Clustering Based on Eye Tracking Data for Depression Recognition

Minqiang Yang, *Member, IEEE*, Chenlei Cai, and Bin Hu, *Senior Member, IEEE*

***Abstract*—The attention-based approach would be a good way of detecting depression, assisting medical diagnosis, and treating the patients at risk earlier. In this article, a new approach of recognizing depression is proposed, which avoids eye movement event identification and directly performs clustering based on eye tracking data to obtain regions of interesting (ROIs), and then conducts depression recognition modeling. Based on these, a novel spatiotemporal clustering algorithm was proposed, i.e., ROI Clustering with Deflection Elimination, which takes the noisy data into consideration to better describe attention patterns. On the data set with 45 depression patients and 44 healthy controls, the proposed algorithm achieved the best classification accuracy of 76.25%, which has the potential to provide methodological reference on the assessment of mental disorders based on eye movements.**

***Index Terms*—Depression, gaze points, ordering point to identify the cluster structure (OPTICS), regions of interesting (ROIs), spatiotemporal clustering.**

I. Introduction

DEPRESSION is one of the most common mental disorders, which is estimated by the World Health Organization (WHO) to become the first cause of the burden of disease worldwide by 2030 [1]. Depression directly results in poor physical health states [2], [3]. It affects physiological function and even leads to the reduction of productivity [4]. The worst result is that some patients finally choose to end their lives, and with up to 850 000 suicides caused by depression reported every year [5]. Despite the efforts devoted to recognizing and treating depression, the prevalence of depression keeps on rising, particularly in younger people. However, the methods of depression diagnosis almost exclusively rely on the symptom severity reported by the patients or clinical judgments, which has a wide range of subjective biases [6], [7]. Inspired by this, researchers have extensively carried out research on emotional objective assessment and multimodal fusion [8] based on text, behavioral data and physiological signals [9], and further, extensively investigated quantitative assessment research on affective disorders of major depression and bipolar [10], [11], [12]. Brain cognitive models and tools developments might also help pathological study of mental disorder [13], [14].

The latest developments in eye tracking technology make it a powerful and popular tool that can be used to collect fine-grained temporal measurements of cognitive processes [15], [16]. Because of this superiority, eye tracking has been adopted in more and more fields, such as visual science [17], marketing [18], and human–computer interaction [19], [20]. Cognitive impairment is one of the most common core symptoms of depression. Relevant studies have shown the possibility to analyze the difference on response inhibition, attention, memory, and other cognition aspects between depression patients and healthy controls through eye movement data [21]. Li et al. [22] used fixation task, saccade task, and free-view task to reveal some significant differences in eye movement indicators between depression patients and healthy controls, such as patients with depressive disorder showed more fixations, shorter fixation durations, more saccades, and longer saccadic lengths.

For a long time, the research on attentional bias has been a hot spot for depression recognition based on eye movement. A large number of studies have shown that depression patients have a negative attentional bias. Eizenman et al. [23] used a free-view task to investigate the negative attentional bias of depression patients, the result of which showed that compared with the healthy controls, depression patients gazed at negative pictures longer, and were difficult to disengage from negative stimuli. Gotlib et al. [7] and Goeleven et al. [24] used emotional faces to draw similar conclusions, that depression patients had a significant attentional bias toward sad faces. Researchers also investigated the attention distribution for positive stimulus, but there were not consistent conclusion so far. Goeleven et al. [24] showed that the attention of depression patients in positive stimulus pictures was not significantly different from that of healthy control, but Duque and Vázquez [25] found that depression patients were lacked of positive attentional bias.

Regardless of inconsistency of research conclusions, the effectiveness of attention distribution for depression

Manuscript received 24 May 2022; revised 28 October 2022; accepted 14 November 2022. Date of publication 18 November 2022; date of current version 11 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFA0706200; in part by the National Natural Science Foundation of China under Grant 61632014 and Grant 61627808; and in part by the Natural Science Foundation of Gansu Province, China, under Grant 22JR5RA488. *(Corresponding author: Bin Hu.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee the Third People's Hospital of Guangyuan, under application No. GJWLSP2020012, and performed in line with the Declaration of Helsinki.

Minqiang Yang and Chenlei Cai are with the School of information Science and Engineering, Lanzhou University, Lanzhou 730000, China (e-mail: yangmq@lzu.edu.cn; caichl19@lzu.edu.cn).

Bin Hu is with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: bh@bit.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCDS.2022.3223128.

Digital Object Identifier 10.1109/TCDS.2022.3223128

recognition has been proved by many studies. Quantitative measures of attention distribution are often calculated from the eye movement behaviors under emotional or cognitive experimental paradigm recorded by the eye tracking system. The eye movement behavior indicators directly related to attention mainly include fixation, saccade, microsaccade, and smooth pursuit. Among those behaviors, fixation is the most significant indicator which are frequently used for depression recognition. The raw eye movement data are sample points of visual locations collected by infrared devices at a certain frequency, which are post-processed by eye movement event identification algorithms to get behavior event of fixation or saccade. The traditional identification methods, e.g., velocity-based algorithms and dispersion-based algorithms [26], [27], distinguish between fixation and saccade by artificially setting decision thresholds. Recent research uses unsupervised machine learning for eye movement event recognition, such as clustering [28]. The eye movement event analysis of fixation combined with saccade can reflect the dynamic changes of visual attention.

Attention is one of the most significant features in the human vision system (HVS), which represents the ability to focus on a region in a specific scene [29]. Different pictures have different prominent areas, and highly prominent areas may attract more attention than other regions [30]. Those parts of the picture that attract fixations and capture primary attention are called regions of interesting (ROIs) [31]. Fixation points can be used as a measure of regional importance, while ROIs are more direct and holistical way of reflecting the degree of human visual attention to picture scenes [32], which are always used to study the distribution and conversion of visual attention. Because the attention relationship between visual entities conveys a wealth of information, automatic determination of this relationship provides us with the semantic representation of the picture. The ROIs can be extracted from fixation events by a machine learning process or manually outlining [33]. For a specific area, the more fixations that fall within it, the more likely it is to be considered an ROI [34], [35].

Density-based clustering is a very intuitive method of clustering, in which adjacent regions with high density are practiced to form clusters. The method can find clusters of various sizes and shapes and has some noise resistance properties. Therefore, this article used some classic density-based clustering algorithms as baseline methods. Ordering point to identify the cluster structure (OPTICS) by Ankerst et al. [36] improved the oversensitivity of input parameters of earlier methods, which sorts the samples to get a reachability plot which contains hierarchical structure of clustering, and the different levels of clusters correspond to different granularity of attention distribution. Due to the spatiotemporal properties of eye tracking data, this article adopts the spatiotemporal variant of OPTICS, spatiotemporal OPTICS (ST-OPTICS) [37] for further innovative work.

With the motivation of analyzing the attention patterns of depression patients from raw eye movement data, this article proposed a novelty ROI analysis method based on spatiotemporal clustering. Experimental results showed that the proposed method ROIs clustering with deflection elimination (RCDE) achieves the accuracy of 76.25% and proved that the proposed method is more effective in identifying depression than traditional methods. The main contributions of this article are as follows.

1) We proposed a novelty ROIs recognition method, called ROIs eventless clustering (REC), which does not require eye movement event identification. This would be helpful for eye movement-based cognitive or emotional researches.
2) We proposed a novelty spatiotemporal clustering method on the mixed attribute features model in eye tracking data supporting depression recognition, i.e., ROI Clustering with Deflection Elimination (RCDE), which took noisy data into consideration during clustering.
3) We introduced a practical method to measure head movement and uses it for classification modeling, providing ideas for related research based on wearable eye trackers.

