DOI: 10.1111/cgf.14862

COMPUTER GRAPHICS *forum*
Volume 42 (2023), number 6, e14862

Accompany Children's Learning for You: An Intelligent Companion Learning System

Jiankai Qian,[1] Xinbo Jiang,[2] Jiayao Ma,[1] Jiachen Li,[3] Zhenzhen Gao[4] and Xueying Qin[1]

[1]School of Software, Shandong University, Jinan, China
{285407647, 908187229}@qq.com, qxy@sdu.edu.cn
[2]School of Qilu Transportation, Shandong University, Jinan, China
xinbojiang@sdu.edu.cn
[3]State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China
lijiachen93@zju.edu.cn
[4]The Primary School Attached to Shandong Normal University, Jinan, China
332577496@qq.com

Abstract

Nowadays, parents attach importance to their children's primary education but often lack time and correct pedagogical principles to accompany their children's learning. Besides, existing learning systems cannot perceive children's emotional changes. They may also cause children's self-control and cognitive problems due to smart devices such as mobile phones and tablets. To tackle these issues, we propose an intelligent companion learning system to accompany children in learning English words, namely the Intelligent Augmented Reality Educator (IARE)*. The IARE realizes the perception and feedback of children's engagement through the intelligent agent (IA) module, and presents the humanized interaction based on projective Augmented Reality (AR). Specifically, IA perceives the children's learning engagement change and spelling status in real-time through our online lightweight temporal multiple instance attention module and character recognition module, based on which analyses the performance of the individual learning process and gives appropriate feedback and guidance. We allow children to interact with physical letters, thus avoiding the excessive interference of electronic devices. To test the efficacy of our system, we conduct a pilot study with 14 English learning children. The results show that our system can significantly improve children's intrinsic motivation and self-efficacy.*

Keywords: interaction, human–computer interfaces, methods and applications, education, interaction, user studies

CCS Concepts: • Human-centred computing → Interactive systems and tools

1. Introduction

Primary education, as the beginning of the educational process, often requires a significant amount of attention and effort from parents. To achieve the ideal educational effect, parents must devote sufficient time and energy. In the era of artificial intelligence (AI), using multimedia intelligent technologies to assist primary education, such as literacy, has essential significance.

Integrating AI into educational scenarios can assist or offload part of parents' work. The increasing maturity of computer vision and natural language processing provides new tools for education. However, these current AI educational tools can only play an auxiliary role due to their lack of perception of children's emotions and states, which means that parents are still indispensable.

The ARCS [Kel87] theory provides a highly comprehensive and succinct summary of the characteristics of educational behaviour, which believes that the four elements, namely Attention (A), Relevance (R), Confidence (C) and Satisfaction (S), can stimulate students' learning engagement and enthusiasm. Although existing applications can attract children's attention through novel contents and forms, they can neither maintain children's learning engagement nor respond to their emotions. The question of how to better integrate AI with the ARCS theory remains an urgent issue to be addressed.

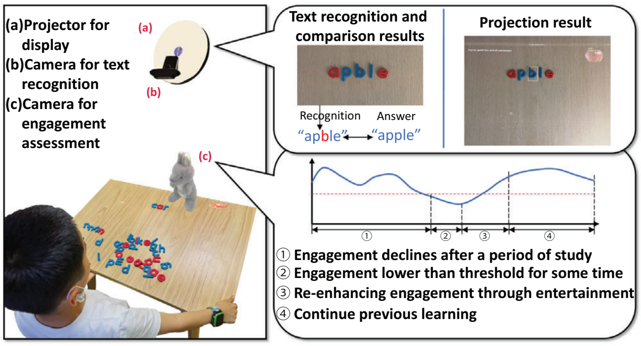
(a)Projector for display
(b)Camera for text recognition
(c)Camera for engagement assessment
(a)
(b)
(c)
Text recognition and comparison results
Recognition
Answer
"apble"
"apple"
Projection result
①
②
③
④
① Engagement declines after a period of study
② Engagement lower than threshold for some time
③ Re-enhancing engagement through entertainment
④ Continue previous learning

Figure 1: *The illustration of IARE. Left: the overall system, which uses the projector (a) to display the AR interface, camera (b) for character recognition, and camera (c) for engagement perception. Right-top: character recognition results and AR guides. Right-bottom: real-time engagement perception feedback for IA.*

Current educational applications tend to be overly reliant on electronic devices, which creates potential cognitive and self-control issues for children. Weiser *et al.* [WB96] proposed the Calm Technology and believed that a good interaction design should minimize users' perceptible invasiveness to electronic devices. As an essential platform for feedback in human–computer interaction in the era of AI, AR can be maximally integrated with the real world to enhance users' sense of reality. Therefore, how to design an effective presentation method for AR and integrate it with AI technology is another urgent problem.

In this work, we study how to integrate AI and AR into educational applications based on ARCS theory to provide companionship and heuristic education, which is fundamentally different from traditional educational methods that only focus on the imparting of knowledge. This model is aimed to young children who need accompanying education, especially pre-school children. We focus on the perception and feedback of children's psychological feelings, aiming to stimulate positive effects on children's psychological level, such as self-efficacy and learning motivation. Pedagogical theories such as ARCS also suggest that primary education itself should be concerned with these aspects. More specifically, we propose an intelligent companion learning system named *Intelligent Augmented Reality Educator (IARE)* for children to learn English words, as shown in Figure 1. Our system followed a participatory design with a domain educator we invited and the concept of the ARCS model.

