Self-supervised Learning for Reading Activity Classification

MD. RABIUL ISLAM, Osaka Prefecture University, Japan and BSMRSTU, Bangladesh
SHUJI SAKAMOTO, Osaka Prefecture University, Japan
YOSHIHIRO YAMADA, Osaka Prefecture University, Japan
ANDREW W. VARGO, Osaka Prefecture University, Japan
MOTOI IWATA, Osaka Prefecture University, Japan
MASAKAZU IWAMURA, Osaka Prefecture University, Japan
KOICHI KISE, Osaka Prefecture University, Japan

Reading analysis can relay information about user's confidence and habits and can be used to construct useful feedback. A lack of labeled data inhibits the effective application of fully-supervised Deep Learning (DL) for automatic reading analysis. We propose a Self-supervised Learning (SSL) method for reading analysis. Previously, SSL has been effective in physical human activity recognition (HAR) tasks, but it has not been applied to cognitive HAR tasks like reading. We first evaluate the proposed method on a four-class classification task on reading detection using electrooculography datasets, followed by an evaluation of a two-class classification task of confidence estimation on multiple-choice questions using eye-tracking datasets. Fully-supervised DL and support vector machines (SVMs) are used as comparisons for the proposed SSL method. The results show that the proposed SSL method is superior to the fully-supervised DL and SVM for both tasks, especially when training data is scarce. This result indicates the proposed method is the superior choice for reading analysis tasks. These results are important for informing the design of automatic reading analysis platforms.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Self-supervised learning, reading detection, confidence estimation, reading analysis, fully-supervised deep learning

ACM Reference Format:
Md. Rabiul Islam, Shuji Sakamoto, Yoshihiro Yamada, Andrew W. Vargo, Motoi Iwata, Masakazu Iwamura, and Koichi Kise. 2021. Self-supervised Learning for Reading Activity Classification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 105 (September 2021), 22 pages. https://doi.org/10.1145/3478088

1 INTRODUCTION

Reading analysis is essential for developing human learning strategies because it is possible to obtain a wide variety of information from reading activities [43]. The development of technologies such as eye-trackers and electrooculography (EOG) glasses means that useful data can be obtained via stationary devices and wearable sensing devices [6, 29]. There are numerous types of aspects of reading that can be analyzed and classified such

Authors' addresses: Md. Rabiul Islam, rabiul@m.cs.osakafu-u.ac.jp, Osaka Prefecture University, Sakai, Japan and BSMRSTU, Gopalganj, Bangladesh; Shuji Sakamoto, sakamoto@m.cs.osakafu-u.ac.jp, Osaka Prefecture University, Sakai, Japan; Yoshihiro Yamada, imenurok@yahoo.co.jp, Osaka Prefecture University, Sakai, Japan; Andrew W. Vargo, awv@m.cs.osakafu-u.ac.jp, Osaka Prefecture University, Sakai, Japan; Motoi Iwata, iwata@cs.osakafu-u.ac.jp, Osaka Prefecture University, Sakai, Japan; Masakazu Iwamura, masa@cs.osakafu-u.ac.jp, Osaka Prefecture University, Sakai, Japan; Koichi Kise, kise@cs.osakafu-u.ac.jp, Osaka Prefecture University, Sakai, Japan.

2474-9567/2021/9-ART105 $15.00
https://doi.org/10.1145/3478088

as reading detection, where the objective is to detect whether the user is reading or not [7, 31]. Other research has tackled issues like identifying the type of text being read, such as reading English or Japanese text [23]. Another classification task is a problem-solving task such as confidence estimation in answering multiple-choice questions (MCQs) [59]. In this paper, we use the term "reading activity" to cover not only the activity of reading plain text but also problem-solving tasks completed via reading.

With the data collected from sensing technologies, reading analysis can be conducted in multiple ways. Traditional machine learning methods have achieved satisfactory results in laboratory settings where features are manually selected, which require feature engineering expertise. In addition, these methods may not produce similar results outside of the lab [19] due to noise which obfuscates features that need to be extracted. Deep Learning (DL), on the other hand, has been successfully applied in the areas of image recognition [28], speech recognition [13], natural language processing [36], human activity recognition [45], and eliminates manual feature engineering. The key to successful DL is to prepare enough labeled samples for training the network. In most fields, accumulating enough labels is a serious issue [44].

The lack of labeled data is also a problem for reading activity classification. Obtaining large curated reading activity datasets is problematic because the annotation costs and time it takes to generate a satisfactory dataset are prohibitive. In addition, the diversity of devices and embedded sensors, variations in specifications regarding sampling rates, and different deployment environments make dataset construction a challenge. For these reasons, it is difficult to apply a fully-supervised DL method in this domain directly.

Self-supervised Learning (SSL) presents a potential solution [18, 41]. This method employs a pretext task [24] for feature learning before training for the task of interest (target task). The pretext task is different from the target task, but the network representation acquired based on it is effective for the target task. Because labeled samples for the pretext task are generated without manual labeling, the network can be trained with a much larger amount of data. In general, this helps to improve classification accuracy. For reading activity classification in particular, SSL can also provide certainty in performance compared to existing technologies where performance is unknown.

In the field of human activity recognition (HAR), some researchers have successfully attempted to employ SSL to solve the issue of the lack of labeled data by employing simple signal transformations to produce the pretext task for sensor data [17, 46]. Their goal is to recognize *physical* human activities by using accelerometers (ACC) and gyroscopes (GYRO) as sensors to distinguish body movements that characterize activities. However, we do not know if similar approaches are effective for recognizing *cognitive* human activities such as reading, where the target task is cognitively intensive with fewer bodily movements, and biological signals such as eye movements are captured by sensors. The fundamental differences between how *physical* and *cognitive* activity data are collected present a challenge to the generalizability of SSL to cognitive activities.

This research aims to clarify how effective SSL is at solving the labeled data issue in the cognitive HAR area of reading activity classification. We propose an SSL method and evaluate it for two different but related reading activity classification tasks placed at two extreme points on the reading activity spectrum. The first one is reading detection, a low-level reading activity relevant to the *quantity* of reading, where periods of reading are identified throughout the day. The unique point is that we attempt fine-grained reading detection, which distinguishes subordinate categories of reading: identification of periods of reading texts in different scripts and layout such as horizontally written English, horizontally written Japanese, vertically written Japanese, and not reading anything. Identification of subordinate categories allows us to obtain detailed information about reading activities. For example, vertically written Japanese often indicates a user is reading material such as novels and newspapers. The second one is confidence estimation: whether the user is confident or not on the answer of MCQs, which is a high-level reading activity relevant to the *quality* of reading. By identifying the confidence in the answer, in addition to its correctness, we can produce better strategies for personalized review [22]. For example, a

correct but unconfident answer needs review to ensure the knowledge can be used in the future. An incorrect and confident answer requires more attention to revise the incorrect knowledge.

Our trials on both quantity and quality of reading activities allow us to obtain a full picture of the effectiveness of the proposed SSL method across the reading activity spectrum. In the evaluation process, we recorded eye movement using EOG glasses for reading detection and eye gaze using an eye-tracker for confidence estimation. We compared the effectiveness of the proposed SSL method by training and evaluating the network for a different number of training samples per class, starting from the availability of all samples per class to 10 samples per class. We used the fully-supervised DL as a comparative method along with support vector machines (SVMs), a traditional machine learning method, as a baseline.

The results show that the proposed SSL method is superior compared to other methods at both tasks, e.g., reading detection and confidence estimation. Specifically, the proposed SSL method demonstrates better performance than the fully-supervised DL except at the largest number of training samples, where the proposed SSL method performs equally well. Although the fully-supervised DL performs worse than SVM with a smaller number of training samples, due to the impact of insufficient training samples, the proposed SSL method does not face this problem; it is always superior. The statistical analysis supports the above statements.

From the results, we conclude that the proposed SSL method is superior to other methods over a wide range of training samples on both tasks and is equal to fully-supervised DL when numerous training samples are available. Thus, we can recommend the SSL method for any size of available training samples. This insight can help system designers, and researchers more efficiently pursue reading activity classification.

The main technical contributions of this paper are as follows:

- We propose an SSL method for cognitive HAR for the cases of reading activity classification. It is a novel contribution because existing methods with SSL in the field of HAR are for physical HAR.
- Reading activity classification needs new sensors, EOG, and eye-tracker, that are not used in physical HAR. Pretext tasks for SSL are newly proposed for dealing with those sensors.
- The task of reading detection is fine-grained, which is more difficult but informative for cognitive human activities. Although a method with traditional machine learning cannot work well for fine-grained reading detection, the proposed SSL has been successful to improve the accuracy.
- We employed in-the-wild datasets for realistic evaluation of the proposed method.

The remainder of the paper is organized as follows: Section 2 presents related work on reading detection, confidence estimation, and SSL. Section 3 presents the proposed SSL method. Section 4 presents the details of datasets, and the data collection framework. In section 5 we present the experimental conditions, results, and discussion. Finally, section 6 presents the conclusion and future work.

This work has been evaluated and approved by the ethical committee of our institute.

2 RELATED WORK

Our work relates to several active research areas, including reading detection, confidence estimation, and SSL. In this section, we describe how our work builds on these fields.

2.1 Reading Detection

Reading detection strategy varies depending on its purpose, and over the past years, researchers have proposed many methods for different kinds of automatic reading detection. For example, they have proposed methods for reading detection as a part of other human activities such as reading in transit [5, 51], in office settings [6], and with talking [20, 21] by exploring eye movements in controlled settings using classical machine learning approaches. In another eye-based activity recognition study [50], authors detected reading with desktop activities such as search and writing by using traditional machine learning methods. Many existing studies attempt to

explore reading activity as an individual activity accomplishing different modes, such as regular reading, detailed reading, skimming, and spell-checking [52]. Researchers used traditional machine learning methods to detect whether the user is reading or skimming [4, 27], reading or searching [7], and reading or not reading [19, 31] in laboratory settings. Recently, Ishimaru et al. [23] proposed a classical machine learning method to classify the language of text segments, English or Japanese, read by the user. They were able to demonstrate an ability to differentiate the language of the text by analyzing eye movement data obtained through an in-the-wild study.