The remainder of this article is organized as follows. Section II introduces the experimental paradigm and data collection process. Section III provides a detailed description of the proposed method, and Section IV shows and discusses the experimental results. Section V concludes this article.

II. Experiment and Data

A. Eye Tracker Equipment

There are two main types of eye trackers, i.e., desktop and wearable eye trackers. The wearable eye tracker does not require the subjects to fix their heads on a chin rest, so that the subjects' behaviors would be closer to the real state. This article adopts a free-browse experimental paradigm to record eye movement behaviors under emotional stimuli through a wearable eye tracker, i.e., Pupil Core by Pupil labs, Germany [38]. Pupil Core equips a world camera with the resolution of 1080P and frame rate of 30 Hz, and two adjustable eye cameras which work at 120 Hz to record the raw eye video. The world camera capture is used to record the vision field of the subject and the predicted relative gaze position in the vision field.

B. Subjects

Patients with major depressive disorder (MDD) were recruited among inpatients and outpatients from the Guangyuan Hospital of Sichuan Province, China. They were diagnosed by the structured Mini-International Neuropsychiatric Interview (M.I.N.I.) and recommended by at least one clinical psychiatrist. All subjects need to fit the criteria of inclusion and exclusion criteria listed in Table I. The local Ethics Committee approved consent forms and study design for Biomedical Research at the Second People's Hospital of Gansu Province and Guangyuan Hospital of Sichuan Province following the Code of Ethics of the World Medical Association (Declaration of Helsinki). Written informed consent was obtained from all subjects before the experiment. A total of 109 subjects including 55 outpatients diagnosed with depression, and 54 healthy controls were recruited. Some invalid data were excluded, such as subjects

TABLE I
INCLUSION AND EXCLUSION CRITERIA OF EYE TRACKER EXPERIMENTS FOR DEPRESSION RECOGNITION

	depression patients	healthy controls
Inclusion Criteria	1.Between the age of 32-48 years old, male or female 2.Primary school or higher education level 3.All MDD patients received a structured Mini-International Neuropsychiatric Interview (M.I.N.I.) 4.The Patient Health Questionnaire-9 item (PHQ-9) [40] score of subjects was greater than or equal to 5 5.No psychotropic drug treatment having been performed in the last two weeks 6.A normal or corrected-to-normal vision	1.Between the age of 32-48 years old, male or female 2.Primary school or higher education level 3.No psychotropic drug treatment having been performed in the last two weeks 4.A normal or corrected-to-normal vision
Exclusion Criteria	1.The one has mental disorders or brain organ damage 2.The one has a severe physical illness and extremes suicidal tendency 3.Subjects were abused or dependent on alcohol or psychotropic drugs in the past year 4.Women who were pregnant and in lactation or taking birth control pills	1.A personal or family history of mental disorders 2.The one has mental disorders or brain organ damage 3.The one has a severe physical illness and extremes suicidal tendency 4.Subjects were abused or dependent on alcohol or psychotropic drugs in the past year 5.Women who were pregnant and in lactation or taking birth control pills

TABLE II
DEMOGRAPHIC DATA OF THE SUBJECTS

Group	Number	Age (Mean ± SD)	Gender
Depression Patients	45	Female: 35.65 ± 11.95 Male: 31.73 ± 10.69	Female: 34 Male: 11
Healthy Controls	44	Female: 35.06 ± 12.06 Male: 33.56 ± 11.96	Female: 25 Male: 19

who failed to calibrate, or the subjects which did not comply with the requirements of the experimenter. A data set of 45 depression and 44 healthy controls (Table II) was used. The previous study by Wang et al. [39] had proved that the emotional feelings of females and males in most pictures were similar, and the correlation was high, so we ignore the gender imbalance of our data set.

C. Stimuli

Depression patients the healthy controls have different reactions when watching emotional stimulus [41]. Depression patients are relatively numb to positive stimulus, but sensitive to negative stimulus. Researchers found when depression patients face three stimuli of positive, neutral, and negative, they would have attention bias on negative stimuli [7]. Therefore, our study used the above-mentioned three types of pictures stimuli with different emotions. The whole paradigm consists of two blocks, each block consists of four trials, with a total duration of 245 s. Each trial contains two parts: one is the annotation focus, and the other is the stimulus pictures playback. In each trial, a total of five pictures of the same type of stimulus are played, each picture displays for 5 s.

The pictures were selected from the International Affective Picture System (IAPS) which was compiled and formulated by the National Institute of Mental Health (NIMH) American Center for emotion and attention [42]. IAPS involves a wide range of contents and can comprehensively induce all kinds of human emotions. And it has been widely used in psychology, neurophysiology, brain cognitive science, and other fields, especially in the research of emotion and attention [43], [44], [45]. The sufficient variability of the picture scores indicates that it is feasible to select pictures based on these variables. In general, current ratings should allow researchers to use these norms for research purposes, especially in research dealing with the

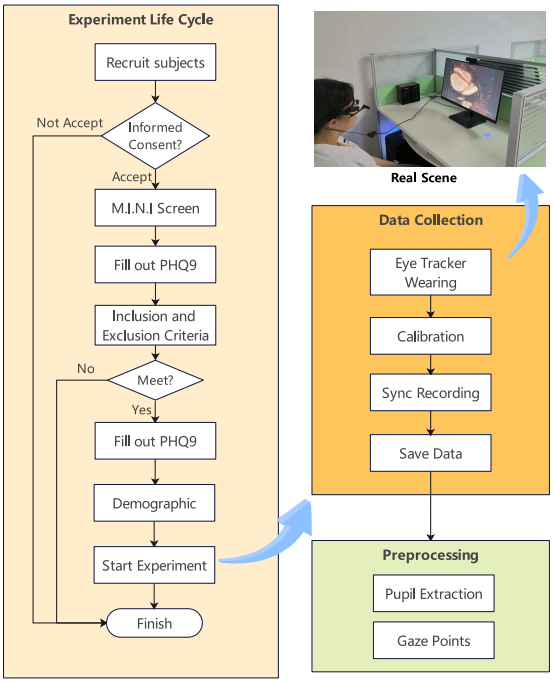


Fig. 1. Eye tracking experiment procedure.

relationship between emotion and cognition [46]. In this study, we finally selected 40 pictures, in which 10, 20, and 10 pictures of negative (valence: 2.95 ± 1.62, arousal: 5.35 ± 2.24), neutral (valence: 7.43 ± 1.48, arousal: 4.33 ± 2.27), and positive (valence: 5.03 ± 1.15, arousal: 2.91 ± 1.97), respectively. In the whole stimuli video, each picture is played for 5 s, and the resolution is $1024 * 768$. The brightness, contrast, and color of the monitor are all set uniformly. While the picture is presenting, the subjects only need to watch the picture, and the eye tracker collects the eye movement data of the subjects. The subjects have no prior knowledge of the pictures and have no introduction provided about the pictures before or during the study.