The main contributions of this paper are as follows:

- We propose an intelligent companion learning system named IARE, which seamlessly integrates AR and AI based on the ARCS model. It can stimulate positive psychological feedback in children, and the pilot study shows its efficacy.
- We develop an intelligent agent (IA) module to imitate and substitute the role of parents in word learning scenarios of primary education, which perceives the changes in learning engagement and spelling status to regulate the learning process.
- We implement a learning scenario called exploring mode in our system, which induces children to conduct discovery-style unintentional learning and stimulates their interest and motivation. Combined with the other three modes and IA adjustor, children can maintain their learning engagement.

2. Related Work

AI and AR are already playing an important role in many fields, including primary education. This section will present related work from the education application of AI, AR and the ARCS model.

2.1. The application of AI in education

The idea of teaching students according to their aptitude was put forward thousands of years ago, which requires teachers to pay attention to the differences among students of different ages and the same ages. This process requires many labour costs, and the persons undertaking primary education may be parents without professional pedagogical knowledge, making it challenging to implement this idea. The development and maturity of AI technology represented by computer vision [YCLC19, GLY*19, Oso19] and natural language processing [ZL22] have enabled computers to accurately perceive and analyse the learning status of each student [WYG*19, WZW*19]. Therefore, the educational idea of teaching students in accordance with their aptitude can be realized.

The intelligent tutoring system [MH20] is one of the critical researches of AI in education. The basis of the intelligent tutoring system is to establish a learner model [AKS20, CLL*15, PPŘ*17] to evaluate the level of knowledge. In the education process, intelligent auxiliary methods can effectively help learners to perform efficient learning, such as text simplification [BDCM20], detection of their confused state [CEF20a], detection of their engagement [GMB*20], identification of their needs [CLP*20] and intelligent recommendation [CEF20b].

It is necessary but cumbersome to evaluate learners' learning progress at different stages. Establishing an automated evaluation method [Cad20, CWB*17, GSV19] to liberate teachers is necessary for catering to learners' abilities. Among the many problems in the learning process, student dropout is the most complicated and negative thing. AI for detection and early intervention [HMMH21] can effectively reduce the dropout rate.

With the development of deep learning networks, it becomes possible to perceive the states and even events of the real world from images and videos. Scene text detection and recognition technology [LCS*20, ZCL*21, NNT*21] can accurately identify various texts with rich presentation forms. Learning engagement assessment technology [MJXQ21, WYWH20, LXQ22] can perceive the engagement degree throughout the learning process. Advances in these areas provide the foundation for higher-level AI educational tools. However, these technologies are still unable to replace the role of educators in the educational process.

2.2. Human–computer interaction of AR in education

Applications used to assist education, especially primary education, should have simple and convenient interaction methods. AR can enhance users' sense of experience and make it easier to use and acquire knowledge, so it can be a natural way to present feedback in primary educational applications. AR allows users to interact with virtual objects in the real world [CCH16, LLCH20] by using location information or real-world objects as visual markers and uses digital information superposition technology to integrate dynamic and interactive digital content into the real-world environment [CYH14, DD14, ZSHC14]. Thus, AR can be used to teach some difficult concepts and enable learners to solve complex problems [TMBF17, CYH14].

Augmented Reality (AR) has been widely applied in the field of education, with language learning being one of the most prominent areas, particularly for young children. In language learning, AR has been used to label real objects in foreign vocabulary acquisition [VNL*17], teach structural concepts of language [DLSC20] and assist in developing children's phonetic awareness [LKWW16]. Despite the significant achievements and applications of AR in various educational domains, its usage has primarily focused on enhancing user experience and often involves high complexity, requiring children to use it under adult guidance. To the best of our knowledge, there is currently no existing research exploring the deep integration of AI and AR technologies specifically tailored to provide intelligent and companion-based learning experiences for children learning independently in primary education settings, which is highly meaningful.

2.3. The ARCS model

Using AI and AR to accompany and guide children to learn requires scientific theoretical guidance. Hoffman [Hof96] believes that a truly effective learning system must not only stimulate but also maintain learners' motivation. Aiming at this problem, Keller *et al.* [Kel87] proposed a learning process design model to stimulate and maintain learning motivation, namely the ARCS learning motivation model. They believe that using the four elements of Attention, Relevance, Confidence, and Satisfaction can motivate students to keep learning. Drawing and sustaining learners' attention is the first and most crucial step to stimulating and sustaining learning motivation. Relevance includes purpose-oriented relevance and process-oriented relevance. The former depends on the urgency of the learner for his learning purpose, and the latter emphasizes the pleasant feeling of the learning process. Confidence means that the learner can build up the confidence to complete the corresponding task after learning to motivate him further. Satisfaction refers to learners' sense of achievement for their learning outcomes which can maintain their motivation to learn. These four elements are indispensable for maintaining learning motivation.

3. Overview of IARE

The IARE can seamlessly integrate AI and AR through the ARCS model to provide children with a humanized and companion language learning experience. The key to the system is to simulate the role of children's parents through the IA, which is the focus of our design. Therefore, the IA is designed to provide timely feedback according to children's learning ability and state, especially to respond to emotional situations such as attention.

