Neurocomputing 444 (2021) 390–398

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep-learning-based reading eye-movement analysis for aiding biometric recognition

Xiaoming Wang, Xinbo Zhao *, Yanning Zhang

National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, PR China

ARTICLE INFO

Article history:
Received 28 September 2019
Revised 14 May 2020
Accepted 21 June 2020
Available online 24 November 2020

Keywords:
Eye tracking
Eye-movement model
Deep-learning
Biometrics
Identity authentication
Reading eye-movement

ABSTRACT

Eye-movement recognition is a new type of biometric recognition technology. Without considering the characteristics of the stimuli, the existing eye-movement recognition technology is based on eye-movement trajectory similarity measurements and uses more eye-movement features. Related studies on reading psychology have shown that when reading text, human eye-movements are different between individuals yet stable for a given individual. This paper proposes a type of technology for aiding biometric recognition based on reading eye-movement. By introducing a deep-learning framework, a computational model for reading eye-movement recognition (REMR) was constructed. The model takes the text, fixation, and text-based linguistic feature sequences as inputs and identifies a human subject by measuring the similarity distance between the predicted fixation sequence and the actual one (to be identified). The experimental results show that the fixation sequence similarity recognition algorithm obtained an equal error rate of 19.4% on the test set, and the model obtained an 86.5% Rank-1 recognition rate on the test set.

1. Introduction

Biometric recognition technology is widely used in e-commerce, electronic products, and network security. Reading eye-movement is a human behavior with the biometric characteristics of convenience, security, universality and collectability [1]. Thus, users can be identified by comparing human eye-movement trajectories.

Based on measurements of eye-movement trajectory similarity, the existing eye-movement recognition technology extracts the measurable features of the eye-movement trajectory, including the fixation duration and the lengths of the saccades. Other technologies compare the eye-movement trajectories using more complicated space and time information of the eye-movement. The existing technologies are usually used to obtain the similarity measurement value (or values) of eye-movement trajectories. However, the characteristics of the stimuli are not considered.

Related studies on reading psychology have shown that human eye-movements during reading are significantly different between individuals [2–4], but the same individual exhibits a certain similarity. This suggests that human eye-movement is unique and stable to some extent and can be used in biometrics [5,6]. Fig. 1 shows the fixation sequences as ten subjects read the same text.

In this study, multiple-input deep neural networks were utilized to learn the reading eye-movement behaviors and construct a computational model for reading eye-movement recognition (REMR). The model can learn the features of the stimuli (reading materials) and the eye-movement trajectory. The model can fully simulate human eye-movement after training and can be applied for user identification by comparing the predicted and actual (to be identified) fixation sequences. Combined with other biometric recognition technologies, this technology can be used as a supplementary tool for the existing identity authentication methods to realize multi-factor identity authentication.

The main contributions of the present paper are as follows.

1. A biometric recognition technology is proposed based on reading eye-movement. Accounting for the stimuli (reading materials) and scanning path, this technology uses fewer handcrafted features to obtain effective recognition by utilizing the deep-learning characteristics of automatic feature extraction. As a result, the model obtained an 86.5% Rank-1 recognition accuracy on the test set.

* Corresponding author.
E-mail addresses: wxmgo@163.com, xmwang@mail.nwpu.edu.cn (X. Wang), xbozhao@nwpu.edu.cn (X. Zhao), ynzhang@nwpu.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.neucom.2020.06.137
0925-2312/

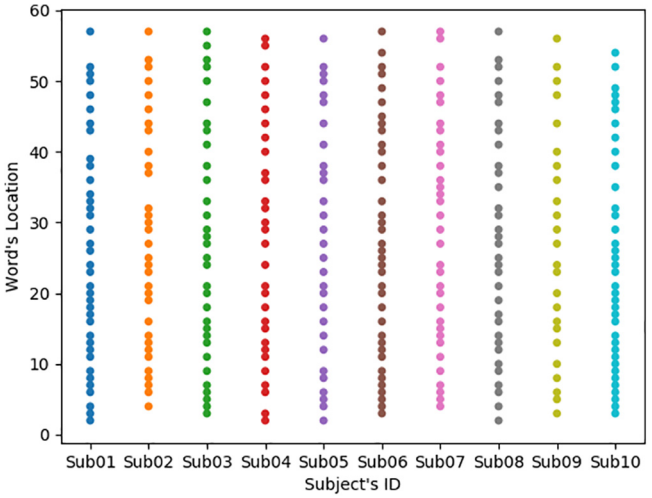

Fig. 1. Fixation sequences as ten subjects read the same text. The horizontal axis represents the subject's ID, and the vertical axis indicates the word's location labels in the text (60 words total). The dots in the figure represent the subjects' fixation location distribution while reading. The fixation sequences are significantly different when different subjects read the same paragraph.

2. A REMR computational model based on deep-learning is proposed. This model uses a deep neutral network to generate the predicted fixation sequence and measure the similarity distance between the predicted and actual fixation sequences to identify the subject.
3. An algorithm is presented for evaluating the fixation sequence similarity. The algorithm uses dynamic time warping (DTW) to measure the similarity between two fixation sequences, and the proposed algorithm obtained an equal error rate (EER) of 19.4% on the test set.

2. Related work

The biometric study of eye-movement stems from the early study of scanpath theory, in which the word "scanpath" refers to the space path formed by an orderly fixation and saccadic sequence. In 1971, Noton and Stark [7] found that the general scanpath followed by a subject during the first viewing of a pattern was repeated in the initial eye-movements of roughly 65% of subsequent viewings, and the scanpath for specific stimuli varied from person to person.

The 2004 paper by Kasprowski and Ober [8,9] was, as far as we know, the first study that applied eye-movement to biometric research. Referring to a method commonly used in the voice recognition, they conducted an eye-movement biometric recognition test and obtained an average false positive rate (FPR) of 1% and false negative rate (FNR) of 23% based on a dataset of nine subjects. In 2011, Holland and Komogortsev [10] started to study complex eye-movement patterns (CEM-P) and made use of averages and aggregate features, including fixation counts, the average fixation duration, the average vector saccade range, the average horizontal saccade range, the average vertical saccade range, the average vector saccade speed, the average vector peak saccade velocity, the velocity waveform indication, the scanpath length, the convex hull area of the scanpath, the region of interest, inflection point counts, the coefficient of the relationship between the range and time of duration, and the coefficient of the relationship between the range and peak velocity. They accomplished this using a Gaussian kernel and linear combination. During testing, they obtained an EER of 27% from 32 subjects. In 2013, Holland and Komogortsev [11] achieved an EER of 16.5% using modified CEM-P technology. This is the lowest EER reported thus far. In 2012, Rigas et al. [12] applied graphics-based matching technology (similar to face recognition) to comparisons of eye-movement location labels and compared the minimum spanning trees using the multivariate Wald–Wolfowitz runs test. In this method, an EER of 30% was achieved on a dataset of 15 subjects. In 2015, Cantoni et al. [13] proposed a type of gaze analysis technique (GANT), in which the eye-movement model was constructed using the fixation time and regression counts on the different gaze points of different subjects, and the similarity between the two records was found using the Frobenius norm of the density map. The study of computational eye-movement models has been mainly based on psychology methods [14–16], traditional machine learning methods [17–21], and neural network learning methods [22–25].