The whole flow of experiment and real scene is shown in Fig. 1, the experimental video playback process is shown in Fig. 2, the vision captured by world camera is shown in Fig. 3. The experiment was arranged in a quiet and clean room. Subjects sat 70–80-cm away from the screen, no head fixation required. Subjects were introduced to watch the whole stimulus video on a comfortable chair. Before the official start

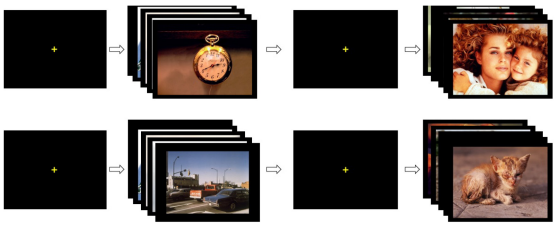


Fig. 2. Paradigm process.

of the experiment, calibration was required to ensure that the gaze position could be predicted.

D. Preprocessing

In our experiment, subjects sat in front of the computer screen without fixing the head on a chin-rest, so that the subject's visual field might change as the subject's head move. The gaze points predicted by the eye tracking system are the relative position of the visual field. To calculate the relative position of the gaze point on the stimulus image, it is necessary to know the position of the stimulus image in the visual field. The display and stimulus pictures in the field of view are located by target recognition technology to obtain the relative position of the eyes relative to the screen. We use single-shot detector (SSD) [47] with the backbone network of Visual Geometry Group (VGG-16). We labeled "screen" and "stimulus" on 300 samples as the training data set, the samples are the pictures captured by the world camera on the eye tracker. This SSD model adds several feature layers to the end of a base network, which predict the offsets to default boxes of different scales and aspect ratios and their associated confidences [47]. The input picture is used for features extraction by the VGG network and four additional convolution layers.

The predicted result is shown in Fig. 3, the inner target box detects the stimulus picture position, and the outer target box detects the monitor position. Through a series of processes, we obtained the central point coordinate of the final detected boxes. On this basis, we corrected the coordinates of the gaze points accordingly and removed the invalid points that fall off the screen. The head movement traces of each subject in the whole recording video were also calculated, which were useful to recognizing depression proven by [48].

E. Feature Extraction

The definition of the basic terms related to eye movement data can refer to [49]. Among them, the eye tracking data is the original eye movement data collected by the eye tracker. The fixations are always generated by the software shipped with eye tracker with the thresholds of velocity or dispersion. The saccade describes the rapid eye movement from one fixation point to another. ROI is created from the semantic information of visual stimuli, usually marked as a whole or part of an object.

The method proposed in this article does not identify specific eye movement events, so the eye movement event-based features defined in this article are no longer applicable. However, in order to apply some features that can reflect the distribution of attention, this article regards the lower level clusters generated by hierarchical density clustering as fixation and the higher level clusters as ROIs. Thus, 12 ROI-related features shown in Table III, which are still available in our proposed method, were used in this article to further machine learning modeling. Regarding the head movements, we used the head angle, quadrant, and distance between two head movement coordinates as features.

TABLE III
EYE MOVEMENT FEATURES EXTRACTED IN THIS ARTICLE

Group	Eye Movement Features
ROI Information	First gaze time
	First gaze duration
	Gaze number
	Percentage of gaze points in the current ROI in a trial of all gaze points
	The start time of the last gaze point entering the current ROI
	The X,Y coordinates of the last fixation in the current ROI
	The sum of the number of gaze points in the ROI
	The sum of all gaze points duration
	The duration of the first gaze event in the current ROI
	The X,Y positions of the first fixation in the current ROI

III. METHOD

The traditional method of extracting ROIs was to replace the subjects' saccade path with ROIs sequences. In which the fixations were first obtained, then clustering algorithms or manually marking were conducted to get ROIs. This article draws on the idea of clustering gaze points to generate gaze regions and ignores the different types of eye movement events. We propose a novel approach to get the attention distribution, i.e., REC, which directly perform clustering on the raw eye tracking data for further classification modeling. In addition, we propose a variant of REC under our special data set, i.e., ROI Clustering with Deflection Elimination (RCDE). The core ideas of REC and RCDE are shown in Fig. 4. This section first introduces some basic concepts and definitions, and then presents the proposed method.

A. Concepts and Definition

We adopted some concepts and terms from ST-OPTICS [36], [37], [50] can be defined as follows.

Definition 1 (Object): Spatiotemporal information database "object," which contains all the spatiotemporal data points that are needed to be clustered

$$\begin{aligned} \text{Object} &= \{O_1, O_2, \ldots, O_i, \ldots, O_N\}, (0 \le i \le N) \\ \text{O}_i &= (\text{Identifier}_i, X_i, Y_i, T_i, \text{AF}_i) \\ \text{AF}_\text{i} &= \left(\text{AF}_{c_i}, \text{AF}_{d_i}, \text{AF}_{o_i}\right) \end{aligned} \tag{1}$$

where N is the number of spatiotemporal objects, and O_i is the ith spatiotemporal object. $Identifier_i$ is the identity of the object; X_i and Y_i are the x- and y-axis coordinates of the spatiotemporal samples; T_i is the timestamp; and AF_i is the attribute features set of the spatiotemporal samples, which consists of ordered, disordered, and continuous variables attribute features.

(a) (b)

Fig. 3. Screen and stimulus picture recognition result. (a) and (b), respectively, show the recognition effect under different stimulus pictures.

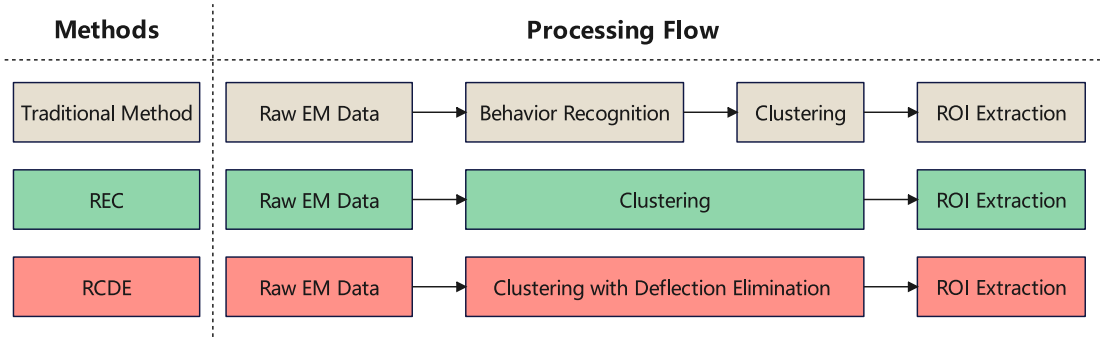


Fig. 4. Core idea of REC and RCDE.

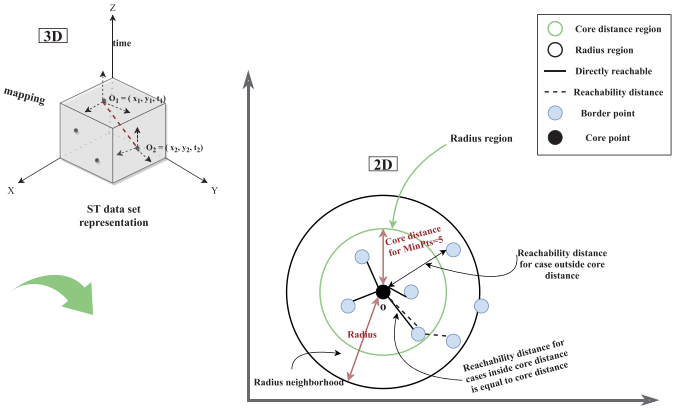


Fig. 5. Basic concepts of the algorithm.

