge and Barry D Waterhouse. 2003. The locus coeruleusâĂŞnoradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain Research Reviews* 42, 1 (2003), 33–84. https://doi.org/10.1016/S0165-0173(03)00143-7

[11] Zhijian Chen and Nelson Cowan. 2005. Chunk limits and length limits in immediate recall: a reconciliation. *Journal of experimental psychology. Learning, memory, and cognition* 31, 6 (11 2005), 1235–49. https://doi.org/10.1037/0278-7393.31.6.1235

[12] John Daugman. 2004. How Iris Recognition Works. In *IEEE Transactions on Circuits and Systems for Video Technology*. Vol. 14. https://doi.org/10.1109/TCSVT.2003.818350

[13] Pan Du, Warren A Kibbe, and Simon M Lin. 2006. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 22, 17 (2006), 2059–2065.

[14] Maria K. Eckstein, BelÃĺn Guerra-Carrillo, Alison T. Miller Singley, and Silvia A. Bunge. 2017. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience* 25 (2017), 69–91. https://doi.org/10.1016/j.dcn.

2016.11.001
[15] Wolfgang Fuhl, Thomas Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2015. Excuse: Robust pupil detection in real-world scenarios. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 39–51.
[16] Kunihiko Fukushima and Sei Miyake. 1982. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*. Springer, 267–285.
[17] Sanyam Garg, Abhinav Tripathi, and Edward Cutrell. 2016. Accurate eye center localization using Snakuscule. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016* (2016). https://doi.org/10.1109/WACV.2016.7477673
[18] Alaa Hilal, Bassam Daya, and Pierre Beauseroy. [n. d.]. Hough Transform and Active Contour for Enhanced Iris Segmentation. ([n. d.]). https://www.ijcsi.org/papers/IJCSI-9-6-2-1-10.pdf
[19] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. 2015. TabletGaze: unconstrained appearance-based gaze estimation in mobile tablets. *arXiv preprint arXiv:1508.01244* (2015).
[20] Shamsi T Iqbal, Xianjun Sam Zheng, and Brian P Bailey. 2004. Task-evoked pupillary response to mental workload in human-computer interaction. *Extended abstracts of the 2004 conference on Human factors and computing systems CHI 04* (2004), 1477. https://doi.org/10.1145/985921.986094
[21] Amir-Homayoun Javadi, Zahra Hakimi, Morteza Barati, Vincent Walsh, and Lili Tcheang. 2015. SET: a pupil detection method using sinusoidal approximation. *Frontiers in neuroengineering* 8 (2015).
[22] John S Kafka. 2016. Psychoanalysis and the Temporal Trace. *Time and Trace: Multidisciplinary Investigations of Temporality* (2016), 197.
[23] Daniel Kahneman and Jackson Beatty. 1966. Pupil Diameter and Load on Memory. *Source: Science, New Series* 154, 3756 (12 1966), 1583–1585. http://www.jstor.org/stable/1720478http://www.jstor.org.proxy.libraries.smu.edu/stable/pdfplus/10.2307/1720478.pdf?acceptTC=truehttp://about.jstor.org/terms
[24] Koray Kara, Dursun Karaman, Uzeyir Erdem, Mehmet Ayhan Congologlu, Ibrahim Durukan, and Abdullah Ilhan. 2013. Investigation of Autonomic Nervous System Functions by Pupillometry in Children with Attention-Deficit/ Hyperactivity Disorder Investigation of autonomic nervous system functions by pupillometry in children with Attention-Deficit/Hyperactivity Disorder. *Bulletin of Clinical Psychopharmacology* 23, 1 (2013). https://doi.org/10.5455/bcp.20121130085850
[25] Canan Karatekin, David J Marcus, J W Couperous, and Jane W Couperus. 2007. Regulations of cognitive resources during sustained attention and working memory in 10-year-olds and adults. *Psychophysiology* 44, 1 (1 2007), 128–144. https://doi.org/10.1111/j.1469-8986.2006.00477.x
[26] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. 1988. Snakes: Active contour models. *International journal of computer vision* 1, 4 (1988), 321–331.
[27] Vahid Kazemi and Josephine Sullivan. 2014. One Millisecond Face Alignment with an Ensemble of Regression Trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.13140/2.1.1212.2243
[28] Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam:. *International Conference on Learning Representations (ICLR2015)* (12 2015). https://doi.org/10.1145/1830483.1830503
[29] Jeff Klingner. 2010. Measuring cognitive load during visual tasks by combining pupillometry and eye tracking. *Perspective* May (2010), 130.
[30] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. 2008. 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 (2008), 69. https://doi.org/10.1145/1344471.1344489
[31] Jaehan Koh, Venu Govindaraju, and Vipin Chaudhary. [n. d.]. A Robust Iris Localization Method Using an Active Contour Model and Hough Transform. ([n. d.]). https://pdfs.semanticscholar.org/4709/a9e2920f4083264f04e94c71463b528af128.pdf
[32] Bruno Laeng, Marte Ørbo, Terje Holmlund, and Michele Miozzo. 2011. Pupillary Stroop effects. *Cognitive processing* 12, 1 (2 2011), 13–21. https://doi.org/10.1007/s10339-010-0370-z
[33] Daniel Lafond, René Proulx, Alexis Morris, William Ross, Alexandre Bergeron-Guyard, and Mihaela Ulieru. 2014. Hci dilemmas for context-aware support in intelligence analysis. In *Adapt. 2014, Sixth Int. Conf. Adapt. Self-Adaptive Syst. Appl.* 68–72.
[34] Yann LeCun et al. 1989. Generalization and network design strategies. *Connectionism in perspective* (1989), 143–155.
[35] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[36] Dongheng Li, David Winfield, and Derrick J Parkhurst. 2012. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. (2012). https://pdfs.semanticscholar.org/db1d/7f94e91feea0a0e0b2f4563f2d05b0338732.pdf
[37] Irene E. Loewenfeld. 1993. *The pupil: Anatomy, physiology, and clinical applications*. Wayne State University Press. Google Scholar, Detroit, MI.
[38] George A. Miller. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63, 2 (1956), 81–97. https://doi.org/10.1037/h0043158
[39] Shwetak Patel. 2008. Infrastructure Mediated Sensing. August (2008), 274. http://hdl.handle.net/1853/24829

