Information Engineering Express
International Institute of Applied Informatics
2025, Vol. 11, No. 1, IEE816

Contrastive Learning for Fine-Grained Reading Detection

Md. Rabiul Islam [*] , Andrew W. Vargo [†] , Motoi Iwata[†] ,
Masakazu Iwamura[†] , Koichi Kise[†]

Abstract

Reading is a cognitive activity that we perform aiming at various purposes, such as gaining knowledge and entertaining ourselves, with different scripts and layouts. Therefore, automatic reading detection gives useful information about users' reading activities. Deep learning enables automatic feature extraction and model creation but needs large-sized labeled data. The self-supervised learning devised to overcome this limitation work as non-contrastive self-supervised learning (SSL) and contrastive self-supervised learning (contrastive learning). Although SSL is well explored for reading analysis, contrastive learning is not still well explored. This paper explores contrastive learning that works in several ways. A Simple Framework for Contrastive Learning of Visual Representations (SimCLR) is one way that has attracted much attention in many research domains because of its superior performance. We explore SimCLR for the cognitive activity recognition task of fine-grained reading detection employing electrooculography datasets. These datasets describe eye movements that have been recorded for in-the-wild condition. The obtained results are compared against SSL and supervised baselines. The results show that, for an equal number of training samples, the SimCLR method obtains a maximum performance gain of 3.02 and 3.96 percentage points compared to the two baselines, respectively. Besides, SimCLR shows the best performance for large-sized data with a data efficiency of about 80%, whereas SSL shows the best performance for small-sized data. The analysis conducted in this paper shows a direction for researchers and system designers to employ self-supervised learning for automatic reading detection.

Keywords: Contrastive learning, Reading detection, Self-supervised learning, SimCLR.

1 Introduction

When a human reads, they are decoding a series of symbols to obtain the intended meaning from these symbols [1]. Humans perform different types of non-mutually exclusive reading activities, such as reading for knowledge acquisition or reading for entertainment. Therefore, reading is an important source of information that shapes our minds and is key for both the development and maintenance of cognitive ability [2]. So that, understanding

[*] BSMRSTU, Gopalganj, Bangladesh
[†] Osaka Metropolitan University, Sakai, Japan

and improving daily reading habits through reading behavior monitoring can provide several benefits, including improved reading quality, increased vocabulary, and logical thinking [3]. For instance, just as people are encouraged to be physically fit by monitoring step counts [4], tracking the read volume in a day has the potential to motivate them to read more. Researchers have devised multiple ways to measure the daily read volume, and one idea to quantify reading is by estimating the number of read words [5]. Another idea is reading detection, which aims to differentiate periods of reading from all other activities [6]. A challenge is that since reading is one of the main activities of our daily lives, reading materials are numerous and have different scripts and layouts. Therefore another important area of study is in identifying subordinate reading categories, e.g., fine-grained reading detection.

There is a wide range of diversity in how writing is presented in the world. To start, there is a wide range of writing systems, from scripts to pictographic systems. For example, the Japanese writing system follows two conventions that are horizontal and vertical. The horizontal writing moves from left to right with multiple downward rows without spaces between words. On the other hand, vertical writing moves from top to bottom having multiple columns from right to left [7]. According to the conventions, reading Japanese vertical text usually includes the user reading novels or newspapers. Besides, in Japan, reading English text means the user is reading something technical materials such as scientific papers if the user is a science student. Finally, reading Japanese horizontal includes all other text materials. Therefore, instead of just differentiating reading from all other activities, fine-grained reading detection gives more valuable information about reading activities.

Reading activities are reflected by reading behavior, such as a reader having difficulty in understanding the contents of a document is characterized by low reading speed and frequent rereading. Therefore, a fundamental way of reading detection is by analyzing reading behavior. Reading detection by analyzing reading behavior can be conducted in multiple ways [8, 9] whereas most of the existing technologies employ classical machine learning in laboratory settings. But, these technologies suffer from multiple issues, including poor performance in real-world applications beyond the controlled laboratory settings (in-the-wild).

A large number of machine learning techniques have been proposed to improve the accuracy in real-world applications. The increasingly popular machine learning algorithm is deep learning (DL) [10] which enables simultaneous automatic feature extraction and model creation employing noisy in-the-wild data. DL attracted much attention in the last decade by successfully solving many tasks in various domains with superior performance [11, 12]. Among all its successes, the major drawback is that it is the world's most data-hungry algorithm that needs to prepare large-sized labeled data, that captures diverse behavior, to attain peak performance. In most domains, accumulating enough labels is a serious issue [13]. This is because in-the-wild study is labor-intensive, financially expensive, and needs significant investment. These limitations prohibit generating labeled datasets with satisfactory size. The lack of large-sized labeled data is also a problem for reading analysis such as reading detection, reading quality classification, and read word count. Therefore, it is not easy to directly apply DL in reading analysis.

The self-supervised learning [14] is a promising way to solve challenges posed by the over-dependence of DL algorithms on labeled data that can be separated into two task types, pretext task and target task. The pretext task is formatted by employing an automated process to synthesize training data from unlabeled data and solved to pre-train the model. An example of a pretext task is the identification of transformations (for example, noise addition and permutation) applied to data samples. After pre-training, the model is fine-tuned for the target task training (reading detection). Based on the pre-training self-supervised

learning work in two ways [15]; noncontrastive self-supervised learning (SSL) [14] and contrastive self-supervised learning (contrastive learning) [16, 17].

The SSL technique arranges the pretext task of classification by automatically generating labels (pseudo labels) employing unlabeled data. There are many ways to create classification tasks by generating pseudo labels such as transformation prediction and masked prediction [18]. An enhanced objective function solved in classification assists the model in learning robust feature representation, which is needed to solve the target task.

On the other hand, contrastive learning arranges a pretext task by generating positive data samples coming from similar distributions and negative data samples coming from dissimilar distributions. The model learns features by solving a contrastive task of maximizing agreement, i. e., minimizing and maximizing the distance between positive and negative data samples, respectively. Based on contrastive task generation, contrastive learning work in multiple ways [19]. The simple framework for contrastive learning of visual representations (SimCLR) [20] is one framework proposed in the computer vision domain. It uses data augmentation for generating positive and negative data samples to create pretext tasks. In the activity recognition domain, researchers have also adopted the SimCLR method with effective performance for tackling the lack of large-sized labeled data issues by employing simple signal transformations for data augmentation [21, 22].

