2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), pp. 66–73.
DOI 10.1109/AIVR46125.2019.00019

Using Eye Tracked Virtual Reality to Classify Understanding of Vocabulary in Recall Tasks

Jason Orlosky
Osaka University
Augusta University
orlosky@lab.ime.cmc.osaka-u.ac.jp

Brandon Huynh
Osaka University
University of California Santa Barbara
huynh.brandon@lab.ime.cmc.osaka-u.ac.jp

Tobias Höllerer
University of California Santa Barbara
holl@ucsb.edu

***Abstract*—In recent years, augmented and virtual reality (AR/VR) have started to take a foothold in markets such as training and education. Although AR and VR have tremendous potential, current interfaces and applications are still limited in their ability to recognize context, user understanding, and intention, which can limit the options for customized individual user support and the ease of automation.**

This paper addresses the problem of automatically recognizing whether or not a user has an understanding of a certain term, which is directly applicable to AR/VR interfaces for language and concept learning. To do so, we first designed an interactive word recall task in VR that required non-native English speakers to assess their knowledge of English words, many of which were difficult or uncommon. Using an eye tracker integrated into the VR Display, we collected a variety of eye movement metrics that might correspond to the user's knowledge or memory of a particular word. Through experimentation, we show that both eye movement and pupil radius have a high correlation to user memory, and that several other metrics can also be used to help classify the state of word understanding. This allowed us to build a support vector machine (SVM) that can predict a user's knowledge with an accuracy of 62% in the general case and and 75% for easy versus medium words, which was tested using cross-fold validation. We discuss these results in the context of in-situ learning applications.

***Index Terms*—virtual reality; eye tracking; memory; cognition; pupillometry; classification**

I. Introduction

The problems associated with traditional augmented and virtual reality (AR and VR) are well known [1]. These include localization and mapping, image reproduction, and latency, to name a few [23]. Solving these problems means that we can display virtual objects or information that are perceptually indistinguishable from real, physical items. However, AR has tremendous potential not only to add or modify content, but to enhance vision, memory, and even cognition. Quite a bit of literature exists on this topic, going back to the beginnings of AR [20] [1], but research on automatic assessment of user cognition is still limited in many ways.

One specific research area with great potential is that of learning enhancement. Lack of education across the globe is also still a significant problem. As a step towards improving education, our goal is to build an automated education framework that supports in-situ learning through AR and VR. As one step in this process, we need to better determine when an individual understands a particular concept and to what level. With respect to language learning, we need to recognize when a user remembers a particular word in a given context. To do so, we hypothesize that eye tracking can be used to classify a user's level of understanding of a particular event or concept when combined with context. In addition to observations of the tendencies of the eye during learning tasks, we evaluate a variety of different eye metrics to help with the classification of this kind of understanding.

More specifically, our system makes use of an eye tracked VR environment as a test bed. Using metrics including eye and head movement, pupillary response, and focus duration, we can to a certain degree classify the moment a user knows or is having trouble recalling a particular concept. We also hypothesize that we can determine the level of understanding or extent to which someone is able to recognize a particular concept based on the amplitudes and irregularities in some of these metrics. Most other work attempts classification of general cognitive activities over time, such as that by Henderson et al. or Marshall et al. [12][18]. Our research differs from most prior studies in that we are evaluating the understanding of short-term, individual events as part of a specific context. Understanding events on a shorter time scale is important for learning interfaces since humans often learn new words or concepts in a matter of seconds.

Another contribution of this paper is the VR environment and series of experiments that will help benchmark suitable algorithms and reveal more about the physiological processes that occur during recall and understanding. Within our VR environment, we designed a series of word memory and tasks that should facilitate a certain amount of cognitive load. In comparison with previous studies that classify tasks based on viewing of images like that of Henderson et al. [12], we use an interactive environment to more closely resemble interactions with in-situ objects or tasks.

Results from our experiments show that fixation time, eye movement, and pupil size had the highest correlation to perceived word difficulty. Using these metrics, we were able to achieve a rate of 62.8% (known/recalled vs. unknown/forgotten) when trying to classify all easy, medium, and hard words, and 75.6% classification accuracy when considering only easy and medium difficulty words.

To summarize, our contributions in this paper include:

- the design of a VR environment and word recall tasks to help test understanding/memory,
- an experiment evaluating the most relevant eye metrics and subsequent analysis and classification of this kind of short term memory event, and
- discussion of how this kind of classification can be used to benefit next-generation learning interfaces.

II. Related Work

Related work is primarily focused on cognitive state classification, with many sub-areas focusing on virtual, mixed and augmented reality applications.

