Real-Time Estimation of Eye Movement Condition Using a Deep Learning Model

Akihiro Sugiura[1](✉), Yoshiki Itazu[2], Kunihiko Tanaka[1],
and Hiroki Takada[2]

[1] Gifu University of Medical Science, Seki, Gifu 454-0822, Japan
asugiura@u-gifu-ms.ac.jp
[2] Fukui University, Fukui 910-8507, Japan

Abstract. In this study, we conducted a basic investigation involving the discrimination of eye movement condition (peripheral and central vision) using deep learning techniques. The subjects were 6 males aged 21–23 years. They watched two three-minute videos for central vision and peripheral vision in a random order for a total of eight sessions (four sessions each). The subjects wore an eye movement measurement device, and their eye movements (viewing angles) during the viewing of each video were continuously. From the time series data for eye movement, with four different lengths (0.5 s, 1 s, 2 s, 3 s) and shift length of 0.5 s, short time series data for each 3 min was obtained in sets of 350, and the data were utilized for deep learning and its evaluation. For the deep learning model, input nodes according to data length were placed in the input layer. For the middle layer, seven to eight units were put in place that brought together the one-dimensional convolution layer, the batch-normalization layer, normalized linear function, and the max-pooling layer. The output layer consisted of the fully-connected layer, sigmoid function, and multi-class cross-entropy. As a result, the accuracy of the discrimination was improved as the data length increased, and it was possible to determine the condition with an accuracy of over 90% if the eye movement data was at least one second.

Keywords: Deep learning · Eye movement · Classification · Peripheral vision · Central vision · Convolutional Neural Network

1 Introduction

Dementia is a syndrome in which the capacity for memory, thought, behavior, and daily life activities is lowered. There are 50 million people with dementia worldwide, and 10 million new cases each year [1]. The symptoms of dementia can be broadly classified into "core symptoms" and peripheral symptoms called "behavioral and psychological symptoms of dementia" (BPSD) [2, 3]. Core symptoms include complete or partial loss of short-term or episodic memory, impaired orientation, executive dysfunction, and impairment of understanding and judgment in general. In BPSD, on the other hand, behavioral and psychological symptoms (peripheral symptoms) occur that are a complex combination of changes in mental status caused by core symptoms (strong anxiety and confusion, low self-esteem, etc.) and factors such as the surrounding environment, the

C. Stephanidis et al. (Eds.): HCII 2021, LNCS 13095, pp. 132–143, 2021.
https://doi.org/10.1007/978-3-030-90963-5_11

responses of others, and one's own experiences and personality. These symptoms include agitation, violence and verbal abuse, refusal of care, depression, anxiety, apathy, wandering, delusions, and hallucinations. These diverse symptoms develop and become severe as the condition progresses. Although dementia occurs primarily in the elderly, the symptoms show that it is not a normal phenomenon of aging. Dementia is a major cause of disability and dependency on care, and comprehensive measures are required, because it has physical, psychological, and socioeconomic effects for the person concerned, as well as the caregiver, the individual's family, and society. Because there is currently no cure for dementia, initiatives on the preventive side are particularly important.

Mild cognitive impairment (MCI) has received much attention in recent years [4, 5]. MCI is a condition that involves problems with some cognitive functions but does not interfere with daily life. There is thus concern over delayed detection and treatment. The percentage of people who progress to dementia from MCI is 10% per year [6], and it is estimated that about 40% of people will transition to dementia over 5 years. However, it has been suggested that early detection and treatment can potentially suppress the progression of MCI, with the possibility of recovery. Initiatives addressing lifestyle habits, such as diet and exercise, and social participation activities that are effective in maintaining and promoting cognitive function are expected to help [7, 8].