Previous work on activity recognition using SimCLR is for the class of "physical human activity recognition." This recognizes physical activities employing data produced from a series of observations collected from a set of body-worn sensors for physical movement. Another class of activity recognition is "cognitive human activity recognition," which recognizes activities that are related to human mental processes by employing data captured from biological signals. Reading is a cognitive human activity. In reading analysis, the lack of labeled data is a serious issue, and researchers adopted SSL to tackle it [23]. However, the usefulness of contrastive learning in this field is not well known, and we do not know if different signal transformations will be more effective for data augmentation.

In this work, to answer these questions, we take fine-grained reading detection, as a representative task of reading analysis, which is the classification of four classes: reading English (ENG), Japanese vertical (JV), and Japanese horizontal (JH) texts, and not reading anything (NR). We evaluate the SimCLR method, as a surrogate of contrastive learning, for this task employing seven signal transformations for data augmentation with a combination of two. We employ electrooculography (EOG) data describing users' eye movements and record for in-the-wild condition to evaluate the SimCLR method. We compare obtained results against the SSL and pure supervised (supervised) baselines.

The obtained results show that for a significant number of signal transformation pairs with a wide range of 100% to 20% of available labeled data (5,340), the SimCLR method outperforms the SSL and supervised baselines. Besides, the SimCLR method shows a maximum performance gain of 3.02 and 3.96 percentage points compared to the SSL and supervised baselines, respectively, when an equal number of labeled data is used for training both. The results also show a data efficiency of about 80% for the SimCLR method against both baselines. It means that the same performance was obtained for the SimCLR method and baselines with 20% of the data and 100% of the data, respectively. Further analysis shows that the SimCLR method performed best compared to the baselines for enough data samples although SSL is superior when data is scarce. The detailed results show directions to apply contrastive learning to pursue the best performance based on the amount of available labeled data that makes it practical to get accurate user reading behavior for giving feedback to motivate and improve users' reading behavior.

The remainder of this paper is composed as follows. Section 2 includes a literature review that is divided into three subsections to cover three perspectives of this paper self-supervised learning, contrastive learning, and reading detection. Section 3 presents the methodology with a detailed explanation of the employed self-supervised learning techniques. Section 4 includes the description of the employed datasets. Section 5 presents experimental protocols and outcomes of the experiments with a detailed comparative discussion of SimCLR, SSL, and supervised learning. Finally, we concluded the paper with a future direction in Section 6.

2 Literature Review

The research work presented in this paper is related to three fields; SSL, contrastive learning, and reading detection. In this section, we describe how our work builds on these fields.

2.1 Self-Supervised Learning

The SSL utilizes pretext tasks that need domain expertise which requires some semantic understanding. This technique generates discriminative and predictive models that generally measure how good is the classifier by calculating loss in the output space. These methods have shown great promise across many domains [15]. In physical activity recognition, a number of SSL methods have been proposed for time-series data. A multi-task SSL method is proposed by Saeed et al. [14] by creating pretext tasks by employing signal transformations to the unlabeled data. Another SSL method is proposed in [24] for learning effective representations from sensory data to mitigate the lack of large and labeled data showing that the SSL is effective. A similar SSL method is proposed by Taghanaki et al. [25] for boosting performance by pre-training the network by generating a pretext task of predicting the values in future time steps. Likewise, the masked reconstruction pretext task is proposed by Haresamudram et al. [26] for self-supervised pre-training using time series data. Tang et al. [27] increased the internal and external diversity of training data to learn more generalizable features by combining multi-task SSL and teacher-student self-training. Although most of the existing SSL methods are for physical activity recognition, researchers take a pioneering step to apply SSL for cognitive activity recognition, and Islam et al. [23] proposed an SSL method for reading activity classification employing simple signal transformations for pre-training the model.

2.2 Contrastive Learning

Contrastive learning [19] measures how good the representation is and calculates loss in the representation space. This technique teaches the model to identify data samples coming from similar distribution and dissimilar distribution called positive and negative data samples, respectively. The model, therefore, is pre-trained by learning the features by solving a contrastive task that maximizes agreement [28]. Recently, contrastive learning has attracted much attention because of the enhanced performances that can be performed in numerous ways [19]. Hadsell et al. [29] introduced a contrastive loss function with the goal to maximize the distance between dissimilar pairs and minimize the distance between similar pairs. Haresamudram et al. [30] have employed contrastive predictive coding to learn the temporal structures of sensor data. Schroff et al. [31] introduced a triplet loss in feature extraction that consists of anchor images, positive images of the same image, and negative images of

a different image. The goal is to minimize the distance between the anchor and positive images and maximize the distance between the anchor and negative images. Chen et al. [32] have employed siamese networks for contrastive learning. Chen et al. [20] propose a contrastive learning framework called SimCLR, which learns representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. The SimCLR has attracted much attention by superior performance and adopted in various research domains. The SimCLR has also been adopted in physical activity recognition and health to tackle the lack of large-sized labeled data issues. Tang et al. [21] have explored SimCLR in human activity recognition where they employed signal transformations for data augmentation using sensory data. Wang et al. [22] have also applied SimCLR to human activity recognition where authors evaluated the conventional signal transformations for data augmentation and proposed a new one that achieved good performance. Khaertdinov et al. [33] combine the SimCLR framework, with a transformer-based encoder for sensor-based human activity recognition. Shah et al [17] have applied SimCLR to human activity recognition for clinical outcomes. Although contrastive learning and SimCLR have been well explored for different purposes including physical activity recognition, it is not still well explored for cognitive activity recognition.

2.3 Reading Detection

Reading is a cognitive human activity [1]. Reading analysis, as a representative of cognitive activity recognition, varies depending on the purpose. In the past years, researchers devised methods for automatic reading detection to assist readers. Bulling et al. [8] proposed a method to detect reading as a part of other human activities. Strukelj et al. [9] proposed a method to accomplish different modes of reading. The goal is to differentiate regular reading, through reading, skimming, and spell-checking. On the other hand, Landsmann et al. [6] proposed a method for reading detection, reading versus not reading. The lack of large and labeled data in this research domain forces to carry out most of the studies by using classical machine learning, except for a small portion employing DL [34]. The application of SSL for reading analysis has shown ways to tackle the lack of large-sized labeled data issues [23] in cognitive activity recognition. Contrastive learning has shown a potential solution to adopt DL in many domains including physical activity recognition using small-sized labeled data. But to the best of our knowledge, it has not been as of yet well explored for cognitive activity recognition with the exception of the preliminary work [35]. This study, therefore, aimed to take fine-grained reading detection as an example of cognitive human activity and explore SimCLR contrastive learning for it.