Definition 2 (Radius-Neighborhood): Radius-neighborhood of a spatiotemporal point p is defined by

$$\{p, q \in \text{Object} | \text{S\_dist}(p, q) \leq \text{S\_threshold} \ \& \ \text{T\_dist}(p, q) \leq \text{T\_threshold}\} \tag{2}$$

where $\text{S\_dist}(p, q)$ is the Euclidean distance for two points p and q, the schematic of core concepts shows in Fig. 5.

Definition 3 (Core Object): An object is a core object if the number of objects in its radius-neighborhood is not less than the predefined minimum number (*MinPts*).

Definition 4 (Core Distance): Assuming that o is the core object, in its radius-neighborhood, the core distance is the shortest distance between itself and its *MinPts*th nearest neighbor object

$$cd(\text{Object}) = \begin{cases} \text{undefined}, & \text{if } |N_\varepsilon(o)| < \text{MinPts} \\ \text{dist}\left(o, N_\varepsilon^{\text{MinPts}}(o)\right), & \text{if } |N_\varepsilon(o)| \geq \text{MinPts} \end{cases} \tag{3}$$

where $N_\varepsilon^{\text{MinPts}}(o) \in N_\varepsilon(o)$ represents the *MinPts*th nearest neighbor object of o.

Definition 5 (Density-Reachable): An object p is density-reachable from the object q with respect to spatiotemporal, if there is a chain of objects $p_1, \ldots, p_n$, $p_1 = q$ and $p_n = p$, such that p_{i+1} is directly density-reachable from p_i with respect to spatiotemporal, for $1 \leq i \leq n$, $p_i \in$ Object.

Definition 6 (Reachability Distance): $p, o \in$ Object, the reachable distance of p with respect to o is defined as

$$rd(p, o) = \begin{cases} \text{undifined}, & \text{if } |N_\varepsilon(o)| < \text{MinPts} \\ \max\{cd(o), \text{dist}(o, p)\}, & \text{if } |N_\varepsilon(o)| \geq \text{MinPts}. \end{cases} \tag{4}$$

Specifically, when the point o is the core point, and $rd(p, o)$ represents the minimum radius-neighborhood.

B. Mixed Attribute Features Modeling

This section mainly summarizes the mixed attribute features modeling and the calculation of *COD*.

1) Part 1 (Attribute Features Representation): There are three types of variables, i.e., ordered categorical, disordered categorical and continuous variables, whose feature sets are represented as follows.

1) *Disordered Categorical Variables:*

$$\text{AF}_d = \left\{af_{d_1}, af_{d_2}, \ldots, af_{d_i}, \ldots, af_{d_n}\right\}. \tag{5}$$

AF_d is a feature set of n disordered categorical variables in two spatiotemporal objects, and af_{d_i} is the ith feature of disordered categorical variable, $i \in [1, n]$.

2) *Ordered Categorical Variables:*

$$\text{AF}_o = \left\{af_{o_1}, af_{o_2}, \ldots, af_{o_i}, \ldots, af_{o_n}\right\}. \tag{6}$$

AF_o is a feature set of n disordered categorical variables in two spatiotemporal objects, af_{o_i} is the ith feature of ordered categorical variable, $i \in [1, n]$.

3) *Continuous Variables:*

$$\mathrm{AF}_c = \left\{af_{c_1}, af_{c_2}, \ldots, af_{c_i}, \ldots, af_{c_n}\right\}. \tag{7}$$

AF_c is a feature set of n continuous variables in two spatiotemporal objects, af_{c_i} is the ith feature of continuous variable, $i \in [1, n]$.

2) Part 2 (Calculation of Attribute Feature Similarity): The similarity algorithms corresponding to categorical variables and continuous variables are as follows.

1) *Similarity of Categorical Variables: OD* is the result about the similarity of categorical variables

$$OD = O \cap D = \begin{cases} 0, & O = 0 \text{ or } D = 0 \\ 1, & O = D = 1. \end{cases} \tag{8}$$

The result about the similarity of attribute characteristics of disordered categorical variables (D)

$$D = \begin{cases} 0, & D_p \neq D_q \\ 1, & D_p = D_q \end{cases} \tag{9}$$

where p and q are two objects. Because there is no relationship between the eigenvalues of this kind of attributes, we can only consider whether the corresponding attribute features of two spatiotemporal objects are equal or not. If it is equal, the similarity is 1; if it is not equal, the similarity is 0. The similarity of ordered categorical variables (ΔO)

$$O = \begin{cases} 0, & \Delta O \leq \Delta O\_\text{threshold} \\ 1, & \Delta O > \Delta O\_\text{threshold} \end{cases} \tag{10}$$

where $\Delta O_$threshold is the threshold of ordered categorical variables. Besides, the Gower coefficient can be used to calculate the similarity of ordered categorical variables. The closer the Gower coefficient is to 1, the higher the similarity

$$\Delta O\left(\mathrm{AF_{OP}}, \mathrm{AF_{OQ}}\right) = \frac{1}{n}\sum_{i=1}^{n} \delta\left(af_{op_i}, af_{oq_i}\right) \tag{11}$$

where $i \in [1, n]$

$$\delta\left(af_{op_i}, af_{oq_i}\right) = 1 - \frac{\left|af_{op_i} - af_{oq_i}\right|}{\mathrm{Range}_i}$$
$$0 \leq \delta\left(af_{op_i}, af_{oq_i}\right) \leq 1.$$

Range_i is the range of the ith attribute eigenvalues of ordered categorical variables in spatiotemporal objects.

2) *Similarity of Continuous Variables:* The similarity of continuous variables (ΔC)

$$\mathrm{C} = \begin{cases} 0, & \Delta C > \Delta C\_\text{threshold} \\ 1, & \Delta C \leq \Delta C\_\text{threshold} \end{cases} \tag{12}$$

where the ΔC_threshold is the threshold.

Besides the Euclidean distance is used to calculate the feature similarity distance of the ordered categorical variables of two spatiotemporal objects, and the smaller the distance is, the more similar it is

$$\Delta C\left(\mathrm{AF}_{cp}, \mathrm{AF}_{cq}\right) = \sqrt{\sum_{i=1}^{n}\left(af_{cp_i} - af_{cq_i}\right)^2}, \mathrm{i} \in [1, n]. \tag{13}$$

Finally, our improved deflection elimination algorithm in wearable eye tracker scene also takes into account the fusion