Decreased visual function is associated with symptoms of MCI. In an epidemiological study conducted in the United States on 635 older adults with cognitive impairment such as Alzheimer's disease, the patients were followed for more than 8 years in relation to visual function, and it was found that those in the favorable visual acuity group had a 63% lower risk of developing dementia [9]. In particular, visuospatial cognitive impairment is a common symptom observed in Alzheimer's dementia, and this shows that there is potentially a strong association between cognitive impairment and visual decline. It is therefore assumed that the promotion of strategic eye movement for visual function improvement is effective, because controlling decreases in visual function can potentially contribute to the prevention of dementia. Meanwhile, there remains the problem of whether the eye movement required by researchers for video observation is being carried out reliably. Visual confirmation of objects is broadly classified into the process carried out by the central visual field (hereafter, central vision) and the process carried out by the peripheral visual field (hereafter, peripheral vision), and eye movement also depends on the features of the visual confirmation method. Because the condition of eye movements changes easily depending on which visual confirmation method is used, it becomes important to monitor the state of eye movements in real time during observation. Therefore, in this study, we conducted a basic investigation involving discrimination of eye movement condition (peripheral and central vision) using deep learning technology.

2 Materials and Methods

2.1 Visual Stimulation and Eye Movements

Two videos were prepared for the experiment to efficiently elicit specific eye movements. Figure 1(a) shows a still image of the video for the central vision, and Fig. 1(b) shows a still image of the video for the peripheral vision. These videos were produced using computer graphics (CG) software. Multiple dots were arranged at random positions in the video, and movement of the entire image was realized by moving the CG camera in the space in a sinusoidal movement in a horizontal direction and a vertical direction at 0.25 Hz. As for the difference between the two videos, it was only the presence or absence of a gaze point. The yellow dot in (a) is the gaze point, which moved in the same way as the blue dot in the periphery. By continuing to follow this yellow dot, the gaze point was always captured in the central visual field. In contrast, in the video for peripheral vision, the subjects' gaze was not fixed, and the entire image would always be captured mainly in the peripheral visual field since subjects consciously tried to see the entire screen.

When visual confirmation of the entire screen of uniformly moving images is carried out with visual pursuit and central vision, unique eye movements can be observed. If the gaze point is continuously captured by the central vision, the line of sight is always fixed to the gaze point, and eyeball movement is dependent on the movement of the gaze point. In video (a), it is assumed that smooth pursuit eye movement of sinusoidal movement of 0.25 Hz is performed, because the gaze point is moving sinusoidally left/right and up/down at 0.25 Hz. In contrast, in video (b), given that visual stimulation has been input with the entire screen moving as one, it is assumed that there is slow eye movement (slow phase) and eye movement called optokinetic nystagmus (OKN), in which rapid eye movements (rapid phase) are repeated in reverse to the slow phase for resetting. Given that different eye movements are performed in central and peripheral vision, they can be distinguished by their characteristics. Meanwhile, in order to carry out discrimination of eye movements in real-time during video observation, which was the subject of our investigation, it is necessary to carry out discrimination from very short duration measurement results, and it can be said that evaluation using statistical attributes and analysis of frequency can be difficult. Consequently, in this study, we attempted a discrimination method that differed from conventional methods, using machine learning (deep learning technology).

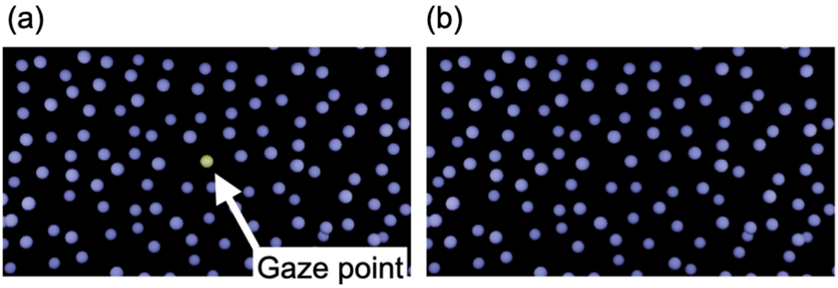

Fig. 1. Experiment video images: (a) Video image for central vision, (b) video image for peripheral vision2

2.2 Deep Learning

Deep learning is a machine learning technique that has attracted attention in recent years. Machine learning using an artificial neural network with multiple layers in the middle layer (Fig. 2) is called deep learning. In deep learning, there are learning methods that include supervised learning, unsupervised learning, and reinforcement learning, but in this study, we will only explain supervised learning, as only supervised learning was used in our study.