3 Method

We explore the SimCLR method, with the necessary modifications to apply, for fine-grained reading detection that consists of SimCLR pre-training and target task training, as shown in Figure 1. Reading detection distinguishes between reading and not reading. In this study, we implement reading detection as a classification task by dividing the users' reading and not reading activities into short segments and then classifying them into fine-grained classes as reading ENG, JV, JH texts, and NR.

3.1 SimCLR Pre-training

The SimCLR pre-training, as shown in the upper part of Figure 1, is done employing the unlabeled EOG data and a constraint by applying data augmentation. The SimCLR pre-training works based on the constraint that the extracted features coming from two augmented data samples originating from the same data sample must be similar, and those originating from separate data samples must be dissimilar. Therefore, we divide the unlabeled continuous time-series EOG data into short data segments. We then group the EOG data segments into batches of size 1024. From each original batch (b), we generate two slightly dissimilar copies of ($\dot{b}$ and $\ddot{b}$) by applying a pair of signal transformations twice with slightly different random parameters. The augmented copies of the batch are sent to an encoder whose outputs, ($\dot{h}$ and $\ddot{h}$), are then sent to a projection head that outputs two feature vectors ($\dot{z}$ and $\ddot{z}$). Finally, we compute NT-Xent contrastive loss [20] employing these two feature vectors with maximizing the agreement which maximizes the similarity between augmented copies of data segments originating from the same data segment and minimize the similarity between augmented copies of data segments originating from the separate data segments.

We employ seven signal transformations [21, 23] to generate transformation pairs for data augmentation which are scaling, noise addition, negate, time-flip, channel shuffle, permutation, and time-warp. In short, scaling is done by multiplying the values of the data segment with a random number coming from a normal distribution with a mean of 1 and standard deviation of 0.1, noise addition is done by adding Gaussian noise with a mean of 0 and standard deviation of -0.05 to the values of the data segment, negate is done by inverting the values of the data segment which is multiplying with -1, time-flip is done by horizontal-flip, i.e., by reversing the time axis of the data segment, channel shuffle is done by permuting the channels of the data segment, permutation is done by applying permutation along the time axis of the data segment, and time-warp is done by stretching and warping the data segment along the time axis.

The encoder network consists of CNN layers. We employ four 1D convolutional layers with the number of filters of 16, 32, 64, and 96, respectively, with kernel sizes of 32, 24, 16, and 8, respectively. We add a dropout layer after each convolutional layer and a Global max-pooling layer after the last dropout layer. The projection network consists of three fully-connected layers that carry 128, 64, and 32 units, respectively. We employ Adam and relu for the optimization and activation, respectively. In Figure 1, we have shown the same encoder and projection networks twice for easy understanding, but both share the same network parameters.

3.2 Target Task Training

After SimCLR pre-training, we conduct the target task training, as shown in the lower section of Figure 1 employing the segmented labeled EOG data. We fine-tune the pre-trained encoder and re-train it by replacing the projection head with a classifier network consisting of a fully-connected layer carrying 4 units with linear activation. We use the same hyperparameters as used for the SimCLR pre-training.

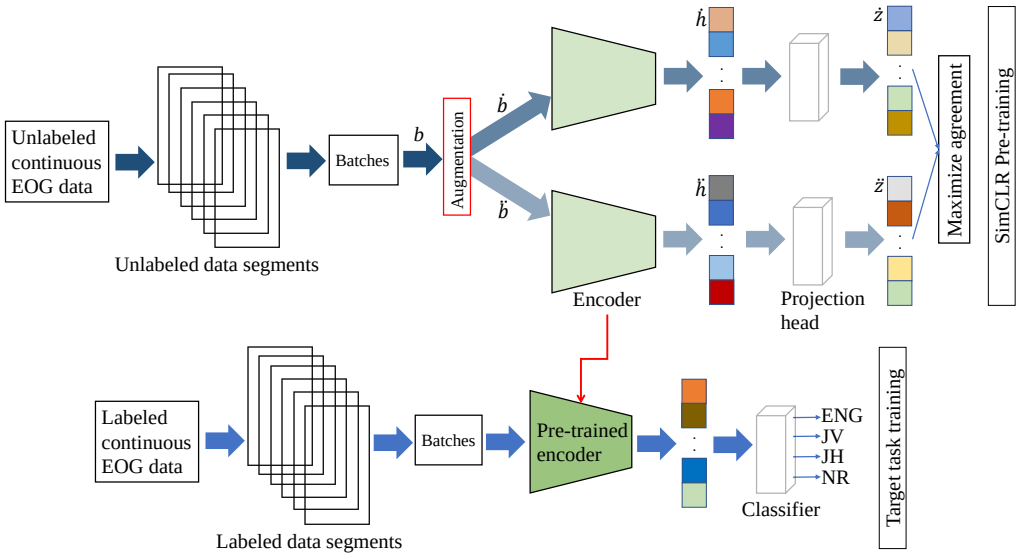


Figure 1: The SimCLR method which consists of two steps, the pre-training and target task training, where the pre-training comprises four basic elements; data augmentation, encoder, projection head, and agreement maximization based on contrastive loss.

4 Datasets

Two main organs that we use while reading are the eye and brain. The human eyes are the main organ that we use to acquire information while we read and the brain process it to extract the meaning of it. We, therefore, can capture reading behavior by measuring eye movements. In addition to the eye reading behavior is slightly described by head and body movement.

In this study, we employ data describing eye movements recorded using J!NS-MEME glasses, an eye-wear device shown in Figure 2, with a sampling frequency of 100 Hz. The J!NS-MEME carries three sensors; EOG, gyroscope, and accelerometer. The EOG measures potential generated due to the eye movements, whereas the gyroscope and accelerometer measure head and body movements. Although J!NS-MEME carries three sensors, we used data coming only from the EOG sensor. This is because a previous study [36] shows that only EOG data describing eye movements are sufficient to describe reading behavior. The EOG data is recorded as two channels describing the horizontal and vertical eye movements. We use two datasets, an unlabeled EOG dataset and a labeled EOG dataset.

