Universal Access in the Information Society (2019) 18:551–563
https://doi.org/10.1007/s10209-019-00678-7

LONG PAPER

Beyond engagement: an EEG-based methodology for assessing user's confusion in an educational game

Yun Zhou[1] · Tao Xu[2] · Shaoqi Li[2] · Ruifeng Shi[2]

Published online: 22 July 2019

Abstract
Confusion is an emotion, which may occur when the learner is confronting inconsistence between new knowledge and existing cognitive structure, or reasoning for solving the puzzle and problem. Although confusion is not pleasant, it is necessary for the learner to engage in understanding and deep learning. Consequently, confusion assessment has attracted increased interest in e-learning. However, current studies have targeted no further than engagement detection and measurement, while there is lack of studies in investigating cognitive and emotional aspects beyond engagement in the context of game-based learning. To quantify confused states in logic reasoning in game-based learning, we propose an EEG-based methodology for assessing the user's confusion using the OpenBCI device with 8 channels. In the complicated context of game play, it is difficult, and sometimes impossible, to collect the ground truth of the data in real tasks. To solve this issue, this work leverages cross-task and cross-subject methods to build a classifier, that is, training on the data of one standardized cognitive test paradigm (Raven's test) and testing on the data of real tasks in game play (Sokoban Game). It provides a new possibility to create a classifier based on a small dataset. We also employ the end-to-end algorithm of deep learning in machine learning. Results showed the feasibility of this proposal in the task variation of the classifier, with an accuracy of 91.04%. The proposed EEG-based methodology is suitable to analyze learners' confusion on the long game-play duration and has a good adaption in real tasks.

Keywords Educational game · Game-based learning · Electroencephalography (EEG) · Assessment · Confusion · Machine learning

1 Introduction

Educational games refers to games designed for a primary purpose of pedagogy rather than pure entertainment, which provides learners plenty of opportunities to put into practice what they have learned in a game-based context, supports the development of cognitive, practical, and social skills [18, 22, 29, 33], and involves the activities of problem-solving and competition. Although digital learning provides students an environment supporting them to develop cognitive skills, there is still a lack of understanding about how games foster such skills. Finding an effective methodology that detects affective and cognitive processes of students during game learning can pinpoint whether these games offer the settings with the appropriate difficulty levels, satisfy requirements of learning, and achieve the initial pedagogical objectives.

Assessment after learning in a game-based environment often concentrates on the outcome or performance [1, 51]. Such assessment methods may neglect changes during the learning process, which are mainly related to cognitive and affective processes. In recent years, physiological measures have drawn attention to game play assessment with respect to cognitive aspects [48]; most notably, EEG-based measurements have potentials and benefits distinguished from others [49]. EEG technology provides a direct means for internal

✉ Yun Zhou
zhouyun@snnu.edu.cn

✉ Tao Xu
xutao@nwpu.edu.cn

Shaoqi Li
lsqylxq@163.com

Ruifeng Shi
sruifeng@126.com

[1] School of Education, Shaanxi Normal University, Xi'an 710062, People's Republic of China

[2] School of Software, Northwestern Polytechnical University, Xi'an 710072, People's Republic of China

brain activity detection, revealing patterns of the internal states of the brain, compared with other recognition methods of cognitive states which are based on facial expression patterns using computer vision technology. Also, it has a good temporal resolution, which can be used for real-time detection in real tasks without distracting students, as compared with self-reports. Due to these bright prospects, EEG-based recognition methods of mental states have gained increased focus. With the emergence of portable commercial head-mounted devices for raw EEG data acquisition like Emotiv EPOC, NeuroSky MindWave, and OpenBCI 3D printed devices, such technology has been considered to measure and analyze the learner's engagement in educational games.

Engagement is regarded as a mediator between students' emotions and their achievement, which is categorized into five types including cognitive, motivational, behavioral, cognitive-behavioral, and social-behavioral engagement [32]. It is considered as the cognitive process related to attention (cognitive engagement) [43] and also as a main motivation indicator [37, 38]. Studies on such EEG-based evaluation tools so far have mainly targeted engagement (motivational engagement) detection [12, 13], while there is a lack of studies in investigating cognitive and emotional aspects beyond engagement.

Among those cognitive skills that the game supports, rules induction and reasoning skills are advanced skills, crucial for deep learning [2]. Confusion is an emotion, which is provoked due to the cognitive disequilibrium in learning [14]. It occurs when the individual's current cognitive structure is inconsistent with the new coming information [10], or when s/he is unable to move further while doing rule-based reasoning or solving a puzzle [36]. Once the learner fails to solve the puzzle and stays in confusion for a long time, s/he will fall into frustration and then boredom. Although confusion is unpleasant, it can foster the individual to engage in a high level and reflect. It has been proved that learners who are confused would be more vigilant and process the material at deeper levels of comprehension than learners who are not confused [25]. Therefore, measuring confusion paves the way for monitoring learner's internal reaction in the process of solving the problem and could be used to inform the design of educational games or game-based learning when adjusting the setting of difficulty appropriately.

When measuring confusion in educational games using machine learning techniques, there will be one issue that one may encounter: it is difficult, and sometimes impossible, to make the ground truth, namely class labels, of the data in real tasks, when the user is playing the game. The answers in the self-report or questionnaire are often used to obtain class labels. For example, in the work of [27, 30, 50], answers of the Self-Assessment Manikin [5] were used to give the ground truth, or in another common way, the ground truth was obtained based on types of stimuli. These methods of collecting labels are more appropriate for short-time tasks in standardized experiments rather than long-time real tasks like in educational games. In educational game play, it is common for learners to solve the puzzle spending thirty minutes or even more time. It is inappropriate to assign a label to a piece of EEG data during such long time. One of the possible solutions is to segment the data into small pieces of few seconds and assign one label to each piece of data. However, letting participants to segment the process of puzzle solving and identify their cognitive states for each piece of data would not only increase the labor cost to assign labels and check, but also introduce more errors. The assigned labels in this way may not be the ground truth. In an attempt to solve these issues, in this study, we addressed the key research question as follows: Is it feasible to leverage a cross-task and cross-subject method to build the classifier for detection of confusion in the educational game or game-based learning?

In this paper, we propose a novel EEG-based methodology to detect confusion in the context of game-based learning and demonstrate its advantages. We collect EEG time-series from the OpenBCI device with 8 channels and leverage a cross-task and cross-subject method to build a classifier based on machine learning, that is, training on the data collecting from one standardized cognitive test paradigm (Raven's test) and testing on the data from real tasks in the game play (Sokoban Game). Results showed the robustness of this proposal in the task variation of the classifier, with the accuracy up to 91.04%. The proposed EEG-based methodology is suitable to detect learners' confusion on the long game-play duration.

Addressing the problems as we stated above, the main contributions of this study are as follows:

- This work proposes detecting confusion states of students in the play of digital educational games, and we discuss the necessity and significance of confusion detection in educational game play. The EEG-based methodology that we proposed can recognize students' confusion in game-based learning when they are doing logic reasoning;
- With respect to the assessment of the game, this work proposes using EEG-based technology, revealing the internal states of brain directly and having a potential of supporting real-time detection due to the good temporal resolution;
- This work proposes leveraging a cross-task and cross-subject method to build a classifier based on machine learning, that is, training on the data collecting from one standardized cognitive test paradigm (Raven's test) and testing on the data from real tasks in the game play (Sokoban Game). Results showed the robustness of this proposal in the task variation of the classifier. Furthermore, end-to-end learning is proposed to decode confusion states from raw EEG data, which offers the benefit

of not considering handcrafted features. In this paper, we also describe how to design and train deep learning with convolutional neutral networks.

