Towards Reading Trackers in the Wild: Detecting Reading Activities by EOG Glasses and Deep Neural Networks

Shoya Ishimaru
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
Shoya.Ishimaru@dfki.de

Koichi Kise
Osaka Prefecture University
Sakai, Japan
kise@cs.osakafu-u.ac.jp

Kensuke Hoshika
Osaka Prefecture University
Sakai, Japan
hoshika@m.cs.osakafu-u.ac.jp

Andreas Dengel
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
Andreas.Dengel@dfki.de

Kai Kunze
Keio University
Yokohama, Japan
kai.kunze@gmail.com

UbiComp/ISWC'17 Adjunct , September 11–15, 2017, Maui, HI, USA
ACM 978-1-4503-5190-4/17/09.
https://doi.org/10.1145/3123024.3129271

Abstract
Reading in real life occurs in a variety of settings. One may read while commuting to work, waiting in a queue or lying on the sofa relaxing. However, most of current activity recognition work focuses on reading in fully controlled experiments. This paper proposes reading detection algorithms that consider such *natural readings*. The key idea is to record a large amount of data including natural reading habits in real life (more than 980 hours from 7 participants) with commercial electrooculography (EOG) glasses and to use them for deep learning. Our proposed approaches classified *controlled reading* vs. *not reading* with 92.2% accuracy on a user-dependent training. However, the classification accuracy decreases to 73.8% on *natural reading* vs. *not reading*. The results indicate that there is a strong gap between controlled reading and natural reading, highlighting the need for more robust reading detection algorithms.

Author Keywords
Eye movement; electrooculography; reading; quantified self; convolutional neural network; recurrent neural network

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

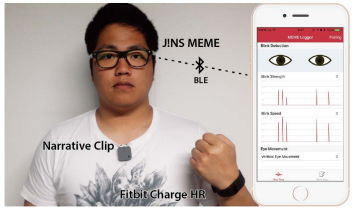


Figure 1: The recording setup in this work. Participants wore commercial sensors every day for more than two weeks.

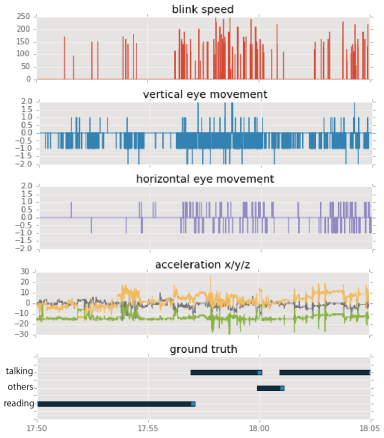


Figure 2: An overview of the sensor signals exported by J!NS MEME. The ground truth is annotated by the participant at the end of the day with reviewing images of Narrative Clip.

Introduction

Reading is an integral part of our daily lives, as texts remain one of the vital sources of information and knowledge even in this age of multimedia. The cognitive benefits of reading (e.g. better vocabulary skills) and the benefits of increased reading volumes, are well explored in the fields of education and cognitive science [5, 17]. However, we know still little about reading habits of people. Building measurement tools for daily reading can help understand those habits better. The motivation of this work is to improve people's cognitive abilities by developing an application that motivates people to read more on a daily basis.

As one can be motivated to exercise more by wearing a pedometer, quantifying the activity is the first step to changing one's habits [15]. Some common physical activities (e.g., walking and sleeping) can be recognized and measured by body-mounted sensors. However, recognizing reading activity is still a challenging task because the body movements during reading are limited, making it hard to apply general motion-sensor based approaches. One of the most promising directions for reading detection is the use of eye movements. There is extensive research on eye movements during reading in psychology, cognitive science, and pervasive computing [16, 4].

However, as far as we know, most of the previous work has involved specific reading behaviours during controlled experiments and has not covered natural reading activities in a real life setting. Reading in real life occurs in a variety of settings, involving various devices and document layouts that would result in irregular eye movements. Our assumption is that there is a substantial difference between controlled reading and natural reading, meaning that the reading detection methods that work in labs may not necessarily be usable in the wild.