A. Cognitive State Classification and Memory Interfaces

In addition to augmented virtual content itself, the timing and presentation of content to the user is of equal importance. A user likely wouldn't want to be interrupted with a language learning word reminder if he or she were trying to read a map while navigating abroad, and someone wouldn't need to see virtual annotations for every single object in the environment if he or she were currently visually searching for the library. To help understand context and user state, eye tracking has long been used in a number of different applications such as neurology, childhood development, and clinical research [7]. Extraction of user state and analysis of scene context are already established areas in themselves. For example, some of the more common states that can be detected algorithmically include:

- Search (navigation or object search) [12]
- Trauma [22]
- Concentration (on a particular object or target) [18]
- Reading [4] [2] [3]
- Null (minimal or no activity) [26]

Tsai et al. [27] evaluated a number of different eye metrics to ascertain cognitive load. Blink frequency, blink duration, fixation frequency, fixation duration, pupil diameter, and horizontal vergence were measured as metrics for studying cognitive load on a minute-to-minute basis for the purposes of assessing driver behavior.

Most current algorithms only evaluate cognitive activity on a scale of several minutes and do not provide information about short-term events. To solve such problems, we are exploring and evaluating logical combinations of cognitive state recognition via eye tracking, scene analysis, and AI to facilitate more effective enhancement of human vision through AR. In particular, the experiments conducted in this paper focus on the understanding and classification of short term memory events.

We propose the use of cognitive state-based augmentation, which takes into account the user's current state and actions, environmental context, and prior temporal links to these augmentations (much like paths or links to nodes in neural networks). Much like similar contextual computing approaches, our framework is designed to improve the recall and relevance of information for the user by taking cognitive state and memory into account in more depth. The first and most important part of this framework is to be able to classify understanding, which is the focus of the experiments described in this paper. Simply put, we first need to determine the user's state and state of content in the environment. Subsequently, we can use those states to recall existing augmentative information and update a database of user states and environments with new relevant information. Through such an interface, we can improve both the timing and relevance of AR content, and thereby user performance.

B. Context in Learning and Cognition

In the past, sensing and augmentation have traditionally been separate fields. For example, sensor systems that attach to the body such as AutoSense, the wearable device developed by Matthews et al. [8], solutions based on eye-tracking [18], [25] or pupillometry [19] have been developed for recognizing a number of different mental or cognitive states. On the other side of the spectrum, AR researchers have been focused on topics such as image reproduction [1], latency reduction, context recognition, and content management [17]. However, the resulting systems are often kept separate for the purpose of conducting in-depth research in a single field.

Some recent progress has been made in the way of developing AR interfaces that are more 'intelligent' [25][28]. Even so, we still need better ways to merge content into a user's world so that he or she can interact with the most relevant augmentation based on the current context, yet safely engage with the surrounding environment. One requirement for this kind of seamless interface is that both the user's cognitive state and environment must be taken into account when managing, interacting with, or displaying content. Most current applications require the user to start an application and interact with an interface manually, which often times detracts from attention to surroundings and potential hazards. This paper explores the automation of classifying understanding as a fundamental building block for learning interfaces and applications.

C. Applications Research

Cognitive state recognition has resulted in some interesting practical applications. For example, in 2002, Duric et al. proposed the used of eye tracking and cognitive state recognition for the control of human computer interfaces [6]. More recent work by Toyama et al. took this idea a step further and used both gaze and cognitive state recognition to control the display and brightness settings of virtual content inside an optical see-through HMD [26]. Virtual reality has also often been studied as a method for cognitive rehabilitation or monitoring [10] [24]. We even have the ability to assist clinical diagnosis with the use of eye tracking and virtual reality systems, such as that by Cifu et al. [5].

Several recent and highly relevant works by Karolus et al. showed that language proficiency can be inferred just by fixation duration when reading an excerpt [15], [16]. Though the approach is only applicable at the sentence level, the

Fig. 1. Images showing the HTC Vive with integrated pupil labs tracker (left) and a screenshot from the video stream with the tracked pupil ellipse (right), which is a customized version of open-source tracking software [14].

results help motivate algorithms that can extract higher levels of understanding for more specific situations or temporally shorter events. In addition, future display-integrated classifiers need to have a closer connection to the user's cognitive states and goals if we are to keep humans in the loop.

Building on top of this prior work, we set out to 1) create a better framework for integrating cognitive state into AR and VR interfaces, and 2) test this approach by implementing a VR test bed with which to study and classify understanding via eye movements.

III. System Setup

Our system makes use of a number of different software libraries, including drift-resistant eye tracking, the Shark machine learning library, and a word-search environment developed in Unity designed to facilitate our experiment. We next describe the hardware and software we implemented in order to carry out our experiments.

A. Hardware and Software

The first step in this framework is to track the user's eyes and visual state. To do so, we used a small form factor Pupil-Labs Dev IR camera that can be integrated into a variety of different AR and VR devices. For this test, we fitted and tested the camera in an HTC Vive, as shown in Figure 1. Images were taken at a resolution of 640 x 480 at a frame rate of 30Hz.