The rest of the paper is organized as follows: We briefly review the assessment of educational games, theories of confusion and machine learning approaches in EEG-based classification in Sect. 2. We discuss the proposed methodology and the architecture of building the classifier for detecting confusion states in educational games based on end-to-end learning approach in Sect. 3 and the design and setup of the experiment in Sect. 4. We present and discuss the results in Sect. 5, and the limitations and challenges of EEG-based brain–computer interfaces (BCI) in education in Sect. 6. In Sect. 7, we conclude proposing future research objectives.

2 Related work

In this section, we summarize and discuss the research work inspiring our proposal of detecting confusion states in educational game using EEG-based methodology and machine learning. First, we present studies on genres of educational games, skills that games offer, and the assessment in game-based learning. Then, we discuss the recent research on the theories of confusion, existing detection methods, and difficulties in confusion induction. Finally, we survey EEG-based detection methods and the endeavors in transferring advances from machine learning to EEG analysis.

2.1 Assessment of educational games

Educational games encompass a variety of genres that can be categorized based on the levels of psychological engagement [7]. Rapid response games, involving low levels of psychological engagement, are well suited for automated skills training through repeated practice. However, the goal of a large amount of educational games, involving high levels of psychological engagement, is learning of cognitive skills [33]. These games provide the activities supporting deep learning [17], including reasoning, problem-solving, and decision making. An effective educational game must align with learning goals, activities, feedback, interfaces, and the desired instructional outcomes [7]. Once the game elements are antagonistic to the learning objectives, the intended learning will not occur. To ensure a good design requires not only validated measures of learning outcomes, but also assessment methods to detect changes during the learning process in order to determine which design elements work best, when, and why. Therefore, rather than a final outcome or performance, learning emotions and cognitive aspects measurement and assessment can expose changes during the learning process.

Different methodologies revolving around the measurement and assessment of emotions and cognitive aspects in learning have been proposed and studied, and include five types [49] according to the measuring techniques: self-reported measures, observer's reports, behavior detection, interactions, and physiological measures.

Self-reported measures and observer's reports are subjective methods, while behavior detection, interactions, and physiological measures are objective methods [49]. The self-reported measures are commonly based on the questionnaire, subjectively collecting learners' attitudes, opinions, thoughts, etc., which are filled out during or after tasks. When collecting the data during tasks, these measures might interrupt the playing or performing tasks. Among those scales used in self-reports for assessing emotions, Self-Assessment Manikin (SAM) [5] scales have been widely used to assess emotions on the affective valence and arousal dimensions. SAM is a nonverbal design assessment based on pictorial rating, the studies of which found that subjects selected the emotion level faster and more directly using SAM than verbal scales. Observer's reports refer to reporting the learner's affective or cognitive states through the observation of another person rather than the learner, which are usually based on reading facial expression of the learner from videos [20]. Most of self-reported measures and observer's reports are the assessment after learning in a game-based environment, which may neglect changes and variation related to cognitive and affective processes during the learning process.

Behavior detection methods refer to recognition of facial expression, gestures and postures, speech and voice, eye tracking and gaze, etc. In e-learning and educational games, facial expressions recognition based on computer vision technologies is one of the most important measures and has been used commonly. Facial expressions expose the internal patterns of brain activity and are considered to have connection with emotions and cognitive states. For example, Whitehill et al. [47] studied whether human observers can reliably judge engagement from the face and analyzed the signals that observers use to make these judgments. They explored approaches for automatic recognition of engagement from students' facial expressions and found that automated engagement binary classification (two levels: high and low) performed with comparable accuracy to humans. Interactions methodologies are based on the analysis of interactions of learners, such as typing speed and semantic analysis of assignment.

Physiological measures are based on the recording and analyzing physiological signals, which detect emotions and cognitive states in an objective way. Physiological measures like electroencephalograph (EEG), near infrared (NIR), galvanic skin response (GSR), blood volume pulse (BVP) have good temporal resolution, while functional magnetic resonance imaging (fMRI) technology has good spatial resolution with respect to detecting and analyzing mental states. Physiological measures are leveraged commonly with self-reporting questionnaires. The data of self-reporting questionnaires include three facets: (1) to be used as ground truth when processing physiological signals [28, 45, 50], (2) to be compared with recognition results of physiological methods [45], and (3) to be used together with physiological signals for analyzing [6]. Wang et al. [45] employed Massive Open Online Courses video clips as the confusion stimuli to build a classification model to classify whether the student is confused or not when watching the course material, with an accuracy around 60%. Results from the questionnaire showed that the stimuli were supposed to be confusing but participants found them not confusing, which might be one of main reasons that the classification model was not well performed. In [6], Chanel et al. proposed an approach to maintain player engagement by adapting game difficulty according to the player's emotions assessed from physiological signals. They analyzed the questionnaire responses, EEG signals, peripheral signals (including GSR, BVP, heart rate, chest cavity expansion, and skin temperature) of the players playing a Tetris game at three difficulty levels and obtained a classification accuracy of 63% with fusing two EEG signals and peripheral signals.

2.2 Confusion in learning

Understanding confusion in learning theoretically and empirically has been the focus in recent years. In some related research work, confusion has been considered as an emotion. D'Mello et al. regarded confusion as a knowledge or an epistemic emotion that occurs during complex learning tasks [10]; the similar argument can also be found in [32, 40]. A few works considered confusion as non-affective feeling although it has characteristic feeling or experiential aspects [8]. A learner may experience confusion as an affect response to the cognitive processing of information [32]. It occurs as feedback when the individual is unable to move further. One instance of confusion is that existing cognitive structure is inconsistent with the new information [10]. Another common instance is that the individual cannot infer the rules when doing rule-based reasoning or solving a puzzle [36]. Instances or scenarios of confusion show difference [10]; thus, the stimulus should be carefully designed to conform to the target instances or scenarios.

On the one hand, although confusion is unpleasant, it can foster the individual to engage in a high level of learning and reflect profoundly. It has been proved that learners who are confused would be more vigilant and process the material at deeper levels of comprehension than learners who are not confused [25]. On the other hand, once the learner fails to solve the puzzle and stays in confusion for a long time, s/he will fall into frustration and then boredom [9]. Therefore, measuring confusion paves the way for monitoring the learner's internal reaction in the process of solving the problem and could be used for informing the design of educational games when adjusting the setting of difficulty appropriately.

In the study of confusion detection, confusion induction is a daunting task, which should be considered carefully. The appropriate induction determines the success of classification. Although researchers endeavored to induce the emotion accurately, the gap between pre-assigned stimulus and induced emotions still exists. In the work of [45], Wang et al. employed Massive Open Online Courses video clips as stimulus to evoke confused or non-confused states of students. After the experiment, they found stimulus materials were supposed to be confusing but participants found them not confusing. Besides, the observers who gave the labels for training classifier were not formally instructed. All of these may lead to inaccuracy of classification.

Besides, with regard to the type of the stimulus materials, pictures, sounds, video clips, interactive items are used to evoke emotions or cognitive states. Standardized and non-standardized databases of movie clips have been constructed such as [28] for general emotion induction like joy, amusement, or fear, although the number is still a few. To evoke learning emotions, tests, pedagogical contents, pictures, sounds, and courses video clips have been leveraged. In [26], pedagogical content was used to trigger confusion, frustration, anxiety, curiosity in one-to-one expert tutoring sessions. Wang et al. [45] used selected online courses video clips to trigger confused and non-confused states in learning. In [25], four computer learning environments that were well designed with AutoTutor have been developed to elicit confusion in learning.