We believe that developing less obtrusive optical eye trackers is key to achieving reading quantification in real life settings. In this regard, using commercially available electrooculography (EOG) glasses seems promising since they are relatively light, visually familiar (looking like conventional eyewear) and have sufficient battery life for all-day use. Their cost is also relatively low, making them suitable for conducting large-scale data recording [1]. In this work, we record natural reading activities using commercial EOG glasses (see Figure 1) and evaluate the accuracy of the detection algorithm in the wild.

All of the work described in this paper was done with the permission of the research ethics committee of the graduate school of engineering, Osaka Prefecture University.

Approaches

We propose three types of reading detection approaches in this research. The first is a manual feature extraction based approach. In this approach, we analyze the data obtained from the devices to find characteristic sensor patterns during reading, and select the features for manual classification. The second and third approaches are automatic feature extraction based. We designed a convolutional neural network (CNN) and recurrent neural network with Long short-term memory (LSTM) for classifying the raw data. They extract best features by training with large-scale data. This section describes the sensing device and details of the three reading detection methods.

J!NS MEME

We utilize J!NS MEME[1] for the sensing. The device is equipped with three electrodes for eye movement detection and a 6-axis internal measurement unit (IMU) for head movement detection. It is developed by JIN CO., LTD. The company

[1] https://jins-meme.com/en/

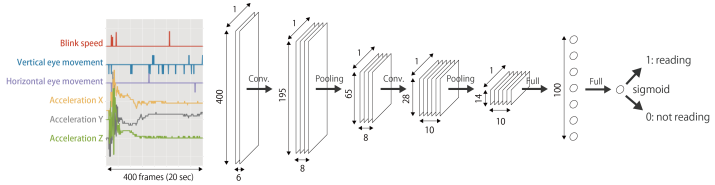


Figure 3: The CNN architecture for the reading detection

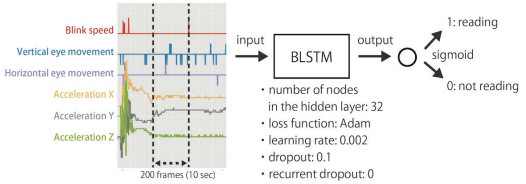


Figure 4: The LSTM architecture for the reading detection

Table 1: Features for the SVM based reading detection approach

1	freq. of eye blinks	mean
2		STD
3	freq. of eye move up	mean
4		STD
5	freq. of eye move down	mean
6		STD
7	freq. of eye move left	mean
8		STD
9	freq. of eye move right	mean
10		STD
11	raw signal of acc. x	mean
12		STD
13	raw signal of acc. y	mean
14		STD
15	raw signal of acc. z	mean
16		STD

has released two models of the device: the developer's version and the academic version. We used the former for this research, as it's widely available to consumers. J!NS MEME calculates basic eye movements (blink speed, blink strength, two-step strength of up/down/left/right eye movements) internally on the device itself as shown in Figure 2, and stream them with IMU data to a smartphone via Bluetooth Low Energy. The sampling rate is 20 Hz. Battery run time is 18 hours, which is sufficient for gathering data all day during the day. Data are recorded on the iOS application, MEMELogger[2] and sent to a hosted server every day.

Approach 1: SVM Based Reading Detection

We employ 16 statistical features from the sensor signals of J!NS MEME (10 features from eye movements and 6 from head movements) as shown in Table 1. The frequencies of eye blinks and eye movements are calculated as inverse values of the duration between two blinks or eye movements. Acceleration x, y and z are raw signals. We created samples with the window size of 60 seconds. After the data were normalized and whitened, we calculated the mean and standard deviation for each of the sensor values in the window. Support Vector Machine (SVM) with RBF kernel is used for learning and classification in this approach. After the classification, we applied majority voting for 5 minutes' worth of data to smooth out the results.