Eye-movement biometric technology has been developed for more than ten years and is still in an early and exploratory stage [13]. All the above eye-movement biometric technologies contrast two given scanpaths by extracting the easily measurable eye-movement space and time information without taking the stimuli into account. Taking advantage of deep-learning techniques that can extract data features automatically [26], a type of biometric recognition technology based on reading eye-movement is proposed in this paper that accounts for the stimuli (reading materials) and scanning path to obtain effective identification results while requiring fewer eye-movement features.

3. Method

3.1. Problem setting

Experimental results of eye-movement and reading have indicated that eye-movement during reading is goal-oriented and discrete [4]. This means that the saccade is non-random in selecting a visual target and that the saccade target points to a particular word rather than a specific distance. Based on this notion, there are many candidate words in the latent saccade period, and each word has a chance to be selected as the target of subsequent saccades. In this work, the probability of each word in the text being fixated on is set by the eye-movement model without any constraints. The text and fixation sequence are elaborated on below.

R represents a set of readers, and a single reader is denoted as $r \in R$. T represents text whose word order is $(w_1, \ldots, w_n)$. For each $r \in R$, we generate a fixation point sequence F_{RAW} based on each word in T and assume that F_{RAW} obeys the following relation:

$$F_{\mathrm{RAW}} \sim p(F_{\mathrm{RAW}}|T, r), \tag{1}$$

where $p(F_{RAW}|T, r)$ is the distribution of eye-movement patterns when a particular subject reads a piece of text. For example, the text "Kate quivered and went to the window" is expressed as T = (Kate, quivered, and, went, to, the, window). A possible fixation sequence F_{RAW} is recorded as (Kate, quivered, and, Kate, quivered, went, the, window), and the corresponding location sequence is (1, 2, 3, 1, 2, 4, 6, 7). The area is divided by the rectangular region in which the word is located, and the set of region locations corresponding to F_{RAW} is {1, 2, 3, 4, 6, 7}. For all the regions, if there is a fixation on the region, the region is marked as 1; otherwise it is marked as 0. The area tag sequence is represented by IA, and the corresponding IA is (1,1,1,1,0,1,1). The number of elements in IA is the same as that in T. Using the method described above, we refined our target by excluding regressions from the fixation data.

In the case of identity recognition based on the reading eye-movement model, it is assumed that M is a computational eye-movement model, and the given text sequence is T. When a reader r reads text T, the sensor acquires the fixation sequence F_{RAW} of the eye-movement. The following formula is deduced:

$$r = \arg\max_{r \in R} p(R|M, T, F_{RAW}). \quad (2)$$

To facilitate the data processing of the model, the *IA* sequence is used instead of the F_{RAW} sequence, and a linguistic feature sequence $A_1 \ldots A_n$ (n ∈ N) corresponding to T is introduced. The length of A_n is the same as that of T. Therefore, the following formula is deduced:

$$r = \arg\max_{r \in R} p(R|M, T, IA, A_1 \ldots A_n)(n \in N). \quad (3)$$

A reading eye-movement recognition computational model based on deep-learning was constructed to solve the above problem. The model takes advantage of deep-learning techniques that extract the features automatically.

3.2. Framework of REMR computational model based on deep-learning

The computational model takes the text sequence, text-based linguistic feature sequence, and actual fixation sequence as inputs, and it outputs the results by measuring the similarity distance between the predicted fixation sequence and the actual one. The framework of the model is presented in the Fig. 2.

The text sequence output processing goes through three layers. The first layer is the word embedding layer, in which the model transforms the word into the sparse high-dimensional one-hot vector and maps the one-hot vector into a dense low-dimensional word vector. The second layer is a bidirectional long short-term memory (LSTM) layer. It saves the previous information in the text sequence for later use. The third layer is a dense layer,

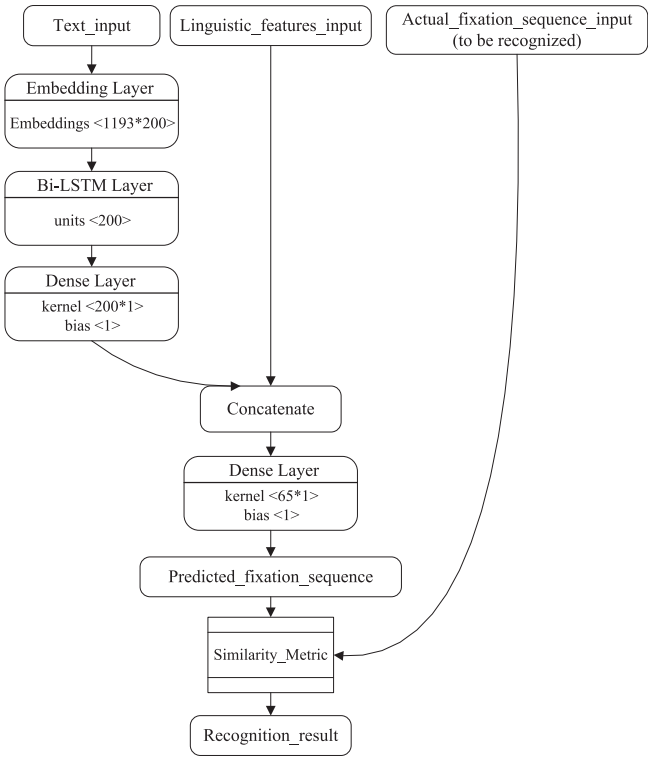

Fig. 2. Framework of the reading eye-movement recognition (REMR) computational model based on deep-learning. The model first processes the input text through word embedding, the bidirectional LSTM, and dense layers, after which it outputs a vector and merges the vector with the text-based linguistic feature sequence (digitized linguistic features with the same length as the input text) to learn. Through the above procedures, the model generates the predicted fixation sequence. Finally, the model outputs the results by comparing the similarity distance between the predicted reading eye-movement fixation sequence and the actual one (to be identified). The actual fixation sequence consists of zeros and ones with the same length as the input text. The output of the model is the subject's ID in the dataset.

which serves as the output layer of the text sequence processing. The model merges the generated vector with the text-based linguistic feature sequence vector after completing the input text processing and feed to a dense layer to generate the predicted fixation sequence. At this point, the task of predicting the fixation sequence is completed, and the similarity distance between the predicted and actual fixation sequences is measured for identity recognition.