2.3 EEG-based detection methods

EEG, as one representation of the brain's electrical activities, has been widely used to measure activities or states like working memory [15, 23], engagement [43], happiness [28], or stress [39]. In e-learning, engagement and motivation are the states that have been investigated mostly. In a survey of portable EEG technology in educational research [48], twenty-two papers were coded and discussed, and all except one revolved around the attention or motivation recognition for all five research topics, including interactive behavior, reading context, e-learning, presentation patterns of learning materials, and edutainment. Pekrun and Linnenbrink-Garcia [32] discussed academic emotions and student engagement, and summarized five types of engagement, that is, cognitive, motivational, behavioral, cognitive-behavioral, and social-behavioral engagement. Cognitive engagement (like attention) refers to the process of selectively concentrating on a discrete aspect of information [2]. Engagement has been considered as the indicator of motivation (also named as motivational engagement). For example, in [13], Ghergulescu et al. proposed a real-time EEG-sensor-based learner motivation analysis methodology in game-based learning. Engagement is not the only aspect that should be focused. However, limitation exists due to a lack of studies in investigating other aspects beyond engagement. There is considerable room for studies and advances with respect to assessment of cognitive and affective aspects like confusion, the importance of which is stated in Sect. 2.2.

In the last two decades, machine learning techniques for the analysis of EEG signals have been regarded as novel tools, which allow extracting features from EEG data and then performing classification or prediction [21]. As EEG being complicated in nature, many machine learning methods have been involved in this domain. In many EEG-based systems, machine learning techniques work as a central component and meet the requirements of neuroscience of brain signals decoding. The classic methods contain three steps: the preprocessing data, time–frequency analysis, and classification [21]. The goal of classification is to find the relationship between EEG and brain activities. Among classic machine learning techniques, support vector machine (SVM), which has good performance and suitability for small sample size, has been used as one of the most common tools for the classification of EEG signal [27, 30, 46]. In [46], Wang et al. designed and built a system of positive and negative emotions recognition using EEG signals. Their system achieved an accuracy up to 78.41% using SVM. The power spectrum feature, wavelet feature, and nonlinear dynamical feature were extracted and principal component analysis (PCA), linear discriminant analysis (LDA), and correlation-based feature selector (CFS) methods were used to reduce dimensions. SVM has been proved to be effective in EEG signals classification; however, the step of a priori feature extraction and selection is inevitable. In brain signal decoding, not all relevant features can be foreseen clearly, especially for confusion, which is an emerging research topic in the direction of brain decoding and has not been fully studied.

Deep learning methods or deep neural networks have greatly improved the performance of supervised learning in many domains like speech recognition and visual object recognition, which “allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction” [24]. Due to the great success in recognition tasks within a wide range of applications, these methods have gained great interest by researchers to address problems in EEG signals decoding and classification. Deep learning with convolutional neural networks (CNNs) proposed by LeCun et al. [24] can learn inherent patterns from data and objects automatically through end-to-end learning rather than employing prior extracted features. Differing from traditional machine learning methods, CNN provides a possibility to directly jump to the third step of EEG classification as stated above. End-to-end deep learning [24] method learns from the raw data and replaces multiple steps with just a single neural network, reducing the process of feature extraction.

The EEG signals are complex and weak, intertwined with noises; thus, it is hard to directly uncover the underlying essence from raw EEG data using traditional methods. However, the end-to-end method can map raw data directly to objectives. To take advantages of this, increased work adopt end-to-end CNN to analyze the EEG raw data, with decoding problems covering imagined movement classification [35, 41, 44], mental load recognition [4, 19], cognitive performance [16], memory prediction [42], seizures detection [3], etc. In their work [4] of modeling cognitive events from EEG data, Bashivan et al. transformed EEG data into topology-preserving multispectral images and trained a deep recurrent-convolutional network to learn representations from images. Their empirical evaluation on the cognitive load classification task showed significant improvements in classification accuracy. Hajinoroozi et al. [16] proposed a channel-wise convolutional neural network (CCNN) to predict driver’s cognitive states related to driving performance using EEG signals. Results showed that CCNN and CCNN variation achieved robust and improved performance.

3 Proposed methodology for confusion assessment in game-based learning

In this section, we present the core of our proposed EEG-based confusion detection methodology at first and then discuss the major components in details.

3.1 The core of the methodology

This paper proposes a novel noninvasive portable EEG-based learner confusion analysis methodology that is to be used in game-based learning systems or educational game context. The methodology is built upon four major components, including the experiment of confusion evoking, data collecting from EEG acquisition device, data preprocessing, and classifier building. Among these four major steps, the experiments of confusion induction and classification model building are essential. Figure 1 shows the core of our proposed EEG-based confusion detection methodology.

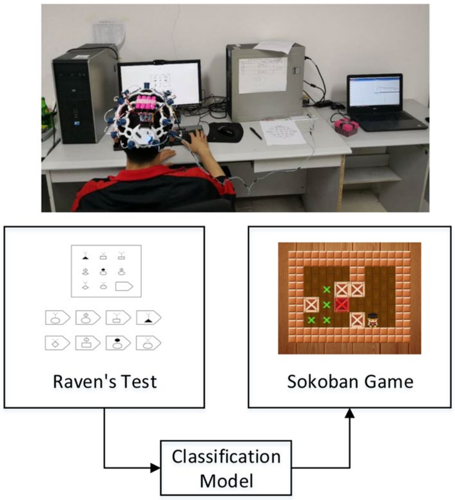


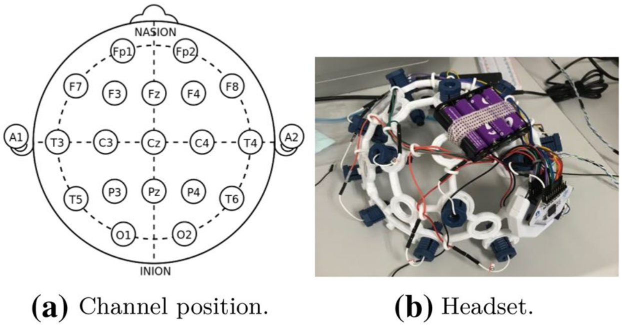
Fig. 1 The core of the EEG-based confusion detection methodology: training on the data from one standardized cognitive test paradigm (Raven's test) and testing on the data from real tasks in the game play (Sokoban Game)

3.2 Major components

Figure 2 shows the major components of building the confusion states classifier for an educational game based on end-to-end deep learning algorithm in machine learning.

Components 1: Experiment design and set-up The experiment will be described in Sect. 4. The experiment serves for data acquisition, preprocessing, and classifier building and delivers the data to the latter components.