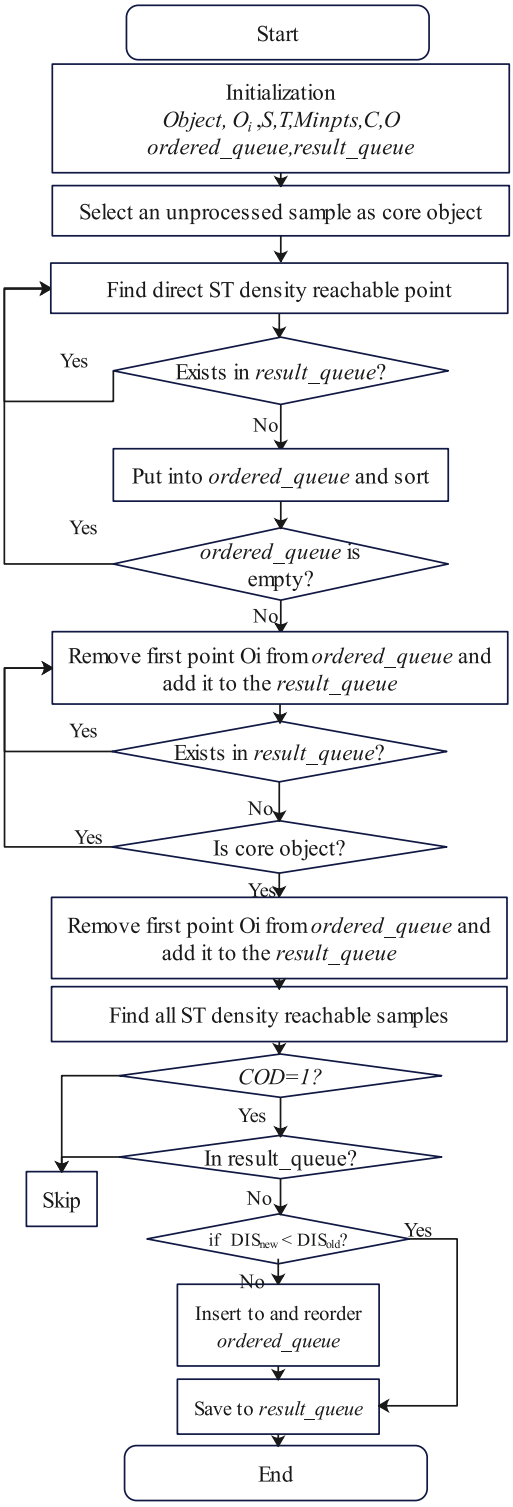


Fig. 6. RCDE framework.

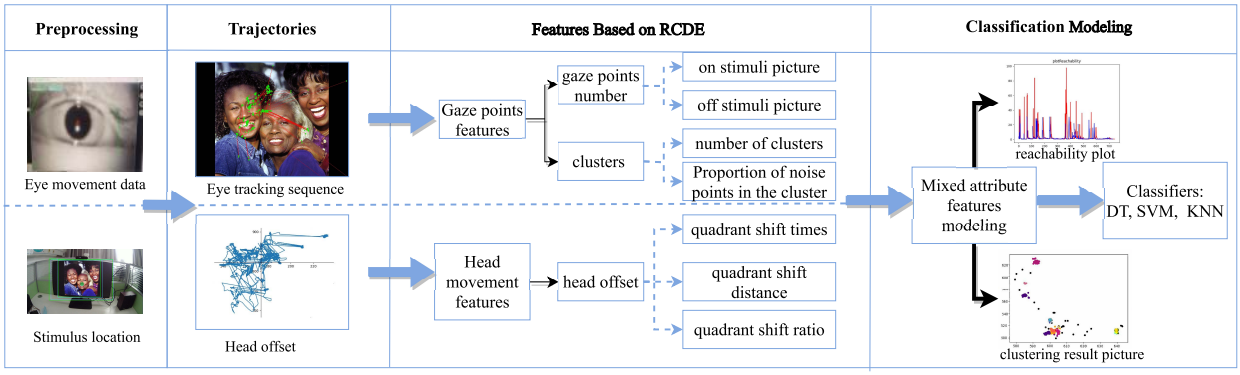


Fig. 7. Workflow of classification modeling with RCDE.

of different attribute features such that

$$\mathrm{COD} = C \cap OD = \begin{cases} 0, & C = 0 \text{ or } OD = 0 \\ 1, & C = OD = 1 \end{cases} \tag{14}$$

when the COD $= 1$, the features of the two objects are similar.

C. ROI Eventless Clustering and ROIs Clustering With Deflection Elimination

Existing eye-movement-based depression recognition research often requires behavior recognition to obtain eye movement events [26], [51], [52], such as fixation and saccade, and then extract eye movement events features as needed, such as statistical features or higher level semantic features [53], [54], [55], [56]. As one of the most frequent used high level semantic feature to indicate attention distribution or other cognitive pattern, ROIs are always extracted by two steps, one is eye movement event identification, and the other is to determine the regions by the eye movement events of fixation. This would obviously increases the computational complexity and brings more uncertainty, which is not conducive to later learning tasks. Therefore, fixation and ROI-based features are computations of different levels of visual distribution, which motivates us to explore the idea of extracting high-level features of visual distribution directly based on raw eye-tracking data. The method we proposed, which does not require eye movement event recognition, is called ROI Eventless Clustering (REC). The core idea of REC is shown in the middle part of Fig. 4.

The eye movement data has a large amount of noise data due to the influence of factors, such as the subjects' low compliance with instructions, blinking, and the instability of the gaze tracking algorithm. Those noisy data might affect the eye movement identification and following depression recognition learning. In order to avoid the influence of noise on the depression recognition task, this article proposes the RCDE method based on REC to eliminate noise during the clustering process. RCDE constructs a mixed-attribute feature model by combining head movement features in the process of extracting the subject's gaze ROIs. Then, the constructed model is used to calculate the similarity of different features, and the similarity results are sorted, so as to facilitate the deflection elimination of the subjects' fixation data, and obtain the corresponding ROIs.

The framework of RCDE is shown in Fig. 6. RCDE finally generated clustering objected data structure from raw gaze samples. The clusters generated by RCDE are hierarchical regions, we regarded the higher level of clusters, i.e., bigger clusters as ROI, and regarded the lower level of clusters as fixations, i.e., smaller clusters. Those hierarchical information can be extracted from the reachability plot.

At the stage of preparation, five parameters should be set, i.e., S_threshold, T_threshold, *MinPts*, C_threshold, and O_threshold, which are separately spatiotemporal thresholds. First, an unprocessed sample point as the core object is selected and its core distance and direct density reachability are found. If it exists in result queue, the other points according to their reachable density distance are sorted. If the ordered queue is not empty, then go back to the first stage. Otherwise, removing the first point from the ordered queue for expansion and save it to the result queue. Second, if the expansion point does not exist in the result queue, but it is the core point, finding its all direct density reachability points and calculating the similarity of attribute features (*COD*). If $COD = 1$, it is not in result queue and its direct density reachability point has been in the ordered queue. If $\mathrm{Dis}_{\mathrm{new}} < \mathrm{Dis}_{\mathrm{old}}$, reorder the ordered queue. Otherwise, the point is inserted, and the ordered queue is reordered.

Based on proposed methods, the classification modeling process of RCDE can be summarized as Fig. 7, which includes data preprocessing, feature extraction, and classification. To investigate the performance of our proposed methods, we tried REC and RCDE, respectively, in the feature extraction and classification phases.