As Figure 1 shows, our system runs on a computer with a projector and two webcams connected to it. The webcams were used to capture video images for perceiving children's spelling and engagement status. The projector presents the learning content and the feedback given by the system on the table. Children can use the physical letters to make words and see augmented overlays displayed on the table directly over the letters, which provides cues to spelling. This spatial AR interaction method based on the projector minimizes children's contact with electronic devices.

As Figure 2 shows, IA is the core of our system and regulates the entire learning process. Details about IA will be explained in Section 4 INTELLIGENT AGENT.

3.1. Four modes in our IARE

Children learn new words through various forms of hints and practice spelling in learning mode and review the learned words in the form of a spelling test in testing mode. Entertaining mode, also named spell then match, in which children spell words to eliminate corresponding pictures, helps children stimulate their motivation and review some words.

The learning content of the above three modes is personalized provided by IA. Besides, IA will record and analyse the performance information of children in learning and testing modes, further updating the personalized database. For details, please see Section 4.2 Performance Statistics Analysis and Personalized Learning Content Database of Section 4.4.2.

In exploring mode, children are free to place physical letter blocks. When a word is detected (unrelated letters are allowed in the word, as long as the letters that make up the word are in the correct order), IA will frame the word through the projector, and provide feedback of the audio and corresponding image of the word. This mode can induce children to 'discover' words, help children to conduct discovery unintentional learning, and improve children's interest and learning motivation.

Under the control of IA, our system can automatically switch between modes to maintain good companionship and affinity. The four modes and their switch principle are as shown in Figure 3. The specific rules are described in Reward and Mode Switch of Section 4.4.3.

3.2. Design purpose

IARE breaks down learning difficulty by providing periodic rewards and giving marking hints, to establish children's optimistic expectations for success and self-efficacy (C), and stimulate learning motivation; IARE contains multiple modes, perceives changes in children's attention, and switches between modes when necessary, which provides enough changes to maintain children's attention (A) and interest; IARE provides age-appropriate learning content to ensure that children's learning goals are related to the knowledge they have mastered (R); IARE provides children with the opportunity to apply the words they have learned in the entertaining mode, making it easier to affirm their learning outcomes to increase

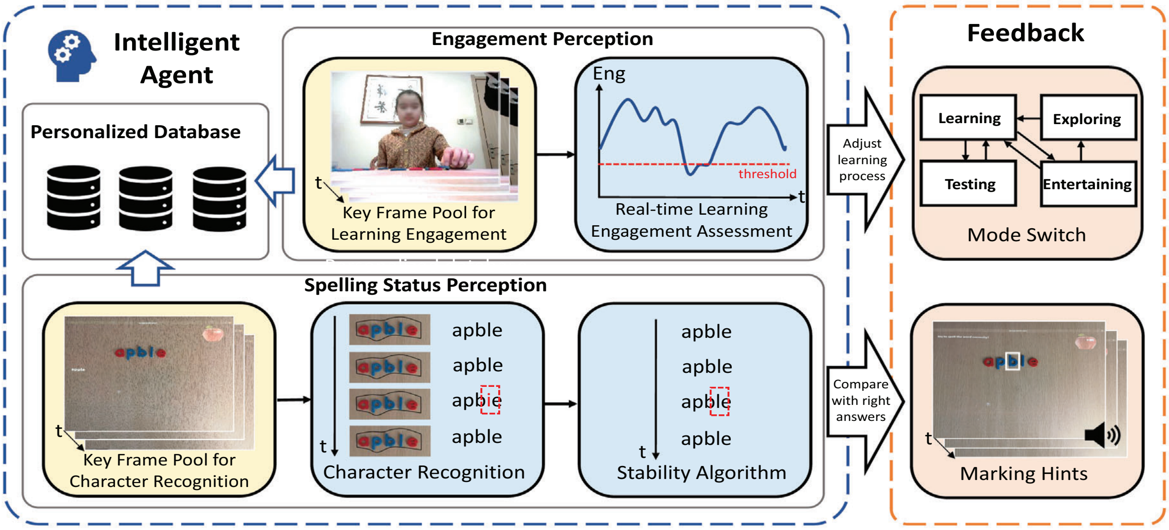

Figure 2: *The detailed structure of IARE with IA as the core. The main functions of IA include real-time engagement perception, real-time spelling status perception and performance analysis. IA controls IARE to switch between modes to maintain children's attention by real-time engagement perception. IA gives appropriate hints through real-time spelling status perception, and the picture shows the situation of spelling error correction in learning mode. IA provides a personalized learning content database by analysing children's performance (*e.g. *the time spent spelling a word).*

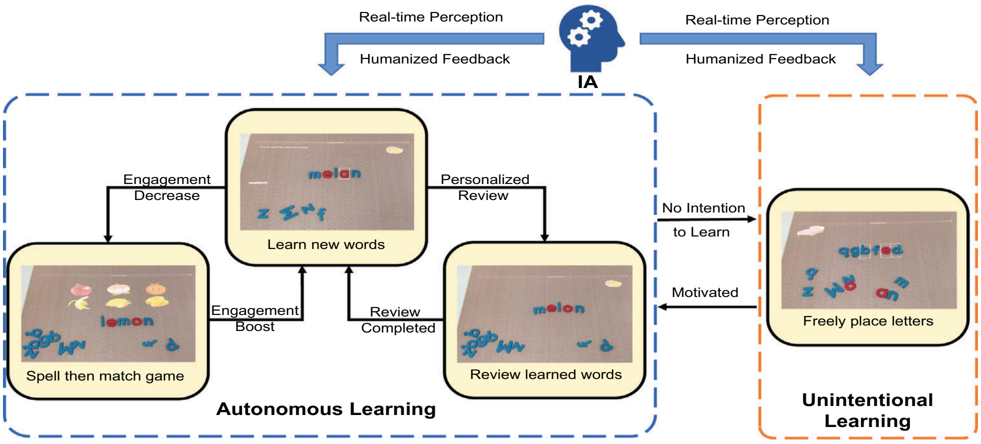

Figure 3: *Four modes and mode switch principle. Under the control of IA, our system automatically switches between modes to adjust the learning process and maintain children's learning engagement.*