The labeled EOG dataset was recorded and reported by Ishimaru et al. [37]. This dataset has also been used by Islam et al. [23]. The labeled EOG dataset is recorded employing ten native Japanese university students as the user. The data is recorded for two days and each user wore the J!NS-MEME glasses for 12 hours per day. The users read ENG, JV, and JH texts each for one hour per day, and did not read anything for the remaining period of the day. The users also wore the Narrative Clip that brings a camera to take frontal images. These images are used for labeling data. We first pre-processed the recorded EOG data and then divided it into short segments of 30 seconds with an overlap of 15 seconds. Each segment contains 3000 time steps with the dimension of 3000×2. There exist noise in some data segments due to poor skin-electrode contact. We removed noisy data segments using noise judgment criteria [23]. After discarding noisy segments, the number of data

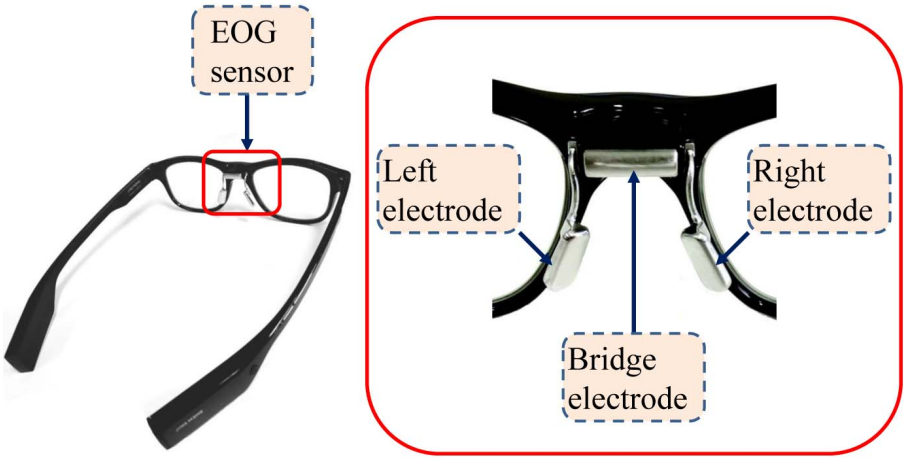


Figure 2: JINS-MEME glasses carry an EOG sensor consisting of left, right, and bridge electrodes that used for data recording.

segments for ENG, JV, JH, and NR is 5340, 5798, 5792, and 32708, respectively.

The unlabeled EOG dataset employed in this study is recorded and reported by Islam et al. [23]. This dataset was also recorded in the same way described for the labeled EOG dataset except for the collection of the data labels. A total of 52 users were employed for recording the data. We pre-possessed, segmented, and discarded noisy data segments in an above-mentioned way. The total number of unlabeled EOG data segments is 177,921.

Both datasets were recorded without imposing any restrictions and conditions except for the guidelines to obtain data that perfectly reflects natural behavior. The users behaved as needed and maintained a convenient distance from the texts. Therefore these datasets are considered "in-the-wild." This accumulated naturalistic data enables us a more realistic evaluation of methods to validate their applicability in real environments.

5 Evaluation and Results

5.1 Experimental Protocols

The aim of the experiment is to evaluate the performance of the SimCLR method that includes SimCLR pre-training and then target task training. We conducted the SimCLR pre-training using the unlabeled EOG data. In SimCLR pre-training, we made a 7×7 matrix of possible signal transformation pairs using seven signal transformations as we explained in Section 3. We conducted experiments for all 49 signal transformation pairs. For each pair of signal transformations, we applied two transformations sequentially, one after another, to the values of the data segment except for the case where the same signal transformation (diagonal) creates the transformation pair by repeating it and, in this case, we applied the transformation to the value of data segment only once. After the SimCLR pre-training for each pair of signal transformations, we generated the target task model by re-training it employing labeled EOG data.

We use SSL and supervised methods as a baseline to measure the proficiency of the SimCLR method. In the case of SSL, we adopted the method described in [23] and reproduced results employing the same unlabeled and labeled EOG datasets and signal transformations

Table 1: Prediction accuracy for SimCLR method (in percentage) for 7×7 signal transformation pairs. The diagonal entries represent employing a single transformation twice. The orange (italic) and violet (bold) texts outperform only the supervised baseline and both SSL and supervised baselines, respectively.

(a) 100% training samples

1st transformation \ 2nd transformation	scale	noise	negate	time-flip	channel shuffle	permute	time-warp
scale	**57.95**	**59.45**	*56.47*	*56.43*	**60.38**	**58.24**	**58.30**
noise	**58.07**	**58.23**	**59.76**	**58.49**	**58.82**	**57.71**	**58.41**
negate	**59.70**	**59.01**	*57.16*	*57.14*	**57.43**	**57.69**	**58.69**
time-flip	**58.37**	**59.44**	55.83	**58.24**	*56.68*	**57.73**	**59.23**
channel shuffle	**57.85**	*56.86*	56.01	**59.66**	**57.57**	**58.64**	**59.48**
permute	**58.59**	**57.67**	**57.81**	**57.47**	**58.73**	**58.56**	**57.45**
time-warp	**59.02**	**57.84**	**58.11**	**57.47**	*57.24*	**58.13**	**58.43**

(b) 50% training samples

1st transformation \ 2nd transformation	scale	noise	negate	time-flip	channel shuffle	permute	time-warp
scale	*56.51*	**57.82**	*56.26*	**57.54**	**58.14**	**57.72**	**59.10**
noise	**57.97**	**58.26**	**57.77**	**57.65**	**57.45**	**58.48**	**58.70**
negate	**58.18**	**58.22**	*56.76*	*56.63*	**57.74**	**58.22**	**58.48**
time-flip	**58.64**	**57.99**	**57.57**	**58.41**	**58.30**	**57.97**	**58.25**
channel shuffle	*57.00*	**57.32**	*56.60*	**57.69**	**58.33**	**58.47**	**58.45**
permute	**58.22**	**57.58**	**58.59**	**57.91**	**58.57**	**58.10**	**57.76**
time-warp	**57.79**	**58.85**	**57.73**	**57.74**	**58.88**	**57.97**	**57.52**