Components 2: Data acquisition We record the learner's confusion while s/he doing the logic reasoning in the Raven's test and Sokoban Game, leveraging the OpenBCI EEG data acquisition device. The noninvasive medical EEG data acquisition devices are expensive and not easy to use, requiring using conductive paste to stick EEG electrodes directly to the skin. It usually takes from 20 minutes to one hour to wear the cap, depending on the numbers of channels that would be used. Besides, it requires the cable to transmit the data. These limitations keep these devices away from being used in the game play. Other types of devices like Emotiv, NeuroSky and OpenBCI are portable and use the Bluetooth or Wi-Fi to transmit the data to the computer. Among them, OpenBCI, is an open-source brain–computer interface (BCI) device, the software and hardware of which can be modified and developed as needed, providing more opportunities for researchers. In this work, we use OpenBCI Cyton board with the 3D printed headset to acquire raw EEG data (as shown in Fig. 3), and the data is delivered to the computer via Bluetooth. The neuro-headset featured 8 channels (Fp1, Fp2, C3, C4, T5, T6, O1, and O2) plus 2 references (A1 and A2) based on the 10-20 format. The trigger function and hardware were implemented to segment the data.



(a) Channel position. **(b)** Headset.

Fig. 3 Data acquisition device

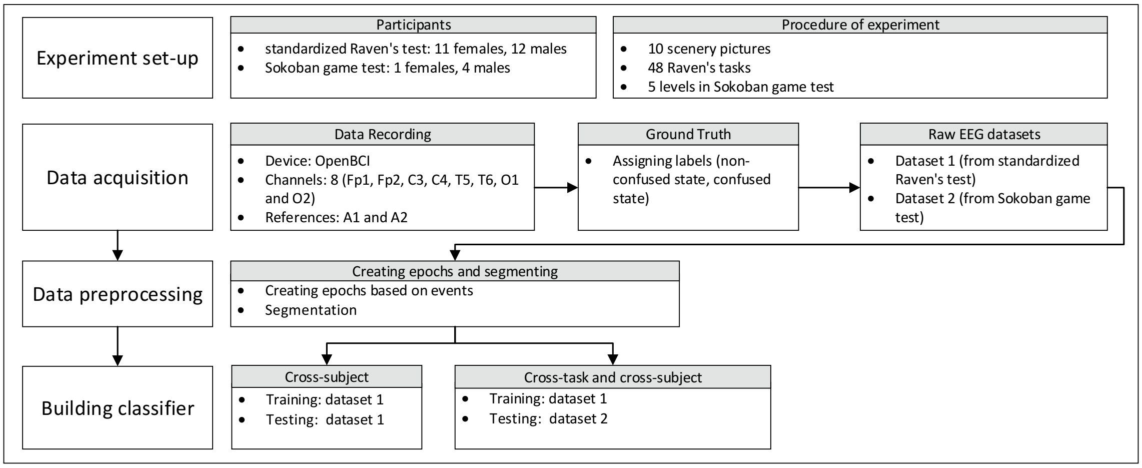

Experiment set-up
Participants
standardized Raven's test: 11 females, 12 males
Sokoban game test: 1 females, 4 males
Procedure of experiment
10 scenery pictures
48 Raven's tasks
5 levels in Sokoban game test
Data acquisition
Data Recording
Device: OpenBCI
Channels: 8 (Fp1, Fp2, C3, C4, T5, T6, O1 and O2)
References: A1 and A2
Ground Truth
Assigning labels (non-confused state, confused state)
Raw EEG datasets
Dataset 1 (from standardized Raven's test)
Dataset 2 (from Sokoban game test)
Data preprocessing
Creating epochs and segmenting
Creating epochs based on events
Segmentation
Building classifier
Cross-subject
Training: dataset 1
Testing: dataset 1
Cross-task and cross-subject
Training: dataset 1
Testing: dataset 2

Fig. 2 Major components of building the confusion states classifier for educational game based on end-to-end learning algorithm of deep learning in machine learning

Components 3: Data preprocessing Before processing the data, it is necessary to segment the data into desirable pieces and assign the data to train and to test. The segmentation, assignment, and implementation will be discussed in Sect. 5.

Components 4: Building Classifier In spite of its importance, detecting confusion states of game-based learning is hard to be implemented in practice. Trials of game-based learning usually take some time to complete and introduce diverse electromyography (EMG) artifacts into EEG data due to moving arms for interaction. This makes the preprocessing step complicated and requires extra experiments to remove such diverse EMG noises, that is, valid data are difficult to obtain. In the experiment to evoke confusion states based on standardized Raven's test, subjects were asked to select the right answer by using only one finger. This can reduce artifacts in maximum. Besides, it is easy to record and provide the benchmark to evaluate knowledge-free confusion states of logic reasoning. The patterns and features hidden in the EEG signals of confusion states in different kinds of activities are similar, while the differences are the artifacts produced by various activities. The basic idea of transfer learning [31] of deep learning is that early layers usually represent generic features, while later layers describe specific features. Due to the relative small size of the EEG dataset, it is unsuitable to use transfer learning directly. Inspired by transfer learning, if the learning model can benefit from the basic experiment in good condition (standardized Raven's test) at first, it will make a good performance in experiment in a complex condition (Sokoban Game).

Based on this idea, we propose a methodology leveraging cross-task and cross-subject methods to build classifiers based on end-to-end learning with convolutional neural networks (ConvNets). Cross-task refers to training on the data collecting from one standardized cognitive test paradigm (Raven's test) and testing on the data from real tasks in the game play (Sokoban Game). Cross-subject refers to dividing subjects into three groups (that is, group one, group two, and group three) and letting group one take part in the first task (Raven's test), group two take part in the second task (Sokoban Game), and group three take part in both tasks (Raven's test and Sokoban Game). Since the task based on the Sokoban Game is complex, the valid data cannot be easily obtained. Compared with group one, groups two and three are relatively small. We attempt to find confusion states of group two based on the model trained using the data of group one and group three. The main idea contains two steps. The first step is to build the datasets on the experiment of standardized Raven's test and the game, respectively. The second is to build the learning model based on the training data containing all of the data from the Raven's test and part of the data from Sokoban Game. In short, our approach trains a model on the mixed data from standardized Raven's test and Sokoban Game test and provides a prediction of confusion states on the Game test. It uses a few of labeled data from the experiment of the Sokoban Game to achieve a good performance based on the cross-task and cross-subject model in the educational game.

The learning model is built based on ConvNets. The basic structure is as shown in Fig. 4. One hidden layer is constructed by three kinds of parts: convolution, activation, and pooling. EEG data collecting by the OpenBCI are considered as the input, and the confused or non-confused states (two confusion states) are considered as the output.

4 Experiment design and setup

In this section, we present the experiment design and setup, including participants' basic information, two experiments, stimuli of Raven's test and Sokoban Game, and the procedure.

4.1 Stimulus design

Overall, the whole experiment is composed of two parts: the experiment using Raven's test as stimuli, and the experiment in which participants play the Sokoban Game, as shown in Fig. 5.