satisfaction with their learning process (S); IARE encourages children to carry out discovery learning in the exploring mode and encourages children to observe the relationship between words and objects in the entertaining mode, which will stimulate children's thinking and give higher cognitive benefits.

4. Intelligent Agent

The IA is the core of the system and plays the role of the brain and heart. Its main task is to perceive changes in the real world, that is, to perceive the changes in children's spelling status and engagement, and to conduct performance analysis so as to provide humanized feedback and help.

4.1. Spelling status perception

Children's spelling is done by placing physical letters. In order to give appropriate feedback and help, IA needs to accurately perceive children's spelling of words in real-time. This requires the character recognition model to detect the character position and recognize the spelling accurately in real-time to determine whether the current spelling needs a hint. If it is needed (*i.e.* there is a spelling error or children unintentionally discover a word in exploring mode), IA hints at the right time through the projector based on the actual position of the corresponding physical letter blocks (please see Marking Hints of Section 4.4.1).

4.1.1. *Character detection and recognition*

Selecting a suitable character recognition algorithm is the first step for the real-time perception of spelling status. The existing character recognition algorithms are mainly divided into video text recognition algorithms and image text recognition algorithms. The main task for video text recognition is text tracking in natural scenes, and the running speed of the existing video text detection networks [FYZL21, WWSS18, YHP*21] is far below the requirements of real-time operation. In our indoor scene of spelling words, the position of the characters remains the same after being placed. Motion blur and illumination change rarely significantly impact recognition, though there are occasional hand occlusions. So lightweight real-time image text recognition algorithms are more suitable for our needs.

In order to accurately recognize the spelling of the current frame, the following should be taken into account. (1) Children can move physical letter blocks in different places such as desktops and whiteboards, so the background of character detection and recognition is diverse. (2) Physical letter blocks are also diverse, and they vary in size, colour, material and font. (3) When children place physical letter blocks, they may not be placed in a horizontal line, but in a curved line.

Considering the above situation, we chose ABCNet [LCS*20] as the character recognition model used by IARE. ABCNet is a real-time lightweight end-to-end scene text detection and recognition model. The core idea is to use parameterized Bezier curves to adaptively fit text of any shape. The model predicts the position of the control points of the parameterized Bezier curve through the network, which not only achieves efficient text detection, but also extracts and corrects the text area features through the Bezier Align based on the Bezier curve, effectively removing the redundant background for subsequent recognition. Using ABCNet, the accuracy and speed requirements for single-frame image character recognition are basically met.

4.1.2. *Stability algorithm*

In our learning scenes, frame-by-frame character recognition is unnecessary given the speed children spell words. In order to further improve the efficiency of spelling status perception and reduce the system load, we propose a stability algorithm (SA) to judge the timing of text detection and recognition.

When the recognition results of consecutive k frames ($k = 3$ in the experiment) are consistent, and the confidence levels all exceed the set threshold t ($t = 0.8$ in the experiment), the algorithm determines that the video stream has entered a 'stable state' and takes the kth frame as the 'stable frame'. The video input will no longer be detected and recognized for text in the stable state but to determine whether to leave the stable state. The algorithm compares the difference between the stable frame and the input video frame in the stable state, using *PSNR* (peak signal-to-noise ratio) [ASS02] as the indicator. When the image difference is significant enough, the algorithm determines that the video stream is out of the stable state, indicating that the children continue spelling words, and sends the subsequent video input into the ABCNet until the next stable state is entered.

This algorithm enables IA to invoke the character recognition model only when necessary, significantly improving the performance of our system in practical use. Besides, according to the continuity of consecutive frames in the recognition sequence, the algorithm votes by the majority, thereby eliminating the recognition errors of individual frames caused by factors such as hand occlusion to a certain extent.

4.2. Performance statistics analysis

IA records information about children's performance while children are using the learning and testing modes, including the word's spelling date in both modes, the word's spelling time, the number of times the word was misspelled, and the total number of times the word was spelled in the testing mode. Then, IA weights these data to obtain children's mastery coefficient w of the words, preparing for generating personalized learning content. The formula for w is

$$w = k_1 \cdot \frac{t_1}{t_2} + k_2 \cdot \frac{n_1}{n_2}, \tag{1}$$

where k_1 and k_2 are hyperparameters with values of 0.4 and 0.6, respectively, t_1 is the time spent by the user to spell the word in the last test, t_2 is the maximum time allowed ($t_2 = 120$ s in the experiment), n_1 is the number of times the word was misspelled in the testing mode, and n_2 is the total number of times the word was spelled in the testing mode. When the word was misspelled, $t_1 = t_2$. The value of w is between (0, 1], and a larger value of w represents a lower mastery degree.