[2]https://itunes.apple.com/en/app/memelogger/id1073074817

Approach 2: CNN Based Reading Detection

An overview of the CNN architecture is shown in Figure 3. The network receives raw sensor values from J!NS MEME as inputs, and classifies the gaze activity as either reading or not reading. For the input layer, 6 maps with a size of 400×1 were created from 400 frames of 6 sensors' values, including blink speed, vertical eye movement, horizontal eye movement, acceleration x, y, and z. To increase the number of training samples, we employed different input window size (20 seconds) compared to SVM. There is no overlap between windows. The network has two convolution layers, each followed by a pooling layer. For the first convolution layer, the approach utilize a filter with size 12×1 with step 2 that exports 8 maps. Since the convolution is done without zero-padding, the window goes from 400 to 195. Then the approach utilize an max pooling with a stride of 3 to the 8 maps, thus maps with size 65×1 are exported. The same process with filtering size 11×1 and max pooling stride 2 are applied for the second convolution and pooling. Finally, 10 maps with size 14×1 are fully linked to 100 units, and fully linked to the output channel with 2 units: reading or not reading. Activation functions are rectified linear units (ReLU). We employ dropout with dropping rate 0.5 in each pooling and full connecting.

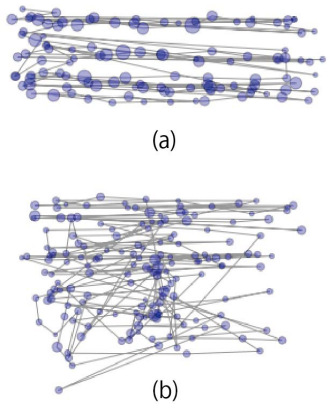


Figure 5: Eye gazes during a minute of (a) controlled reading and (b) natural reading. The data were collected by Tobii eyeX and classified into fixations (circles) and saccades (lines).

Approach 3: LSTM Based Reading Detection
By utilizing the advantage of the characteristics of time series data, we have also designed the network architecture including LSTM [9]. The input shape and parameters of the architectures are described in Figure 4. The parameters of the network were selected by random search. Since our purpose is to quantify reading activities and give feedback to a user later in the same way with physical activity tracker, a realtime analysis is not necessarily required. Therefore Bidirectional LSTM is utilized to precede high accuracies.

After the both of classifications, we apply majority voting for 5 minutes of data (as we did in the SVM approach) to smooth the results.

Data Recording

We asked 7 participants to record their habits using the following commercial sensors: J!NS MEME, Fitbit Charge HR, Narrative Clip and Tobii eyeX (see Figure 1). Note that Fitbit Charge HR and Tobii eyeX are not used in this experiment. All of the participants were college students studying computer science, who worked on computers most of time. They used the tracking/recoding devices during the day and charged them while they slept for more than two consecutive weeks. The dataset contains 22 hours of controlled reading, 427 hours of natural reading, 156 hours of social interactions and 375 hours of other activities.

Natural Reading Activity
We did not place any limit on the participants' activities. Therefore various types of reading activity are included in the dataset. Participants, for example, read texts on computers, smartphones, e-book readers, as well as paper. Browsing web pages and typing on a computer were also labeled as natural reading.

Controlled Reading Activity
To record enough labeled reading activities, we also conducted a controlled experiment. We prepared 60 documents and asked the participants to read them from beginning to end. They read 15 English documents on paper, 15 English documents on a screen, 15 Japanese documents on paper, and 15 Japanese documents on a screen. Reading on paper was recorded with J!NS MEME, and reading on a screen was recorded with J!NS MEME and Tobii eyeX. We did not prohibit them from reading back during the recording, but most of them read documents continuously without vertical movements. Figure 5 represents the example of the difference of eye movements while natural reading and controlled reading.