The framework of the reading eye-movement recognition computational model based on deep-learning is described in detail later. The neural network components (layers) and the similarity measurement for the fixation sequence will also be discussed.

3.3. Word embedding layer

The word embedding layer implements the process of text vectorization. The input is a sequence of numbers, and the output is a list of vectors. The neural network model cannot receive raw text as input, and it can only process numeric tensors. It is therefore necessary to digitize and vectorize the text. The processing steps are as follows:

(1) Read the text data.
(2) Build the token.
- The size of the dictionary is set to be greater than or equal to the number of unique words in the corpus.
- Sort all the words that appear in the corpus based on the number of occurrences.
- The established dictionary is of the form {'the': 1, 'to' 2, 'a': 3, 'and': 4, 'of': 5, ' ': 6, 'in': 7, 'that': 8, 'was': 9, 'is': 10, ...}.
- The dictionary is used for the vocabulary-to-digital conversion. For example, "the" is converted to 1, and "to" is converted to 2.

(3) Use the token to convert the "text sequence" to the "digital sequence".
(4) Left-pad all the "digital sequence" data, making it 60 digits long. Since a "digital sequence" is subsequently converted to a "vector sequence" and sent to the deep-learning model for training, its length must be fixed. In the experiment, the lengths of the text sequence, fixation sequence, and text-based linguistic feature sequence are padded to 60. The text and the text-based linguistic feature sequences are left-padded with the number 0, and the fixation sequence is left-padded with the number 1.
(5) Use the embedding layer to convert "digital sequence" to "vector list".

In the steps described above, the vocabulary is converted to a number, but the numbers are semantically unrelated. To make the vocabulary relevant, the words are mapped into a vector of multidimensional space. The semantically similar vocabulary vectors are close in distance in the multidimensional geometric space. Finally, the "vector list" could be sent to the deep-learning model for training.

Along with other parameters, the embedding vector is trained as a parameter of the network. Dyer et al. [27] proved that word embedding plays a crucial role in improving the performance of the sequence label.

3.4. Long short-term memory network layer

The LSTM layer is used to extract high-level linguistic information between words in a sentence. It accepts the vector list from the previous layer as an input and feeds the processed vector to the output layer. A LSTM network was designed to address the gradient disappearance and is the first structure to introduce the gate

mechanism [28]. The LSTM structure decomposes the state vector into two parts: one part is the "memory unit" and the other part is called the "running memory." The memory unit is designed to hold memory and gradient information over time while being controlled by the smoothing function of the micro-gate simulation, which is known as the analog logic gate [29]. This gate mechanism enables the gradient associated with memory to remain high for an extended time span [30]. The algorithm proposed in this paper learns the representations of the text sequence, processed by the embedding layer, in both directions by making using of bidirectional long short-term memory (bi-LSTM). It is composed of forward and backward LSTMs and is commonly used for modeling context in natural language processing [31].

3.5. Merged module layer

To introduce domain expert knowledge in the fixation sequence prediction task, in addition to inputting the text sequence, we also hope to be able to input some linguistic features (such as the word length and frequency) that have an effect on reading eye-movement fixation. Thus, the fixation sequence prediction tasks presented in this paper require a multi-mode input, and the input must include at least one text sequence and a text-based linguistic feature sequence. One simple approach involves these two kinds of data being used simultaneously to train two independent recurrent neural network models, after which a weighted average of the two predictions is calculated. However, this method is not optimal because it cannot extract the correlated information of the two sequences, and redundancy may exist in the extracted information. A better approach is to use a model that can examine all available inputs simultaneously and jointly learn a more accurate data model [32]—a model with multiple input branches (see Fig. 3).

For convenience, we selected only one linguistic feature. In Section 4.2, we will discuss which effective linguistic feature to choose. The merged input model uses the text sequence $t_{1:n}$ and the text-based linguistic feature sequence $a_{1:n}$ as inputs, converts them into vector representations, and splices them to obtain the corresponding input representation $x_{1:n}$:

$$x_i = [t_i; a_i] \tag{4}$$

The merged module layer feeds the merged vector to the next dense layer and outputs the predicted fixation sequence composed of zeros and ones, where 1 indicates that the word at this position is fixated and 0 indicates that the word at this position is skipped.

3.6. Method for fixation sequence similarity measurement

Based on dynamic time warping (DTW) [33], the similarity measure of two fixation sequences is determined. First, a distance matrix $dp[i][j]$ containing the distances between the fixation sequences a and b is constructed based on dynamic programming; $dp[i][j]$ refers to the square of the similarity distance between $a[0{:}i]$ and $b[0{:}j]$, as follows:

$$dp[i][j] = \begin{cases} (a[0]-b[0])^2 & i=0,\ j=0 \\ (a[0]-b[j])^2 + dp[0][j-1] & i=0 \\ (a[i]-b[0])^2 + dp[i-1][0] & j=0 \\ (a[i]-b[j])^2 + \min(dp[i-1][j],\ \ dp[j-1][i],\ \ dp[i-1][j-1]) & i,j>0 \end{cases} \tag{5}$$

In this formula, $dp[\text{len}(a)-1][\text{len}(b)-1]$ is the square of the similarity distance, and the extraction of $dp[\text{len}(a)-1][\text{len}(b)-1]$ is the similarity distance between two fixation sequences. The optimal path of the distance matrix $dp[i][j]$ from the upper-left corner to the lower-right corner is the required similarity distance. However, the algorithm does not calculate the distances of the periodic sequences very well, especially when two sequences have similar periods, but one of the sequences is obtained by the translation of the other sequence. The eye-movement sequences of a subject reading the same text at different times may be such a sequence with a similar cycle [3]. Thus, we obtain a penalty coefficient α and multiply this by the original distance to get the updated distance. The procedure is as follows.

1. The first step is to determine the longest common substring of the two fixation sequences *seq*1 and *seq*2, and the length is recorded as *a*.
2. Since *seq1* and *seq2* are numerical sequences, the standard deviations of *seq1* and *seq2* are first obtained when finding the longest common substring, and the larger one is set as the maximum standard deviation (STD). For any i, j, we use the formula $|seq1[i] - seq2[j]| <$ STD instead of the formula $seq1[i] = seq2[j]$ to indicate that they are part of the common substring.
3. Finally, the penalty coefficient is obtained: $\alpha = 1 - \frac{a \times a}{len(seq1) \times len(seq2)}$. Thus, the longer the longest common substring of two numerical sequences is, the smaller the penalty coefficient is. In this way, if α is multiplied by the distance of the original algorithm, the updated distance will be smaller.