4.3. Engagement perception

In order to perceive children's learning engagement in the process of spelling words in real-time and accurately, this paper proposes a more lightweight learning engagement evaluation model that can achieve faster online evaluation based on the temporal multiple instance attention module proposed by Ma *et al.* [MJXQ21]. The overall structure of our model is shown in Figure 4.

The temporal multiple instance attention module is mainly used to obtain the mapping relationship between low- and high-level features. Here, we use it to obtain the relationship between the underlying video features and video s composed of these features. For the underlying video frames, their features often have a strong temporal relationship, so a bidirectional Long Short-Term Memory (LSTM) [HS97] network is applied at the bottom of the model to obtain the relationship between them. Then, based on a score calculation branch using the sigmoid activation function and another score calculation branch using the Tanh activation function [KK92], the contribution score of each frame feature to the higher-level video segment features is obtained. The attention obtained based on this score is weighted with each video frame feature to obtain the feature representation at the video segment level. Finally, based on the

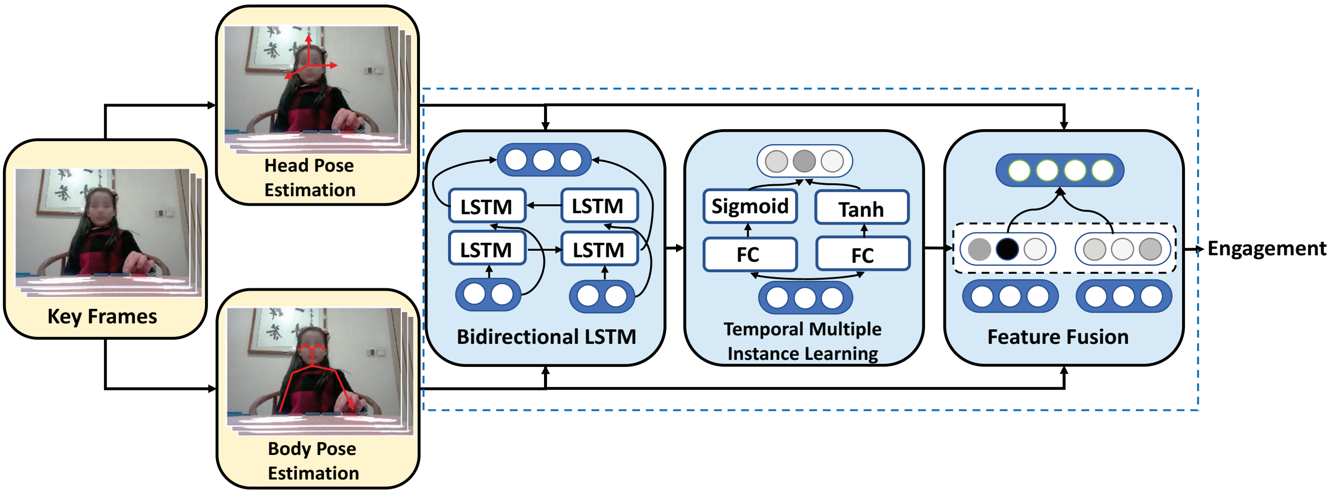

Figure 4: *Overall structure of our lightweight attention model.*

trainable feature fusion module, the video segment features obtained from different underlying features are further fused. The final learning participation evaluation result can be obtained through a simple regression based on this fusion feature.

Unlike [MJXQ21], our model only uses head pose and body pose as the underlying features. This is mainly because the facial keypoint model often generates many missing values that are difficult to fill in our scene. This problem is caused by the fact that children's hands often cover their faces, and they often turn their heads at a large angle. We found through experiments that the lack of too many facial keypoint features leads to a decrease in the accuracy, so the model proposed in this paper discards this feature. In addition, since the video data used in this paper is annotated segment by segment, we only use the structure of the bottom modules in [MJXQ21] without using the top modules. In order to further speed up the model, we also reduce the hidden layer dimension and feature dimension involved in the model.

Compared to Ma *et al.* [MJXQ21], our model is about nine times faster while ensuring accuracy in our scene, meeting the real-time requirements of engagement perception. The experimental results can be seen in Section 6.

4.4. Humanized feedback

Based on the results of spelling perception, engagement perception and performance statistics analysis, IA gives children appropriate feedback.

4.4.1. *Marking hints*

There are two situations for marking hints. One is when children misspell a word in the learning mode, and the other is when children unintentionally discover a new word in the exploring mode. In either case, to give a hint through the projector, the coordinates of the physical letter block in the projector's coordinate system must be obtained.

IA compares the recognition result with the specified word in the learning mode. If there exist wrong characters, IA obtains the position of the wrong letter blocks in the picture through the character detection result (that is, the position of the letter blocks in the camera coordinate system). In the exploring mode, IA searches for actual words in the character recognition sequence. If a word is implied in the character sequence, IA also obtains the position of the letter blocks that make up the word in the picture. Then, IA converts the coordinates of the letter blocks from the camera coordinate system to the projector coordinate system and frames the letter blocks through the projector to give marking hints.

In addition, to leave time for children to think independently, IA will not bother at the beginning of word spelling but give marking hints only when the number of spelling characters reaches three.

4.4.2. *Personalized learning content database*

Children's learning contents in learning, testing and entertaining modes are personalized and customized by IA based on children's performance. We store the learning contents customized for each child in a personalized database.