(c) 20% training samples

1st transformation \ 2nd transformation	scale	noise	negate	time-flip	channel shuffle	permute	time-warp
scale	*54.45*	**56.14**	*54.94*	*54.91*	*55.15*	*55.13*	**57.03**
noise	**55.62**	**55.64**	**56.00**	*55.16*	**55.99**	**56.83**	**56.07**
negate	*55.03*	*54.67*	*54.34*	*54.69*	*54.64*	**55.46**	**56.28**
time-flip	*54.44*	**55.96**	53.08	**55.71**	*53.80*	**56.03**	**55.75**
channel shuffle	**55.35**	*55.00*	*55.10*	*54.84*	*55.22*	**56.82**	**55.36**
permute	**57.29**	**55.72**	**56.07**	**56.42**	**56.89**	**55.99**	*55.26*
time-warp	**55.68**	**55.71**	**56.73**	**55.53**	**56.68**	**56.55**	**56.24**

(d) 0.1% training samples

1st transformation \ 2nd transformation	scale	noise	negate	time-flip	channel shuffle	permute	time-warp
scale	27.60	*28.67*	*29.44*	27.82	*29.57*	*28.49*	*29.62*
noise	*30.56*	*29.21*	28.19	*29.56*	27.48	*30.16*	28.23
negate	*29.00*	*28.50*	*28.92*	27.59	28.32	*30.43*	*28.73*
time-flip	28.25	27.90	*28.89*	27.60	27.92	*28.90*	28.39
channel shuffle	*30.97*	27.86	27.72	*28.82*	27.78	*29.47*	*31.34*
permute	28.09	*28.83*	*32.35*	*29.34*	*29.89*	*29.68*	*29.44*
time-warp	*30.07*	*29.58*	*29.52*	28.04	*30.74*	*29.16*	28.43

Table 2: Prediction accuracy (in percentage) for baselines, SSL and supervised, with Sim-CLR (max accuracy) method.

Method	Training samples			
	100%	50%	20%	0.1%
SimCLR (max accuracy)	60.38	59.10	57.29	32.35
SSL	57.36	57.06	55.28	40.57
Supervised	56.42	55.77	53.51	28.46
Chance rate	25.00	25.00	25.00	25.00

used in the SimCLR method. We generated a classification task by applying all signal transformations to the unlabeled EOG data. After that, we pre-trained the model by solving a classification task of predicting the applied transformation to the data segment whereas the SimCLR method optimized the contrastive objective. After pre-training, same as SimCLR, we fine-tuned the network and re-trained it using the labeled EOG data. For the supervised method, we used the model with the same architecture described for the SSL method and trained it for the target task employing the same labeled EOG dataset without pre-training.

There exist class imbalance, as we discussed in Section 4, in the labeled EOG dataset, and we removed it by down-sampling the majority classes to 5,340 samples, which is the smallest number, by random selection. Therefore, the chance rate for a four-class classification target task is 25%. We evaluated the performance of all methods employing the labeled training data segments of 100% (all available), 50%, 20%, and 0.1% training samples for the target task training. This allows us to evaluate performance for a wide range of available

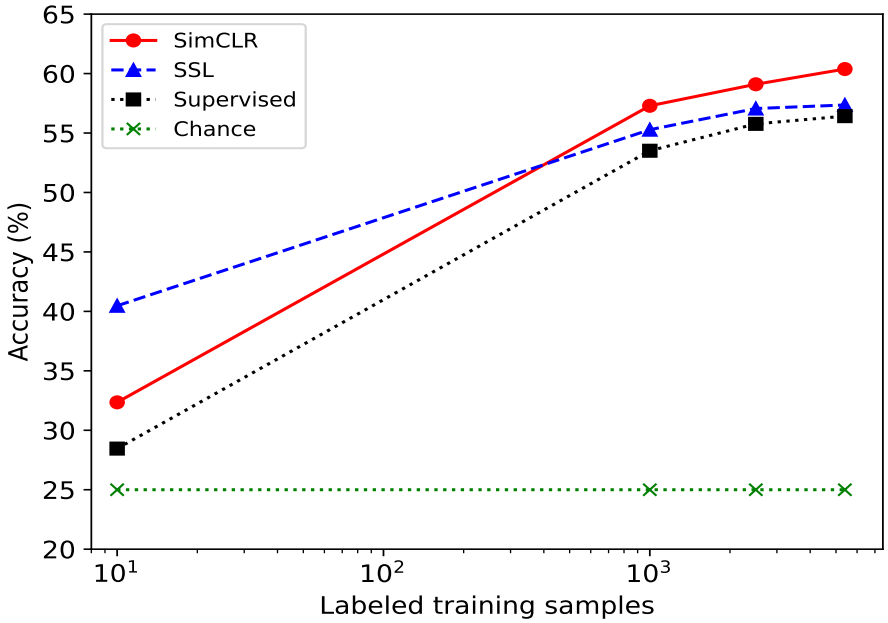


Figure 3: Comparison between SimCLR and baseline methods.

data. The model was trained using data coming from nine out of ten users and tested on the data coming from the remaining one that evaluated methods in a user-independent way. The prediction accuracy is used as the evaluation metric.

5.2 SimCLR Outcomes

We reported the obtained results in accuracy for the SimCLR method in Table 1 for a batch size of 32 and a learning rate of 0.003 where Table 1(a)-(d) report the results for 100%, 50%, 20%, and 0.1% training samples, respectively. In addition, we reported the results of baseline methods, SSL and supervised, along with obtained maximum SimCLR outputs in Table 2. The orange (italic) and violet (bold) texts in Table 1 represent that the SimCLR method outperforms only the supervised baseline and both baselines, SSL and supervised, respectively. The results show that the SimCLR method outperforms baseline methods for a wide range of 100%, 50%, and 20% training samples for a vast majority of signal transformation pairs employed in SimCLR pre-training for data augmentation. On the other hand, for training samples around 0.1%, the SSL method performed well compared to the SimCLR method, although the SimCLR outperformed the supervised method for a significant number of signal transformation pairs.

The performance of the SimCLR method varies based on the signal transformation pairs and the number of training samples employed for data augmentation in SimCLR pre-training and target task training, respectively. The noise addition, permutation, and time-warp signal transformations along with other signal transformations performed well for 100%, 50%, and 20% training samples counting the number of outperforms. On the other hand, the permutation along with other signal transformations performed well for 0.1% training samples. Another important point is that the signal transformation pairs (diagonal) created using the same transformation performed quite well although the best performance is obtained for the signal transformation pair of different transformations.