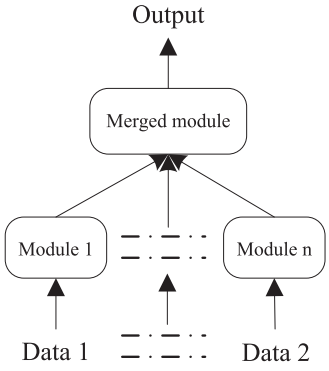

Fig. 3. Multi-input neural network model.

4. Experiments

4.1. Experimental environment and dataset

The experimental environment was Python3.7 + Keras2.2.4 + TensorFlow1.13. All the code was released on GitHub at https://github.com/wxmgo/eye_movement_in_reading/. In addition to being interesting to readers in the fields of biometrics and machine learning, the public posting of the code on Github will allow others to easily utilize or modify this method for their own problems.

We obtained the dataset from the Provo Corpus [34]. The corpus is open and can be downloaded from the Open Science Framework (https://osf.io/sjefs). This eye tracking corpus contains the eye-movement data of 84 native English speakers who have read 55 text passages, including online news articles, popular science magazines, and public domain fiction. These text passages had an average length of 50 words (39–62 words). None of the participants had participated in the previous experiment and none had read the material. Eye-movements were recorded at a frequency of 1000 Hz with an SR Research EyeLink 1000 Plus eye tracker (with a spatial resolution of 0.01°) (further details are reported elsewhere [34]).

Generally, function words (words that have little lexical meaning or have ambiguous meanings and express grammatical relationships between other words within a sentence, e.g., *in*, *why*, and *then*) are more easily skipped by readers than content words (words that name objects of reality and their qualities, e.g., *dog*, *snow*, and *young*), because function words are generally shorter in length and have no practical meaning. However, in the Provo corpus, 16% (4605/27430 = 16.78%) of the content words are also skipped, because they have shorter word lengths or a higher frequency. Word length and frequency are important linguistic factors that affect reading saccades [3].

The experiment used data from all the subjects (the 1st–84th subjects), where each subject read 55 text passages. To make the test results more objective, it was necessary to isolate the training and test sets. These two corpora did not overlap.

4.2. Linguistic feature extraction

Evidence from psychology studies suggests that fixation and saccade patterns are driven by high and low information at the visual levels [35]. The word length, frequency, and predictability are known to be the linguistic features that affect reading saccades [36]. Among these, the word length is low-level visual information, and the frequency and predictability are high-level language information [37]. Related studies have also found that the effect of high-level language information on reading saccades is much smaller than that of low-level visual information [38]. For example, shorter words in the text are more likely to be skipped than longer words. To analyze this phenomenon quantitatively, Fig. 4 shows the relationship between the word length and skip frequency in the Provo Corpus, and the word length and skip frequency were normally distributed. In addition, previous studies [23–25] have also reached the same conclusion, i.e., the influence of the word length on the reading eye-movement is most significant. Therefore, the linguistic features of the data used in the experiment contained only one type of low-level visual information—the word length.

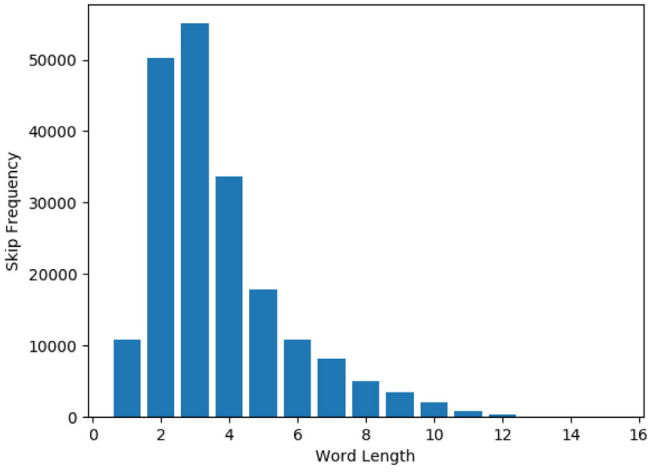

Fig. 4. Relationship between the word length and skipping frequency.

4.3. Data preprocessing

4.3.1. Serialization of data

The data in the corpus were vocabulary-based and did not reflect the linguistic relations between the vocabulary. Therefore, serialization processing is required to merge vocabulary-based data into sequence-based data. The first five rows of data after the conversion are shown in Table 1.

The IA column in the table is a sequence of zeros and ones, where 0 indicates that the area where the word was located was skipped over by the eye, and 1 indicates that fixation occurred in the area where the word was located.

To facilitate the one-hot processing of the data label, all the Subject_IDs in the table were decremented by 1 to ensure that Subject_ID began with the integer 0. The 0 in the table corresponds to the subject numbered 1.

4.3.2. Vocabulary-to-vector conversion

Since only numbers are accepted by the deep-learning model, the model must convert the vocabulary sequence into a digital sequence first. The experiment uses the *Tokenizer* module provided by *Keras* [39]. The digital sequence is then converted into a vector sequence in the process of network training.

4.3.3. One-hot encoding of subject's ID

To not imply big or small sizes of the subject's ID (0–83) during the network training process, the subject's ID must be one-hot encoded to be represented in a discrete form. One-hot encoding of the subject's ID was performed by using the *to_categorical* method provided by *keras.np_utils* [39]. The converted vector had a total of 84 numbers, one of which was 1, and the rest were 0. The one-hot coding of the training data label is shown in Fig. 5. There was a total of 550 lines with ten numbers in each line. Only the first and last three rows are listed in the figure.

4.4. Network optimization algorithm

In this work, we used the root mean square propagation (RMSProp) algorithm [40] to train the neural network model. This optimizer is a top choice for training recurrent neural networks [41]. RMSProp implicitly applies simulated annealing. In moving toward the minimum, RMSProp automatically reduces the learning step so as not to skip over the minimum. The loss function uses a categorical cross-entropy function.

4.5. Similarity distance of fixation sequence statistics

We trained the model with the eye-movement data of a subject using the fixation sequence similarity measurement described in Section 3.6 and obtained the predicted fixation sequence when the subject reads new text. We then compared the predicted fixation sequence with the actual fixation sequence. Fig. 6 shows the contrast similarity diagrams of the fixation sequence when the 1st subject read text sample #46 (48 words). To unify the length of the input sequence of the neural network, the text sample #46 is left-padded by zeros to a length of 60. When there are no words that certainly should not to be fixated, this is reasonable.

From the fixation distributions in Fig. 6, we concluded qualitatively that the model prediction exhibited a high accuracy. To quantitatively measure the similarity of the two sequences, we used the method mentioned in Section 3.6 to obtain the sequence distance matrix (Table 2).