The literature on reading detection suggests a strong relationship between eye movements and reading activities. Moreover, the existing body of work provides evidence that eye movement plays a crucial role in reading detection, making it compelling to explore the development of eye-based reading detection systems. However, most of the existing methods occur in laboratory settings and use classical machine learning approaches, with the exception of some preliminary work that applies DL methods [9, 19], and pays very little attention in classifying the scripts and layout of texts read by users. Although classical machine learning approaches can produce satisfactory results in laboratory settings, they may not do so in-the-wild [23].

2.2 Confidence Estimation in Answering Multiple-Choice Questions

MCQs are fundamental forms of assessing knowledge, ability, and performance [14] and are popular since they offer quick and objective scoring. However, random guessing can result in correct answers. A user may answer correctly and know the subject matter, but a user may guess the correct answer without understanding the subject matter [32]. The user may also be skilled in answering MCQs correctly [37]. Moreover, a user may answer correctly even though the user is confused by other options [39], or the user may be confident even though the answer is incorrect. Therefore, only correctness does not indicate understanding, which is a significant drawback of MCQ assessment [48] whereby the user's understanding of the subject material may be misunderstood [8]. Therefore, an assessment system that provides feedback is essential for both users and instructors. Since it is not possible to manually track users, there is a need to develop automatic confidence estimation.

Researchers have proposed some methods for automatic confidence estimation. Tsai et al. [55] analyzed user's visual attention spans when solving MCQs by using eye-tracking under laboratory settings and with the application of a traditional machine learning method. The results show that successful problem solvers focus and spend more time examining relevant factors than irrelevant ones, while unsuccessful problem solvers spend more time decoding the problem, and have difficulty in recognizing the relevant factors. Yamada et al. [59] proposed a method to classify whether a user is confident or not when answering MCQs through manually selected features from the eye gaze recorded with an eye-tracker in a controlled environment and applied a traditional machine learning approach.

2.3 Self-supervised Learning

In the past decade, the development and application of DL have successfully solved many problems in the field of ubiquitous computing [15], health [16], well-being [34], and similar. Most of the methods use fully-supervised DL approaches that need large and carefully labeled data that is unfeasible in most of the research domains, because of the tremendous cost required to manually label data. To overcome the innate limitations of the fully-supervised DL approaches, researchers introduced several unsupervised methods [30, 35].

Recently, researchers proposed a DL technique called SSL [1, 2, 12] that encourages a network for representation learning [33] by using unlabeled data and their automatically generated *labels*. It is desirable to obtain precise labels without human supervision, but automatic labeling is not possible in many tasks. Hence, a kind of SSL that utilizes an auxiliary task called pretext tasks to obtain better feature representation is used. Generally speaking, a pretext task should meet the following requirements; the labels of unlabeled data for the task can be automatically generated, the pretext task is somehow related to the target task, and the pretext task is of appropriate difficulty.

Representative pretext tasks include classification [11], image repair [42], image patch alignment [10], and jigsaw puzzle [40]. SSL which employs a pretext task has various applications in different research domains [3, 38, 49]. In physical HAR tasks, researchers applied SSL for learning representations by solving pretext tasks using a large amount of wearable sensor data that aims to enhance performance [33]. A technique is proposed in [46] by generating pretext tasks by utilizing simple signal transformations. A similar technique is proposed in [47] for representation learning from raw sensory data to mitigate data limitations and shows that method is effective even for small sized data. Likewise, masked reconstruction is proposed as a viable self-supervised pre-training objective for time series data of HAR [17]. Similarly, an SSL method is proposed for enhancing performance by pre-training the network by predicting the values of sensor signals in future time-steps [53] and their findings show that the SSL technique is effective in boosting the performance.

Inspired by the successful application of the SSL technique to address the issue of insufficient labeled data, specifically in physical HAR with sensor data, we explore the generalized efficacy of SSL for eye movement sensory data for cognitive activities such as reading.

3 PROPOSED METHOD

We propose an SSL method for reading activity classification using sensor data, as shown in figure 1 that consists of two stages. The first stage shown in the upper parts of figures 1(a) and (b) is self-supervised pre-training consisting of solving the pretext task. The second stage shown in the lower parts of figures 1(a) and (b) is target task training by fine-tuning the pre-trained base network using labeled sensor data.

We implemented the proposed SSL method on two different but related reading activity tasks: reading detection and confidence estimation in answering MCQs. The reason is that reading activities are distributed on a wide spectrum. For example, some activities are related to the quantity of reading, while other activities involve the quality of reading, such as understanding and confidence. To investigate the applicability of the proposed SSL method, we apply it to tasks in these two categories: reading detection that differentiates reading periods from all activities and confidence estimation in answering MCQs. These two activities are also recorded using different devices: EOG glasses for reading detection and an eye-tracker for confidence estimation. We consider that the proposed method is generalizable if it works for both tasks.

3.1 Reading Detection

Reading detection aims to differentiate periods of reading from all other activities. This is implemented as a classification task; the user activities are divided into short segments and then classified into one of the predefined classes of activities. We do not just classify "reading" and "not reading", but include fine-grained classes. We classify segments into one of the following four classes: reading English text (EN), reading Japanese horizontal text (JH), reading Japanese vertical text (JV), and not reading (NR). Japanese text can be written horizontally or vertically. Japanese horizontal writing is written left to right with no spaces between words. In the vertical writing system, characters are read from top to bottom, going right to left [57]. The reason we employ these fine-grained classes is that we can obtain detailed information about the user's reading activities. For instance, class JV often indicates that the user is reading novels or newspapers, class EN suggests that the user is reading something technical (scientific materials if the user is a science student), and class JH includes various materials.

The devices employed to measure reading detection are EOG glasses that generate EOG data of eye movements, and ACC and GYRO data from the movement of the EOG glasses themselves. From the EOG signals, we obtained data of horizontal and vertical eye movements. ACC and GYRO data consist of x, y, and z components. In total, we collected eight different kinds of data. The details of data recording are described in section 4.1.

3.1.1 Self-supervised Pre-training. Self-supervised pre-training involves learning the representation of signal by using a pretext task to understand the fundamental characteristics of the signal. Researchers proposed pretext

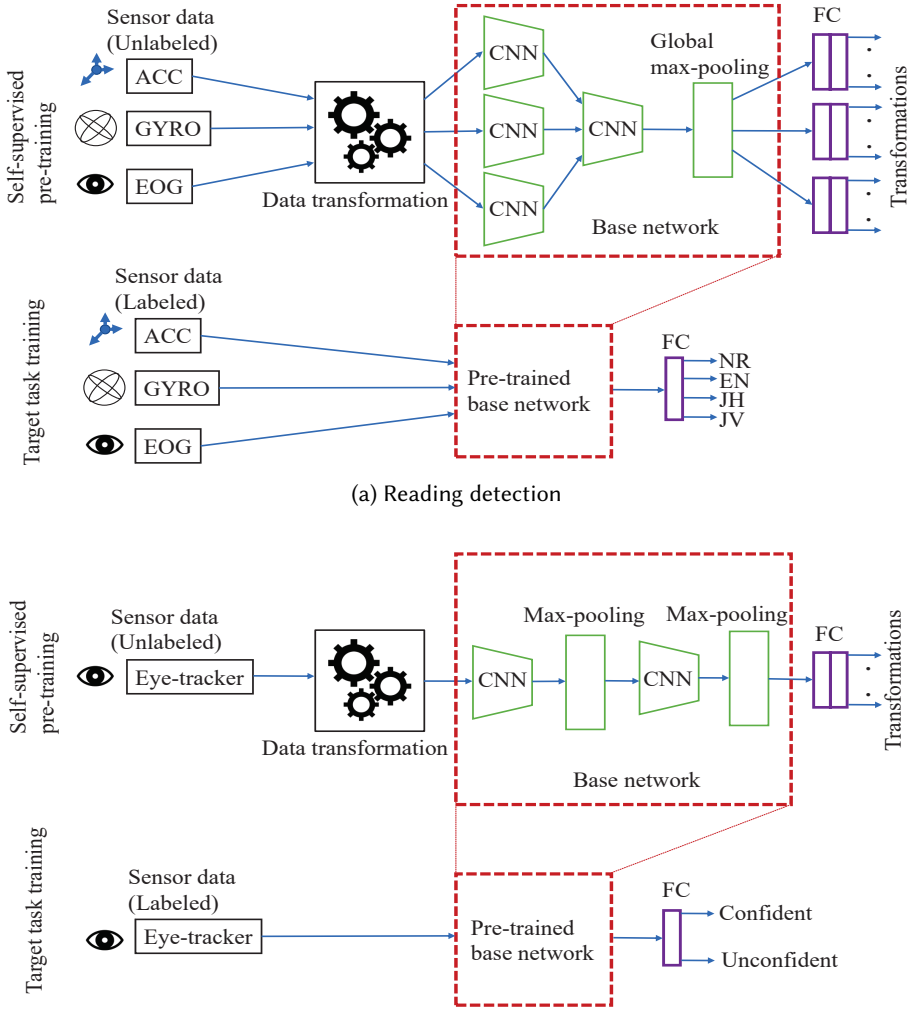
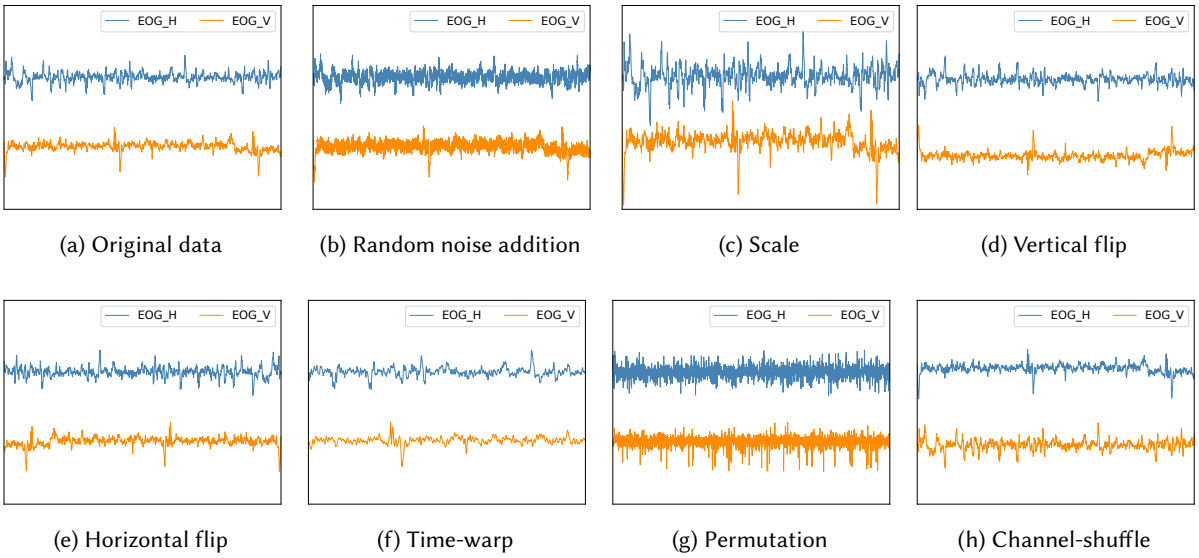