This implementation was run on a PC with a core i7 processor and NVIDIA GTX 1070 graphics card. The 3D interface and fixation detection and head movement metrics were written in C#. Our experiment environment was designed with Unity (v2017.1). The eye tracking, including metric detection algorithms and drift correction, was written in C++, and all inter-program communication and writes/reads were conducted via synchronized text file. Lastly, we used the Shark v3.1.0 Machine Learning library for our support vector implementation [13], which was chosen for its C++ implementation and wide range of available classifiers.

B. Eye Tracking Framework

We already have a robust, custom eye tracking implementation, the original version of which was written by Itoh et al. [14]. Our code is developed in-house, aside from the utilization of OpenCV libraries, which gives us direct access to the calibration framework and ellipse fitting code that is unavailable in commercial trackers such as the FOVE or Tobii. One major update to our existing framework for use in our experiments

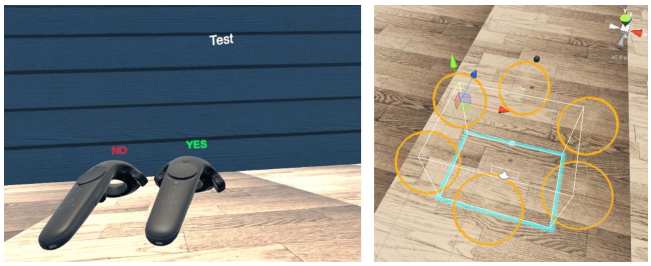

Fig. 2. Images showing the virtual reality environment from the participant's perspective with a test word and the yes/no controllers (left) and experiment environment showing spawn points (centers of the six spheres) and gaze regions (highlighted orange boundaries) used to delineate fixations.

is a drift-correction model, which updates the eyeball position over time to account for shifts of the VR display on the user's face. During initial tests of this interface, dynamic head movement resulted in shifts of the display during use, so this drift correction algorithm was necessary to ensure tracking over the duration of the experiment. In addition to the raw eye gaze vectors, we have also implemented a number of different detection mechanisms to assist with classification, including blink, saccade, and fixation detection. Exact definitions for each of these metrics are listed in the experiments section. Head movement data via internal sensors, in our case from the HTC Vive's IMU, is also included to help achieve a more accurate classification.

In addition to traditional eye tracking using point-of-gaze techniques and detection of subconscious movements, we also use pupillometry to help classify a user's mental state. In contrast to gaze-based interaction, the size, shape, and contraction of the pupil itself can reveal different characteristics of a user's cognitive processes. Relative size can represent sensitivity to light, changes in size can show cognitive workload, and irregular shape may be representative of cognitive decline. In addition to cognitive processing, evidence exists that shows the pupil is also tied to memory [21]. Integrating pupil analysis into the overall eye tracking framework can further enhance our classification, as we later show through experiments.

IV. Experiments

Our experiment was primarily designed to answer the following questions: 1) What eye movement responses are evoked during cognitive recall tasks? 2) Is it possible to use these metrics to determine the extent to which a user understands a particular concept? 3) How can we make use of context in combination with machine learning to classify this understanding over short periods of time?

A. Method

To answer these questions, we wanted to employ an easily testable recall task in which we had a well established ground truth. Though many different types of understanding and memory exist, word understanding provided a good opportunity for us to effectively test the extent to which we can classify responses to an individual concept in a short period of time. As such, we developed the eye-tracked VR environment as

shown in Figure 2, where users viewed words (annotations) and determined whether or not they understood the meaning of the word by pressing yes/no triggers on hand-held controllers. While viewing and responding to a series of words, our interface recorded and monitored a variety of performance and eye tracking data.

B. Participants

A total of 16 individuals (mean age of 31.8, Stdev 8.16, range from 24 to 50) participated in the experiment. The participants came from a wide range of language backgrounds. All were non-native English speakers but had some English language ability, which also ensured a large distribution of answers. Moreover, any classification we achieve needs to be culture- and language- independent, so the larger variety of language abilities benefited results. This experiment was approved by Osaka University IRB SA2016-2, and all participants signed a consent form prior to starting the experiment.

C. Materials

As shown in the bird's eye view on the right of Figure 2, the environment consisted of a room in which 6 spawn points were distributed around the participant. The centers of the spheres in the image represent the spawn points, but were not visible during the experiment. Each sphere was approximately 1.2 meters (Unity units) from the participant and had a radius of 0.5 meters and defined a fixation point encompassing a field of view of approximately 23 degrees. Each participant stood at the center point, which was equidistant to the spawn points during the tasks, but searching for words during the task resulted in the distance to the words changing slightly over time.

In total, we presented 30 words to each participant. We initially selected three difficulty rankings (10 words each), since we wanted to have words that were definitely known, words that were borderline, and words that would definitely not be known by all participants. The difficulties were selected by choosing words from different occurrence levels in English texts, as determined by the Google Ngrams database over the last 10 years [11]. Difficulties were divided into easy, medium, and hard categories, which were separated by occurrence ranges of 0.02% - 0.0011%, 0.001% - 0.0001%, and below 0.0001% occurrence, respectively.