IV. Results and Discussion

In order to analyze and verify the experimental results, we used traditional method as baseline. REC adopted classical density-based clustering algorithm. The classification models were trained by decision tree (DT), support vector machine (SVM), and *k*-NearestNeighbor (kNN) algorithms with ten-fold cross validation. It can be seen from Table IV that with the same features and classification methods, the REC-based method is generally higher than the traditional method. RCDE got the highest classification accuracy of 76.25%. Fig. 8 gives intuitive indicators comparison chart of various methods based

TABLE IV
ACCURACY, *F*1 SCORE, RECALL, AND PRECISION OF THE ORIGINAL METHOD AND SOME USED CLUSTERING ALGORITHMS

		Classifiers (%)	Accuracy (%)	F1 score (%)	Recall (%)	Precision (%)	AUC (%)
Traditional Method		DT	58.06	57.20	59.00	60.33	60.81
		SVM	62.36	60.43	61.00	69.45	70.50
		KNN	54.58	50.25	49.00	55.00	58.75
REC with Density-based Clustering Algorithms	DENCLUE [57]	DT	55.83	59.10	69.00	55.21	58.69
		SVM	59.52	55.01	56.25	59.68	65.80
		KNN	54.45	46.76	44.50	53.67	60.44
	CURD [58]	DT	55.69	63.51	77.50	56.30	50.31
		SVM	59.86	56.75	58.00	65.00	66.13
		KNN	53.47	47.35	45.00	55.17	61.25
	DBSCAN [59]	DT	58.89	57.66	62.50	59.88	57.25
		SVM	59.03	58.23	62.00	64.37	63.38
		KNN	53.47	46.28	44.50	60.95	59.25
	OPTICS [36]	DT	61.67	61.67	64.00	62.22	62.56
		SVM	69.17	68.43	72.50	72.79	71.75
		KNN	54.58	57.24	65.50	53.00	56.50
	ST-DBSCAN [50]	DT	62.36	60.63	58.50	68.57	64.75
		SVM	70.63	68.43	72.50	72.79	71.88
		KNN	55.56	49.67	48.00	55.17	57.56
	ST-OPTICS [37]	DT	68.19	61.79	55.00	72.50	71.06
		SVM	**74.03**	**73.48**	**75.00**	**75.17**	**74.13**
		KNN	63.75	65.60	70.00	63.00	68.31
RCDE		DT	62.50	59.08	56.00	67.67	63.00
		SVM	**76.25**	**76.20**	**79.50**	**76.83**	**75.39**
		KNN	67.22	70.24	79.50	63.38	66.94

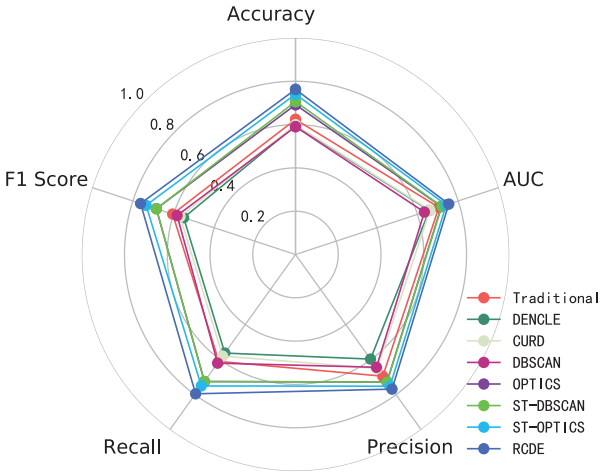


Fig. 8. Comparison of various methods based on SVM on multiple evaluation indicators.

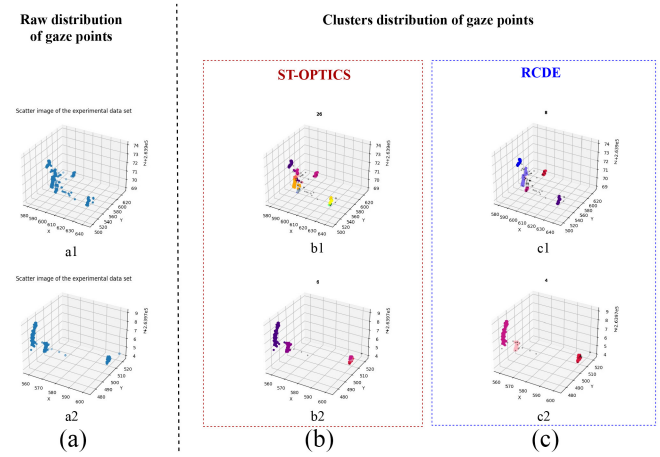


Fig. 9. Scatter picture of the experimental data set. (a) Raw distribution of samples, and (b) and (c) are the clustering plot of ST-OPTICS and RCDE, respectively. In which the *x*- and *y*-axes are the *x* and *y* position coordination of gaze points, and the *z*-axis is the time coordination of gaze points. Besides, the number at the top of the figure shows the number of clusters (except for the noise cluster), i.e., the number of ROI. In addition, in the picture, a cluster is marked with one color, where the black dots represent the identified noise points.

TABLE V
DETAILED PARAMETERS OF THE CLASSIFICATION ALGORITHM

Classifier	Parameters Optimization Setting
DT	parameters = 'splitter':('best', 'random'), 'criterion':("gini", "entropy"), "max_depth":[*range(1,10)], 'min_samples_leaf':[*range(1, 50, 5)]
SVM	c_range = [0.0001, 0.001, 0.01, 0.03, 0.1, 1, 10, 20, 50, 75, 100], gamma_range = [0.0001, 0.001, 0.01, 0.1, 1, 10, 20, 30, 100], param_grid = ['clf__C': c_range,'clf__kernel': ['linear'],'clf__C': c_range,'clf__gamma': gamma_range, 'clf__kernel': ['rbf']]
KNN	param = "n_neighbors": range(3, 15)

on SVM. The detailed parameters setting of classifier are shown in Table V.

As an unsupervised learning process, the clustering results in this article have no ground truth, and some internal indexes that measure the clustering performance cannot intuitively reflect the identification effect of the region of interest. Therefore, this article adopts a graphical qualitative comparison to visually show the clustering differences between the two methods. We established a 3-D coordinate system to visualize our gaze points data. Fig. 9 shows the distribution of gaze points in a stimuli picture before and after clustering. In this figure, Fig. 9(a) is the raw distribution of gaze points, Fig. 9(b) is the clusters' distribution obtained by ST-OPTICS,

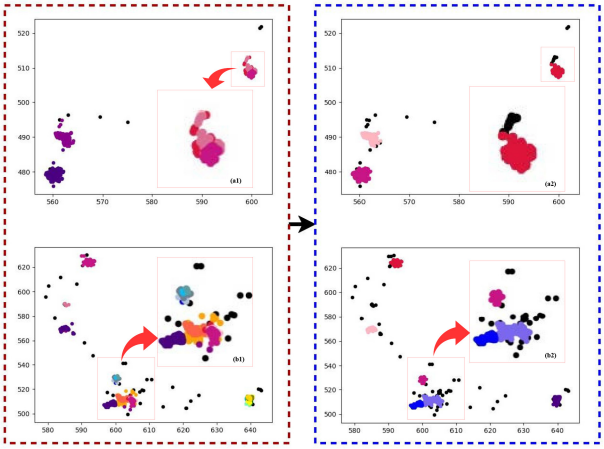


Fig. 10. Clustering results comparison of ST-OPTICS (red box) and our RCDE (blue box) algorithm.

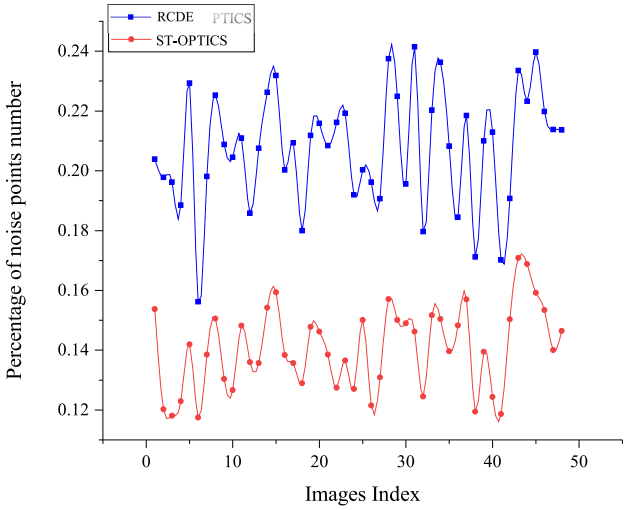


Fig. 11. Noise identified comparison of ST-OPTICS (red) and RCDE (blue). The figure shows the ratio of noise points number to all points number.

and Fig. 9(c) is the clusters' distribution obtained by RCDE algorithm. In order to more intuitively observe the advantages of our algorithm for noise detection, we further map the 3-D scatter picture to the 2-D plane, which can be seen in Fig. 10. In which we can directly observe that our algorithm has improvement in noise points detection. Furthermore, we made statistics on the proportion of noise points in Fig. 11, we can see that RCDE detected more noisy samples.