Fig. 1. Proposed SSL method for (a) reading detection where not reading (NR), reading English text (EN), reading Japanese horizontal text (JH), and reading Japanese vertical text (JV) and (b) confidence estimation. Both methods consist of two stages; self-supervised pre-training of the base network, for representation learning using unlabeled sensor data and target task training by fine-tuning the pre-trained base network by using a small amount of labeled sensor data.

tasks for this purpose in different domains. In HAR, the eight signal transformations are proposed by Saeed et al. [46] for ACC and GYRO sensor data; noise addition, scale, rotation, vertical flip, horizontal flip, permutation, time-warp, and channel-shuffle. Noise addition means the addition of random noise, scale means changing the magnitude of the samples within a window, by multiplying with a randomly selected scalar m, rotation means rotating samples about z-axis within a window in 90° anti-clockwise direction, vertical flip means a reflection of the window samples about x-axis achieved by multiplying by -1, horizontal flip means a reflection of the



(a) Original data (b) Random noise addition (c) Scale (d) Vertical flip

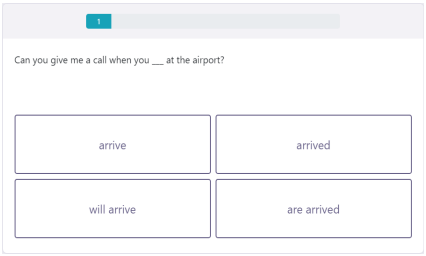
(e) Horizontal flip (f) Time-warp (g) Permutation (h) Channel-shuffle

Fig. 2. Transformed EOG data, (a) shows an original data segment (no transformation applied), and (b) to (h) are the transformed versions of (a). These transformations constitute the pretext task for self-supervised pre-training in the reading detection task.

window samples about y-axis, permutation means randomly perturbs the events within a window, time-warp means locally stretches or warps a time-series by an amount of n through a smooth distortion of time intervals between the values, and channel-shuffle means randomly shuffle the axial dimensions. For the pretext task, we also employ the same set of transformations for ACC and GYRO. For EOG, seven transformations, excluding rotation, are employed because rotation is not meaningful. Figure 2 shows the transformed EOG data.

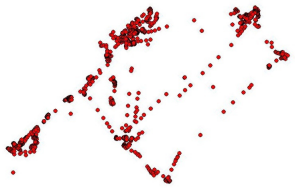
The red-dashed rectangle in the upper part of figure 1(a) is the base network trained by solving the pretext task. It consists of three Convolutional Neural Network (CNN) blocks for EOG, ACC, and GYRO data, a CNN block (128 units) that concatenates three CNN layers, and a global max-pooling layer. Each CNN block of EOG, ACC, and GYRO data consists of three 1D CNN layers. The numbers of units and kernel sizes in the CNN layers are 32, 64, and 96, and 24, 16, and 8, respectively. We applied batch normalization after each CNN layer, and a dropout layer after the global max-pooling layer. Finally, we added three classifiers at the end of the base network. Each classifier consists of two fully connected (FC) layers with 256 and 512 units, respectively. We use ReLU, softmax function, and Adam as the activation function, output layer, and optimizer, respectively. Finally, we solved the eight-class (not + seven transformed) and the nine-class (not + eight transformed) transformation recognition tasks for EOG data, and ACC and GYRO data, respectively.

3.1.2 Target Task Training. The final step is the target task training. For the fine-grained reading detection, we have four target classes as mentioned above: reading EN, reading JH, reading JV, and NR. We create a fine-grained reading detection network by retraining the pre-trained base network with a supervised approach, as shown in the lower part of figure 1(a). Unlike freezing the pre-trained base network, the proposed SSL method fine-tunes the base network. This is because higher performance is achieved. The FC layer in the target task training has 1024 units. We use the same activation function, optimizer, and output layer as used in the self-supervised pre-training. We select all hyperparameters in self-supervised pre-training and target task training by a preliminary experiment.



(a) MCQ format that consists of two parts; the problem statement, a sentence with a blank, and four choices in rectangles. The user is asked to select one correct choice.

(b) Eye gaze image



(c) Rotation

(d) Reflection about y-axis

(e) Reflection about x-axis

Fig. 3. (a) MCQ format used for data recording of confidence estimation, (b) is the actual eye gaze represented as an image, and (c) to (e) are transformed versions of (b), where the red dots represent the eye gaze points. Horizontal and vertical directions of (b) represent the x and y axes, respectively.

3.2 Confidence Estimation

Confidence estimation in answering MCQs involves classifying whether the answer is produced with confidence or not. By knowing the confidence in addition to the correctness of answers, we can offer some strategies for personalized learning. For example, correct but unconfident answers may be obtained by chance. In order to make the knowledge reliable, the question should be reviewed. Note that it is possible only if we know the confidence. Incorrect but confident answers often indicate wrong knowledge. The same incorrect answers would be reproduced when the user answers the same questions. In order to avoid this issue, it is necessary to revise the incorrect knowledge. We can warn the user of the revision if we know the confidence.

The format for how we handled MCQs in this paper is shown in figure 3(a); the user is asked to select one correct choice. We used an eye-tracker to record raw eye gazes while answering MCQs. In general, eye tracking data are converted into fixations and saccades [59] for further processing. But in our case, we employ the raw eye gaze data without conversion. Unlike the classification of fixed length segments in the reading detection, the amount of eye gazes varies from segment (MCQ) to segment (MCQ). Researchers in various domains have transformed 1D time-series data into 2D images to solve the classification task using CNN [25, 56, 58]. This direction helps with handling the variability in the amount of sensor data; therefore, we transform the eye-tracking data into images by plotting eye gaze graphically, as shown in figure 3(b). The red circles represent eye gaze points, and the black parts are the accumulation of black borders from the red circles. The x and y axes of the screen coordinate belongs to the horizontal and vertical directions of the eye gaze image, respectively.

3.2.1 Self-supervised Pre-training. We employ three image transformations as shown in figures 3(c)-(e) for the confidence estimation. The rotation is to apply 45° anti-clockwise rotation to the original image. Reflection about x and y axes mean the transformation of each pixel at (x, y) to $(x, -y)$ and $(-x, y)$, respectively. Therefore, we solve the four-class transformation recognition (not + three transformed) problem in pre-training.

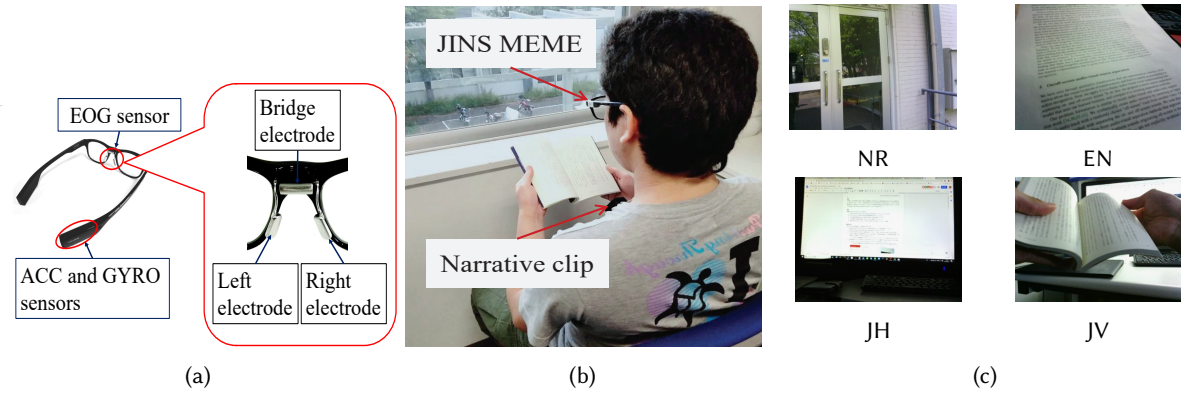


Fig. 4. Data recording for reading detection, (a) JINS MEME EOG glasses, (b) a user reading text documents wearing narrative clip takes frontal images, and JINS MEME EOG glasses records eye movement, and (c) Images taken by the Narrative clip.