The entire experiment was broken up into two tasks, 15 ordered words and 15 randomized words, picked randomly from the list of 30 words for each participant, but ensuring that an even distribution of difficulties between the two tasks. The first 15 words were broken up into three groups, sequentially going from easy to medium to hard, with presentation of each of the 5 words in each group being randomized. The second set of 15 words was completely randomized.

This was done to ensure that when we try to classify understanding, we can make the classification regardless of the order in which the words are viewed by the user. These task orders were not explained to the participants to avoid bias, and the experiment proceeded as a single session with all 30 words appearing one after another. Upon making a YES/NO selection for understanding the word, each subsequent word appeared at a spawn point at least two points away from the previous target to ensure that participants would have to search for the next word. This also prevented false positive fixations on the following target.

D. Procedure

The first step in the process was to calibrate the eye tracker. To do so, we implemented a 5-point calibration in which the user gazed at and verbally confirmed five separate display-relative points. The user was finally asked to confirm the accuracy of the calibration by re-focusing on a central confirmation point. If the calibration was off by more than two or three degrees as determined by the experimenter, the 5-point calibration was then re-conducted and re-confirmed.

Once the calibration was satisfactory, the participant was instructed to look around the environment until he or she spotted a word (annotation), as shown on the left of Figure 2. Upon viewing the annotation, the participant needed to determine whether or not he or she understood the meaning of the word. He or she then had to press the trigger on the corresponding YES or NO controllers, which were labeled in VR with the corresponding answers as also shown on the left of Figure 2. The participant continued searching for words and pressing the triggers to select an answer until all 30 words had been found. All participants completed the experiment, including the calibrations, in less than 20 minutes. However, three participants were excluded from eye gaze analysis since two participants accidentally unplugged the eye tracker when moving and another participant did not have stable eye tracking due to the poor fit of the display and camera.

Again, our goal is to determine when the user understands a particular word within this interface. Our ground truth was obtained by user selections of whether they did or did not know the meaning of the given word, i.e., when a YES or NO selection was made via the controller buttons. Using the timing of these selections, we measured exactly what happens to the eye just before the event occurs and compared YES to NO selections. The real challenge was to be able to accurately separate these two responses via eye and head movements alone, so we recorded the following metrics (with the corresponding definitions below) during the selection tasks:

- **Focus / Fixation duration:** The amount of time in ms that the participant spends gazing at the word, defined as the moment the participant's eye gaze enters the fixation region, to time of answer selection.
- **Saccade frequency:** The number of saccades per second from the start of fixation to answer selection.
- **Blink frequency:** The number of blinks per second from the start of fixation to answer selection.
- **Pupil Size:** The average pupil size from the start of fixation to answer selection.

- **Pupil Deviation (Dilation/Constriction):** The average total change (summed from frame to frame) in pupil size from the start of fixation to answer selection.
- **Eye movement per second:** The average total change (summed from frame to frame) in gaze position from the start of fixation to answer selection.
- **Head Roll:** The average total change (summed from frame to frame) in the roll of the head along the z-axis from the start of fixation to answer selection.

To obtain these metrics, we saved all data from the time the participant's gaze entered the fixation area (orange borders in the right image of Figure 2) for the visible annotation to the time they made a selection so that we had data leading up to and at the moment of the answer. This allowed us to analyze what happened to the eyes in proximity to the selection event, and gave us ground truth to train a support vector machine (SVM) to classify the user's state of understanding.

These particular metrics were tested based on prior research that shows that the frequency or irregularity of these gaze metrics (saccades, focus, and pupil response) are evident of memory access or cognition [9] [27]. We further refined these into temporal characteristics by computing the number of occurrences in the time spent gazing at a particular annotation.

E. Results

1) Word Difficulty Analysis: We first verified that the method of rating word difficulty in our experiment was correlated to the number of participants who knew a particular word, i.e., perceived difficulty. To test this, we mapped a rank of the words by difficulty (in order of occurrence) to the number of YES answers per participant. Figure 3 shows a graph representation of this mapping. A statistical analysis revealed a Pearson Correlation with $R(28) = .8983, p < 0.01$, meaning word difficulty was well correlated with participant knowledge of words. Our next steps were to 1) find which of these metrics were statistically significant, 2) determine which of the significant results, if any, were correlated to the participant's level of knowledge of a word and 3) determine if machine learning could be used to classify (and predict) participant answers based on eye and head movement alone.

2) Focus Time and Head Movement: Next, we wanted to determine whether head movement or the time spent focusing on a particular word differed by answer. The statistical analysis was conduced using a mixed effects model in R. For the binary outcome of YES or NO, we tested each of the metrics listed above for differences in means, while including participants as a random effect in the model. Within this model, type III ANOVA using Satterthwaite's method and a separate Pearson correlation between each metric and word difficulty were computed and reported where applicable.