Table 3: The SimCLR method's performance gain that is calculated based on Table 2.

(a) For both SimCLR and baseline methods are trained for an equal number of training samples.

Baselines	Training samples used for both SimCLR and baseline			
	100%	50%	20%	0.1%
SSL	3.02	2.04	2.01	-8.22
Supervised	3.96	3.33	3.78	3.89

(b) For SimCLR and baseline methods are trained for different numbers of training samples and only 100% training samples, respectively.

Baselines (samples used)	SimCLR (training samples used)			
	100%	50%	20%	0.1%
SSL 100%	3.02	1.74	–0.07	–25.01
Supervised 100%	3.96	2.68	0.87	–24.07

Finally, we show a comparison between SimCLR and baseline methods as shown in Figure 3 generated based on Table 2. The SimCLR and SSL methods, which belong to self-supervised learning, are never inferior compared to the supervised method. On the other hand, for enough labeled training samples, the SimCLR is superior to SSL although for small-sized labeled data the SSL is superior to SimCLR. Therefore, with respect to the data-hungry, SSL is least hungry than SimCLR. On the other hand, if we are not intending to get a relatively good performance only with a very small number of samples, it is always better to use SimCLR. Another advantage is that SimCLR is also better than supervised with enough amount of labeled data. The results show a path for researchers to select the best model depending on the signal transformation pairs and available labeled training samples.

5.3 Study of Performance Gain and Data Efficiency

We also explored the SimCLR method by analyzing the performance gain and data efficiency. We calculated performance gain as the difference between the outputs (accuracy) for the SimCLR and baseline methods. Besides, the data efficiency represents whether or not the SimCLR method performs well with fewer training samples compared to the baseline methods and is measured by taking performance gain as a parameter.

Table 3 reports the performance gain that we calculated based on Table 2. We measured the performance gain for two cases. In the first case, we employed an equal number of data for training both methods, SimCLR and baseline, and Table 3(a) reports the results. The results show that we obtained a maximum performance gain of 3.02 and 3.96 percentage points compared to the SSL and supervised baselines, respectively. In the second case, we employed 100%, 50%, 20%, and 0.1% training sample cases for the SimCLR method, whereas 100% training samples for the baseline methods and Table 3(b) reports the results. The results show that an almost equal performance is obtained when the SSL and SimCLR methods are trained by employing 100% and 20% training samples, respectively. Therefore, the SimCLR method is data-efficient by about 80%. On the other hand, when the supervised and SimCLR methods are trained by employing 100% and 20% training samples, respectively, a performance gain of 0.87 is obtained. Therefore, a data efficiency of more than 80% is obtained for the SimCLR method.

The obtained performance gain and data efficiency show that the SimCLR pre-training help in learning discriminative features that, in turn, help to achieve superior performance in the target task by improving class-level prediction.

Table 4: Dependency of outcomes on batch size and learning rate, here #Outperform de-notes the total number of signal transformation pairs for which the SimCLR method out-performs two baselines; SSL and supervised.

(a) Dependency on batch size

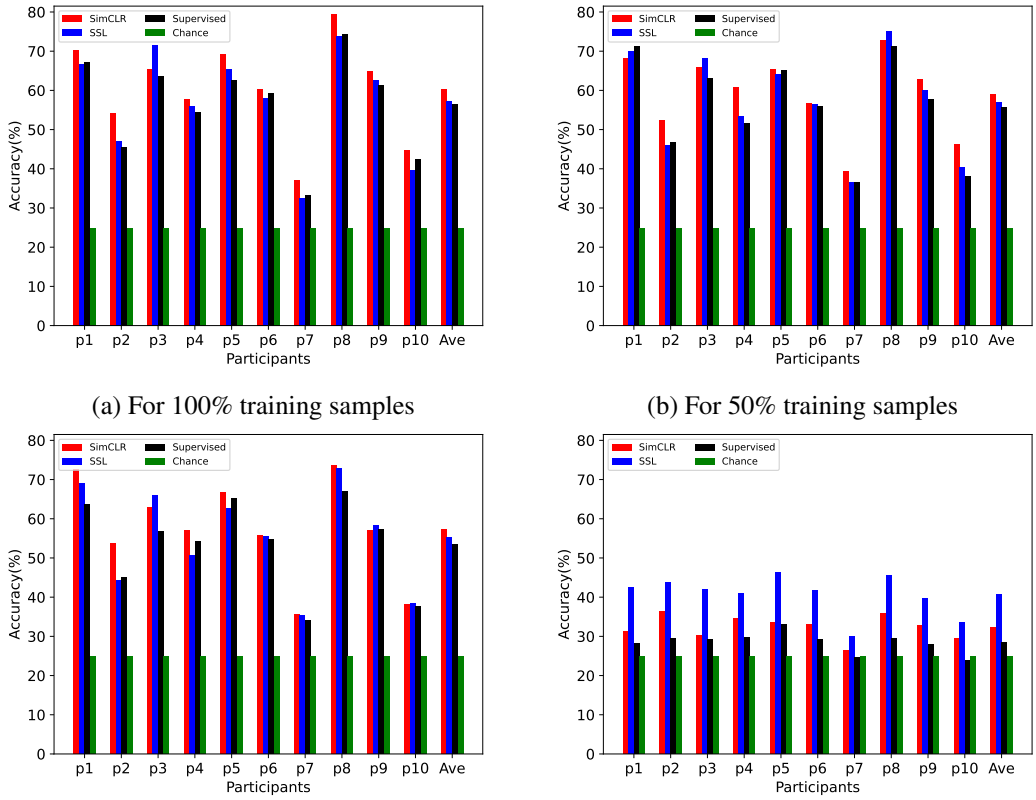
Data size	Parameter	Batch size				
		16	24	32	64	128
100%	#Outperform	27	34	40	37	29
	Max accuracy	58.88	59.90	60.38	59.73	59.27
50%	#Outperform	22	39	43	40	17
	Max accuracy	58.48	58.91	59.10	58.77	58.15
20%	#Outperform	25	38	30	28	11
	Max accuracy	57.00	56.76	57.29	57.16	56.09
0.1%	#Outperform	0	0	0	0	0
	Max accuracy	33.08	32.74	32.35	30.78	30.78