The red-dashed box in the upper part of figure 1(b) shows the base network that consists of two CNN blocks and a max-pooling layer after each block. We add a dropout layer after the second max-pooling layer followed by a flatten layer. Each CNN block consists of two 2D CNN layers. The numbers of units of CNN layers are 8 for the first CNN block and 16 for the second CNN block. The kernel size is 3×3 for all CNN layers. We added a batch normalization after each CNN layer. At the end of the base network, we add a classifier consisting of two FC layers, and the number of units of both layers is 36. We use ReLU, softmax function, and SGD as the activation function, output layer, and optimizer, respectively. The input image size is $64 \times 64 \times 3$, where 3 shows the RGB channels.

3.2.2 Target Task Training. After the pre-training, the target task training is performed by replacing the classifier with an FC layer of 64 units. Same as with reading detection, unlike freezing the pre-trained base network, the proposed method fine-tunes the base network and retrains it using a labeled dataset as shown in the lower part of figure 1(b) that performed better. We designed the target task as a binary classification: confident or unconfident. We used the same input image size, activation function, optimizer, and output layer used in the self-supervised pre-training task. In addition, all hyperparameters are selected by a preliminary experiment.

4 DATA COLLECTION

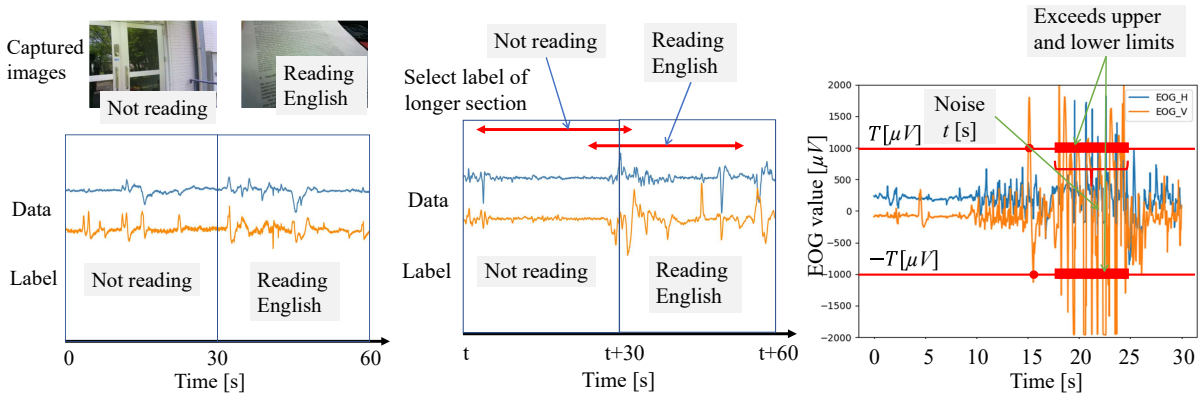
4.1 Reading Detection Datasets

We used a labeled dataset and an unlabeled dataset for reading detection recorded using JINS MEME EOG glasses [26]. This is an eye-wear device developed by JINS, which equips EOG, ACC, and GYRO sensors as shown in figure 4(a). The EOG sensor consisting of left, right, and bridge electrodes, as shown in figure 4(a), records the potential change due to eye movement in the horizontal and vertical directions. The JINS MEME is also equipped with a three-axis ACC sensor and a three-axis GYRO sensor. Although reading behaviors are mainly described by eye movements, they are often completed with slight head movements. Therefore, it is also useful to capture these movements by using ACC and GYRO. The sampling rate of EOG, ACC, and GYRO sensors is 100 Hz.

4.1.1 Labeled Dataset. We employed the labeled dataset, which was first introduced by Ishimaru et al. [23]. Ten (male) Japanese university students were recruited. All are native Japanese, and the age range is 18 to 25 years old. Each participant wore the JINS MEME glasses for about 12 hours a day for two days and was asked to read EN, JV, and JH texts for about 1 hour for each in a day and not to read anything for the rest of the time

Table 1. Recording duration [minute] of labeled data for reading detection, where D1 and D2 represent day 1 and day 2, respectively.

	p1		p2		p3		p4		p5		p6		p7		p8		p9		p10	
Activity	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
NR	474	466	447	499	490	429	499	567	476	283	379	409	204	358	538	496	476	307	494	509
EN	54	100	73	96	74	61	67	61	63	74	82	115	91	67	62	55	59	58	49	62
JH	97	71	70	74	101	117	72	53	73	64	86	60	127	53	63	58	75	58	90	68
JV	74	89	101	75	71	67	60	60	83	116	73	72	65	74	66	73	113	60	65	75



(a) Labeling a non-overlapped segment (b) Labeling an overlapped segment (c) Detecting EOG noisy segment

Fig. 5. Data labeling and noisy segment detection, (a) and (b) shows data labeling using captured images, apply a label to a segment that the corresponding image represents, and (c) EOG data segment with burst noise and noise judgment criteria.

shown in figure 4(b). Participants also had a small camera called narrative clip on their clothing to take frontal images every 30 seconds shown in figure 4(c). At the end of the day, participants were asked to generate data labeling information using a labeling tool and captured images. Then images were discarded to protect privacy. As a reward, each participant received a gift card worth 5,000 JPY per day. Except for the above instructions, no restrictions were imposed during data recording. Thus, the dataset can be regarded as "in-the-wild." The summary of the participant's activities is shown in table 1.

The data were prepared as follows. First, the EOG, ACC, and GYRO data were split into segments of size 30 seconds slid by 15 seconds. Figure 2(a) shows an example of the EOG segment. After that, each segment is labeled. Since the length of a segment and the frame rate of the narrative clip are the same, one label is assigned to one segment as shown in figure 5(a). However, in exceptional cases two labels may overlap. This is because the frame rate of the narrative clip fluctuates. In these cases, the most overlapping label is selected, as shown in figure 5(b).

EOG data sometimes suffer from bursts of noise, as shown in figure 5(c), due to the poor electrode contact to the skin. To discard noisy segments, we set the noise judgment criteria as EOG values above T $[\mu V]$ or below $-T$ $[\mu V]$ and exist for continuous t seconds or more to be considered as noise. We set $T = 1,000$ $[\mu V]$ and $t = 2$ seconds based on the preliminary analysis. Also, the sensor's data deviate from the reference value. Therefore, we corrected it by subtracting the average value of segments from each data segment. Finally, the number of segments in the labeled dataset after noise removal is 32,708, 5,340, 5,792, and 5,798 for NR, EN, JH, and JV, respectively.

(a) Tobii 4C eye-tracker (b) Computer screen and Tobii 4C eye-tracker (c) User answering MCQs

Fig. 6. Data recording proceedings for confidence estimation, (a) eye-tracker used, (b) computer screen with an example question and eye-tracker fixed at the bottom, and (c) eye-tracker recording user's eye-gaze while answering MCQs.

4.1.2 Unlabeled Dataset. We recruited 13 male Japanese university students who are native Japanese and whose age range is 18 to 25 years old. Each participant wore a JINS MEME device for three to eight days during their daily life. The measurement time is about 20 to 60 hours per person and in total 676 hours. Each participant received a gift card worth 5,000 JPY per day. In addition, we employed an unlabeled dataset that is recorded when 39 participants (volunteer) attended presentations at an international conference. Because there is no restriction, these datasets are also considered "in-the-wild." The unlabeled data totals about 1,359 hours. We prepared the unlabeled dataset in the same way as the labeled one. We discarded noisy segments using the noise judgment criteria with parameters $T = 1,000$ $[\mu V]$ and $t = 0.01$ seconds that are also selected by the preliminary analysis and corrected the reference value. Finally, the number of segments in the unlabeled dataset is 177,921.

4.2 Confidence Estimation Datasets

We also use a labeled and an unlabeled datasets for confidence estimation. We recorded both datasets using the Tobii 4C pro-upgraded eye-tracker [54], shown in figure 6(a), a stationary eye-tracker of sampling rate 90 Hz.

4.2.1 Labeled Dataset. We recruited 20 Japanese university students (14 males and 6 females) to generate the labeled dataset. All participants are native Japanese, and the age range is 18 to 25 years old. The experiment was conducted in first day (two hours), third day (one hour), and fifth day (two hours). The experimental procedure is shown in figure 6. The participants read and answered as many four-choice English grammatical questions as they could, selected from a randomized question pool. Right after answering each MCQ, participants were asked to assess the confidence behind their answer that constitutes the label. The eye gaze represents the participant's behavior during the answering process and does not include the labeling process. As a reward, each participant received a gift card worth 1,000 JPY per hour. Table 2 shows the summary of the dataset. The dataset includes a serious skew in the number of confident and unconfident answers due to differences in English ability among the participants. While we recorded the data in a classroom setting, we did not impose restrictions with the exception of the guidelines, and participants acted on their own accord. Therefore this dataset is also considered "in-the-wild."

4.2.2 Unlabeled Dataset. We recruited 80 Japanese high school students who are native Japanese and whose age range is 16 to 17 years old. We recorded the unlabeled data following the experimental procedure described for labeled data except collecting user's confidence in the answer for the full dataset where each participant read and answered four-choice English vocabulary questions. We asked the users to assess their confidence in the answer

Table 2. Summary of the labeled dataset for confidence estimation.

Participant	No. of MCQs answered		
	Confident	Unconfident	Total
s1	361	108	469
s2	398	55	453
s3	415	103	518
s4	420	14	434
s5	390	175	565
s6	68	458	526
s7	222	253	475
s8	263	272	535
s9	348	87	435
s10	210	180	390

Participant	No. of MCQs answered		
	Confident	Unconfident	Total
s11	159	271	430
s12	296	243	539
s13	325	140	465
s14	263	253	516
s15	174	2	176
s16	556	109	665
s17	202	354	556
s18	316	260	576
s19	306	140	446
s20	304	135	439

once in every five MCQs to build a ground truth, but we do not use them. All participants worked voluntarily and received no remuneration. The total number of segments (MCQs) in the unlabeled dataset is 57,460. Both labeled and unlabeled datasets were processed and converted into images as described in section 3.2.