First and foremost, the average time spent on NO answers was 2753.69 ms vs. 1731.28 ms for YES answers. T-tests using Satterthwaite's method confirmed a significant difference in means for time $F(1, 347.07) = 14.402, p < 0.001$. As a follow-up, we compared the time taken for answers in the ordered set of words versus the random set of words. This

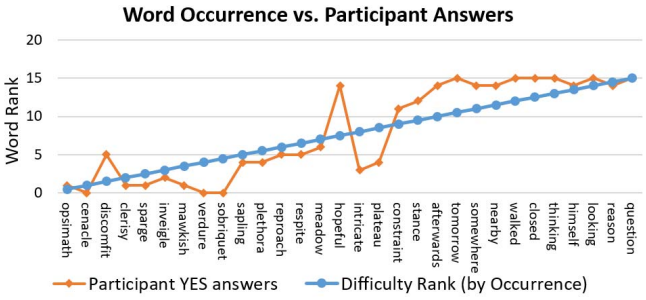

Fig. 3. Graph showing the number of YES responses (in orange) out of 15 possible, along with words ranked by increasing occurrence (in blue) according to Google's ngram viewer [11]. This shows that the difficulty rankings were well correlated to the number of correct responses per word, with a Pearson correlation of $R(28) = .8983, p < 0.01$.

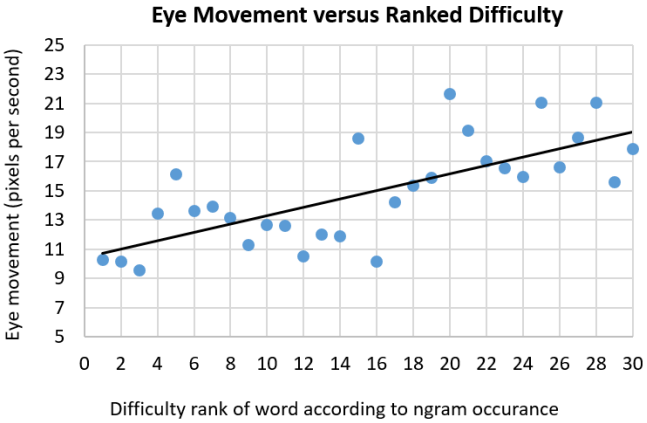

Fig. 4. Graph showing the average eye movement in pixels per second against word difficulty ordered by increasing difficulty. This resulted in a Pearson correlation with $R = .8983, P < 0.01$.

effect was not significant, $F(1, 367.29) = 2.02, p < 0.155$, with the average times being 2396 ms for ordered answers and 2027 ms for random answers. Secondly, we wanted to see whether average time spent had a correlation to the number of known words and difficulty rank. Time and the number of YES answers for a particular word were not strongly correlated, with $R(28)$ = 0.291, p = 0.153, and $R(28)$ = -0.351, p = 0.082, for ranked difficulty. Though time data can help us classify YES/NO understanding, it may not help establish the level of understanding of the word.

The next metric we explored was head movement. In particular, we analyzed head roll since several participants were observed cocking their heads to the side when thinking during the experiment. The average clockwise (from the participant's perspective) roll angle (abs value) was 3.61 degrees for NO answers versus 3.82 degrees for YES answers and counter-clockwise roll was 2.95 degrees for NO answers versus 2.81 degrees for YES answers. Neither of these were significant, with $F(1, 352.21) = 0.550, p = 0.458$ and $F(1, 328.16) = 0.013, p < 0.908$, respectively.

3) Eye Based Metrics: The second and perhaps more interesting set of metrics we used were those pertaining to eye movements and pupillometry. We analyzed saccades, blinks,

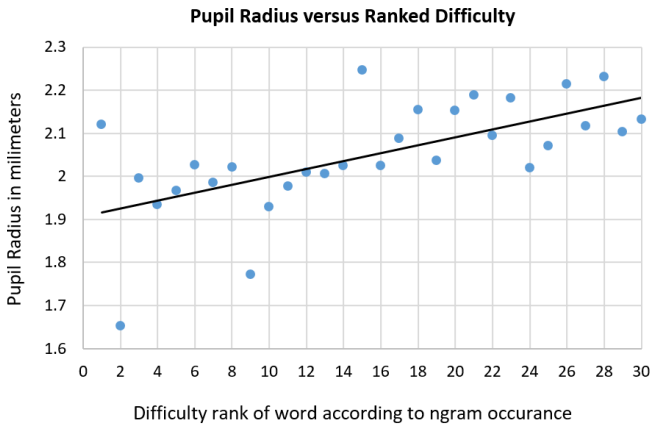

Fig. 5. Graph showing the average pupil radius in millimeters versus ranked word difficulty, ordered by number of correct answers. Pupil radius and word difficulty were well correlated, $R(28) = 0.720, p < 0.01$.

eye movement, and pupil size and movement. The most significant measures were the pupil related metrics and eye movement, which were also well correlated to difficulty.