Experiment 1: Raven's Test To build the classification model, we adopt Raven's Matrices family of tests to evoke

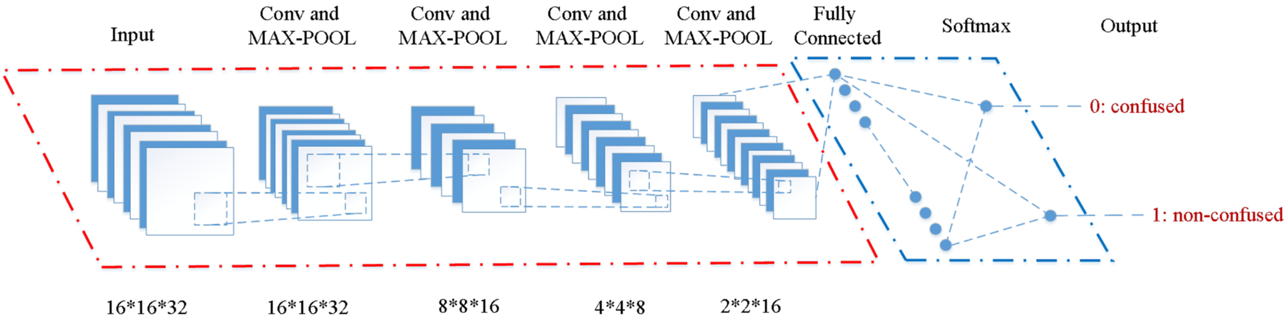


Fig. 4 The main structure of ConvNets

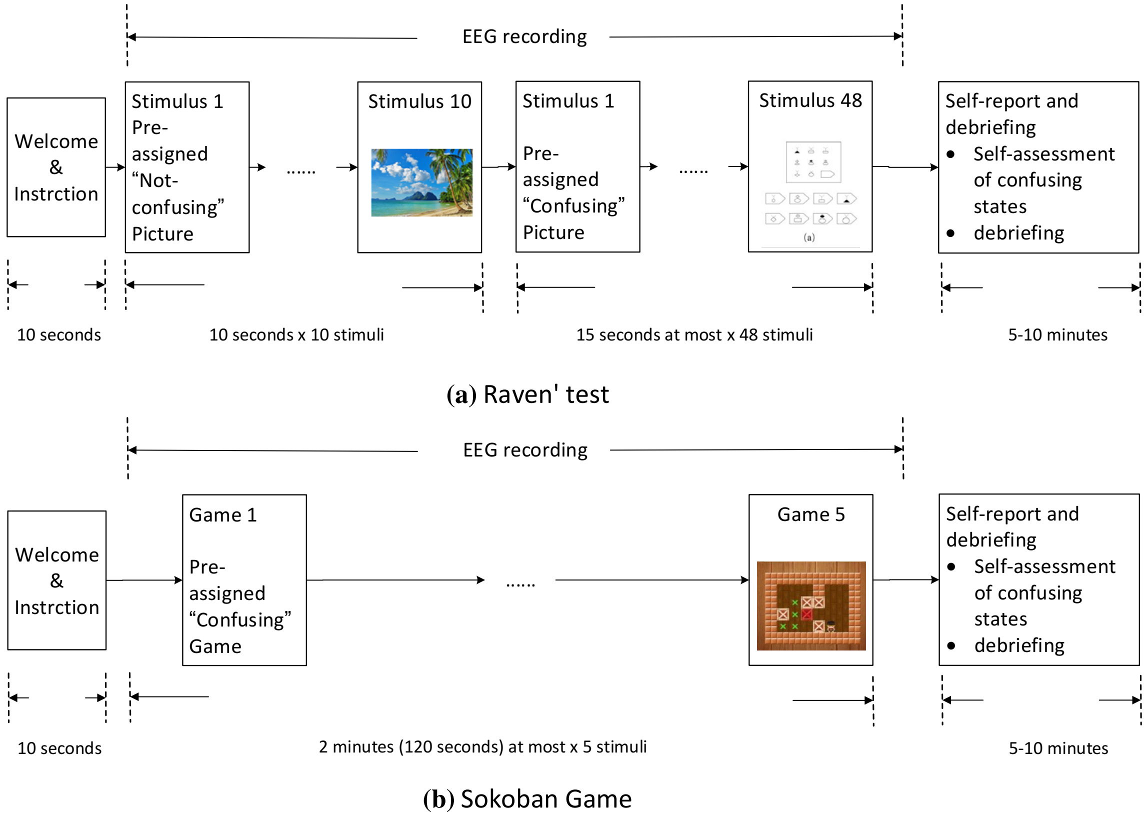

EEG recording
Welcome & Instrction
Stimulus 1 Pre-assigned "Not-confusing" Picture
......
Stimulus 10
Stimulus 1 Pre-assigned "Confusing" Picture
......
Stimulus 48
Self-report and debriefing
• Self-assessment of confusing states
• debriefing
10 seconds
10 seconds x 10 stimuli
15 seconds at most x 48 stimuli
5-10 minutes
(a) Raven' test
EEG recording
Welcome & Instrction
Game 1 Pre-assigned "Confusing" Game
......
Game 5
Self-report and debriefing
• Self-assessment of confusing states
• debriefing
10 seconds
2 minutes (120 seconds) at most x 5 stimuli
5-10 minutes
(b) Sokoban Game

Fig. 5 The experiment design and the procedure

confusion and then obtained the EEG data. Raven's Progressive Matrices (RPM) is a family of standardized intelligence tests, in which a matrix of figures is presented with one entry missing, and the correct missing entry is expected to be selected from a set of answer choices. It is a nonverbal group test typically used in educational settings to measure the taker's abstract reasoning ability [34] and is administered to the groups ranging from 5-year-olds to the elderly. The original test of Raven's matrices consists of increasingly difficult pattern matching tasks, which has little dependency on language abilities. In this experiment, we selected 48 matrices and changed the presenting order to meet the requirement of our experiment. Currently, three versions of RPM have been published. They are the original standard progressive matrices (SPM), advanced progressive matrices (APM), and the colored progressive matrices (CPM). In SPM, there are five groups of tests, named from A to E, each containing 12 tests. The levels of difficulty correspond to the alphabetical order, that is, group A is the easiest one, while group E is the hardest one. The tests in our experiment were selected from SPM (E group of SPM, containing 12 tests) and APM (all tests of APM, containing 36 tests). Therefore, we conducted totally 48 tests to induce confusion. In each test item, the subject was asked to identify the missing element and complete a pattern. The pattern that was used in this experiment was in the form of a 2×2 or 3×3 matrix, as shown in Fig. 6. In the reasoning test, the confusion would decrease by time. Thus, we restricted the presentation time of tests within 15 seconds, regardless if the problem was solved or not. Then, the next stimulus would be presented. In this experiment, we assumed that the stimuli, 10 scenery pictures, would be not-confusing and the stimuli selected from Raven's tests would be confusing.

Experiment 2: Sokoban Game Sokoban (also called warehouse keeper) is a type of transport puzzle game, in which the player pushes boxes or crates around in a warehouse, trying to get them to storage locations. The puzzle solving process requires searching and building logic reasoning strategies. Otherwise, before inference strategies have been

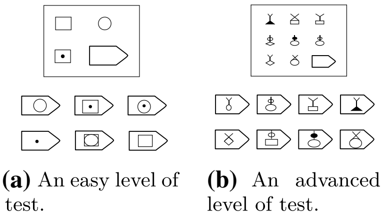


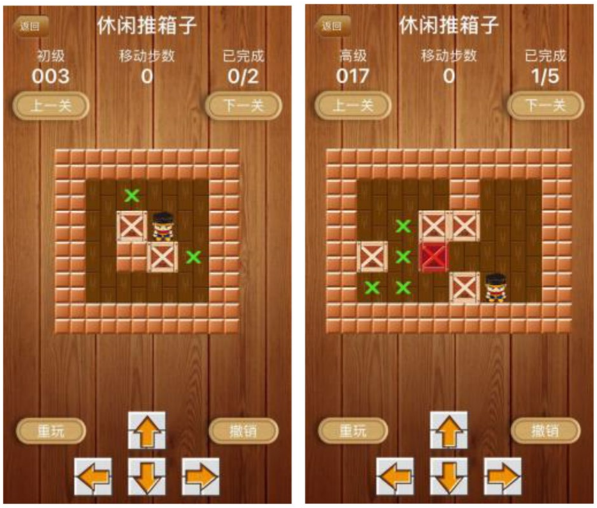
Fig. 6 A 3×3 example test created for demonstration purposes using the similar rules to those from the Raven's progressive matrices tests