In supervised learning, a machine learns based on pre-specified correct answer labels, thereby optimizing the neural network model, and a model is constructed in which predictions are carried out as response values for a dataset. Conventional single-layer neural network models involve entries being made, manual extraction of attributes from certain data, and the use of attributes and data sets with correct answer labels. In contrast, in deep learning, the extraction of attributes is also carried out in the neural network, so manual extraction of attributes becomes unnecessary. In this study, eye movement angle data and viewing methods (central vision, peripheral vision) as correct answer labels were used as datasets.

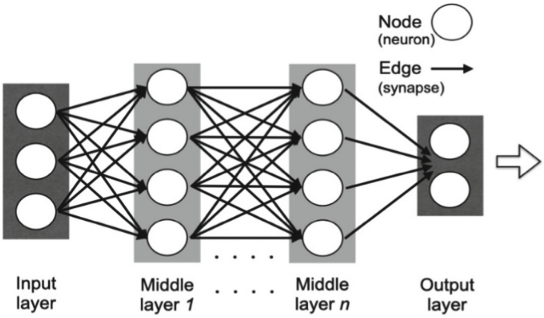

Fig. 2. Schematic representation of a deep neural network

2.3 Procedure and Design

Six healthy 21-to-23-year-old males (with vision correction used if necessary) were included in the study. Figure 3 shows a subject viewing the video. Subjects viewed peripheral and central vision movies on a 42-in. monitor with their chin on a chin rest 1 m away from the monitor, under an environment of approximately 10lx of ambient illuminance. After prior practice using central vision and peripheral vision with the experiment videos, a total of eight three-minute central vision videos and peripheral vision videos (four each) were observed in random order with small breaks in between. To record eye movements during video viewing, an eye movement measurement device based on the scleral reflection method (Manufactured by Takei Scientific Instruments Co., Ltd.) was attached in front of the subject's right eye. Eye movements (viewing angles) during each video viewing were continuously recorded at 500 Hz for the horizontal/vertical and left/right directions.

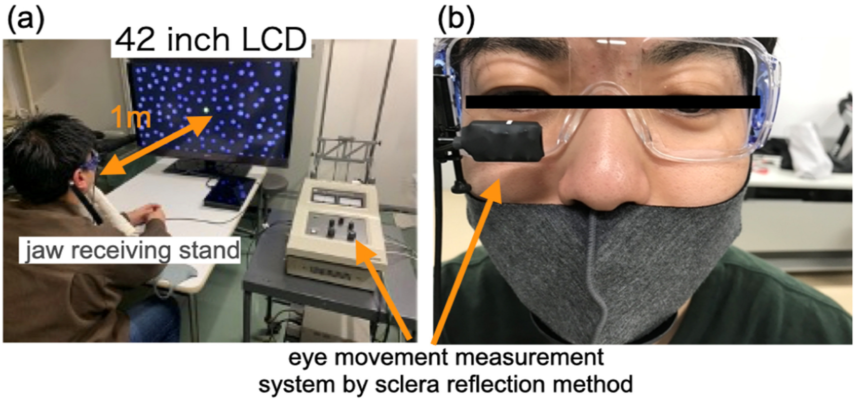

Fig. 3. Experiment set-up and eye movement measurement4

To verify the classification accuracy depending on different data lengths, we acquired short-duration series data in sets of 350 from the 3-min time series data, with data lengths fixed at 4 types (0.5 s, 1 s, 2 s, and 3 s) and shift lengths fixed at 0.5 s from the measured 3-min time series data of eye movement. In other words, 1,400 data (equal numbers for both peripheral and central vision) were acquired for each data length per person.

The Neural Network Console Ver. 1.57 (manufactured by SONY) was used for the deep learning in this study. The deep neural network utilized in this experiment is shown in Fig. 4, with a model using data length of 3 s. In the input layer, nodes are placed according to the amount of time series data (1,500 for 3 s). After sequence transformation using Reshape as pre-processing, processing is taken over by the middle layers. The middle layer consists of a total of 32 layers of 8 units, consisting of the convolution layer, the normalization layer, the activation function, and the max-pooling layer. Details of each layer are described below.