(b) Dependency on the learning rate

Data size	Parameter	Learning rate				
		0.0001	0.0005	0.001	0.002	0.003
100%	#Outperform	0	3	18	45	40
	Max accuracy	57.13	58.09	58.42	59.85	60.38
50%	#Outperform	0	4	10	40	43
	Max accuracy	55.91	57.42	57.49	58.83	59.10
20%	#Outperform	0	3	8	23	30
	Max accuracy	53.22	56.27	56.19	56.65	57.29
0.1%	#Outperform	0	0	0	0	0
	Max accuracy	27.25	28.57	29.42	31.15	32.35

5.4 Study of Outcomes' Dependency on Batch Size and Learning Rate

We also studied the dependency of the performance of the SimCLR method upon two hyperparameters, batch size and learning rate. To explore the dependency, we trained the SimCLR model for the target task with 100%, 50%, 20%, and 0.1% training samples for different batch sizes for the learning rate of 0.003 and learning rates for the batch size of 32. We set the following two criteria to explore the performance of the SimCLR method; the number of signal transformation pairs for which the SimCLR method outperformed the baseline methods, and the obtained maximum accuracy among all employed signal transformation pairs. We reported the obtained results in Tables 4(a) and 4(b). The obtained results report that the SimCLR method is robust for a wider range of smaller batch sizes, although the best result is obtained for a batch size of 32. On the other hand, in the case of the learning rate, the SimCLR method's worst performance is obtained for the low learning rate and improves with increasing it and produces the best performance for the learning rate of 0.003. Therefore, in general, for smaller batch sizes and larger learning rates, the SimCLR method performs best.



(a) For 100% training samples

(b) For 50% training samples

(c) For 20% training samples

(d) For 0.1% training samples

Figure 4: Results of the study of outcomes' sensitivity to users' behavior.

5.5 Study of Outcomes' Sensitivity to Users' Behavior

We also take a closer look at the performance of each method for individual users which helps to investigate the effect of users' reading behavior on the model's performance. We conducted a user-independent evaluation, therefore, this study carries important insight. The results, in Figure 4, show that the performance, across users, is highly imbalanced for 100%, 50%, and 20% training sample cases, although it is relatively balanced for 0.1% training sample case. Moreover, the advantage of the SimCLR method over baselines depends on the user's reading behavior. The test accuracy for all methods is quite insufficient for some users which may happen because of the following possible reasons. Firstly, for some users, the reading behavior may be quite different. Our analysis shows that the type of data segments for different classes for some users are quite different from other users. Therefore, the model failed to learn compelling features in training it using data from other users. Secondly, in-the-wild nature of the dataset may affect the model performance. The reading detection dataset has been recorded for in-the-wild condition. Our analysis shows that some not reading behavior is misclassified into reading. This is because users may make some unintentional mistakes, such as reading something unintentionally while data is being recorded for not reading class because we do not impose any restrictions on their behavior and act on their necessity in daily life. Thirdly, noisy data segments may affect the model performance. We further analyzed the nature of data that affects the classification performance of the model. The analysis shows that the dataset contains many noisy data

segments because of the poor electrode contact with the skin. Although we discarded these noisy segments as discussed in Section 4, it is not possible to discard all of these noisy segments considering the size of the dataset. The data segment that contains noise and cannot be removed by the noise judgment used in this study is mostly, because of this noise pattern, judged in a different class. Altogether the task is quite difficult in these cases, and test accuracy is terrible, which significantly affected the average performance we reported in this study, although test accuracy is verily good for many users.

5.6 Discussion

In the current era, deep learning is the key machine-learning tool that belongs to a variety of deep-learning techniques. So researchers are always wondering about which technique they should select for the dataset of their problem. One of the most promising deep learning techniques is self-supervised learning. In this paper, we explored the SimCLR, one of the contrastive self-supervised learning techniques, for a cognitive activity recognition task of reading detection by comparing the results with the SSL, another self-supervised learning technique, taking supervised learning as the baseline. The results of our experiments show that self-supervised learning is absolutely superior for any number of data samples compared to supervised learning. This superior performance of self-supervised learning is due to the ability of this technique to extract useful features in pre-training using the unlabeled data samples which are abundant. This in turn enhances the efficiency and the generalization capability of the self-supervised learning model. On the other hand, none of the two self-supervised learning techniques, SimCLR and SSL, are absolutely superior for any number of data samples with respect to one another. For a large number of data samples, SimCLR is superior and the opposite is true for a small number of data samples. So the size of the data samples is the most significant factor in achieving the best results using self-supervised learning. We also explored the SimCLR technique by measuring the data efficiency. The outcomes of the experiment show that the SimCLR is about 80% and more than 80% data efficient compared to the SSL and supervised learning techniques, respectively. To validate the effectiveness of the SimCLR for large data samples and SSL for small data samples, we also explored the SimCLR technique from the point of view of the dependency of outcomes on the signal transformations applied for self-supervision purposes, data efficiency, batch size, and user reading behavior. The results of all these factors validate the performance dependency on the size of data samples.

6 Conclusion

The self-supervised learning devised to tackle the lack of large-sized labeled data is conducted in two ways SSL and contrastive learning. The SimCLR method, a contrastive learning technique, reports an excellent performance in recent studies in handling the lack of labeled and large-sized data issues. This method has been evaluated in many research areas, including physical activity detection. Our study is one of the pioneers in exploring the SimCLR method for cognitive activity recognition like reading detection. We explored the SimCLR method for reading detection employing a large number of signal transformation pairs and compared it against the SSL and supervised baselines. The results show that for a vast majority of signal transformation pairs and a wide range of available labeled training data, the SimCLR method outperforms baseline methods with excellent data efficiency and performance gain. The detailed analysis shows that the SimCLR method is superior

for only large-sized labeled data and needs more computations (49 times compared to the SSL in this study). On the other hand, SSL is superior for small-sized labeled data and needs fewer computations. The obtained results and analysis carried out in this study show a path to achieve peak performance by applying self-supervised learning regardless of the available training data.

Future work includes studies to verify the effectiveness and suitability of signal transformations in pre-training employing different combination sizes such as three and also exploring the SimCLR method for other reading analysis tasks such as confidence and correctness estimation of tasks performed via reading.