Table 1
The initial five lines of data after serialization.

Text_ID	Text	Word_Length	IA	Subject _ID
1	[‘are’ ‘now’, ‘rumblings’, ‘that’, ‘Appl…	[3, 3, 9, 4, 5, 5, 4, 6, 3, 5, 5, 5, 6, 3, 7, …	[1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, …	0
2	[‘days’, ‘later’, ‘the’, ‘British’, ‘ast…	[4, 5, 3, 7, 10, 7, 0, 3, 9, 3, 11, 7, 5, 7, 3…	[1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …	0
3	[‘agree’,‘that’, “California’s”, “three’…	[5, 4, 12, 5, 7, 3, 6, 3, 3, 4, 2, 1, 9, 8, 3,…	[1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, …	0
4	[‘was’, ‘in’, ‘a’, ‘bad’, ‘temper,’, ‘fo…	[3, 2, 1, 3, 6, 3, 3, 3, 6, 3, 9, 8, 2, 3, 7, …	[0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, …	0
5	[‘Darling’, ‘quivered’, ‘and’, ‘went’, ‘t…	[7, 8, 3, 4, 2, 3, 6, 2, 3, 8, 8, 3, 6, 3, 3, …	[1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, …	0

```
array([[1., 0., 0., ..., 0., 0., 0.],
       [1., 0., 0., ..., 0., 0., 0.],
       [1., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 1.],
       [0., 0., 0., ..., 0., 0., 1.],
       [0., 0., 0., ..., 0., 0., 1.]], dtype=float32)
```

Fig. 5. One-hot coding of subject’s ID.

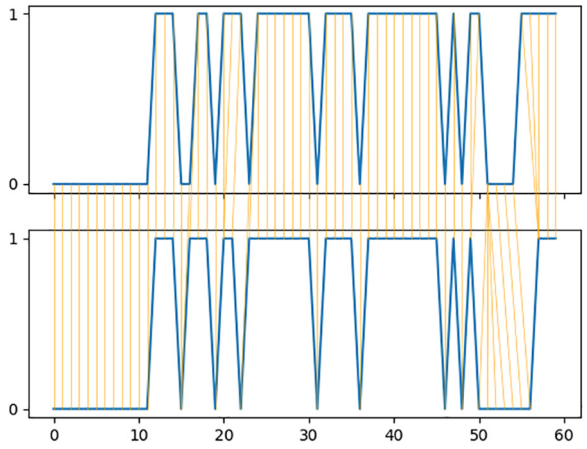

Fig. 6. Contrast similarity diagrams between the predicted and actual fixation sequences. The upper graph is the fixation sequence predicted by the model, and the lower graph is the actual (to be identified) fixation sequence. The abscissa is the word location label in the same text (60 words). The ordinate of 1 indicates that the word in the location was observed, and 0 indicates that the word in that location was skipped.

The similarity distance is the optimal path in the distance matrix from the upper-left corner to the lower-right corner. The coordinates of the optimal path in the matrix are as follows: [(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (7, 7), (8, 8), (9, 9), (10, 10), (11, 11), (12, 12), (13, 13), (14, 14), (15, 15), (16, 15), (17, 16), (17, 17), (18, 18), (19, 19), (20, 20), (21, 20), (22, 21), (23, 22), (24, 23), (24, 24), (25, 25), (26, 26), (27, 27), (28, 28), (29, 29), (30, 30), (31, 31), (32, 32), (33, 33), (34, 34), (35, 35), (36, 36), (37, 37), (38, 38), (39, 39), (40, 40), (41, 41), (42, 42), (43, 43), (44, 44), (45, 45), (46, 46), (47, 47), (48, 48), (49, 49), (50, 49), (51, 50), (51, 51), (51, 52), (51, 53), (52, 54), (53, 55), (54, 56), (55, 57), (56, 57), (57, 57), (58, 57)]. The optimal path of the predicted and the actual fixation sequences when the 1st subject read text sample #46 can be plotted, as shown in Fig. 7. The similarity distance between the two fixation sequences was 1.7321.

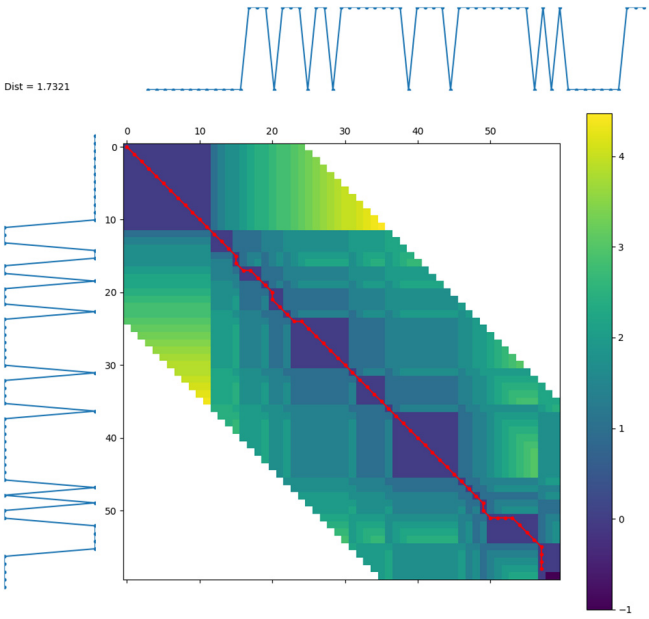

Fig. 7. Optimal path between the two fixation sequences.

The statistical data of the similarity distances between the predicted and actual fixation sequences of the 1st–84th subjects in the Provo Corpus were obtained using the fixation sequence similarity measurements described in Section 3.6. The results are shown in Table 3, where, for example, A1–P1 denotes the similarity distance of the 1st subject between the predicted and actual fixation sequences. The minimum similarity distance was 0.9004 in row 17, column 49, and the maximum value was 2.3403 in row 41, column 5. Therefore, the threshold value of the fixation sequence similarity distance ranged from 0.9004 to 2.3403. Within this range, we set a fixed step and used the evaluation metric to evaluate

Table 2
Distance matrix between predicted and actual fixation sequences.