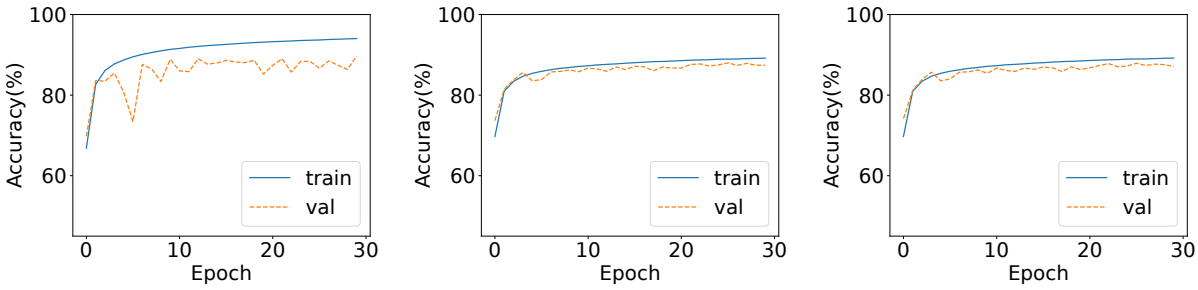
5 EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Reading Detection

5.1.1 Experimental Conditions. The purpose of the experiment is to evaluate the performance of the proposed SSL method for reading detection as compared to the fully-supervised DL and SVM. We are interested in the change of the performance as a function of the number of labeled training samples. For the fully-supervised DL, we simply use the proposed SSL method for the target task without the pre-training; all the structure and parameters are identical to the proposed SSL method. Thus the results show the effectiveness of pre-training by using an unlabeled dataset. For SVM, the method described in [23] is used. We calculated the mean and variance of vertical and horizontal components of EOG data and three axes components of ACC and GYRO data as features and finally selected ten out of sixteen features using a hill climbing feature selection method.

The training of each method was performed as follows. For the proposed SSL method, we first applied the pre-training by using the transformed sensor data. In data transformation, we defined the parameters m and n as described in section 3.1.1 as $5 \leq m \leq 10$ and $n = 2$, respectively. We selected a transformation, including no transformation, and applied it to the segment to produce pre-training data. Because only one transformation was applied to each segment, the number of transformed unlabeled samples is equal to the number of original unlabeled samples. We applied each transformation equally so that the chance rates for the EOG, ACC, and GYRO data are 12.5%, 11.1%, and 11.1%, respectively, where the first one is eight-class, and the latter two are nine-class classification. After this, the target task training was applied. As described in section 4.1.1, the number of available labeled samples is different for each class. We simply took all 5,340 samples (the smallest number) of EN and downsampled other classes to have them match in size. Thus the chance rate is 25%. The same data were also employed for training the fully-supervised DL and SVM. We changed the number of labeled training samples per class in the order of 10, 50, 100, 500, 1,000, and 5,340.

All of the above methods were evaluated in user independent Leave-One-Participant-Out cross-validation way. In training, 20% of the training data were selected for validation, and the rest were employed for learning.



(a) EOG (b) ACC (c) GYRO

Fig. 7. Training and validation accuracy curves of the self-supervised pre-training experiment for reading detection.

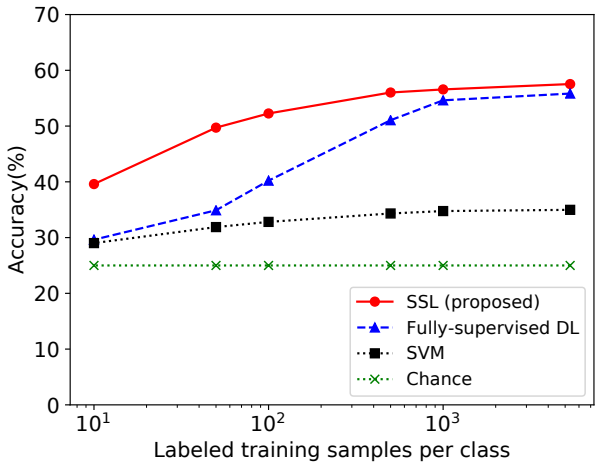


Fig. 8. Results of the reading detection in accuracy. It describes the dependency of the test accuracy on the number of labeled samples per class as 10, 50, 100, 500, 1,000, and 5,340 that evaluates the performance for a wide range of labeled samples.

5.1.2 Results of Pre-training. Training and validation accuracy curves for EOG, ACC, and GYRO data are shown in figure 7. The average test accuracy is 93.2% for EOG, 95.5% for ACC, and 95.3% for GYRO. This high test accuracy primarily indicates that the network was trained well. For EOG, the difference between the training and validation accuracy is slightly high, indicating the tendency of the network to overfit. Thus, the test accuracy for EOG data is also lower than that of the ACC and GYRO. This happened because we could not remove all the noisy segments from the EOG dataset, considering the size of the dataset.

5.1.3 Results of the Target Task. Figure 8 shows the reading detection result in accuracy. It describes the change of average test accuracy for the number of labeled training samples per class. The proposed SSL method performs best regardless of the number of training samples, and the proposed SSL method is more advantageous than the fully-supervised DL when the number of labeled training samples is smaller. This indicates the effectiveness of SSL. As compared to SVM, the fully-supervised DL performs much better when the number of labeled training samples is larger. However, this advantage disappears when the number of training samples decreases, showing the key limitation of the fully-supervised DL. On the other hand, the proposed SSL method is always much better

Table 3. Repeated Measures ANOVA test result for reading detection.

Parameter	Training samples per class					
	10	50	100	500	1,000	5,340
F value	77.24	61.39	63.55	50.13	43.92	34.03
p value	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*

*p<0.01

Table 4. Post-Hoc Paired T-test result for reading detection.

Pair	Parameter	Training samples per class					
		10	50	100	500	1,000	5,340
Proposed SSL and fully-supervised DL	t value	11.30	10.69	11.87	4.79	3.02	1.15
	p value	0.000*	0.000*	0.000*	0.001*	0.014	0.278
Proposed SSL and SVM	t value	9.88	8.49	8.65	7.60	6.68	6.09
	p value	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
Fully-supervised DL and SVM	t value	0.68	1.88	4.26	6.70	6.69	6.04
	p value	0.516	0.093	0.002*	0.000*	0.000*	0.000*

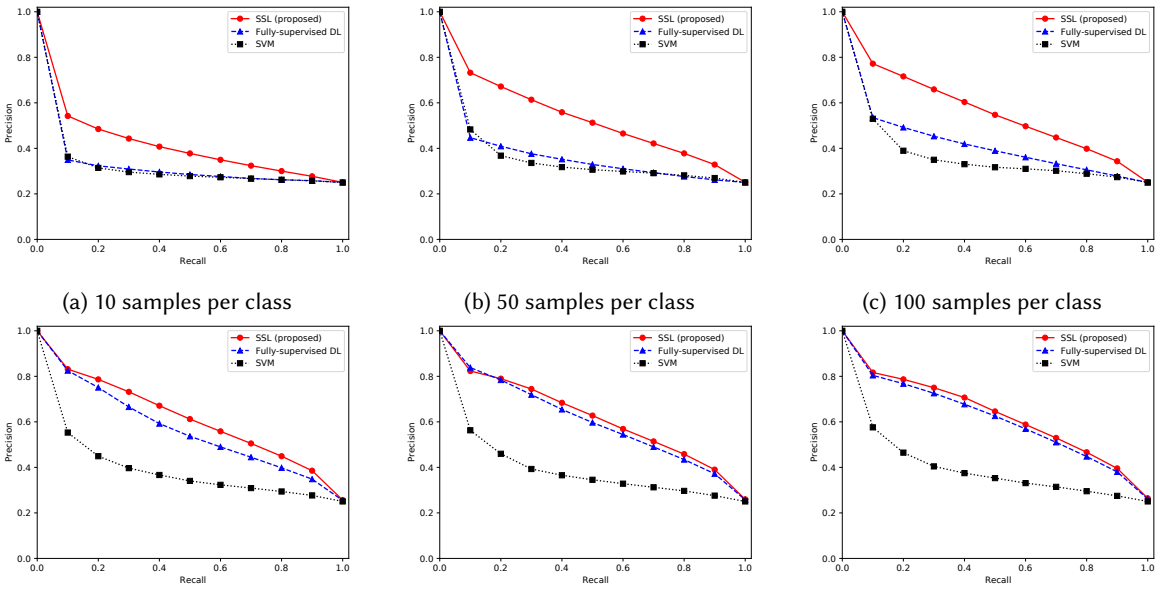
*p<0.0033

than SVM and is never inferior to the fully-supervised DL. Therefore, we can always recommend to use the proposed SSL method.

To investigate the significance of test accuracies, we applied statistical analysis. A one-way repeated measures ANOVA test was first applied to test accuracy. The null hypothesis is that the population means of all three methods are equal. Table 3 shows the results. The null hypothesis was rejected with the significance level $p < 0.01$. The post-hoc paired t-test was applied for further analysis. A null hypothesis here is that the population mean of one method is equal to that of another method. Because we have three methods, we conducted three t-test experiments for all combinations. Multiple comparisons were mitigated with a Bonferroni correction. With the correction applied, significance is found at $p < 0.0033$ (0.01/3). Results are shown in table 4. When comparing the proposed SSL method and the fully-supervised DL, the proposed SSL method is statistically significantly better, with the exception of the 1,000 and 5,340 training samples cases. For the comparison with SVM, a statistically significant difference is shown for all cases for the proposed SSL method. For the comparison between the fully-supervised DL and SVM, there is no significant difference for smaller sample sizes (10 and 50).

We also evaluated the performance of the methods with recall and precision. Figure 9 shows the recall-precision curves for all three methods. The results show that SVM is always worst, and the proposed SSL method is always best for all recall levels. Besides, the performance of the fully-supervised DL depends on the number of labeled training samples. If the number of labeled training samples increases, then it approaches the proposed SSL method, and if the number of labeled training samples is small, then it approaches the SVM. The most important point is that for all recalls, the advantage of the proposed SSL method is clear.