- Convolution layer: The convolution layer performs convolutional operations on the input time series data and extracts attributes. In our study, 64 processing results were obtained, because the size of the convolution kernel was set as 9, and a filter with 64 different parameters was applied to the input.
- Batch normalization layer: Batch normalization is a method used to make artificial neural networks faster and more stable through normalization of the input layer. Specifically, it is implemented to prevent vanishing gradient problems (where the gradient of the error function becomes 0) and exploding gradient problems (where the gradient diffuses).
- ReLU (Rectified Linear Unit [activation function]): The activation function transforms the summation of input signals into output signals and has data available to facilitate attribute learning. It is responsible for determining how the summation of input signals is output (fired) and for neuronal firing. ReLU is expressed in Eq. (1), where x is the input and y is the output.

$$y = \begin{cases} x(x \geq 0) \\ 0(x < 0) \end{cases} \tag{1}$$

- MaxPooling (Pooling Layer): This is often applied after the convolution layer and performs information compression using down sampling to deform the data into an easily-handled shape. In addition, down-sampling also helps to reduce the location dependence of attributes. MaxPooling outputs the largest data in the kernel-size area. In this study, the kernel size was set to 2, and each time pooling was carried out, the data were halved.

These four layers make up one unit with seven to eight units as middle layers. The output layer is composed of a fully-connected layer, a sigmoid function as an activation function, and a multi-class cross-entropy. In the learning process, learning was performed with a batch size of 64 and an epoch number of 100.

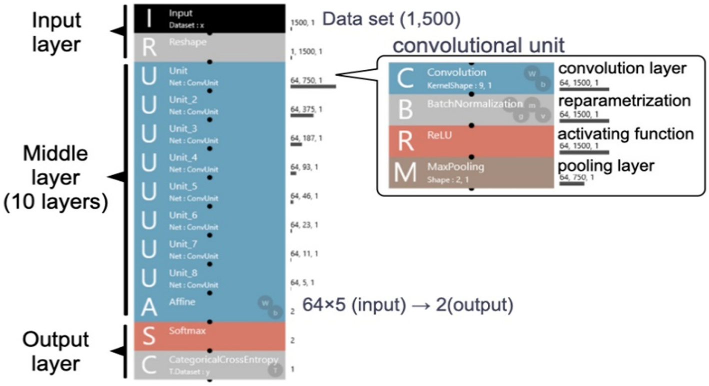

Fig. 4. Details of the deep neural network model used in this study

For learning and evaluation, the Neural Network Console was installed on a desktop PC (Windows 10 Professional, CPU: CORE i7-7700 3.6 GHz, Main memory: 16 GB, GPU: GeForce RTX 2070 SUPER 8 GB [NVIDIA]) and the subject-specific leave one out method was applied (Fig. 5). The leave one out method uses the datasets for all subjects for both learning and assessment. Five of the six subjects were utilized for learning, and the remaining one was utilized for evaluation. This was carried out sequentially on a subject-by-subject basis to obtain assessment results for six people. After that, final evaluation results were obtained by combining these results, to take into account the effect of individual differences.

2.4 Evaluation

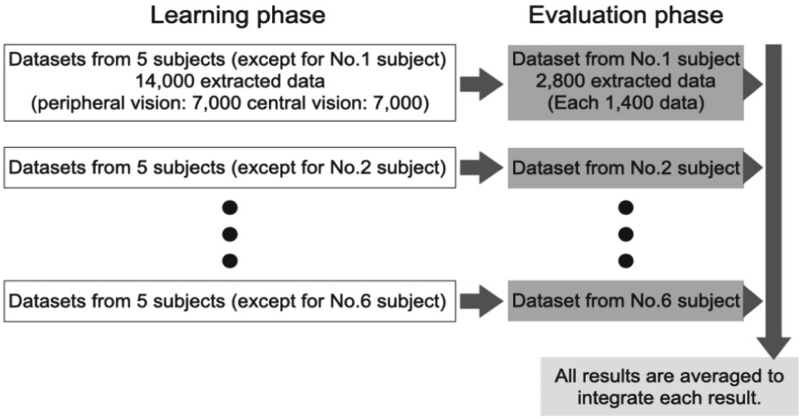
Learning phase
Evaluation phase
Datasets from 5 subjects (except for No.1 subject)
14,000 extracted data
(peripheral vision: 7,000 central vision: 7,000)
Dataset from No.1 subject
2,800 extracted data
(Each 1,400 data)
Datasets from 5 subjects (except for No.2 subject)
Dataset from No.2 subject
Datasets from 5 subjects (except for No.6 subject)
Dataset from No.6 subject
All results are averaged to
integrate each result.