In the learning mode, we refer to the primary school English textbooks to classify the learning contents. When children use it for the first time, IA will automatically assign the learning contents of the corresponding difficulty according to the children's age. After the learning contents of the current difficulty level is mastered, IA will automatically assign learning objectives of the next level of difficulty for children.

In the testing mode, IA sorts the words learned by children into a list from low to high according to the mastery degree (assessed by the mastery coefficient defined in Section 4.2), words with a low mastery degree are reviewed first to improve efficiency. Then IA compares the spelling date of the word with the Ebbinghaus curve and removes words that do not fit the curve from the list. Finally, IA selects the top 10 in the list as the content of the testing mode this time.

In the entertaining mode, most of the content is composed of simple words whose pictures are also interesting. At the same time, IA will also select some words with a low mastery degree into the entertaining mode, so that children can review some words while having fun through games.

4.4.3. *Rewards and mode switch*

When children is active in the learning process, IA will give rewards to improve children's sense of achievement and motivate children to develop a further learning desire (Rewards are presented in various forms such as voice, pictures and animation. Please see the supplementary material and demonstration videos for details); when children's engagement is perceived to be declining in the learning mode, IA will control the system to switch to other modes at an appropriate time to re-attract children's attention and continue the previous learning after their attention returns. Figure 3 shows the detailed process.

The switch between modes is interspersed in the whole process system operation, and is controlled by IA in real-time. When children start learning, IA will first consult the personalized database to determine whether the current user has any content that needs to be reviewed. If so, IA will control the system to enter the testing mode first, and then enter the learning mode to learn new words after the test is completed. Otherwise, IA will control the system to enter the learning mode directly. In the learning mode, if IA finds that children's engagement score is lower than the threshold t for a certain period (the engagement score ranges from 0 to 1, and the threshold t is set to 0.5 in the experiment), IA will give a voice remind. If not working, IA will control the system to switch to entertaining mode. When children break through the match game in the entertaining mode, and the engagement score exceeds the threshold t, IA will give voice reward and control the system to switch back to the learning mode and continue the previous learning.

If children's engagement in the entertaining mode is still shallow, which indicates that they have no intention to learn, IA will control the system to switch to the exploring mode. In this mode, IA induces children to carry out discovery-style unintentional learning by accompanying them to place letter blocks at will, until their engagement returns or IA controls system automatically exits.

5. Implementation

As Figure 1 shows, our system runs on a computer with a projector and two webcams connected to it. The webcams were used to capture video images for perceiving children's spelling and engagement status.

Since IA involves multiple deep neural networks, we use python to implement the system and use PyQt5 to design and develop the interface for each mode. Since the system needs to run in real-time, multi-processing is used to realize the parallelism of the neural networks in the temporal multiple instance attention module. Character recognition module, data exchange and mode switch between different modes are realized through multi-threading and semaphore.

The character recognition network runs in a sub-thread (QThread). The sub-thread will perform corresponding processing according to the recognition result and the current mode, such as comparing the recognition sequence with the correct word, giving error correction prompts in the learning mode and using a dictionary to find words in the recognition sequence in the exploring mode.

The temporal multiple instance attention module runs in the background, sending the generated engagement scores to the specified queue Q. In order to realize the switch between different modes, we instantiate a sub-thread (QThread) for each mode, which reads the data in the queue Q. If the engagement score satisfies the switch condition, the sub-thread will issue a switch signal (pyqtSignal), the main process responsible for system operation receives the signal and switches to the specified mode. We use 30 frames of images to generate an engagement score. Starting from the 31st frame, we take one frame every 10 frames to replace the earliest frame in the previous 30 frames and calculate a new score, further improving the efficiency of engagement evaluation and reducing the operating load of our system.

The key libraries used in our system are pyqt 5.6.0, pytorch 1.8.0, cudatoolkit 10.2 and python 3.7. Our system has achieved the effect of real-time operation on Intel(R) Core(TM) i7-10700K 3.80GHz CPU, GeForce RTX 2080s GPU and 16GB RAM, attributed to the SA and lightweight engagement evaluation model mentioned above.

6. Experiment

The experiments were executed on the same environment used in implementation.

To verify the effectiveness of the SA on the video data, we tested the character recognition module without and with the SA on spelling videos of the 20 words that will be used in our user study. The mean speed and speed improvement indicators are shown in Table 1. More detailed results can be seen in the supplementary material.

We trained the attention model of Ma *et al.* [MJXQ21] and our attention module with the same 567 video clips as the training set, and compared the performance of the two modules with the same 141 video clips as the test set. We collected the videos by ourselves and annotate them clip by clip. The length of each video clip is 10 s. Figure 5 shows three video clips with different labels. The performance of Ma *et al.* [MJXQ21] and our attention model on the test set is shown in Table 2. Besides, since the neural network extracts the features of head pose and body pose only related to the participants' actions and angles, the attention module we train and test with adult videos still works well with children's videos in the user study.

Table 1: *Comparison of the speed of ABCNet with or without the Stability Algorithm we proposed in Section 4.1.2.*

Method	FPS	Speed improvement
ABCNet	4.27	-
ABCNet with the SA	**9.58**	**124.27**%

Bold values represent the fastest image processing speed and the speed improvement compared to the case where the Stability Algorithm is not used.