	0	**1**	**2**	**3**	**4**	**5**	…	**55**	**56**	**57**	**58**	**59**	**60**
0	0.0	0.0	0.0	inf	inf	inf	…	inf	inf	inf	inf	inf	inf
1	0.0	0.0	0.0	0.0	0.0	0.0	…	inf	inf	inf	inf	inf	inf
2	0.0	0.0	0.0	0.0	0.0	0.0	…	inf	inf	inf	inf	inf	inf
3	inf	0.0	0.0	0.0	0.0	0.0	…	inf	inf	inf	inf	inf	inf
4	inf	0.0	0.0	0.0	0.0	0.0	…	inf	inf	inf	inf	inf	inf
..	…	…	…	…	…	…	…	…	…	…	…	…	…
56	inf	inf	inf	inf	inf	inf	…	1.000000	1.000000	1.000000	0.0	0.0	0.0
57	inf	inf	inf	inf	inf	inf	…	1.414214	1.414214	1.414214	0.0	0.0	0.0
58	inf	inf	inf	inf	inf	inf	…	1.732051	1.732051	1.732051	0.0	0.0	0.0
59	inf	inf	inf	inf	inf	inf	…	2.000000	2.000000	2.000000	0.0	0.0	0.0
60	inf	inf	inf	inf	inf	inf	…	2.236068	2.236068	2.236068	0.0	−1.0	−1.0

Table 3
Similarity distance between predicted and actual fixation sequences.

Text_ID	A1-P1	...	A4-P4	A5-P5	A6-P6	...	A48-P48	A49-P49	A50-P50	...	A84-P84
Text1	2.1146	...	1.6636	2.0180	1.1102	...	1.4689	1.5603	2.2537	...	1.0884
Text2	1.7833	...	2.1302	1.0207	2.2017	...	1.2655	2.1932	0.9979	...	1.9809
...	...	...	...	...	...	...	...	...	...	...	...
Text16	1.5061	...	2.1687	0.9028	1.8789	...	2.0043	0.9441	1.7238	...	1.2789
Text17	2.0578	...	1.3679	1.5973	2.0742	...	2.2683	**0.9004**	1.1440	...	1.9521
Text18	2.1306	...	1.1812	2.2985	0.9013	...	1.8768	1.9175	2.0462	...	1.3511
...	...	...	...	...	...	...	...	...	...	...	...
Text40	1.7501	...	2.0273	0.9965	1.0456	...	1.7956	1.9455	0.9636	...	1.7510
Text41	1.5221	...	1.2680	**2.3403**	1.3968	...	2.2948	1.3463	1.9741	...	1.0029
Text42	1.2543	...	2.2644	1.7617	1.1004	...	1.6325	1.1004	1.9905	...	2.2753
...	...	...	...	...	...	...	...	...	...	...	...
Text55	1.3334	...	1.1547	2.1070	1.2138	...	1.7226	1.5274	1.0292	...	1.3723

The underlined values indicate the maximum and minimum values in the table.

the influence of different thresholds in turn. The results are discussed in the next section.

5. Results and discussion

5.1. Evaluation metric

The proposed method was evaluated by the Rank-1 (R1) accuracy rate and the equal error rate (EER). The Rank-1 accuracy rate is the ratio of the total number of correct recognitions to the number of samples. The EER is the value when the false acceptance rate (FAR) and the false rejection rate (FRR) are equal.

The false acceptance rate is the ratio that is considered to be the same subject when the similarity distance of different eye-movement sequences is greater than the given threshold during the testing of the biometric recognition reading eye-movement on the eye tracking corpus. More simply, it refers to the ratio of taking the mismatched eye-movement sequence as the matched sequence.

The false rejection rate is the ratio that is considered to be different subjects when the similarity distance of different eye-movement sequences is smaller than the given threshold during the testing of the biometric recognition reading eye-movement on the eye tracking corpus. It refers the ratio of taking the matched eye-movement sequence as the mismatch sequence.

5.2. Results

Based on the threshold range determined in Section 4.5, we set the threshold values in a fixed step size within this range, with each threshold corresponding to a set of FAR and FRR values. When the FAR and FRR were equal, the common value was the EER, which was 19.4%, and the corresponding threshold value was 1.6071. Fig. 8 shows the detection error tradeoff (DET) curve. The EER is where the curve intersects with the line that passes through the points (0, 0) and (1, 1).

When measuring the Rank-1 accuracy, if the similarity distance between the predicted fixation sequence and the actual one (to be identified) was less than 1.6071, it was considered to be the same subject. The Rank-1 accuracy rate was measured by a 10-fold cross-validation method on the dataset. Since the initial value of the neural network was randomly selected, the experiment was repeated 100 times, and the average Rank-1 accuracy were obtained, which was 86.5%. The experimental results and the standard deviation are given in Table 4.

As is well known, the CEM-P method [11] obtains the highest EER and Rank-1 accuracy among the existing reading eye-movement biometric recognition technologies, and the graph-based matching method [12] makes use of the fewest features.

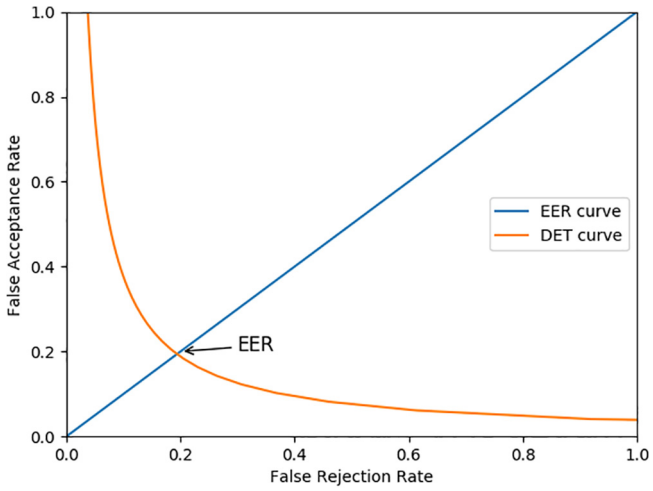

Fig. 8. Detection error tradeoff curve.

Table 4
Comparison of results.

Metrics	Graph-based matching	CEM-P	Our model
R1	70%	82.6%	86.5% ± 2.96%
EER	30%	16.5%	19.4%
Handcrafted features	4	12	1

As shown in Table 4, the proposed method used fewer handcrafted feature to obtain the values of the Rank-1 accuracy and EER that were similar to those obtained by the CEM-P method.

5.3. Discussion

In our study, there are two reasons that R1 and EER values were similar to those obtained by the CEM-P method while using fewer handcrafted feature. First, the proposed method allows the computer to automatically learn the reading eye-movement features based on deep-learning, and the feature learning is integrated into the process of model construction, thereby reducing the incompleteness caused by artificial design features. Furthermore, the text-based linguistic feature sequence selects a word length sequence that is easily obtained and conforms to human processing of low-level visual information. Meanwhile, LSTM is a type of time-recurrent neural network that can process and predict important events with long interval and time delays in the time sequence, which is in line with human processing of high-level lan-

guage information. The proposed model makes full use of high and low information at the visual level, which is more in line with the practical principle of reading eye-movements.