Acknowledgments

This work was supported in part by JST Trilateral AI Research (JPMJCR20G3), and JSPS Grant-in-Aid for Scientific Research (20H04213, 20KK0235).

References

[1] S. Dehaene, "Reading in the brain: The new science of how we read," Penguin, 2010.

[2] A. E. Cunningham and K. E. Stanovich, "What reading does for the mind," American Educator, vol. 22, pp. 8-17, 1998.

[3] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster, "Eye movement analysis for activity recognition using electrooculography," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 4, pp. 741–753, 2011.

[4] G. J. Welk, J. A. Differding, R. W. Thompson, S. N. Blair, J. Dziura, and P. Hart, "The utility of the digi-walker step counter to assess daily physical activity patterns," Medicine and Science in Sports and Exercise, vol. 32, no. 9, pp. S481-S488, 2000.

[5] O. Augereau, C. L. Sanches, K. Kise, and K. Kunze, "Wordometer systems for everyday life," Proceedings of the ACM on IMWUT, vol. 1, no. 4, pp. 1–21, 2018.

[6] M. Landsmann, O. Augereau, and K. Kise, "Classification of reading and not reading behavior based on eye movement analysis," Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, UbiComp/ISWC '19 Adjunct, pp. 109–112, 2019.

[7] Wikipedia, "Horizontal and vertical writing in east asian scripts," https://en.wikipedia.org/w/index.php?title=Horizontal_and_vertical_wrtting_in_East_Asian_scripts&oldid=984358336, Accessed: December 25, 2022.

[8] A. Bulling, J. A. Ward, and H. Gellersen, "Multimodal recognition of reading activity in transit using body-worn sensors," ACM Transactions on Applied Perception, vol. 9, no. 1, article no. 2, pp. 1–21, 2012.

[9] A. Strukelj and D. C. Niehorster, "One page of text: eye movements during regular and thorough reading, skimming, and spell checking," Journal of Eye Movement Research, vol. 11, no. 1, pp. 1–22, 2018.

[10] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," Proceedings of the Twenty-Fifth International Joint Conference on AI, AAAI Press, pp. 1533-1540, 2016.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Commu. of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

[12] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649, 2013.

[13] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data - AI integration perspective," CoRR, abs/1811.03402, 2018.

[14] A. Saeed, T. Ozcelebi, and J. Lukkien, "Multi-task self-supervised learning for human activity detection," Proceedings of the ACM on IMWUT, vol. 3, no. 2, pp. 1-30, 2019.

[15] S. Albelwi, "Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging," Entropy, vol. 24, no. 4, pp. 1–22, 2022.

[16] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton, "Big self-supervised models are strong semi-supervised learners," ArXiv:2006.10029, 2020.

[17] K. Shah, D. Spathis, C. I. Tang, and C. Mascolo, "Evaluating contrastive learning on wearable timeseries for downstream clinical outcomes," ArXiv:2111.07089, 2021.

[18] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," IEEE Signal Processing Magazine, vol. 39, no. 3, pp. 42–62, 2022.

[19] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," Technologies, vol. 9, no. 1, 2021.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," Proceedings of the 37th International Conference on Machine Learning, PMLR, vol. 119, pp. 1597-1607, 2020.

[21] C. I. Tang, I. Perez-Pozuelo, D. Spathis, and C. Mascolo, "Exploring contrastive learning in human activity recognition for healthcare," ArXiv:2011.11542, 2020.

[22] J. Wang, T. Zhu, J. Gan, L. Chen, H. Ning, and Y. Wan, "Sensor data augmentation by resampling for contrastive learning in human activity recognition," ArXiv:2109.02054, 2021.

[23] M. R. Islam et al., "Self-supervised learning for reading activity classification," Proceedings of the ACM on IMWUT, vol. 5, no. 3, article no. 105, pp. 1-22, 2021.

[24] A. Saeed, V. Ungureanu, and B. Gfeller, "Sense and Learn: Self-Supervision for Omnipresent Sensors," CoRR abs/2009.13233, 2020.

[25] S. R. Taghanaki and A. Etemad, "Self-supervised wearable-based activity recognition by learning to forecast motion," ArXiv, 2020.

[26] H. Haresamudram et al., "Masked reconstruction based self-supervision for human activity recognition," Proceedings of the International Symposium on Wearable Computers, Mexico (Virtual Event), ACM, pp. 45–49, 2020.

[27] C. I. Tang, I. Perez-Pozuelo, D. Spathis, S. Brage, N. Wareham, and C. Mascolo, "SelfHAR: Improving Human Activity Recognition through Self-training with Unlabeled Data," ArXiv Preprint ArXiv:2102.06073, 2021.

[28] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," Proceedings of the IEEE/CVF Conference on Comp. Vision and Patt. Recog., pp. 2495-2504, 2021.

[29] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1735–1742, 2006.

[30] H. Haresamudram, I. Essa, and T. Plötz, "Contrastive predictive coding for human activity recognition," Proceedings of ACM on IMWUT, vol. 5. no. 2, 2021.

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823, 2015.

[32] X. Chen and K. He, "Exploring simple siamese representation learning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758, 2021.

[33] B. Khaertdinov, E. Ghaleb, and S. Asteriadis, "Contrastive self-supervised learning for sensor-based human activity recognition," 2021 IEEE International Joint Conference on Biometrics, pp. 1-8, 2021.

[34] L. Copeland, T. Gedeon, and S. Mendis, "Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error," Artificial Intelligence Research, vol. 3, no. 3, pp. 35–48, 2014.

[35] M. R. Islam, A. W. Vargo, M. Iwata, M. Iwamura, and K. Kise. "Evaluating Contrastive Learning for Fine-grained Reading Detection," 12th International Congress on Advanced Applied Informatics (IIAI-AAI), IEEE, pp. 430-435, 2022.

[36] M. R. Islam, A. W. Vargo, M. Iwata, M. Iwamura, and K. Kise, "Exploring Sensor Modalities to Capture User Behaviors for Reading Detection," IEICE Transactions on Information and Systems, vol. 105, no. 9, pp. 1629-33, 2022.

[37] S. Ishimaru, T. Maruichi, M. Landsmann, K. Kise, and A. Dengel, "Electrooculography dataset for reading detection in the wild," 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, pp. 85–88, 2019.