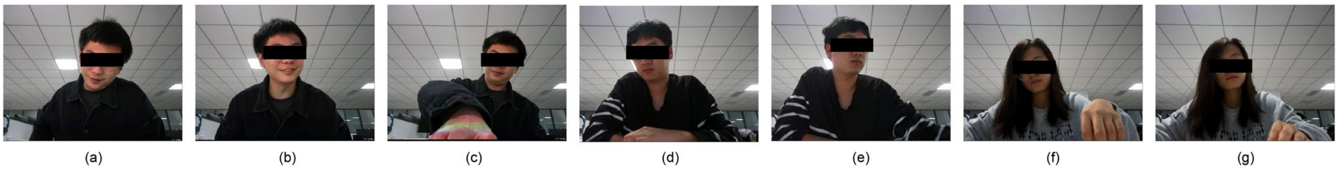

Figure 5: *Three representative video clips, respectively, with label 0.125, 0.5 and 0.875. The frame (a)–(g) are some key frames.*

Table 2: *Comparison of the Temporal Multiple Instance Attention Module. MSE represents the mean squared error between the predicted score and the ground truth. FPS represents the speed at which the two attention modules process video input after they are implemented as online modules through the same multi-process method.*

Method	MSE	FPS
[MJXQ21]	0.026	1.5
Ours	**0.021**	**14.0**

Bold values represent the smallest mean squared error and the fastest video processing speed.

7. User Study

To test the efficacy of our system, we conducted a pilot study with the traditional flashcards group as the control group. In this section, we describe our study and share our results.

7.1. Participants and research design

Given our motivation, pre-school children are optimal in principle. But children should also be able to accept questionnaires to reflect the psychological changes in their learning process, so we chose third-grade primary school students as the subjects of our study, taking into account their age and the research feasibility.

Our study was conducted voluntarily in a primary school, and the sample consisted of 14 Chinese pupils (seven boys and seven girls) who preliminarily mastered the ability to read and recognize English words. The children ranged from 8 to 9 years old. Consent of children and their guardians was obtained for all data collection and possible presentation. We also passed the school's privacy review and ethics committee review.

We designed two groups of experiments where children learn English in different conditions. In group one, children learn with flashcards. In group two, children learn with our IARE. Participants were evenly divided into two groups according to their English level tested in advance. We collected qualitative and quantitative data to analyse the research results. We aimed to investigate whether participants in our system get better psychological feedback. Children in both groups were encouraged to perform better in their own research conditions. They were not informed the hypothesis of the research and were unaware of what was going on in the other group. The domain educator we invited fully participated in the design and implementation of the research and the analysis of the results.

7.2. Procedure

Children were administered a pre-test a week before the research began. The pre-test participant questionnaire included a measure of prior knowledge, intrinsic motivation and self-efficacy. We divided the participants into two groups according to their English level shown in the pretest, ensuring that the overall English level of each group was roughly the same. Words in the experiment were confirmed to be new to the participants. The two groups were trained simultaneously, each in a different classroom, for a total of 25 min. After their respective learning, by flashcards or our system, all children were then given the immediate posttest (enjoyment rating, intrinsic motivation survey, self-efficacy survey and immediate retention test). We assess children's attitudes by calculating the difference between pre-test and post-test self-report measures of intrinsic motivation and self-efficacy [MBGM19]. Finally, 1 week after the research, children were given a delayed retention test to further observe the learning outcome. A 1-week delay for the post-test was chosen since this is a typical window for memory consolidation [FB05].

7.3. Questionnaires

The pre-test participant questionnaire included a measure of prior knowledge, intrinsic motivation and self-efficacy. The prior knowledge scale included 11 items assessing students' prior knowledge of English. The first question asked participants to rate their knowledge of English with the categories: very high, somewhat high, medium, somewhat low and very low. Ten other items assessed their knowledge (*e.g.* 'I know what this word means in Chinese'). The intrinsic motivation for English scale includes five items that were adapted from the intrinsic motivation inventory [Rya82] (*e.g.* 'Activities related to English words learning are fun to perform'). Finally, the self-efficacy scale included five items adapted from [P*91] (*e.g.* 'I'm sure I can handle the most difficult English word'). The intrinsic motivation and self-efficacy items were on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

The post-test participant questionnaire consisted of the same five intrinsic motivation items for learning English words, and the same five questions assessing self-efficacy in English learning as were used in the pretest. Perceived enjoyment was measured with a question adapted from [Tİ15] ('I like to learn English words in this way'). Cognitive benefits were measured with a question adapted from Makransky and Lilleholt [ML18] ('This way of learning makes the memorization and comprehension easier'). The items of perceived enjoyment and cognitive benefits were also on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The post-test also included a retention test, which contains

Table 3: *One-way ANCOVA test results for pre-test-to-post-test change.*

	Group			
Measure	Flash cards	IARE	p	F
Change in motivation	0.09(0.45)	0.71(1.58)	0.001	17.708
Change in self-efficacy	−0.51(0.53)	0.29(0.55)	<0.01	33.354

10 questions asking participants to give the Chinese meaning of the English words learned in the experiment. Each question counts one point. Finally, the post-test encouraged participants to provide qualitative feedback.

The delayed test participant questionnaire consisted of the same 10 retention questions as were used in the post-test.

7.4. Psychological feedback

The participants' psychological feedback was evaluated on four main criteria: intrinsic motivation, self-efficacy, enjoyment and cognitive benefits.