The reachability plot showed a hierarchical structure of clusters, in which the peaks indicate the boundaries between clusters, the valleys indicate the clusters. In Fig. 12, the red curves correspond to the reachability plots obtained by ST-OPTICS, the blue curves correspond to the reachability plots obtained by RCDE. The horizontal and vertical coordinates represent the sample point in the result queue and corresponding reachable distance, respectively. From these figures, we can observe that the blue curves formed finer waves, which contain more attention information. That might help during classification modeling.

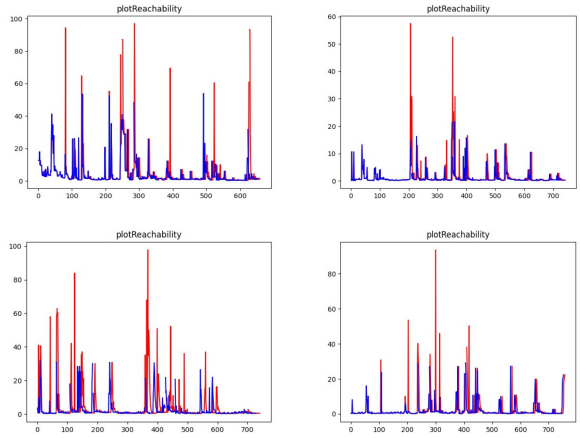


Fig. 12. Comparison figure of reachability plot.

V. Conclusion

With the motivation of analyzing the attention patterns of depression patients from raw eye movement data with considering head movement, this article proposed a novelty ROI analysis method based on spatiotemporal clustering. Experimental results showed that the proposed method REC outperformed the traditional methods, and RCDE achieved the highest accuracy of 76.25%, proved that our proposed methods are effective in identifying depression. From the qualitative and quantitative analysis, RCDE could generate finer and tighter clusters, which would help a better classification accuracy. In the future, we will try multimodal fusion with eye movement, ROI, head movement, and pupil diameter. The raw eye movement video data are also worth for further investigation.

References

[1] G. S. Malhi and J. J. Mann, "Depression," *Lancet*, vol. 392, no. 10161, pp. 2299–2312, 2018.
[2] W. Katon and P. Ciechanowski, "Impact of major depression on chronic medical illness," *J. Psychosomatic Res.*, vol. 53, no. 4, pp. 859–863, 2002.
[3] G. E. Simon, "Social and economic burden of mood disorders," *Biol. Psychiatry*, vol. 54, no. 3, pp. 208–215, 2003.
[4] A. Beck et al., "Severity of depression and magnitude of productivity loss," *Ann. Family Med.*, vol. 9, no. 4, pp. 305–311, 2011.
[5] E. Bromet et al., "Cross-national epidemiology of DSM-IV major depressive episode," *BMC Med.*, vol. 9, no. 1, pp. 1–16, 2011.
[6] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *J. Neurolingus*, vol. 20, no. 1, pp. 50–64, 2007.
[7] I. H. Gotlib, E. Krasnoperova, D. N. Yue, and J. Joormann, "Attentional biases for negative interpersonal stimuli in clinical depression," *J. Abnormal Psychol.*, vol. 113, no. 1, pp. 121–135, 2004.
[8] T. Zhang, S. Li, B. Chen, H. Yuan, and C. L. P. Chen, "AIA-net: Adaptive interactive attention network for text–audio emotion recognition," *IEEE Trans. Cybern.*, early access, Aug. 22, 2022, doi: 10.1109/TCYB.2022.3195739.
[9] T. Zhang, X. Wang, X. Xu, and C. L. P. Chen, "GCB-net: Graph convolutional broad network and its application in emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 379–388, Jan.–Mar. 2022.
[10] J. Shen, X. Zhang, B. Hu, G. Wang, Z. Ding, and B. Hu, "An improved empirical mode decomposition of electroencephalogram signals for depression detection," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 262–271, Jan.–Mar. 2022.