First off, neither saccade nor blink frequency were found to be significant. These were measured by dividing the total number of occurrences over fixation time to establish frequency. No effect was found for saccades, with frequency for YES as 1.99 saccades/sec vs 1.97 saccades/sec for NO, $F(1, 263.39) = .412, p = 0.522$. Moreover, Pearson correlation was equal to R = 0.056, p = 0.882, which was not significant.

Blink frequency was also non-significant, with YES as 0.211 blinks/sec vs 0.221 blinks/sec for NO, $F(1, 174.29) = 0.06, p = 0.807$. Pearson correlation was equal to R = 0.096, p = 0.743.

Some pupillometric measures were significant between answers. Average absolute pupil radius for YES was 2.18 mm versus 1.92 for NO, $F(1, 375.59) = 2.643, p = 0.105$. Pupil radius was well correlated to ranked difficulty, $R(28) = 0.634, p < 0.01$, and to answers, $R(28) = 0.720, p < 0.01$, which is also shown in Figure 5. Pupil deviation was also very significant, resulting in YES answers at 0.478 mm/sec and NO answers at 0.371 mm/sec, which represent the magnitude of any changes in pupil size, $F(1, 376.98) = 27.42, p < 0.0001$.

Finally, the most significant metric turned out to be eye movement, i.e., the average movement per second from the time the participant began gazing at the word from the time they moved on to the next word. Results are separated in to X, Y, and total Euclidean distances. Average magnitude of X velocity (in pixels per second) was 9.276 for NO and 14.422 for YES, for which the difference was significant, $F(1, 362.39) = 43.41, p < 0.0001$. For Y velocity, this was 7.28 for NO answers versus 10.274 for YES answers, for which the difference was also significant $F(1, 353.89) = 15.098, p < 0.0001$. Total Euclidean distance was 12.048 for NO answers versus 18.086 for YES answers, for which the difference was also significant $F(1, 365.04) = 38.591, p < 0.0001$. $F(1, 365.04) = 38.591, p < 0.0001$.

TABLE I
SUMMARY OF ALL STATISTICALLY SIGNIFICANT DATA TO BE USED IN THE SVM CLASSIFIER.

Table Metric	User Selections		
	YES	*NO*	*Unit*
Fixation Time	1.73	2.75	seconds
Eye Movement (Eucl.)	18.086	12.048	pixels/second
Pupil Deviation	0.487	0.371	millimeters/second
Absolute Pupil Size	2.18	1.92	millimeters

Moreover, all movement, including the Euclidean distance from the previous X,Y position to the next yielded significant Pearson Correlations for: X mvt. vs total YES answers: $R(28) = 0.798, p < 0.01$, Y mvt. vs total YES answers: $R(28) = 0.693, p < 0.01$, X mvt. vs difficulty rank: $R(28) = 0.674, p < 0.01$, Y mvt. vs difficulty rank: $R(28) = 0.718, p < 0.01$, Euclidean total of mvt. vs answers: $R(28) = 0.801, p < 0.01$, and finally Euclidean total of Y mvt. vs rank: $R(28) = 0.724, p < 0.01$. A visual representation of eye movement vs. rank is shown in Figure 4. The most significant of these data are summarized in Table I. All of these metrics were then used to develop an SVM for classification of Understood (YES) vs. Not-understood (NO) words.

F. Classifier

For the initial SVM design, we used a single class, supervised linear SVM from the Shark library [13]. We first used all of the available data points (374 user selections) as input to the SVM and ran a full cross-validation. This resulted in a subject-agnostic model with classification accuracy of 62.8% (235 / 374) for any user. Note that a small number of outliers (16) were removed due to issues with the eye tracker disconnecting during the study.

However, further inspection showed that hard words appeared to be more difficult to classify than easy or medium words. We believe this is because hard words tend to have a lower fixation time than medium words, causing the SVM to mix up easy and hard responses.

As such, we re-ran the data to try and classify YES/NO answers for just the subset of words containing easy and medium difficulties. As we hypothesized, the SVM classification improved to 75.6% (198/262 correct). This accuracy will likely be even higher with a personalized classification model, improved eye tracking, and additional training data.

V. DISCUSSION

In the experiment, we were able to achieve between 62.8% and 75.6% accuracy for the user's understanding of a short term word recognition task. However, our SVM classifier utilizes aggregated values within a certain window of time: between when the word enters their field of view up to the point they respond. At the time a user infers meaning in a practical situation, the fixation duration is not as easy to delineate, which may affect the accuracy of classification. Creating a personalized model for each user could help alleviate bad classifications. We have plans to test this strategy in AR

scenarios as immediate future work. In addition, we may be able to use context and image classification to help determine when a user has focused on a particular object.