Next, we look at the effectiveness and differences among pretext tasks employed for reading detection. In between the fully-supervised DL method and SSL method employing the full set of pretext tasks, we selected two choices of using one pretext task (two-class classifications) and removing one pretext task from the full set of pretext tasks. We generated similar curves as shown in figure 8 and calculated the area under curve (AUC) in log scale for comparison. Tables 5(a) and (b) shows the normalized AUC that is defined as $(A - A_c)/(A_s - A_c)$, where A, A_c, A_s, are the AUCs of a method of interest, the chance rate, and the proposed SSL method, respectively. The



(a) 10 samples per class (b) 50 samples per class (c) 100 samples per class

(d) 500 samples per class (e) 1,000 samples per class (f) 5,340 samples per class

Fig. 9. Results of the reading detection task in recall and precision. These are 11 point interpolated recall-precision curves and describe the dependency of the precision to the recall to compare the performance of three methods.

Table 5. Results of analyzing pretext task for reading detection.

(a)

	Noise addition	Horizontal flip	Permutation	Time-warp	Channel shuffle	Scale	Vertical flip
Use	0.723	0.788	0.753	0.730	0.776	0.692	0.550
Remove	0.938	0.872	0.881	0.980	0.910	0.890	0.942

(b)

SSL (proposed)	Fully supervised DL	SVM
1.000	0.716	0.295

results show that no single pretext task performs well where performance is comparable with fully-supervised DL, and the vertical flip performs worse. This means that a single pretext task is not enough for the sensor data. On the other hand, when removing a pretext task, performance is comparable with the full set of the pretext task, and time-warp does not have a significant contribution.

5.2 Confidence Estimation

5.2.1 Experimental Conditions. The purpose of the experiments is the same as for reading detection. In data transformation for the self-supervised pre-training, we selected one of four transformations, including no transformation, as shown in figure 3 and applied it to one unlabeled eye gaze image. Because each transformation was selected equally, the chance rate of the pre-training was 25%.

Although we applied Leave-One-Participant-Out cross-validation for reading detection, it is not appropriate for confidence estimation due to the seriously skewed distribution of labels, as shown in table 2. Note that the number of samples available from each participant s_i is different and in most cases the total number of confident

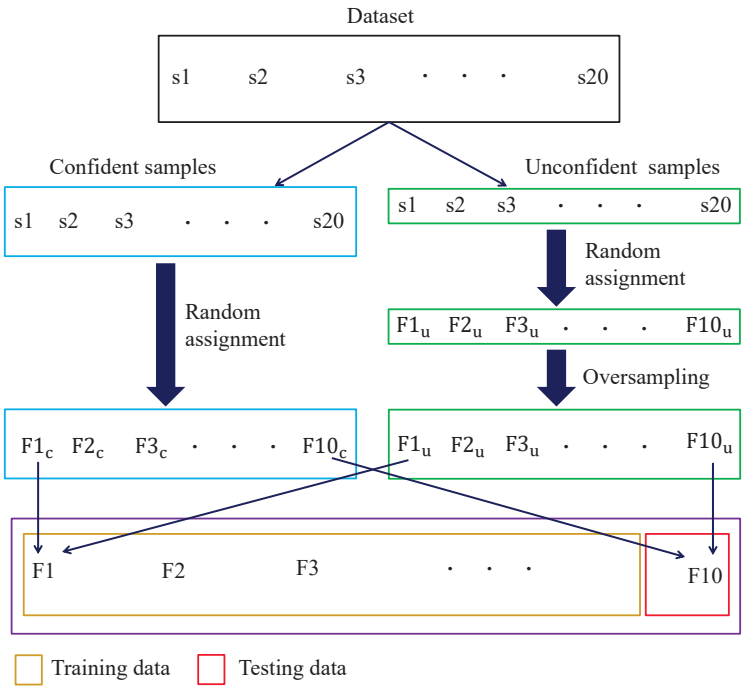
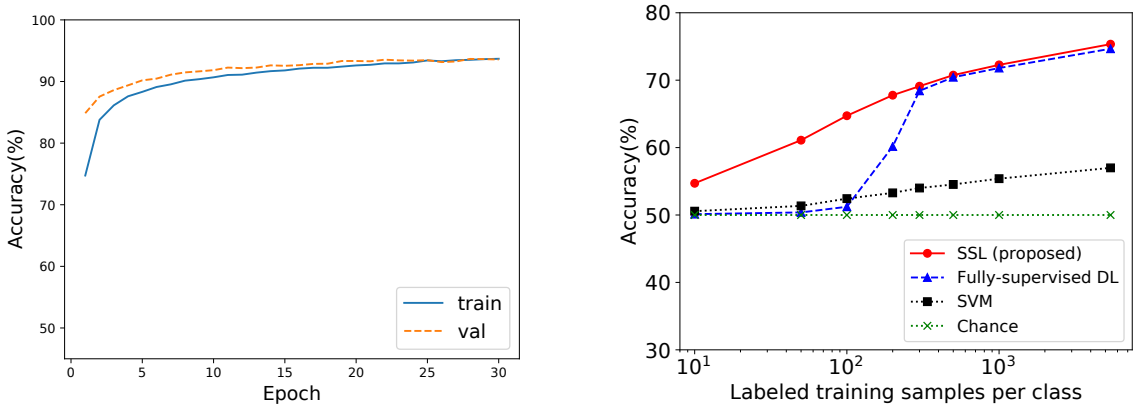


Fig. 10. 10-fold cross-validation evaluation method for confidence estimation. A novel person dependent approach to handle the imbalanced data to make ten folds where the network is trained for all except one fold and tested on the excluded one.

samples is much larger than the unconfident one. We cannot balance the dataset by simply applying over- nor under-sampling. To handle the skew in the data, we used a person dependent approach. Figure 10 illustrates the procedure of data preparation. First, we separated the data into confident and unconfident samples for each participant. For all the confident samples from each participant, we randomly assigned a fold from $F1_c$ to $F10_c$ while keeping the number of samples in each fold as equal as possible. For unconfident samples, we also assigned each sample to one of the 10 folds $F1_u$–$F10_u$ in the same way. In order to make the number of samples in Fi_u equal to that in Fi_c, we applied oversampling by using a 5-degree rotation to randomly selected samples in each fold. Finally, we combined Fi_c and Fi_u to form a fold Fi. The density of data from each participant in all folds is now almost equal. By using the 10 folds of data, we applied the 10-fold cross-validation. In training using nine folds, we randomly selected 80% for learning and the remaining 20% for validation. The chance rate is 50%.

In the case of the fully-supervised DL, we trained the network with the structure and parameters identical to the proposed SSL method using the labeled data. We used SVM as a baseline method using basic statistical features of mean and variance. We calculated and used four features from one sample (MCQ); means and variances along the two axes. These two methods were also evaluated by the 10-fold cross-validation as described above. We changed the number of labeled training samples per class in the order of 10, 50, 100, 200, 300, 500, 1,000, and 5,382.

5.2.2 Results of Pre-training. The average test accuracy of the pre-training experiment was 93.3%; this high test accuracy indicates that the base network was trained well. Figure 11(a) illustrates the training and validation accuracy curves for pre-training. The pre-trained network is good fit since the difference between the training and validation is almost zero.



(a) Result of self-supervised pre-training

(b) Result of target task training

Fig. 11. Results for the confidence estimation, (a) training and validation accuracy curves of the self-supervised pre-training experiment, and (b) result of the confidence estimation task in accuracy. The figure describes the dependency of the test accuracy on the number of labeled training samples per class of 10, 50, 100, 200, 300, 500, 1,000, and 5,382.

Table 6. Repeated Measures ANOVA test result for confidence estimation.

Parameter	Training samples per class							
	10	50	100	200	300	500	1,000	5,382
F value	301.9	1048.2	1641.7	632.3	775.1	1003.1	696.3	615.7
p value	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*	0.00*

*p<0.01

5.2.3 Results of the Target Task. Figure 11(b) shows the results of the confidence estimation target task in accuracy. Tendencies similar to the reading detection results were observed. The proposed SSL method performed the best regardless of the number of labeled training samples. The performance of the fully-supervised DL dropped when the number of labeled training samples was insufficient. SVM was not always the worst, though the performance was limited even with a larger number of labeled samples. From these results, we can always recommend to use the proposed SSL method in the case of confidence estimation.

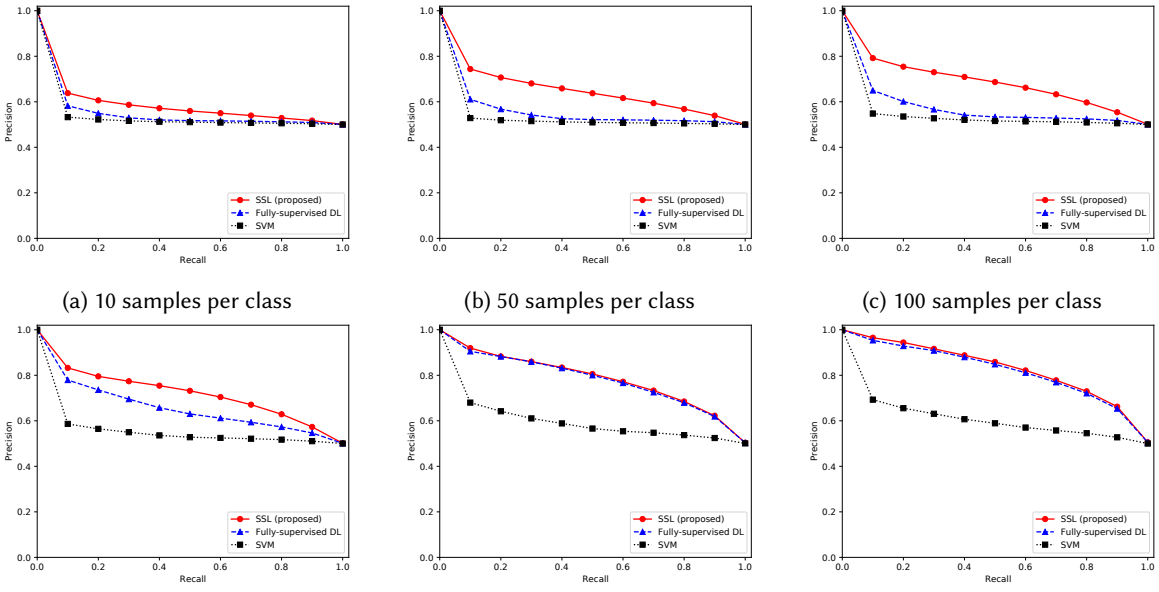
As we did for the results of reading detection, we also applied statistical analysis to the results of test accuracy for the confidence estimation to confirm if differences are significant. We applied the same experimental conditions, hypothesis, and significance level as used in reading detection in section 5.1.3. Tables 6 and 7 show the results. The one-way repeated measures ANOVA test returned significant results. From the results of the post-hoc paired t-test, we have confirmed the following: In the comparison between the proposed SSL method and the fully-supervised DL, we found significant differences except for the cases of 500, 1,000, and 5,382. For the comparison between the proposed SSL method and SVM, all cases show a significant difference. A significant difference was found in all the comparisons between fully-supervised DL and SVM. In cases with a larger number of training samples, the fully-supervised DL worked better than SVM, while the results were opposite with fewer training samples.