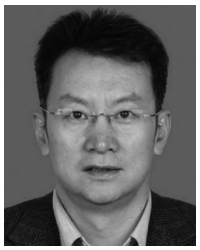
[11] L. He et al., "Deep learning for depression recognition with audiovisual cues: A review," *Inf. Fusion*, vol. 80, pp. 56–86, Apr. 2022.
[12] M. Yang, Y. Ma, Z. Liu, H. Cai, X. Hu, and B. Hu, "Undisturbed mental state assessment in the 5G era: A case study of depression detection based on facial expressions," *IEEE Wireless Commun.*, vol. 28, no. 3, pp. 46–53, Jun. 2021.
[13] E. Q. Wu, Z. Cao, P. Xiong, A. Song, L.-M. Zhu, and M. Yu, "Brain–computer interface using brain power map and cognition detection network during flight," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 5, pp. 3942–3952, Oct. 2022.
[14] E. Q. Wu et al., "Scalable gamma-driven multilayer network for brain workload detection through functional near-infrared spectroscopy," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12464–12478, Nov. 2022.
[15] B. T. Carter and S. G. Luke, "Best practices in eye tracking research," *Int. J. Psychophysiol.*, vol. 155, pp. 49–62, Sep. 2020.
[16] P. S. Holzman, L. R. Proctor, and D. W. Hughes, "Eye-tracking patterns in schizophrenia," *Science*, vol. 181, no. 4095, pp. 179–181, 1973.
[17] B. Catherine, C. Delphine, R. N. Salesse, and R. Stéphane, "Self-face recognition in schizophrenia: An eye-tracking study," *Front. Human Neurosci.*, vol. 10, p. 3, Feb. 2016.
[18] R. J. Jacob and K. S. Karn, "Eye tracking in human–computer interaction and usability research: Ready to deliver the promises," in *The Mind's Eye*. Amsterdam, The Netherlands: Elsevier, 2003, pp. 573–605.
[19] S. Djamasbi, "Eye tracking and Web experience," *AIS Trans. Human-Comput. Interact.*, vol. 6, no. 2, pp. 37–54, 2014.
[20] J. C. Romano Bergstrom, E. L. Olmsted-Hawala, and H. C. Bergstrom, "Older adults fail to see the periphery in a Web site task," *Universal Access Inf. Soc.*, vol. 15, no. 2, pp. 261–270, 2016.
[21] Z. Pan et al., "Cognitive impairment in major depressive disorder," *CNS Spectr.*, vol. 24, no. 1, pp. 22–29, 2019.
[22] Y. Li et al., "Eye movement indices in the study of depressive disorder," *Shanghai Arch. Psychiatry*, vol. 28, no. 6, pp. 326–334, 2016.
[23] M. Eizenman et al., "A naturalistic visual scanning approach to assess selective attention in major depressive disorder," *Psychiatry Res.*, vol. 118, no. 2, pp. 117–128, 2003.
[24] E. Goeleven, R. D. Raedt, S. Baert, and E. Koster, "Deficient inhibition of emotional information in depression," *J. Affect. Disorders*, vol. 93, nos. 1–3, pp. 149–157, 2006.
[25] A. Duque and C. Vázquez, "Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study," *J. Behav. Ther. Exp. Psychiatry*, vol. 46, pp. 107–114, Mar. 2015.
[26] A. Olsen, "The Tobii i-VT fixation filter," Tobii Technol., Danderyd Municipality, Sweden, White Paper, vol. 21, pp. 4–19, 2012.
[27] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye Tracking Res. Appl.*, 2000, pp. 71–78.
[28] R. Zemblys, D. C. Niehorster, and K. Holmqvist, "gazeNet: End-to-end eye-movement event detection with deep neural networks," *Behav. Res. Methods*, vol. 51, no. 2, pp. 840–864, 2019.
[29] A. Nguyen, V. Chandran, and S. Sridharan, "Visual attention based ROI maps from gaze tracking data," in *Proc. Int. Conf. Image Process.*, vol. 5, 2004, pp. 3495–3498.
[30] J. Coddington, J. Xu, S. Sridharan, M. Rege, and R. Bailey, "Gaze-based image retrieval system using dual eye-trackers," in *Proc. IEEE Int. Conf. Emerg. Signal Process. Appl.*, 2012, pp. 37–40.
[31] T.-H. Huang, K.-Y. Cheng, and Y.-Y. Chuang, "A collaborative benchmark for region of interest detection algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 296–303.
[32] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, pp. 5142–5154, 2018.
[33] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, Sep. 2000.
[34] A. Klami, C. Saunders, T. E. de Campos, and S. Kaski, "Can relevance of images be inferred from eye movements?" in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 134–140.
[35] Y. Zhang, H. Fu, Z. Liang, Z. Chi, and D. Feng, "Eye movement as an interaction mechanism for relevance feedback in a content-based image retrieval system," in *Proc. Symp. Eye Tracking Res. Appl.*, 2010, pp. 37–40.
[36] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM SIGMOD Rec.*, 1999, pp. 49–60. [Online]. Available: https://doi.org/10.1145/304182.304187
[37] K. P. Agrawal, S. Garg, S. Sharma, and P. Patel, "Development and validation of OPTICS based spatio-temporal clustering technique," *Inf. Sci.*, vol. 369, pp. 388–401, Nov. 2016.
[38] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Adjunct Publ.*, 2014, pp. 1151–1160.
[39] Y. Wang, J. Xu, B. Zhang, and X. Feng, "Native assessment of international affective picture system among 116 Chinese aged," *Chin. Mental Health J.*, vol. 22, pp. 903–907, Jan. 2008.
[40] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: Validity of a brief depression severity measure," *J. General Internal Med.*, vol. 16, no. 9, pp. 606–613, 2001.
[41] J.-M. Azorin, P. Benhaim, T. Hasbroucq, and C.-A. Possamaï, "Stimulus preprocessing and response selection in depression: A reaction time study," *Acta Psychologica*, vol. 89, no. 2, pp. 95–100, 1995.
[42] B. N. Cuthbert, "International affective picture system (IAPS): Instruction manual and affective ratings," Dept. Center Res. Psychophysiol., Univ. Florida, Gainesville, FL, USA, Rep. A-4, 1999.
[43] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2010.
[44] D. Sabatinelli, M. M. Bradley, P. J. Lang, V. D. Costa, and F. Versace, "Pleasure rather than salience activates human nucleus accumbens and medial prefrontal cortex," *J. Neurophysiol.*, vol. 98, no. 3, pp. 1374–1379, 2007.
[45] D. H. Zald, "The human amygdala and the emotional evaluation of sensory stimuli," *Brain Res. Rev.*, vol. 41, no. 1, pp. 88–123, 2003.
[46] T. M. Libkuman, H. Otani, R. Kern, S. G. Viger, and N. Novak, "Multidimensional normative ratings for the international affective picture system," *Behav. Res. methods*, vol. 39, no. 2, pp. 326–334, 2007.
[47] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[48] S. Alghowinem et al., "Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 478–490, Oct.–Dec. 2018.
[49] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford, U.K.: OUP, 2011.
[50] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial–temporal data," *Data Knowl. Eng.*, vol. 60, no. 1, pp. 208–221, 2007.
[51] V. Krassanakis, V. Filippakopoulou, and B. Nakos, "EyeMMV toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification," *J. Eye Movement Res.*, vol. 7, no. 1, pp. 1–10, 2014.
[52] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vis. Res.*, vol. 45, no. 5, pp. 643–659, 2005.
[53] Z. Pan, H. Ma, L. Zhang, and Y. Wang, "Depression detection based on reaction time and eye movement," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 2184–2188.
[54] A. Lazarov, Z. Ben-Zion, D. Shamai, D. S. Pine, and Y. Bar-Haim, "Free viewing of sad and happy faces in depression: A potential target for attention bias modification," *J. Affect. Disorders*, vol. 238, pp. 94–100, Oct. 2018.
[55] C. M. Bodenschatz, M. Skopinceva, T. Ruß, and T. Suslow, "Attentional bias and childhood maltreatment in clinical depression—An eye-tracking study," *J. Psychiatric Res.*, vol. 112, pp. 83–88, May 2019.
[56] S. Soltani, K. Newman, L. Quigley, A. Fernandez, K. Dobson, and C. Sears, "Temporal changes in attention to sad and happy faces distinguish currently and remitted depressed individuals from never depressed individuals," *Psychiatry Res.*, vol. 230, no. 2, pp. 454–463, 2015.
[57] A. Hinneburg and H.-H. Gabriel, "Denclue 2.0: Fast clustering based on kernel density estimation," in *Proc. Int. Symp. Intell. Data Anal.*, 2007, pp. 70–80.
[58] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," *ACM SIGMOD Rec.*, vol. 27, no. 2, pp. 73–84, 1998.
[59] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Min.*, 1996, pp. 226–231.

Minqiang Yang (Member, IEEE) received the Ph.D. degree in computer science from Lanzhou University, Lanzhou, China, in 2022.

He is an Engineer with the Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University. His current research interests include affective computing, image processing, machine learning, and automatic depression detection.

Chenlei Cai received the bachelor's degree from Beibei Normal University, Shijiazhuang, China, in 2019. She is currently pursuing the master's degree with the School of information Science and Engineering, Lanzhou University, Lanzhou, China.

Her main research interests include affective computing and machine learning.



Bin Hu (Senior Member, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Science, Beijing, China, in 1998.

He was a recipient of many research awards, including the 2014 China Overseas Innovation Talent Award, the 2016 Chinese Ministry of Education Technology Invention Award, the 2018 Chinese National Technology Invention Award, and the 2019 WIPO-CNIPA Award for Chinese Outstanding Patented Invention. He is also the TC Co-Chair of Computational Psychophysiology in the IEEE Systems, Man, and Cybernetics Society (SMC), the TC Co-Chair of Cognitive Computing in IEEE SMC, and the Vice-Chair of the TC 9.1. Economic, Business, and Financial Systems on Social Media at the International Federation of Automatic Control. He is also a Member-at-Large of the ACM China Council and the Vice-Chair of the China Committee of the International Society for Social Neuroscience. He serves as the Editor-in-Chief for IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS and an Associate Editor for IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.