Narrative Clip for Ground Truth Annotation
For the purpose of collecting ground truth, the participants added annotations to all data. They were asked to apply one of the three labels ("reading", "talking", and "other activities") to every 1 minute of data from 0:00 to 23:59. To help with the labeling tasks, we provided each participant with a Narrative Clip [3], a small life-logging camera which can be clipped to one's clothing. Narrative Clip takes a picture every 30 seconds. Participants reviewed the pictures at the end of each day and manually labelled their activities. In order to reduce ambiguities of the labels among participants, we asked them to label activities if pertinent objects (e.g. book, display, person) appeared in more than two consecutive pictures (= one minute). They submitted the annotated pictures after removing some of them for privacy reasons. The reason we asked them to label their activities at the end of each day instead of during the recording is to make the dataset "wild" as much as possible. Regularly asking participants to provide ground truth labels leads to a well annotated dataset but might change their regular behaviors.

[3] http://getnarrative.com/

Evaluation

We evaluated the reading detection approaches on our long-term dataset with user-independent and user-dependent learning. This section presents procedures of the evaluation and classification results.

Experimental Condition

For user-independent learning, training and testing data were separated by leave-one-subject-out cross validation. Samples of one participant were utilized as testing data, and samples of others were utilized as training data.

For user-dependent learning, training and testing data consist of samples from one participant. During our experiment, a new CSV file was created every time when a participant started recording. We shuffled the order of files and divided them to two groups equally. Samples in one groups were utilized as training data and the other were utilized as testing. The reason we employed this way is to prevent carelessly mixing training and testing samples. Applying cross validation with all samples is the easiest way. But it might lead to incorporation of very similar samples into training and test folds in the analysis of time series data [8].

The mean and standard deviation value of results were calculated over all 7 participants. Because the number of samples in each class is unbalanced, "class weight" functions implemented in machine learning frameworks (scikit-learn for SVM based and Keras with TensorFlow for CNN and LSTM based) were utilized during training the model.

Results

Table 2 shows results comparing the SVM, CNN, and LSTM based approaches. The SVM based approach is more accurate than other two approaches to detect controlled reading. Although the differences are small, deep learning approaches performed better to detect natural reading.

	controlled reading		natural reading	
	user-indep.	user-dep.	user-indep.	user-dep.
SVM	**80.7±8.0%**	**92.2±7.2%**	68.5±7.2%	73.1±5.3%
CNN	66.2±20.6%	80.2±12.3%	**69.6±7.1%**	70.0±5.4%
LSTM	74.3±17.5%	90.4±5.8%	67.1±10.1%	**73.8±6.0%**

Table 2: Means and standard deviations of classification accuracies over 7 participants (controlled/natural reading vs. not reading)

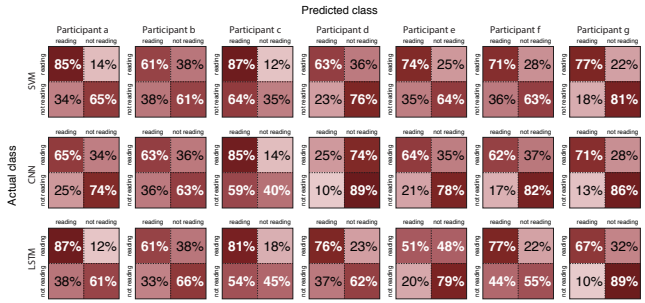
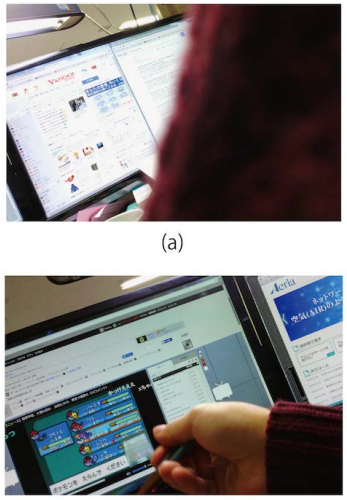


Figure 6: Confusion matrices of natural reading vs. not reading on the user-independent approaches.