Although the proposed method makes use of the features of the stimuli (reading materials) and eye-movement sequence, it is not strictly end-to-end learning, and we needed to label the word length in advance; this is the limitation of this method. Fortunately, the word length feature can be automatically labeled by the computer. Further study is required to evaluate the effects of other linguistic features from multiple angles and solve the problem of automatic labeling of features such as parts of speech.

The proposed biometric recognition method based on reading eye-movement obtained a high recognition accuracy rate on the specified data set, but we believe that it is not a "true" biometric recognition measure and cannot be used alone. "True" biometric identification measures, such as fingerprints and iris patterns, remain more-or-less constant during a person's life, while eye-movement is controlled by cognitive state and oculomotor function, which are relatively unstable. We can easily infer that the eye-movement pattern on the first and subsequent readings of the same text by the same person will be different. Furthermore, there are many factors that affect reading eye-movement, such as the difficulty of the material, the reader's familiarity with the material, the reading environment, the reading goals (some tasks require answering questions after reading, and some only require a readthrough), and even the font size. As a kind of behavioral-feature-based biometric method, eye-movement recognition can never obtain a higher recognition rate or better robustness than physical-feature-based biometric methods.

However, biometric recognition based on behavioral features still has the following unique advantages. 1. Behavioral features are not easily forged. As eye-movement-based recognition uses information that is produced mostly by the brain (so far impossible to imitate), forging this kind of information is much more difficult. However, physical-feature-based biometric methods, such as fingerprint verification or iris recognition, are mostly based on physiological properties of the human body. Therefore, what is needed for proper identification is only the "body" of a person who is to be identified. This makes it possible to identify an unconscious person or, in some cases, a dead person. 2. Biometric recognition methods based on behavioral features can identify people without being perceived. Eye-movement recognition can achieve this using hidden cameras or eye trackers. For physical-feature-based recognition, it is necessary to actively allow the sensor to collect the data from a certain part of the body.

In summary, the combination of behavior-based and physical-feature-based recognition can compensate for the latter's deficiencies and make a system more secure.

6. Conclusion

In this paper, a type of reading eye-movement biometric recognition technology was proposed based on deep-learning. This technology constructs a reading eye-movement recognition (REMR) computational model based on a multi-input deep neural network and identifies human subjects by comparing the predicted and actual fixation sequences. This model is less dependent on the data features and requires less pre-possessing, which makes it attractive for industrial and engineering applications [42]. The results of the simulations show that the Rank-1 recognition rate on the dataset was 86.5%. The experimental results further proved that the proposed method is novel, effective, and superior to other methods in the biometric field. As a kind of behavior (rather than a physical) feature, eye-movement features have not been as accurate in biometric recognition as physical characteristics such as fingerprints and iris patterns. However, behavioral feature-based biometrics is not easily forged and can identify people without being perceived. This gives such methods broad application prospects in the security field after deep integration with physical feature-based biometrics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant number 61871326], the Humanities and Social Science Fund of Ministry of Education of China [grant number 18YJCZH180] , the Natural Science Foundation of Shaanxi Province [grant number 2018JM6116], the Social Science Foundation of Shaanxi Province [grant number 2019M001] , and the Aeronautical Science Foundation of China [grant number 20185153].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neucom.2020.06.137.

References

[1] A.K. Bobak, B.A. Parris, N.J. Gregory, R.J. Bennetts, S. Bate, Eye-movement strategies in developmental prosopagnosia and "super" face recognition, Q. J. Exp. Psychol. (2017), https://doi.org/10.1080/17470218.2016.1161059.
[2] X. Li, P. Liu, K. Rayner, Saccade target selection in Chinese reading, Psychon. Bull. Rev. (2015), https://doi.org/10.3758/s13423-014-0693-3.
[3] K. Rayner, Eye movements in reading and information processing: 20 years of research, Psychol. Bull. (1998), https://doi.org/10.1037/0033-2909.124.3.372.
[4] A. Kennedy, Book review: eye tracking: a comprehensive guide to methods and measures, Q. J. Exp. Psychol. (2016), https://doi.org/10.1080/17470218.2015.1098709.
[5] C. Clifton, F. Ferreira, J.M. Henderson, A.W. Inhoff, S.P. Liversedge, E.D. Reichle, E.R. Schotter, Eye movements in reading and information processing: Keith Rayner's 40 year legacy, J. Memory Lang. (2016), https://doi.org/10.1016/j.jml.2015.07.004.
[6] S.G. Luke, K. Christianson, Limits on lexical prediction during reading, Cogn. Psychol. (2016), https://doi.org/10.1016/j.cogpsych.2016.06.002.
[7] D. Noton, L. Stark, Scanpaths in eye movements during pattern perception, Science (80-) (1971), https://doi.org/10.1126/science.171.3968.308.
[8] P. Kasprowski, J. Ober, Eye movements in biometrics, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) (2004).
[9] R. Kasprowski, Taming the electronic tiger: report on the 2004 ala midwinter symposium on electronic resource management, Ser. Rev. (2004), https://doi.org/10.1080/00987913.2004.10764915.
[10] C. Holland, O. V. Komogortsev, Biometric identification via eye movement scanpaths in reading, in: 2011 Int. Jt. Conf. Biometrics, IJCB 2011, 2011. <https://doi.org/10.1109/IJCB.2011.6117536>.
[11] C.D. Holland, O. V. Komogortsev, Complex eye movement pattern biometrics: Analyzing fixations and saccades, in: Proc. - 2013 Int. Conf. Biometrics, ICB 2013, 2013. <https://doi.org/10.1109/ICB.2013.6612953>.
[12] I. Rigas, G. Economou, S. Fotopoulos, Biometric identification based on the eye movements and graph matching techniques, Pattern Recognit. Lett. (2012), https://doi.org/10.1016/j.patrec.2012.01.003.
[13] V. Cantoni, C. Galdi, M. Nappi, M. Porta, D. Riccio, GANT: Gaze analysis technique for human identification, Pattern Recognit. (2015), https://doi.org/10.1016/j.patcog.2014.02.017.
[14] E.D. Reichle, Computational models of reading: a primer, Lang. Linguist. Compass. (2015), https://doi.org/10.1111/lnc3.12144.
[15] E.D. Reichle, K. Rayner, A. Pollatsek, The E-Z reader model of eye-movement control in reading: comparisons to other models, Behav. Brain Sci. (2003), https://doi.org/10.1017/S0140525X03000104.
[16] R. Engbert, A. Nuthmann, E.M. Richter, R. Kliegl, Swift: a dynamical model of saccade generation during reading, Psychol. Rev. (2005), https://doi.org/10.1037/0033-295X.112.4.777.
[17] M. Barrett, A. Søgaard, Modeling eye movements when reading microblogs, in: Proc. Natl. Conf. Artif. Intell., 2015: pp. 4231–4232.
[18] N. Landwehr, S. Arzt, T. Scheffer, R. Kliegl, A model of individual differences in gaze control during reading, in: EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., 2014.