Similar to the reading detection, we also evaluated the performance of all methods in recall and precision metrics for confidence estimation. Figure 12 shows the recall-precision curves. The same conclusion made for reading detection also holds for this case. For all recalls, the proposed SSL method performs best, and SVM is

Table 7. Post-Hoc Paired T-test result for confidence estimation.

Pair	Param.	Training samples per class							
		10	50	100	200	300	500	1,000	5,382
Proposed SSL and fully-supervised DL	t value	17.32	36.01	81.80	29.88	4.56	1.83	2.26	2.21
	p value	0.000*	0.000*	0.000*	0.000*	0.001*	0.10	0.05	0.055
Proposed SSL and SVM	t value	18.65	34.51	37.61	32.09	30.43	33.67	25.65	26.54
	p value	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
Fully-supervised DL and SVM	t value	-5.01	-5.34	-4.57	14.35	27.04	31.61	27.67	25.47
	p value	0.000*	0.000*	0.001*	0.000*	0.000*	0.000*	0.000*	0.000*

*p<0.0033



(a) 10 samples per class (b) 50 samples per class (c) 100 samples per class

(d) 200 samples per class (e) 1,000 samples per class (f) 5,382 samples per class

Fig. 12. Results of the confidence estimation task in recall and precision. These are 11 points interpolated recall-precision curves and describe the dependency of the precision to the recall to compare the performance of three methods.

always worst. In the case of fully-supervised DL, with enough training samples, the performance is close to the proposed SSL, and performance decreases with decreasing training samples and comparable to SVM for small training samples. The proposed SSL is never inferior to other methods for all recalls. Thus, the proposed SSL is the first choice without any chance of losing the best performance.

Next, we investigate the effectiveness and differences among pretext tasks employed for confidence estimation. We conducted experiments and calculated AUC following the same procedure described for reading detection. The normalized AUC is shown in tables 8(a) and (b). The results show that even a single pretext task is important, and all pretext tasks work almost equally well, although reflection about x-axis and y-axis are most effective.

Table 8. Results of analyzing pretext task for confidence estimation.

(a)

	Reflection about y-axis	Reflection about x-axis	Rotation
Use	1.011	1.007	0.941
Remove	0.967	0.699	0.955

(b)

SSL (proposed)	fully supervised DL	SVM
1.000	0.698	0.211

6 CONCLUSION

Reading behavior is an important cognitive human activity because it's analysis allows users to examine their reading habits which can help with the development of reading strategies. Methods using classical machine learning and handcrafted features may achieve good results in laboratory settings, but may not obtain satisfactory results in-the-wild. DL methods that can solve this issue require a large-sized labeled dataset to extract useful features. However, a large-sized labeled data collections are difficult to obtain. As a step towards tackling this issue and providing robust and feasible reading analysis, we have proposed an SSL method. We evaluated the effectiveness of the proposed SSL method by selecting two reading activities that explore quantity (period) and quality (confidence) of reading, respectively. We evaluated both tasks with the proposed SSL method, the fully-supervised DL, and SVM. The proposed SSL method consists of two stages. In the first stage, we pre-trained a network by solving pretext tasks. In the second stage, we created the target task network by fine-tuning the pre-trained network using labeled data. From the results, we have confirmed that the proposed SSL method performs the best for both reading activity tasks compared to the fully-supervised DL method and SVM for all numbers of training sample cases. Therefore we can recommend to use the proposed SSL regardless of the available number of training samples.

Another important takeaway is that the proposed SSL method provides a level of certainty in performance against existing technologies for both the EOG and eye-tracking data. In the EOG case, we might expect that existing technology works, but this is not known. In addition, the contribution of each pretext task makes it difficult to gauge the performance of existing technology. In the case of eye-tracking, it is unlikely, but not a certainty, that existing technology works due to the nature of the data. Therefore, we proposed new pretext tasks.

Future work includes further improvement in the accuracy of the proposed SSL method by introducing other sensors, as well as its application for other reading activity classification tasks.

ACKNOWLEDGMENTS

This work was supported in part by the JST CREST (Grant No. JPMJCR16E1), JSPS Grant-in-Aid for Scientific Research (20H04213), Grand challenge of the Initiative for Life Design Innovation (iLDi), and OPU Keyproject.

REFERENCES

[1] Pulkit Agrawal, João Carreira, and Jitendra Malik. 2015. Learning to See by Moving. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago, Chile). IEEE, 37–45. https://doi.org/10.1109/ICCV.2015.13

[2] Gregory P. Amis and Gail A. Carpenter. 2010. Self-supervised ARTMAP. *Neural Networks* 23, 2 (March 2010), 265–282. https://doi.org/10.1016/j.neunet.2009.07.026

[3] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E. Petersen, Yike Guo, Paul M. Matthews, and Daniel Rueckert. 2019. Self-Supervised Learning for Cardiac MR Image Segmentation by Anatomical Position Prediction. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention*. Springer, Cham, 541–549. https://doi.org/10.1007/978-3-030-32245-8_60

[4] Ralf Biedert, Jörn Hees, Andreas Dengel, and Georg Buscher. 2012. A Robust Realtime Reading-Skimming Classifier. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (Santa Barbara) *(ETRA '12)*. ACM, New York, NY, USA, 123–130. https://doi.org/10.1145/2168556.2168575

[5] Andreas Bulling, Jamie A. Ward, and Hans Gellersen. 2012. Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors. *ACM Trans. Appl. Percept.* 9, 1, Article 2 (March 2012), 21 pages. https://doi.org/10.1145/2134203.2134205

[6] Andreas Bulling, Jamie A. Ward, Hans Gellersen, and Gerhard Tröster. 2011. Eye Movement Analysis for Activity Recognition Using Electrooculography. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 4 (April 2011), 741–753. https://doi.org/10.1109/TPAMI.2010.86

[7] Christopher S. Campbell and Paul P. Maglio. 2001. A Robust Algorithm for Reading Detection. In *Proceedings of the Workshop on Perceptive User Interfaces* (Orlando, Florida, USA) *(PUI '01)*. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/971478.971503

[8] Nixon Chan and Peter E. Kennedy. 2002. Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple-Choice and "Equivalent" Constructed-Response Exam Questions. *Southern Economic Journal* 68, 4 (April 2002), 957–971. https://doi.org/10.2307/1061503

[9] Leana Copeland, Tom Gedeon, and Sumudu Mendis. 2014. Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research* 3, 3 (Aug. 2014), 35–48. https://doi.org/10.5430/air.v3n3p35

[10] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. 2015. Unsupervised Visual Representation Learning by Context Prediction. In *Proceedings of the IEEE International Conference on Computer Vision* (Santiago, Chile) *(ICCV '15)*. IEEE, USA, 1422–1430. https://doi.org/10.1109/ICCV.2015.167

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *Proceedings of the International Conference on Learning Representations*. ICLR, 1–16. https://openreview.net/forum?id=S1v4N2l0-

[12] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. 2019. Scaling and Benchmarking Self-Supervised Visual Representation Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul, South Korea). IEEE, 6390–6399. https://doi.org/10.1109/ICCV.2019.00649

[13] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (Vancouver, Canada). IEEE, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

[14] Thomas M. Haladyna. 2015. *Developing and Validating Multiple-choice Test Items* (3rd edition ed.). Routledge, New York, NY, USA.