obtained, the learner will stay confused, especially at high levels of difficulty. In this experiment, we prepared an iPad with the Sokoban Game installed for participants to play. This version has two modes: leisure mode and puzzle mode (as shown in Fig. 7). In leisure mode, there are seven grades of difficulty, that is, primary, intermediate, advanced, difficult, professional, master and expert level. In each grade, 200–1000 levels of play are available. The numbers of levels vary though the difficulty of each level in the same grade does not have too much difference. In the puzzle mode, instead of the grade, there are 900 levels and the game level increases. We used leisure mode since it is easier to identify the game levels with the appropriate difficulty. We did a pilot test with three persons to play Sokoban for hours to select game levels. For the novice player, the game levels in the primary and intermediate modes are not that hard to break, which takes 20 s to one minute on average. In the advanced level, it is hard to break in less than 2 min. In the difficult mode, it usually requires 20–30 min to break the puzzle. Therefore, we selected 5 levels in the advanced grade to evoke confusion state in the Sokoban Game.

4.2 Participants

For the first experiment, we used the data of twenty-three subjects, including 11 females and 12 males. Their ages were distributed between 20 and 47 years (Mean = 24.48, SD = 6.36). All subjects had normal or corrected vision and were right-handed. Most of the subjects (60.87%) had college level education. The remaining 34.78% had a higher education level of a master degree or above. One subject had only completed high school. There was a bias toward higher education. All participants were either studying or working in the university. All of the subjects have read and signed the ultimate consent form in single access type version, that is, all data can be shared publicly. Participants were compensated for their time.

In the second experiment, we had five volunteer subjects, including four males and one female. All were novice players of the Sokoban Game. Two of the subjects had participated in the first experiment, while three of them had not. All participants were either studying or working in the university.



(a) An easy Sokoban puzzle to be solved.

(b) A hard Sokoban puzzle to be solved.

Fig. 7 Sokoban game interface

4.3 Procedure

In the first experiment using Raven's test as stimuli, the tester briefly introduced and explained this study. Then, the tester asked for the permission of using the recorded EEG data for the research purpose, and each participant read and signed the consent form. After each participant watched the stimuli, s/he was asked to fill out the questionnaire and explain their choices. As shown in Fig. 5a, each subject was asked to watch the stimuli coded by E-Prime 2.0 [11], including 10 scenery pictures and 48 reasoning pictures. Their responses to reasoning tests were recorded by E-Prime. Then, each subject had to fill out the questionnaire after finishing watching. In this process, EEG data were recorded using a laptop computer and the stimuli were presented via another computer, while the time needed was synchronized through the trigger system developed. After the reasoning task, a questionnaire was asked to fill out, including participants' basic information and their self-assessment of confusion levels for each test.

The second experiment using the Sokoban Game as stimuli was similar to the first one in the procedure. The time needed to play the Sokoban Game is on average longer than that of solving a standard reasoning puzzle game.

5 Results

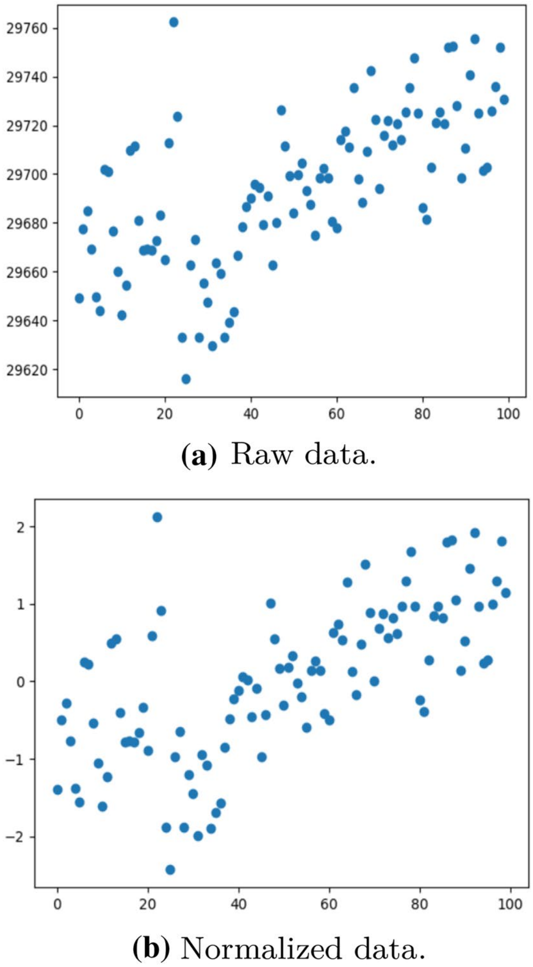
EEG data obtained from the OpenBCI device were sampled at 250Hz. As already mentioned, the data labeled as "confused" came from two sources: Raven's test and the Sokoban Game. By comparison, the non-confused state is defined as the state when subjects watch scenery pictures for ten seconds. The beginning and the end seconds of each piece of data were discarded, because the manipulation action occurred at the beginning of each trial, which brings EMG artifacts. To process, the rest of the data of every subject were merged and then split into small pieces of four seconds.

EEG data have significant differences between subjects. With regard to OpenBCI, the data value of a single channel is around ±10^5. In order to avoid the impact of individual differences in EEG, the normalization process was adopted at first. The raw EEG data were normalized using the Z-score standardization method. This method is based on the normalization and standard deviation of the original data and expressed as:

$$z = (x - \mu)/\sigma \tag{1}$$

Where μ is the mean and σ is the standard deviation.

The comparison of the data before and after normalization by Z-score standardization is as shown in Fig. 8. It can be seen that the result of Z-score standardization is that all data are clustered around 0 with a variance of 1, which reduces individual differences in EEG.



(a) Raw data.

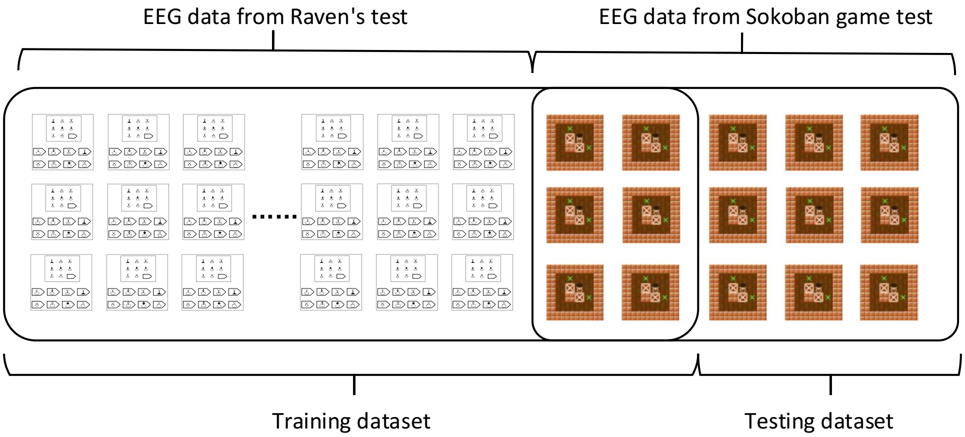
(b) Normalized data.