Fig. 5. Summary of the leave one out method

Several methods have been proposed for evaluating the accuracy of classification. In this study, a typical classification was calculated and evaluated. When positive was set as peripheral vision and negative was set as central vision, the evaluation could be expressed in a confusion matrix, as shown in Table 1.

Table 1. Confusion matrix 7

		Prediction of DL	
		Positive (peripheral vision)	Negative (central vision)
Actual classification	Positive (peripheral vision)	TP (True Positive)	FN (False Negative)
	Negative (central vision)	FP (False Positive)	TN (True Negative)

- Accuracy

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + FN + TN} \tag{2}$$

Accuracy shows the percentage of correctly predicted data. Unlike sensitivity and specificity, it is not restricted to positive and negative data but is the correct response rate calculated for all data and is the most fundamental measure.

- Precision

$$Percision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

Precision is an index to be tested in respect of positives. Because formula (3) does not include FN, none of the missed values are considered, and if all positives are True (i.e., FP is 0), Precision is 100%. In this study, peripheral vision was set as positive and central vision as negative, and Precision represented the correct answer rate for the classification of peripheral vision.

- Recall

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4}$$

Recall is an index to be tested in respect of negatives. Since formula (4) does not include FP, none of the misdetected values are considered, and Recall is 100% if FN is 0. Recall represents the correct answer rate for classification of central vision.

- F-measure

$$F - measure = \frac{2Recall \cdot Precision}{Recall + \mathrm{Precision}} \tag{5}$$

This is the harmonic mean of Precision and Recall. It is often used as an overall evaluation because it is an indicator that balances the characteristics of both contrasting aspects.

3 Results

Firstly, the results of the measurements of eye movements obtained from the experiment are shown in Fig. 6. Figure 6(a) shows the results of central vision (pursuit), showing both the periodic eye movements that occur from the pursuit of the periodic movements of the gaze points and the spiking changes that result from blinking. In contrast, Fig. 6(b) shows the results of peripheral vision, which, unlike Fig. 6(a), shows that minute eye movements caused by OKN are always occurring during video observation. When the results of both measurements are compared, it is possible to discriminate between them easily, because the attributes of the waveforms differ greatly. However, when one part of a measurement result is observed limited to a short time (Fig. 7), the difference in the shape of the waveform can be obvious (Fig. 7(a), (b)), or the forms can be similar and it becomes difficult to discriminate only by visual features (Fig. 7(c)). Therefore, discrimination carried out with deep learning including attribute searches of the data is assumed to contribute to the improvement of the accuracy of the discrimination.

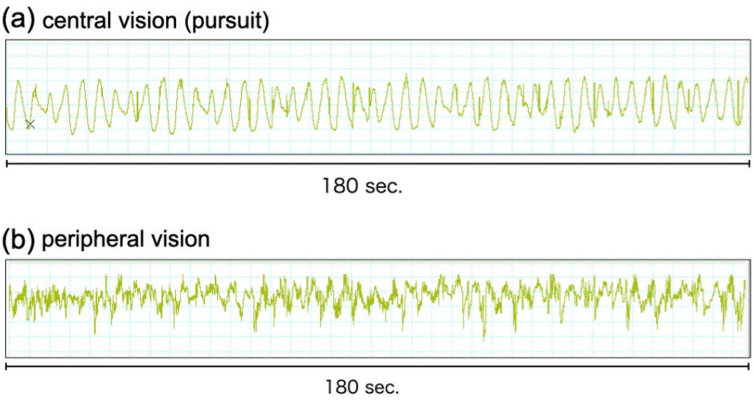

Fig. 6. An example of eye movement measurement results: (a) Central vision, (b) peripheral vision8