[19] H.M. Fehd, A.E. Seiffert, Eye movements during multiple object tracking: where do participants look?, Cognition 108 (2008) 201–209, https://doi.org/10.1016/j.cognition.2007.11.008.
[20] F. Matties, A. Søgaard, With blinkers on: Robust prediction of eye movements across readers, in: EMNLP 2013 - 2013 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., 2013.
[21] K. Bicknell, R. Levy, A rational model of eye movement control in reading, in: ACL 2010 - 48th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., 2010.
[22] M. Hahn, F. Keller, Modeling Human Reading with Neural Attention, in: Proc. 2016 Conf. Empir. Methods Nat. Lang. Process., Association for Computational Linguistics, Austin, Texas, 2016: pp. 85–95. <http://dx.doi.or/10.18653/v1/D16-1009>.
[23] X. Wang, X. Zhao, J. Ren, J. Han, A new type of eye movement model based on recurrent neural networks for simulating the gaze behavior of human reading, Complexity (2019), https://doi.org/10.1155/2019/8641074.
[24] X. Wang, X. Zhao, M. Xia, The Prediction Model of Saccade Target Based on LSTM-CRF for Chinese Reading, in: A. and Z.J. and L.C.-L. and L.B. and Z.H. and Z. X. Ren Jinchang and Hussain (Ed.), Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer International Publishing, Cham, 2018: pp. 44–53. <https://doi.org/10.1007/978-3-030-00563-4_5>.
[25] X. Wang, X. Zhao, Eye movement prediction of individuals while reading based on deep neural networks, J. Tsinghua Univ. (Sci. Technol.) (2019), https://doi.org/10.16511/j.cnki.qhdxxb.2019.26.001.
[26] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, R.X. Gao, Deep learning and its applications to machine health monitoring, Mech. Syst. Signal Process. (2019), https://doi.org/10.1016/j.ymssp.2018.05.050.
[27] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, N.A. Smith, Transition-based dependency parsing with stack long short-term memory, in: ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf., 2015.
[28] K. Greff, R.K. Srivastava, J. Koutnik, B.R. Steunebrink, J. Schmidhuber, LSTM: a search space Odyssey, IEEE Trans. Neural Networks Learn. Syst. (2017), https://doi.org/10.1109/TNNLS.2016.2582924.
[29] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, N.D. Sidiropoulos, Learning to optimize: training deep neural networks for interference management, IEEE Trans. Signal Process. (2018), https://doi.org/10.1109/TSP.2018.2866382.
[30] Y. Goldberg, Neural network methods for natural language processing, Synth. Lect. Hum. Lang. Technol. (2017), https://doi.org/10.2200/S00762ED1V01Y201703HLT037.
[31] Z. Chen, L. Zhang, C. Jiang, Z. Cao, W. Cui, WiFi CSI based passive human activity recognition using attention based BLSTM, IEEE Trans. Mob. Comput. (2018), https://doi.org/10.1109/TMC.2018.2878233.
[32] P. Zhao, Y. Zhang, M. Wu, S.C.H. Hoi, M. Tan, J. Huang, Adaptive cost-sensitive online classification, IEEE Trans. Knowl. Data Eng. (2019), https://doi.org/10.1109/TKDE.2018.2826011.
[33] D. Schultz, B. Jain, Nonsmooth analysis and subgradient methods for averaging in dynamic time warping spaces, Pattern Recognit. (2018), https://doi.org/10.1016/j.patcog.2017.08.012.
[34] S.G. Luke, K. Christianson, The Provo Corpus: a large eye-tracking corpus with predictability norms, Behav. Res. Methods. (2018), https://doi.org/10.3758/s13428-017-0908-4.
[35] S.C. Sereno, C.J. Hand, A. Shahid, B. Yao, P.J. O'Donnell, Testing the limits of contextual constraint: Interactions with word frequency and parafoveal preview during fluent reading, Q. J. Exp. Psychol. (2018), https://doi.org/10.1080/17470218.2017.1327981.
[36] A. Kennedy, J. Pynte, W.S. Murray, S.A. Paul, Frequency and predictability effects in the Dundee Corpus: an eye movement analysis, Q. J. Exp. Psychol. (2013), https://doi.org/10.1080/17470218.2012.676054.
[37] G. Yan, Z. Meng, N. Liu, L. He, K.B. Paterson, Effects of irrelevant background speech on eye movements during reading, Q. J. Exp. Psychol. (2018), https://doi.org/10.1080/17470218.2017.1339718.
[38] A.W. Yu, H. Lee, Q. V. Le, Learning to skim text, in: ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., 2017. <https://doi.org/10.18653/v1/P17-1172>.
[39] J. Moolayil, Learn Keras for Deep Neural Networks (2019), https://doi.org/10.1007/978-1-4842-4240-7.
[40] Y.N. Dauphin, H. De Vries, Y. Bengio, Equilibrated adaptive learning rates for non-convex optimization, in: Adv. Neural Inf. Process. Syst., 2015.
[41] M. Huang, Q. Qian, X. Zhu, Encoding syntactic knowledge in neural networks for sentiment classification, ACM Trans. Inf. Syst. (2017), https://doi.org/10.1145/3052770.
[42] G. Li, Y. Shen, P. Zhao, X. Lu, J. Liu, Y. Liu, S.C.H. Hoi, Detecting cyberattacks in industrial control systems using online learning algorithms, Neurocomputing 364 (2019) 338–348, https://doi.org/10.1016/j.neucom.2019.07.031.

Xiaoming Wang received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2020. He is currently an Associate Professor with Xi'an International Studies University. His research interests include cognitive computing and artificial intelligence.

Xinbo Zhao received the Ph.D. degree from the Northwestern Polytechnical University, Xi'an, China, in 2003. Currently, he is a professor at the School of Computer Science, Northwestern Polytechnical University. His interests include image processing, computer vision, pattern recognition and artificial intelligence. He is the author or co-author of more than 100 scientific papers.

Yanning Zhang received her B.S. degree from Dalian University of Science and Engineering in 1988, M.S. and Ph.D. Degree from Northwestern Polytechnical University in 1993 and 1996 respectively. She is presently a Professor of School of Computer Science and Technology, Northwestern Polytechnical University. She is also the organization chair of ACCV2009 and the publicity chair of ICME2012. Her research work focuses on signal and image processing, computer vision and pattern recognition. She has published over 200 papers in these fields, including the ICCV2011 best student paper. She is a member of IEEE.