Fig. 8 The comparison of raw data and normalized data

We employed two dataset allocations to evaluate the two different parts of our approach, respectively. The first allocation was designed to evaluate whether the end-to-end method can work well on the raw EEG data. The data collected from the experiment of Raven's test were put into this allocation. The data were partitioned into two disjoint sets: the training and the test sets. The model was induced from the training set, and its performance was evaluated on the test set. Of the raw data, 70% were for training, while the rest for testing. We built a ConvNet with five layers, containing four convolutional layers and one full-connected layer, to analyze the EEG data from this allocation. The learning rate was 0.00001. Since the ordinary gradient descent algorithm updates w and b at a slower rate, we used the adaptive moment estimation (Adam) algorithm to optimize. The Adam optimization algorithm speeds up the process of gradient descent and eliminates excessive swings during the gradient descent. According to this evaluation, the accuracy of our approach based on the end-to-end method distinguishing confused and non-confused states reached 96.37%.

The second allocation was designed to evaluate the performance of the cross-task and cross-subject approach, on the basis of the end-to-end classification method. The allocation design is as shown in Fig. 9. The training set consists of the raw EEG data collected from the experiment of Raven's test with 23 subjects and the raw EEG data from the Sokoban Game with two subjects who had already taken part

Fig. 9 The design allocation of cross-task and cross-subject approach



in the Raven's test experiment. The test set consists of the EEG data from the Sokoban Game with three subjects who had not taken part in the Raven's test. The ConvNet shared the same structure as shown in Fig. 7. Finally, in this evaluation, the accuracy level reached 91.04% when distinguishing confused and non-confused states of students in the Sokoban Game playing.

6 Discussion, limitations, and future work

6.1 Discussion

In this study, we have attempted to unveil the relationship between EEG data and confusion states and leverage this relationship to assess confusion in an educational game. Our approach successfully classifies the confused state from non-confused state of students in the logic reasoning in a game, which obtains an average accuracy of more than 90% in the classification performance of within-task and within-subject (using data from the first allocation), and cross-task and cross-subject (using data from the second allocation). First, these results indicated that EEG-based technology can be used to recognize and assess students' confusion in the context of logic reasoning in game-based learning. Second, the results proved that the end-to-end method can extract implicit features from the raw EEG data directly and does not require any preprocessing steps differing from traditional methods. Third, with respect to the same emotion, that is, confusion, for different tasks, the cross-task and cross-subject approach performs well on the dataset of small samples in complex tasks, namely by just using a few labeled data from complex tasks (Sokoban Game) and more labeled data from the standardized task (Raven's test). Our experiment proved the feasibility of leveraging a cross-task and cross-subject method to build the classifier for confusion detection in real tasks of long duration, in an educational game.

6.2 Limitations and future work

Although our work is forward-looking and exploratory, and the findings are promising, there are still limitations. First, the EEG data acquisition device itself becomes the first limitation, the use of which limits the numbers of participants. Although off-the-shelf EEG acquisition devices are available, their design and usability are not as good compared to smartphones. It is not convenient enough to wear such devices over a long period of time in the real game scenarios. Thus, the number of subjects who attend experiments is not big in our study and in other related studies. In addition, the human brain accomplishes different kinds of work with an activation of different parts. In this work, we used eight channels around the scalp to detect the change of EEG. Exploring the activated parts of the brain in function for confusion paves the way toward decreasing the numbers of channels and further making it convenient to wear and feasible to support the application in an educational game.

Second, a two-class classification model is preliminary. In this study, our work focuses on distinguishing only two states, that is, confused or not, and build the classifier. Confusion is complex and dynamic, the states of which are considered as a gradual thinking process. There should be many nuanced states though not two states merely. Once the learner fails to resolve the confusion and stays in a confused state at a high or medium level over a long time, s/he will fall into frustration and then boredom. Therefore, it is worth of studying deeply and unremittingly. Multi-class classification models should be investigated, including defining the levels of confusion like high level, medium level and high level. In a near future, we aim to find these nuanced states of confusion based on EEG data.

Third, logic reasoning is one of the instances that could induce confusion, that is, the individual cannot infer the rules when doing rule-based reasoning or solving a puzzle. Other instances exist, such as the new coming information being inconsistent with existing cognitive structure of the learner. It is different to resolve the confusion for different instances. Therefore, supplementary confusion detection methods to complement the EEG-based method should be considered to distinguish the confusion types. In this way, the confusion states in the educational games context can be profoundly detected, which could be used for building a personalized learning path.

7 Conclusion

Confusion is one of the most important cognitive emotions in learning. It is strongly related to learning efficiency. Our study focuses on confusion detection in the instance of logic inference in educational game. Due to the complexity of EEG data and confusion states, we designed two experiments to arouse confusion in logic reasoning. It extends the approach from the laboratory (Raven's test) to the application (Sokoban Game). Since the size of the dataset of the educational game is too small to train, the end-to-end approach based on cross-task and cross-subject is proposed. It not only provides a way to classify confusion states using raw data directly, but also provides opportunities to build the learning model based on small datasets. Finally, the result achieves 91.04% accuracy to classify confusion in the game play. To conclude, the findings of this research have contributed to our understanding of the relationship between EEG data and confusion states and the potential EEG-based methodology for assessing students' confusion in the context of educational games.

Acknowledgements This work was supported by the National Natural Science Foundation of China (61703259 and 61702417) and the Natural Science Foundation of Shaanxi Province (2017JM6097).