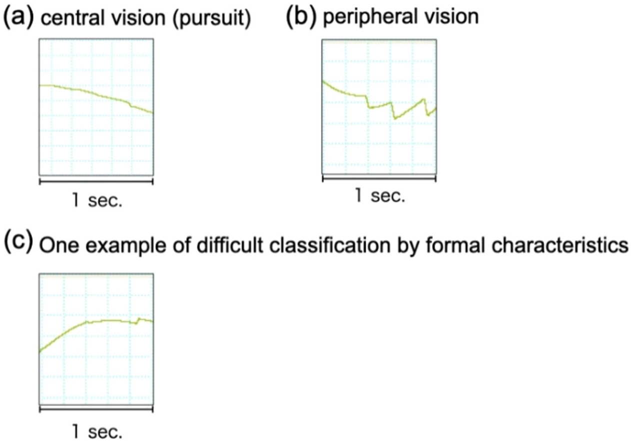

Fig. 7. Short-duration waveform of eye movement measurement: (a) Central vision, (b) peripheral vision, and (c) one example of difficult discrimination from waveform characteristics

Next, Fig. 8 shows the evaluation results of the accuracy of discrimination. Accuracy, Precision, Recall, and F-measure are shown in Figs. 8 (a)–(d), respectively, and all results showed improved accuracy with increasing data length. In particular, accuracy of at least 90% was able to be maintained when the data length was at least one second.

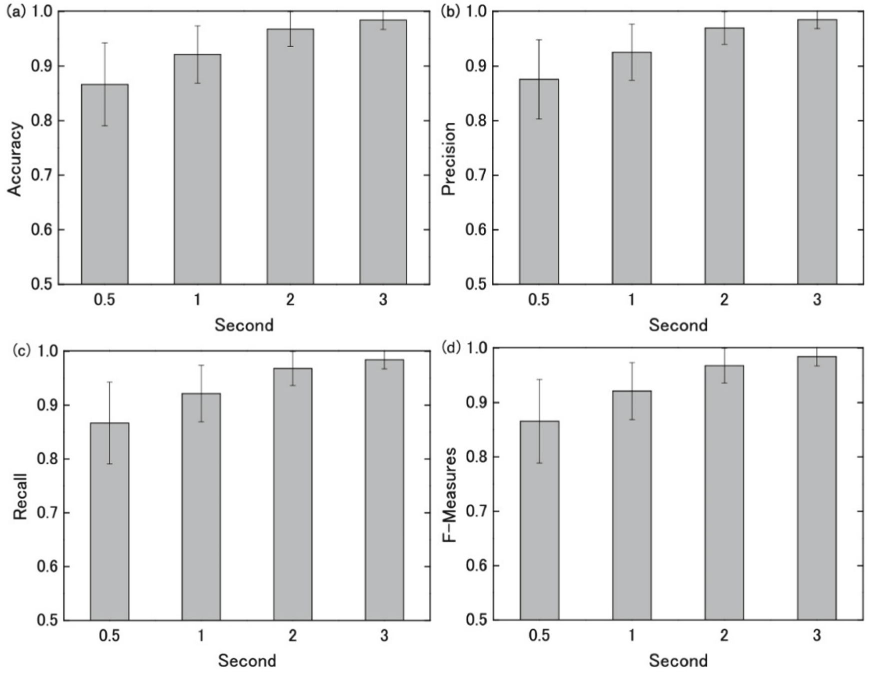

Fig. 8. Evaluation results: (a) Accuracy, (b) Precision, (c) Recall, (d) F-Measure9

4 Discussion

The results on accurate discrimination of eye movement condition showed a discrimination accuracy of 90% or higher if there was a data length of about one second (Fig. 8). This suggests that discrimination of condition can be made in real time with 90% accuracy, if a delay of about one second is accepted. The deep learning model used in this study is a Deep Convolutional Neural Network (DCNN) that has been commonly utilized in classification tasks in recent years. This network model is often applied to images and, in addition, it has been demonstrated that it exerts high accuracy for discrimination [10, 11]. If the convolutional layer of this model is reduced to one dimension, it is possible to apply the same model to one-dimensional time series data such as eye movement and heart rate variability, and as a result, high accuracy of discrimination is expected. Moreover, given that DCNN is a versatile model, it is expected to be applied in many fields, and this study will also potentially provide an example.