Confusion matrices of the natural reading vs. not reading classification on the user-independent approach are shown in Figure 6. For most of the participants, except for Participant *c*, the results show high precision (true positives divided by true positives + false positives) and low recall (true positives divided by true positives + false negatives). This result indicates that there are some reading activities that are still difficult to be detected by the three approaches.



(a)

(b)

Figure 7: (a) False negative: a user is browsing web pages. (b) false positive: a user is watching a video.

Discussion

By reviewing the pictures taken by Narrative Clip, we identified some cases in which an activity can be misclassified. For example, while all the participants labeled Web browsing (See Figure 7-a) as "reading", this activity was sometimes misclassified by the CNN and LSTM as not reading. This may have been caused by the combination of multiple factors, such as the web page layout that combines structured and non-structured texts (e.g., short text passages, banners, ads, etc.) as well as the actions that accompany web browsing, such as clicking on the embedded URIs. An interesting case of false positive occurred when one of the participants was watching a video (See Figure 7-b). The participant himself labeled this activity as "not reading", but our CNN and LSTM based user-independent approach classified it as reading. The participant was watching the video on *www.nicovideo.jp*, a popular video sharing service in Japan, which famously shows many floating subtitles in the videos. This has likely provided some irritations for the classifier.

A major problem we found through this experiment is in labeling ground truth accurately for natural reading. Because the act of reading differs in kind (e.g. reading a paper book, browsing web pages, skimming texts, etc.), classifying activities into the simple two classes (reading vs. not reading) can sometimes be difficult even for humans.

Related Work

Eye Movements and Activity Recognition

Bulling et al. are the first to explore tracking eye movements in natural settings using Electrooculography glasses. They looked into reading (also including reading on the move), as well as other everyday activities and cognitive processes [3, 2]. As a pioneering study, their initial datasets were more controlled than natural, and the study employed relatively obtrusive experimental setups. For example, the participants wore a prototype EOG system (taped around the eyes) or optical eye trackers.

Another study that uses eye movements for analysing reading is done by Kunze et al. [13, 12]. They present some methods for counting the number of words read by the user using mobile eye trackers and EOG systems [14, 10]. For reading detection, however, they rely on artificial reading tasks, and the study is limited to lab experiments.

A study by Ishimaru et al. works at distinguishing several activities on a Google Glass head-mounted computer using only head motion and blink frequency [11]. The target activities include reading, talking, watching videos, etc. The reading tasks in this research are also controlled and do not include natural reading in a daily life setting.

Activity Recognition in the Wild

While increasing reading volume seems to provide substantial cognitive benefits (such as improved vocabulary and critical thinking skills), it is still difficult to determine what constitutes a healthy reading habit [5]. One major reason for that is the lack of tools to quantify reading in a real life setting for a long term.

There are still few long-term datasets in the real environment on reading and other cognitive activities. One major reason is the lack of unobtrusive technology to make long-term tracking possible. There are some datasets contributed by computer vision researchers working on egocentric vision, which are mostly camera recordings, but some also include eye gaze data [7, 6]. Most notably, Steil and Bulling [18] contributed a long-term eye tracking and egocentric vision dataset with over 80 hours of recording in a natural uncontrolled environment.

Position of This Study

The experimental setups used in previous research often involved bulky prototype setups that interfered with the natural flow of the activities, making the users (and bystanders) aware of the fact that they are recorded. Of course, efforts for miniaturization in recent years have produced mobile eye trackers that are far less obtrusive (e.g. Tobii and SMI eye tracking glasses), far less expensive and more open for research (most notably the Pupil Labs eye tracker) compared to the earlier setups. Most of these devices are, however, still quite noticeable, and may not be well suited for wearing in public. The goal of our research is to explore ways to quantify natural reading habits for a long term, using affordable technologies that are less obtrusive and more socially acceptable for daily usage. In this paper, we presented our initial effort towards this goal, providing a dataset recorded with commercially available, truly "wearable" devices that can still give some insights into one's natural reading activities.