References

1. All, A., Castellar, E.P.N., Looy, J.V.: Assessing the effectiveness of digital game-based learning: Best practices. Comput. Educ. **92–93**, 90–103 (2016)
2. Anderson, J.R.: Cognitive Psychology and its Implications. A Series of Books in Psychology, 6th edn. Worth Publishers, New York (2004)
3. Antoniades, A., Spyrou, L., Took, C.C., Sanei, S.: Deep learning for epileptic intracranial EEG data. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6 (2016)
4. Bashivan, P., Rish, I., Yeasin, M., Codella, N.: Learning representations from EEG with deep recurrent-convolutional neural networks. CoRR abs/1511.06448 (2015)
5. Bradley, M.M., Lang, P.J.: Measuring emotion: The self-assessment manikin and the semantic differential. J. Behav. Therap. Exp. Psychiatry **25**(1), 49 (1994)
6. Chanel, G., Rebetez, C., Bétrancourt, M., Pun, T.: Emotion assessment from physiological signals for adaptation of game difficulty. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **41**(6), 1052–1063 (2011)
7. Clark, R.C., Mayer, R.E.: E-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning. Electronic Learning and the Science of Instruction, 3rd edn. Pfeiffer, San Francisco (2011)
8. Clore, G.L., Huntsinger, J.R.: How emotions inform judgment and regulate thought. Trends Cogn. Sci. **11**(9), 393–399 (2007)
9. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learn. Instr. **22**(2), 145–157 (2012)
10. D'Mello, S., Lehman, B., Pekrun, R., Graesser, A.: Confusion can be beneficial for learning. Learn. Instr. **29**(X), 153–170 (2014)
11. E-prime 2.0: http://www.psychology-software-tools.mybigcommerce.com/e-prime-2-0-professional/. Accessed 19 Nov 2018
12. Ghali, R., Ouellet, S., Frasson, C.: Lewispace: an exploratory study with a machine learning model in an educational game. J. Educ. Train. Stud. **4**(1), 192–201 (2016)
13. Ghergulescu, I., Muntean, C.H.: A novel sensor-based methodology for learner's motivation analysis in game-based learning. Interact. Comput. **26**(4), 305–320 (2014)
14. Graesser, A.C., Lu, S., Olde, B.A., Cooper-Pye, E., Whitten, S.: Question asking and eye tracking during cognitive disequilibrium: comprehending illustrated texts on devices when the devices break down. Mem. Cogn. **33**(7), 1235–1247 (2005). https://doi.org/10.3758/BF03193225
15. Grimes, D., Tan, D.S., Hudson, S.E., Shenoy, P., Rao, R.P.: Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08, pp. 835–844. ACM, New York (2008). https://doi.org/10.1145/1357054.1357187
16. Hajinoroozi, M., Mao, Z., Jung, T.P., Lin, C.T., Huang, Y.: Eeg-based prediction of driver's cognitive performance by deep convolutional neural network. Signal Process. Image Commun. **47**, 549–555 (2016)
17. Halpern, D.F., Millis, K., Graesser, A.C., Butler, H., Forsyth, C., Cai, Z.: Operation ARA: a computerized learning game that teaches critical thinking and scientific reasoning. Think. Skills Creat. **7**(2), 93–100 (2012)
18. Ifenthaler, D., Eseryel, D., Ge, X.: Assessment for Game-Based Learning, pp. 1–8. Springer, New York (2012)
19. Jiao, Z., Gao, X., Wang, Y., Li, J., Xu, H.: Deep convolutional neural networks for mental load classification based on EEG data. Pattern Recogn **76**, 582–595 (2018)
20. Kapoor, A., Picard, R.W.: Multimodal affect recognition in learning environments. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA '05, pp. 677–682. ACM, New York (2005). https://doi.org/10.1145/1101149.1101300
21. Kim, M.K., Kim, M., Oh, E., Kim, S.P.: A review on the computational methods for emotional state estimation from the human EEG. Comput. Math. Methods Med. **2013**, 13 (2013). https://doi.org/10.1155/2013/573734
22. Kirriemuir, J., Mcfarlane, A.: Literature review in games and learning. A NESTA Futurelab Research report - report 8 (2004). https://telearn.archives-ouvertes.fr/hal-00190453
23. Klimesch, W.: Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Res. Rev. **29**(2), 169–195 (1999). https://doi.org/10.1016/S0165-0173(98)00056-3
24. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436 (2015)
25. Lehman, B., D'Mello, S., Graesser, A.: Confusion and complex learning during interactions with computer learning environments. Internet High. Educ. **15**(3), 184–194 (2012)
26. Lehman, B., Matthews, M., D'Mello, S., Person, N.: What are you feeling? Investigating student affective states during expert human tutoring sessions. In: International Conference on Intelligent Tutoring Systems, pp. 50–59 (2008)
27. Li, M., Lu, B.L.: Emotion classification based on gamma-band EEG. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1223–1226 (2009)
28. Liu, Y.J., Yu, M., Zhao, G., Song, J., Ge, Y., Shi, Y.: Real-time movie-induced discrete emotion recognition from EEG signals. IEEE Trans. Affect. Comput. **9**(4), 550–562 (2017)
29. Moreno-Ger, P., Burgos, D., Martínez-Ortiz, I., Sierra, J.L., Fernández-Manjón, B.: Educational game design for online education. Comput. Hum. Behav. **24**(6), 2530–2540 (2008). (**Including the Special Issue: Electronic Games and Personalized eLearning Processes**)
30. Nie, D., Wang, X.W., Shi, L.C., Lu, B.L.: EEG-based emotion recognition during watching movies. In: 2011 5th International IEEE/EMBS Conference on Neural Engineering, pp. 667–670 (2011)
31. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191
32. Pekrun, R., Linnenbrink-Garcia, L.: Academic Emotions and Student Engagement, pp. 259–282. Springer, Boston (2012)
33. Qian, M., Clark, K.R.: Game-based learning and 21st century skills: a review of recent research. Comput. Hum. Behav. **63**, 50–58 (2016)
34. Raven, J.: The raven's progressive matrices: change and stability over culture and time. Cogn. Psychol. **41**(1), 1–48 (2000)
35. Ren, Y., Wu, Y.: Convolutional deep belief networks for feature extraction of EEG signal. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 2850–2853 (2014)
36. Rob, B.K., Reilly, R., Picard, R.W.: External representation of learning process and domain knowledge: affective state as a determinate of its structure and function. In: Artificial Intelligence in Education Workshops, pp. 64–69 (2001)
37. Saeed, S., Zyngier, D.: How motivation influences student engagement: a qualitative case study. J Educ Learn **1**, 252–267 (2012)

38. Schunk, D., Pintrich, P., Meece, J.: Motivation in Education: Theory, Research, and Applications. Pearson/Merrill Prentice Hall, Upper Saddle River (2008)
39. Sharma, N., Gedeon, T.: Objective measures, sensors and computational techniques for stress recognition and classification: a survey. Comput. Methods Progr. Biomed. **108**(3), 1287–1301 (2012). https://doi.org/10.1016/j.cmpb.2012.07.003
40. Silvia, P.J.: Confusion and interest: the role of knowledge emotions in aesthetic experience. Psychol. Aesthet. Creat. Arts **4**(2), 75–80 (2010). https://doi.org/10.1037/a0017081
41. Sturm, I., Lapuschkin, S., Samek, W., Müller, K.R.: Interpretable deep neural networks for single-trial EEG classification. J. Neurosci. Methods **274**, 141–145 (2016)
42. Sun, X., Qian, C., Chen, Z., Wu, Z., Luo, B., Pan, G.: Remembered or forgotten?—an eeg-based computational prediction approach. PLOS ONE **11**(12), 1–20 (2016)
43. Szafir, D., Mutlu, B.: Pay attention!: designing adaptive agents that monitor and improve user engagement. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'12, pp. 11–20. ACM, New York (2012)
44. Tang, Z., Li, C., Sun, S.: Single-trial EEG classification of motor imagery using deep convolutional neural networks. Optik Int. J. Light Electron Opt. **130**, 11–18 (2017)
45. Wang, H., Li, Y., Hu, X., Yang, Y., Meng, Z., Chang, K.M.: Using EEG to improve massive open online courses feedback interaction. In: AI-ED Workshop Proceedings, pp. 59–66 (2013)
46. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing **129**(4), 94–106 (2014)
47. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: automatic recognition of student engagement from facial expressions. IEEE Trans. Affect. Comput. **5**(1), 86–98 (2014)
48. Xu, J., Zhong, B.: Review on portable EEG technology in educational research. Comput. Hum. Behav. **81**, 340–349 (2018). https://doi.org/10.1016/j.chb.2017.12.037
49. Xu, T., Zhou, Y., Wang, Z., Peng, Y.: Learning emotions eeg-based recognition and brain activity: a survey study on BCI for intelligent tutoring system. In: Procedia Computer Science. The 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018), vol. 130, pp. 376–382 (2018)
50. Zhou, Y., Xu, T., Cai, Y., Wu, X., Dong, B.: Monitoring cognitive workload in online videos learning through an EEG-based brain-computer interface. Learning and Collaboration Technologies. Novel Learning Ecosystems, pp. 64–73. Springer, Cham (2017)
51. Zhou, Y., Xu, T., Zhu, Z., Wang, Z.: Learning in doing: a model of design and assessment for using new interaction in educational game. In: Zaphiris, P., Ioannou, A. (eds.) Learning and Collaboration Technologies. Learning and Teaching, pp. 225–236. Springer, Cham (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.