[40] Ken Pfeuffer, Jason Alexander, and Hans Gellersen. 2016. Partially-indirect Bimanual Input with Gaze, Pen, and Touch for Pan, Zoom, and Ink Interaction. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 2845–2856. https://doi.org/10.1145/2858036.2858201

[41] Jan L Plass, Roxana Moreno, and Roland Brünken. 2010. *Cognitive Load Theory*. Vol. 55. 286 pages. https://doi.org/10.1016/B978-0-12-387691-1.00002-8 arXiv:arXiv:1011.1669v3

[42] Sohail Rafiqi, Chatchai Wangwiwattana, Ephrem Fernandez, Suku Nair, and Eric C. Larson. 2015. Work-in-progress, PupilWare-M: Cognitive load estimation using unmodified smartphone cameras. In *Proceedings - 2015 IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2015*. https://doi.org/10.1109/MASS.2015.31

[43] Sohail Rafiqi, Chatchai Wangwiwattana, Jasmine Kim, Ephrem Fernandez, Suku Nair, and Eric C. Larson. 2015. PupilWare: Towards pervasive cognitive load measurement using commodity devices. In *8th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2015 - Proceedings*. https://doi.org/10.1145/2769493.2769506

[44] Gerulf Rieger and Ritch C Savin-Williams. 2012. The eyes have it: sex and sexual orientation differences in pupil dilation patterns. *PloS one* 7, 8 (1 2012), e40256. https://doi.org/10.1371/journal.pone.0040256

[45] Kaushik Roy, Prabir Bhattacharya, and Ching Y Suen. 2010. Unideal Iris Segmentation Using Region-Based Active Contour Model. *LNCS* 6112 (2010), 256–265. https://pdfs.semanticscholar.org/a5da/0a5fbfe89bd678d099c504a7d94bce955019.pdf

[46] Wayne J. Ryan, Damon L. Woodard, Andrew T. Duchowski, and Stan T. Birchfield. 2008. Adapting Starburst for Elliptical Iris Segmentation. In *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*. IEEE, 1–7. https://doi.org/10.1109/BTAS.2008.4699340

[47] Lech Świrski, Andreas Bulling, and Neil Dodgson. 2012. Robust real-time pupil tracking in highly off-axis images. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 173–176.

[48] Philippe Thevenaz and Michael Unser. 2006. The Snakuscule. *2006 International Conference on Image Processing* (2006), 1633–1636. https://doi.org/10.1109/ICIP.2006.312658

[49] Warren Tryon W. 1975. Pupillometry: A Survey of Sources of Variation. *Psychophysiology* 12 (1975). https://doi.org/10.1111/j.1469-8986.1975.tb03068.x

[50] Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing* 37, 3 (1989), 328–339.

[51] Richard P Wildes, Jane C Asmuth, Gilbert L Green, Stephen C Hsu, Raymond J Kolczynski, James R Matey, Sterling E McBride, Richard P Wildes, Jane C Asmuth, Gilbert L Green, Stephen C Hsu, Raymond J Kolczynski, James R Matey, and Sterling E McBride. 1994. A system for automated iris recognition. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*. IEEE, IEEE Comput. Soc. Press, 121–128. https://doi.org/10.1109/ACV.1994.341298

[52] Erroll Wood and Andreas Bulling. 2014. EyeTab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 207–210.

[53] Jie Xu, Yang Wang, Fang Chen, and Eric Choi. 2011. Pupillary Response Based Cognitive Workload Measurement under Luminance Changes. 178–185. https://doi.org/10.1007/978-3-642-23771-3{_}14

[54] Beste F Yuksel, Kurt B Oleson, Lane Harrison, Evan M Peck, Daniel Afergan, Remco Chang, and Robert J K Jacob. 2016. Learn Piano with BACh : An Adaptive Learning Interface that Adjusts Task Difficulty based on Brain State. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 5372–5384. https://doi.org/10.1145/2858036.2858388

Received August 2017; accepted October 2017