Both the quality and quantity of data prepared for deep learning are important and are strongly connected to good and poor evaluation accuracy. With regard to the quality of the data in this study, it is worth touching on the blinking that occurred during video observation. The subjects were instructed in advance to refrain from blinking as much as possible during the video observation. The reason for this is that in eye movement measurement, blinking in both central and peripheral vision can be the cause of decreased accuracy of discrimination because it manifests as disturbances shown as spiky waveform changes. However, because it is impossible to stop blinking completely, the accuracy of discrimination is considered to have decreased, especially when the data length was short, because the proportion of the waveform changes occupied by blinking increased with respect to the extracted waveforms. The effect of blinking on eye movement measurement is difficult to remove completely during the measurement phase. For that reason, as one method for accuracy improvement, enabling separate deep learning for the features of blinking is considered as an option to avoid the effect of blinking.

This time, verification was carried out for young subjects, and it is necessary to verify whether the system is also effective for other age groups, because visual function changes according to age [12]. In particular, the elderly experience deterioration of the eye structure including the cornea and lens, as well as reduced function originating in changes in the eyeball and ocular muscles; thus, verifying whether condition discrimination and maintaining the accuracy of discrimination is possible in subjects with clearly reduced visual function seems to be a necessary investigation in order to improve the versatility of the real-time eye movement estimation system.

5 Conclusion

In this study, an investigation was carried out on the discrimination of eye movement from measurement results of short duration using a deep learning model for the purpose of discriminating eye movement condition in real time during video observation. As a result, it was possible to carry out discrimination with accuracy of at least 90% if measurement data of at least one second were measured.

References

1. World Health Organization (WHO): Dementia Homepage. http://www.who.int/news-room/fact-sheets/detail/dementia. Accessed 1 June 2021
2. Cerejeira, J., Lagarto, L., Mukaetova-Ladinska, E.B.: Behavioral and psychological symptoms of dementia. Front. Neurol. **3**, 73 (2012)
3. Seitz, D., Purandare, N., Conn, D.: Prevalence of psychiatric disorders among older adults in long-term care homes: a systematic review. Int. Psychogeriatr. **22**, 1025–1039 (2010)
4. Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E.: Mild cognitive impairment: clinical characterization and outcome. Arch. Neurol. **56**, 303–308 (1999)
5. Petersen, R.C., et al.: Current concepts in mild cognitive impairment. Arch. Neurol. **58**, 1985–1992 (2001)
6. Bruscoli, M., Lovestone, S.: Is MCI really just early dementia? A systematic review of conversion studies. Int. Psychogeriatr. **16**, 129–140 (2004)
7. Kim, K.Y., Yun, J.-M.: Association between diets and mild cognitive impairment in adults aged 50 years or older. Nutr. Res. Pract. **12**, 415–425 (2018)

8. Chandler, M.J., et al.: Comparative effectiveness of behavioral interventions on quality of life for older adults with mild cognitive impairment: a randomized clinical trial. JAMA Netw. Open. **2**, e193016 (2019)
9. Rogers, M.A.M., Langa, K.M.: Untreated poor vision: a contributing factor to late-life dementia. Am. J. Epidemiol. **171**, 728–735 (2010). https://doi.org/10.1093/aje/kwp453
10. Onishi, Y., et al.: Automated pulmonary nodule classification in computed tomography images using a deep convolutional neural network trained by generative adversarial networks. Biomed Res. Int. **2019**, 6051939 (2019)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
12. Meng, Q., et al.: Age-related changes in local and global visual perception. J. Vis. **19**, 10 (2019)