One other source of error from the experiment could be our detection algorithms for blinks and saccades. Several other APIs exist in addition to our custom built detectors, however we do not have a way to benchmark these algorithms against ground truth. Updated eye tracking hardware may alleviate this potential source of error in the future. Moreover, better eye tracking algorithms will also likely reduce the error in calculated pupil size, further improving classification accuracy.

A. Other Findings

One interesting and somewhat counter-intuitive result from our experiments was that medium difficulty words were more easily separable from easy words than were hard words. Through observation and post-experiment discussion with participants, we concluded that it is easy to know when a person doesn't know a word at all since there is no stored memory of the word to recall. Conversely, when a word of medium difficulty is unknown and the participant has either known it and forgotten or had some visual or aural exposure to the word, he or she will have to access his or her memory in more depth and expend more cognitive energy to determine whether or not the meaning is known. This contrasts somewhat with results found by Karolus et al. [16], where increased fixation time was correlated with lower language ability. As such, context (for example reading versus recalling an individual word) seems to play an extremely important role when deciding what data to use for which classifier.

B. Use Cases and Applications for Enhancing Memory and Cognition

To outline one potential example, consider a user who sometimes forgets to take a medication, follow procedural instructions, or interact with an object of significance. To recall an appropriate augmentation for that object at the right time, we need to understand 1) the user's context, 2) the state of that object within its context, and 3) the user's mental state in relation to that object and task.

In other words, we need to determine whether a user would say yes or no to questions such as "do you understand where you are," "do you understand this word," or "are you confused?" Being able to classify even a yes or no to one of these questions can help us display the appropriate navigation interface, word learning annotation, or medication checklist. This model is designed to be linked to a temporal database of augmentations that are associated with cognitive states and learning events over time and could even be used to help with conditions like dementia or memory loss. Much like human memory, these items, along with relevant augmentations, will be encoded into the aforementioned database and linked.

C. Future Work

Despite the existing body of research on cognitive state recognition, we still lack concrete ways to modulate or present virtual content based on the resulting output, especially for short term events. For example, many systems can determine that a user is confused or engaged in visual search over a longer period of time, but very few researchers have focused on how to overlay instructions or augmentations in response to those mental states, let alone the environment.

Our next immediate step as future work is to create a framework that develops a more personalized model for each user over time. In our planned AR framework, new users will undergo a series of basic in-situ learning tasks to initialize the classification model before starting a learning program, and that data will be used to determine an effective aggregation window for each user as well as improve overall classification accuracy. We also intend to explore other methods that may be less prone to such problems, such as time series classification and recurrent neural networks.

This study is also part of a larger project that focuses on the improvement of language learning through AR. We plan to develop a language model and intelligent tutoring system that can help users learn a new language in their own natural environments. We hope to be able to replace flash-card based systems with an automated recognition system that can provide a more intelligent, in-situ, natural way of learning through AR.

VI. Conclusion

In this paper, we set up a virtual environment with foreign language word recall tasks in order to help classify short-term understanding via eye tracking. Through experiments, we found that metrics such as pupil size and eye movement can be used to help classify whether or not a person understands a word within a window of several seconds. We demonstrated that an SVM can produce between 62.8% - 75.6% accuracy in cross-validation, which can be used to help make automated assessments of recall for learning tasks. This work adds to the body of knowledge for cognitive state recognition and can pave the way for new applications in learning and education.

Acknowledgments

This research was funded in part by the United States Office of Naval Research, grant #N62909-18-1-2036. We would also like to thank all of the experiment participants for their time.

References

[1] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. "Recent advances in augmented reality". In: *IEEE computer graphics and applications* 21.6 (2001), pp. 34–47.

[2] D. Beymer and D. M. Russell. "WebGazeAnalyzer: a system for capturing and analyzing web reading behavior using eye gaze". In: *CHI'05 extended abstracts on Human factors in computing systems*. ACM. 2005, pp. 1913–1916.

[3] R. Biedert, J. Hees, A. Dengel, and G. Buscher. "A robust realtime reading-skimming classifier". In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM. 2012, pp. 123–130.

[4] C. S. Campbell and P. P. Maglio. “A robust algorithm for reading detection”. In: *Proceedings of the 2001 workshop on Perceptive user interfaces*, pp. 1–7.

[5] D. X. Cifu, J. R. Wares, K. W. Hoke, P. A. Wetzel, G. Gitchel, and W. Carne. “Differential eye movements in mild traumatic brain injury versus normal controls”. In: *The Journal of head trauma rehabilitation* 30.1 (2015), pp. 21–28.

[6] Z. Duric, W. D. Gray, R. Heishman, F. Li, A. Rosenfeld, M. J. Schoelles, C. Schunn, and H. Wechsler. “Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction”. In: *Proceedings of the IEEE* 90.7 (2002), pp. 1272–1289.