Conclusion and Future Work

In this work, we recorded natural activities in a daily life setting with unobtrusive, commercially available devices. By sacrificing accuracy to a degree, the amount of the dataset reached to more than 980 hours. The recorded data revealed that "natural reading" is a complex activity that includes many factors, as reading plain texts and browsing websites for instance involve different kinds of eye movements. We proposed three approaches to reading detection and found that the deep learning based approaches are superior to the SVM-based approach to detect natural reading activity. By investigating error samples, we have uncovered some of the challenges in detecting natural reading, including how to collect large-scale data with ground truth.

Next we want to explore the large volume of data we gathered but did not use for the purpose of the present study in future work. Such data include recordings of the eye gaze while reading on a screen with Tobii eyeX and the heart rates while reading with Fitbit Charge HR. It should be interesting to see the relationship between J!NS MEME's data and the data obtained by other sensors, and estimate the user's cognitive state such as the level of attention, concentration, and understanding of the contents.

Acknowledgements

This work is supported by JST CREST and JSPS KAKENHI (Grant Numbers: JPMJCR16E1, 17K12728).

REFERENCES

1. Oliver Amft, Florian Wahl, Shoya Ishimaru, and Kai Kunze. 2015. Making Regular Eyeglasses Smart. *Pervasive Computing, IEEE* 14, 3 (2015), 32–43.
2. Andreas Bulling and Daniel Roggen. 2011. Recognition of visual memory recall processes using eye movement analysis. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 455–464.
3. Andreas Bulling, Jamie A. Ward, and Hans Gellersen. 2012. Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors. *ACM Trans. on Applied Perception* 9, 1 (2012), 2:1–2:21.
4. Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. 2011. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence* 33, 4 (2011), 741–753.
5. A.E. Cunningham and K.E. Stanovich. 2001. What reading does for the mind. *Journal of Direct Instruction* 1, 2 (2001), 137–149.

6. Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. 2014. You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video.. In *BMVC*.
7. Joydeep Ghosh, Yong Jae Lee, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1346–1353.
8. Nils Y Hammerla and Thomas Plötz. 2015. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1041–1051.
9. Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
10. Shoya Ishimaru, Kai Kunze, Koichi Kise, and Andreas Dengel. 2016. The wordometer 2.0: estimating the number of words you read in real life using commercial EOG glasses. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 293–296.
11. Shoya Ishimaru, Kai Kunze, Koichi Kise, Jens Weppner, Andreas Dengel, Paul Lukowicz, and Andreas Bulling. 2014. In the Blink of an Eye: Combining Head Motion and Eye Blink Frequency for Activity Recognition with Google Glass. In *Proceedings of the 5th Augmented Human International Conference*. ACM, Article 15, 4 pages.
12. K. Kunze, A. Bulling, Y. Utsumi, S. Yuki, and K. Kise. 2013a. I know what you are reading – Recognition of document types using mobile eye tracking. In *Proceedings of the 2013 ACM International Symposium on Wearable Computers*.
13. Kai Kunze, Hitoshi Kawaichi, Kazuyo Yoshimura, and Koichi Kise. 2013b. The Wordometer–Estimating the Number of Words Read Using Document Image Retrieval and Mobile Eye Tracking. In *2013 12th International Conference on Document Analysis and Recognition*. IEEE, 25–29.
14. Kai Kunze, Katsutoshi Masai, Masahiko Inami, Ömer Sacakli, Marcus Liwicki, Andreas Dengel, Shoya Ishimaru, and Koichi Kise. 2015. Quantifying reading habits: counting how many words you read. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 87–96.
15. Susan Michie, Charles Abraham, Craig Whittington, John McAteer, and Sunjai Gupta. 2009. Effective techniques in healthy eating and physical activity interventions: a meta-regression. *Health Psychology* 28, 6 (2009), 690–701.
16. Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.
17. KE Stanovich and AE Cunningham. 1998. What reading does for the mind. *American Education Journal* (1998).
18. Julian Steil and Andreas Bulling. 2015. Discovery of everyday human activities from long-term visual behaviour using topic models. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 75–85.