[15] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. In *Proceedings of the International Joint Conference on Artificial Intelligence* (New York, NY, USA) *(IJCAI'16)*. AAAI Press, 1533–1540. https://www.ijcai.org/Proceedings/16/Papers/220.pdf

[16] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine* 25, 1 (Jan. 2019), 65–69. https://doi.org/10.1038/s41591-018-0268-3

[17] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L. Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. 2020. Masked Reconstruction Based Self-Supervision for Human Activity Recognition. In *Proceedings of the International Symposium on Wearable Computers* (Virtual Event, Mexico) *(ISWC '20)*. ACM, New York, NY, USA, 45–49. https://doi.org/10.1145/3410531.3414306

[18] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., 15663–15674. https://proceedings.neurips.cc/paper/2019/file/a2b15837edac15df90721968986f7f8e-Paper.pdf

[19] Shoya Ishimaru, Kensuke Hoshika, Kai Kunze, Koichi Kise, and Andreas Dengel. 2017. Towards Reading Trackers in the Wild: Detecting Reading Activities by EOG Glasses and Deep Neural Networks. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) *(UbiComp '17)*. ACM, New York, NY, USA, 704–711. https://doi.org/10.1145/3123024.3129271

[20] 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 Augmented Human International Conference* (Kobe, Japan) *(AH '14)*. ACM, New York, NY, USA, Article 15, 4 pages. https://doi.org/10.1145/2582051.2582066

[21] Shoya Ishimaru, Kai Kunze, Katsuma Tanaka, Yuji Uema, Koichi Kise, and Masahiko Inami. 2015. Smart Eyewear for Interaction and Activity Recognition. In *Proceedings of the Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI EA '15)*. ACM, New York, NY, USA, 307–310. https://doi.org/10.1145/2702613.2725449

[22] Shoya Ishimaru, Takanori Maruichi, Andreas Dengel, and Koichi Kise. 2021. Confidence-Aware Learning Assistant. *CoRR* abs/2102.07312 (2021). arXiv:2102.07312 https://arxiv.org/abs/2102.07312

[23] Shoya Ishimaru, Takanori Maruichi, Manuel Landsmann, Koichi Kise, and Andreas Dengel. 2019. Electrooculography Dataset for Reading Detection in the Wild. In *Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, UK) *(UbiComp/ISWC '19 Adjunct)*. ACM, New York, NY, USA, 85–88. https://doi.org/10.1145/3341162.3343812

[24] Simon Jenni and Paolo Favaro. 2018. Self-Supervised Feature Learning by Learning to Spot Artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, USA). IEEE, 2733–2742. https://doi.org/10.1109/cvpr.2018.00289

[25] Wenchao Jiang and Zhaozheng Yin. 2015. Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks. In *Proceedings of the ACM International Conference on Multimedia* (Brisbane, Australia) *(MM '15)*. ACM, New York, NY, USA, 1307–1310.

https://doi.org/10.1145/2733373.2806333
[26] JINS. 2020. JINS MEME Electrooculography Glasses. https://jins-meme.com/en/. Accessed: Nov 10, 2020.
[27] Conor Kelton, Zijun Wei, Seoyoung Ahn, Aruna Balasubramanian, Samir R. Das, Dimitris Samaras, and Gregory Zelinsky. 2019. Reading Detection in Real-Time. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. ACM, New York, NY, USA, Article 43, 5 pages. https://doi.org/10.1145/3314111.3319916
[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60, 6 (May 2017), 84–90. https://doi.org/10.1145/3065386
[29] Kai Kunze, Yuzuko Utsumi, Yuki Shiga, Koichi Kise, and Andreas Bulling. 2013. I Know What You Are Reading: Recognition of Document Types Using Mobile Eye Tracking. In *Proceedings of the International Symposium on Wearable Computers* (Zurich, Switzerland) *(ISWC '13)*. ACM, New York, NY, USA, 113–116. https://doi.org/10.1145/2493988.2494354
[30] Yongjin Kwon, Kyuchang Kang, and Changseok Bae. 2014. Unsupervised learning for human activity recognition using smartphone sensors. *Expert Systems with Applications* 41, 14 (Oct. 2014), 6067 – 6074. https://doi.org/10.1016/j.eswa.2014.04.037
[31] Manuel Landsmann, Olivier Augereau, and Koichi Kise. 2019. Classification of Reading and Not Reading Behavior Based on Eye Movement Analysis. In *Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, UK) *(UbiComp/ISWC '19 Adjunct)*. ACM, New York, NY, USA, 109–112. https://doi.org/10.1145/3341162.3343811
[32] Paul Lau, Sie Lau, Kian Hong, and Hasbee Usop. 2011. Guessing, Partial Knowledge, and Misconceptions in Multiple-Choice Tests. *Educational Technology & Society* 14, 4 (Jan. 2011), 99–110. https://eric.ed.gov/?id=EJ963283
[33] Frédéric Li, Kimiaki Shirahama, Muhammad Adeel Nisar, Lukas Köping, and Marcin Grzegorzek. 2018. Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors. *Sensors* 18, 2, Article 679 (Feb. 2018), 22 pages. https://doi.org/10.3390/s18020679
[34] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma. 2016. DeepFood: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment. In *Proceedings of the International Conference on Inclusive Smart Cities and Digital Health* (Wuhan, China) *(ICOST 2016)*. Springer, Cham, 37–48. https://doi.org/10.1007/978-3-319-39601-9_4
[35] Martin Längkvist, Lars Karlsson, and Amy Loutfi. 2014. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters* 42, 1 (June 2014), 11 – 24. https://doi.org/10.1016/j.patrec.2014.01.008
[36] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Baltimore, Maryland). Association for Computational Linguistics, 55–60. https://doi.org/10.3115/v1/P14-5010
[37] Cheryl A. Melovitz Vasan, David O. DeFouw, Bart K. Holland, and Nagaswami S. Vasan. 2018. Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course. *Anatomical Sciences Education* 11, 3 (May 2018), 254–261. https://doi.org/10.1002/ase.1739
[38] Chaitanya Mitash, Kostas E. Bekris, and Abdeslam Boularias. 2017. A self-supervised learning system for object detection using physics simulation and multi-view pose estimation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC, Canada). IEEE, 545–551. https://doi.org/10.1109/IROS.2017.8202206
[39] Ross H. Nehm and Leah Reilly. 2007. Biology Majors Knowledge and Misconceptions of Natural Selection. *BioScience* 57, 3 (March 2007), 263–272. https://doi.org/10.1641/B570311
[40] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European conference on computer vision*. Springer, Cham, 69–84. https://doi.org/10.1007/978-3-319-46466-4_5
[41] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. 2018. Boosting Self-Supervised Learning via Knowledge Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, USA). IEEE, 9359–9367. https://doi.org/10.1109/cvpr.2018.00975
[42] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. Context Encoders: Feature Learning by Inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV, USA). IEEE, 2536–2544. https://doi.org/10.1109/cvpr.2016.278
[43] Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of Reading* (2nd ed.). Psychology Press, New York, NY, USA.
[44] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2018. A Survey on Data Collection for Machine Learning: a Big Data - AI Integration Perspective. *CoRR* abs/1811.03402 (2018). http://arxiv.org/abs/1811.03402
[45] Charissa Ann Ronao and Sung-Bae Cho. 2016. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications* 59, 15 (Oct. 2016), 235–244. https://doi.org/10.1016/j.eswa.2016.04.032
[46] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-Task Self-Supervised Learning for Human Activity Detection. *Proceedings of the ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 61 (June 2019), 30 pages. https://doi.org/10.1145/3328932
[47] Aaqib Saeed, Victor Ungureanu, and Beat Gfeller. 2020. Sense and Learn: Self-Supervision for Omnipresent Sensors. *CoRR* abs/2009.13233 (2020). arXiv:2009.13233 https://arxiv.org/abs/2009.13233

[48] Kathleen Siren. 2020. The Best of both Worlds: Expanding the Depth and Breadth of Multiple-Choice Questions. In *Proceedings of the INTED2020 14th International Technology, Education and Development Conference* (Valencia, Spain). Social Science Research Network, Rochester, NY, 7173–7177. https://papers.ssrn.com/abstract=3660997

[49] Boris Sofman, Ellie Lin, J. Andrew Bagnell, John Cole, Nicolas Vandapel, and Anthony Stentz. 2006. Improving robot navigation through self-supervised online learning. *Journal of Field Robotics* 23, 11-12 (Nov. 2006), 1059–1075. https://doi.org/10.1002/rob.20169

[50] Namrata Srivastava, Joshua Newn, and Eduardo Velloso. 2018. Combining Low and Mid-Level Gaze Features for Desktop Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 189 (Dec. 2018), 27 pages. https://doi.org/10.1145/3287067

[51] Julian Steil and Andreas Bulling. 2015. Discovery of Everyday Human Activities from Long-Term Visual Behaviour Using Topic Models. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. ACM, New York, NY, USA, 75–85. https://doi.org/10.1145/2750858.2807520

[52] Alexander Strukelj and Diederick C. Niehorster. 2018. One page of text: Eye movements during regular and thorough reading, skimming, and spell checking. *Journal of Eye Movement Research* 11, 1 (Feb. 2018), 1–22. https://doi.org/10.16910/jemr.11.1.1

[53] Setareh Rahimi Taghanaki and Ali Etemad. 2020. Self-supervised Wearable-based Activity Recognition by Learning to Forecast Motion. *arXiv* (2020). arXiv:2010.13713 https://arxiv.org/abs/2010.13713

[54] Tobii. 2020. Tobii eye-tracker. https://gaming.tobii.com/getstarted/. Accessed: Nov 10, 2020.

[55] Meng-Jung Tsai, Huei-Tse Hou, Meng-Lung Lai, Wan-Yi Liu, and Fang-Ying Yang. 2012. Visual Attention for Solving Multiple-Choice Science Problem: An Eye-Tracking Analysis. *Computers & Education* 58, 1 (Jan. 2012), 375–385. https://doi.org/10.1016/j.compedu.2011.07.012

[56] Zhiguang Wang and Tim Oates. 2015. Imaging Time-Series to Improve Classification and Imputation. In *Proceedings of the International Conference on Artificial Intelligence* (Buenos Aires, Argentina) *(IJCAI'15)*. AAAI Press, 3939–3945. https://www.ijcai.org/Proceedings/15/Papers/553.pdf

[57] Wikipedia. 2020. Horizontal and vertical writing in East Asian scripts. https://en.wikipedia.org/w/index.php?title=Horizontal_and_vertical_writing_in_East_Asian_scripts&oldid=984358336. Accessed: Oct 29, 2020.

[58] Wei Wu and Yuan Zhang. 2019. Activity Recognition from Mobile Phone using Deep CNN. In *Proceedings of the Chinese Control Conference* (Guangzhou, China). IEEE, 7786–7790. https://doi.org/10.23919/ChiCC.2019.8865142

[59] Kento Yamada, Koichi Kise, and Olivier Augereau. 2017. Estimation of Confidence Based on Eye Gaze: An Application to Multiple-Choice Questions. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) *(UbiComp '17)*. ACM, New York, NY, USA, 217–220. https://doi.org/10.1145/3123024.3123138