[7] M. K. Eckstein, B. Guerra-Carrillo, A. T. M. Singley, and S. A. Bunge. “Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?” In: *Developmental cognitive neuroscience* 25 (2017), pp. 69–91.

[8] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. Al’Absi, and S. Shah. “AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field”. In: *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM. 2011, pp. 274–287.

[9] K. Fukuda, J. A. Stern, T. B. Brown, and M. B. Russo. “Cognition, blinks, eye-movements, and pupillary movements during performance of a running memory task”. In: *Aviation, space, and environmental medicine* 76.7 (2005), pp. C75–C85.

[10] P. Gamito, J. Oliveira, C. Coelho, D. Morais, P. Lopes, J. Pacheco, R. Brito, F. Soares, N. Santos, and A. F. Barata. “Cognitive training on stroke patients via virtual reality-based serious games”. In: *Disability and rehabilitation* 39.4 (2017), pp. 385–388.

[11] Google. *Google Ngram Viewer*. 2011. URL: https://books.google.com/ngrams.

[12] J. M. Henderson, S. V. Shinkareva, J. Wang, S. G. Luke, and J. Olejarczyk. “Predicting cognitive state from eye movements”. In: *PloS one* 8.5 (2013), e64937.

[13] C. Igel, V. Heidrich-Meisner, and T. Glasmachers. “Shark”. In: *Journal of machine learning research* 9.Jun (2008), pp. 993–996.

[14] Y. Itoh, J. Orlosky, and L. Swirski. “Eye Tracker Source”. In: (2017).

[15] J. Karolus, P. W. Woźniak, and L. L. Chuang. “Towards Using Gaze Properties to Detect Language Proficiency”. In: *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. ACM. 2016, p. 118.

[16] J. Karolus, P. W. Wozniak, L. L. Chuang, and A. Schmidt. “Robust Gaze Features for Enabling Language Proficiency Awareness”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. 2017, pp. 2998–3010.

[17] T. Langlotz, T. Nguyen, D. Schmalstieg, and R. Grasset. “Next-generation augmented reality browsers: rich, seamless, and adaptive”. In: *Proceedings of the IEEE* 102.2 (2014), pp. 155–169.

[18] S. P. Marshall. “Identifying cognitive state from eye metrics”. In: *Aviation, space, and environmental medicine* 78.5 (2007), B165–B175.

[19] R. Matthews, N. J. McDonald, P. Hervieux, P. J. Turner, and M. A. Steindorf. “A wearable physiological sensor suite for unobtrusive monitoring of physiological and cognitive state”. In: *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. 2007, pp. 5276–5281.

[20] B. Rhodes and T. Starner. “Remembrance Agent: A continuously running automated information retrieval system”. In: *The Proceedings of The First International Conference on The Practical Application Of Intelligent Agents and Multi Agent Technology*. 1996, pp. 487–495.

[21] H. van Rijn, J. R. Dalenberg, J. P. Borst, and S. A. Sprenger. “Pupil dilation co-varies with memory strength of individual traces in a delayed response paired-associate task”. In: *PLoS One* 7.12 (2012), e51134.

[22] U. Samadani, R. Ritlop, M. Reyes, E. Nehrbass, M. Li, E. Lamm, J. Schneider, D. Shimunov, M. Sava, R. Kolecki, et al. “Eye tracking detects disconjugate eye movements associated with structural traumatic brain injury and concussion”. In: *Journal of neurotrauma* 32.8 (2015), pp. 548–556.

[23] D. Schmalstieg and T. Höllerer. *Augmented reality: principles and practice*. Addison-Wesley Professional, 2016.

[24] D. Sonntag, J. Orlosky, M. Weber, Y. Gu, S. Sosnovsky, T. Toyama, and E. N. Toosi. “Cognitive monitoring via eye tracking in virtual reality pedestrian environments”. In: *Proceedings of the 4th International Symposium on Pervasive Displays*. ACM. 2015, pp. 269–270.

[25] K. Takano, N. Hata, and K. Kansaku. “Towards intelligent environments: an augmented reality brain–machine interface operated with a see-through head-mount display”. In: *Frontiers in neuroscience* 5 (2011), p. 60.

[26] T. Toyama, D. Sonntag, J. Orlosky, and K. Kiyokawa. “Attention engagement and cognitive state analysis for augmented reality text display functions”. In: *Proceedings of the 20th international conference on Intelligent user interfaces*. ACM. 2015, pp. 322–332.

[27] Y.-F. Tsai, E. Viirre, C. Strychacz, B. Chase, and T.-P. Jung. “Task performance and eye activity: predicting behavior relating to cognitive workload”. In: *Aviation, space, and environmental medicine* 78.5 (2007), B176–B185.

[28] G. Westerfield, A. Mitrovic, and M. Billinghurst. “Intelligent Augmented Reality Training for Motherboard Assembly”. In: *International Journal of Artificial Intelligence in Education* 25.1 (Mar. 2015), pp. 157–172.