7.4.1. *Internal validity*

The scale of intrinsic motivation and self-efficacy includes multiple items. To ensure the validity of the research results, we calculated their reliability coefficients using Cronbach's alpha. The results indicated that the intrinsic motivation scale had reliability values of 0.810 and 0.866 for the pre-test and post-test in the flashcards group, 0.861 and 0.758 in the IARE group, respectively, and the self-efficacy scale had values of 0.850 and 0.833 for the pre-test and post-test in flashcards group, 0.891 and 0.935 in IARE group.

7.4.2. *Changes in intrinsic motivation and self-efficacy*

Table 3 shows the mean pre-test-to-post-test change (and standard deviations) in intrinsic motivation and self-efficacy rating of the two groups. A one-way analysis of covariance (ANCOVA) with the pre-test score as the covariate and the post-test score as the dependent variable showed that there was a significant difference between the two groups (p= 0.001 for intrinsic motivation and p< 0.01 for self-efficacy). The results showed that the users of our system showed significant improvement in intrinsic motivation and self-efficacy. The results also showed that the users of flashcards showed an apparent decline in self-efficacy. Through the observation of the experimental process, we infer that there are two main reasons. One is that children were at a loss when facing difficult words independently; the other is that it is difficult for children to stay focused for a long time by using flashcards alone. These led to poor performance in the posttest retention test and the decline in self-efficacy. The results supported that participants in our system get better psychological feedback.

7.4.3. *Enjoyment and cognitive benefits*

Table 4 shows the distribution of the two groups' enjoyment and cognitive benefits scores. The results showed that compared with group 1, the scores of group 2 were more concentrated in the high-scoring area, which indicates that participants prefer learning with our system and believe that they get more cognitive benefits.

Table 4: *Enjoyment and cognitive benefits results.*

	Enjoyment		Cognitive benefits	
Score	Flash cards	IARE	Flash cards	IARE
1	14.3%	0	14.3%	0
2	14.3%	0	14.3%	0
3	42.8%	0	42.8%	0
4	0	14.3%	14.3%	28.6%
5	28.6%	85.7%	14.3%	71.4%

Table 5: *One-way ANOVA test results for the immediate and delayed retention test.*

	Group			
Measure	Flash cards	IARE	p	F
Immediate retention test	4.67(2.65)	7.00(1.58)	0.037	5.158
Delayed retention test	3.78(1.86)	5.33(1.32)	0.057	4.193

7.5. User performance evaluation

The immediate and delayed retention test scores (out of 10) were compared for participants of the two groups to measure their performance.

Table 5 shows the mean score (and standard deviations) in immediate and delayed retention tests of the two groups. A one-way analysis of variance (ANOVA) test was used to analyse the difference. As the table shows, in the immediate retention test, the mean score of group 2 (7.00) is 2.33 higher than that of group 1 (4.67), and there was a significant difference between the two groups ($p < 0.05$); in the delayed retention test, the mean score of group 2 (5.33) is 1.55 higher than that of group 1 (3.78); however, the difference between the two groups was not statistically significant, with p slightly higher than 0.05. The lack of significance could be due to participants of group 2 focus more on spelling while interacting with physical letters rather than attending to words' meaning, which is what the retention tests examine. In Section 8, we describe how we plan to change our system design to focus more attention on words' meaning. Overall, the performance of group 2 was obviously better than that of group 1. The results supported that participants in our system get a better performance. And this provides a side view evidence for better psychological feedback in the IARE group.

8. Limitations and Future Directions

The major limitation of our work is that we conducted a small, short-term pilot study on a prototype application. In particular, testing the efficacy of language learning would require a longer-term study with a large population of test subjects. However, due to the COVID-19 pandemic, we will not be able to organize large-scale research in the short term. In addition, in order to reduce exposure, online questionnaires were used for pre-test and delayed test. Participants may cheat, and the data may have certain errors.

We received valuable feedback and helpful recommendations for our future work during the study. For example, most participants said that they liked to learn words by using physical letters, but others said that they spent too much time looking for the specified letters. Our next step is to refine and develop IARE. We plan to move words and their pictures to the same side of the interface in the learning mode, so that children will not ignore the meaning of the words while focusing on the spelling. We also plan to design more conspicuous labels for physical letters so that children can quickly find the letters they want.

The future directions of this work include the introduction of speech recognition and natural language processing technology to analyse the intention of children's responses for more refined guidance; the introduction of a robotic arm to further materialize the IA and its prompts; the migration of IARE to other learning scenarios, such as programming language learning.

9. Conclusion

This paper investigated how to integrate AR and AI to accompany children's learning. We propose an intelligent companion learning system. It introduces an IA to perceive children's attention and learning status to regulate the learning process, trying to stimulate positive effects on children's psychological level. Our pilot study showed that our system could give better psychological feedback, with children's intrinsic motivation and self-efficacy significantly improved and a better retention test result as the side view evidence. Thus, we argue that there is a large potential for IA to replace parents to accompany and guide children's learning, and AI combined with AR based on pedagogical principles can help children improve their self-learning performance.

Acknowledgements

We gratefully acknowledge the anonymous reviewers for their comments to help us to improve our paper, and thank for their enormous help in revising this paper. And we would like to thank all the volunteers from the Primary School Attached to Shandong Normal University who participated in our study. This work is partially supported by the Center-initiated Research Project of Zhejiang Lab (No. 2021NB0AL01), Shandong Provincial Natural Science Foundation (No. ZR2022LZH007), NSF of China (Nos. 62172260, 62202425, 62272399).

References

[AKS20] Abdi S., Khosravi H., Sadiq S.: Modelling learners in crowdsourcing educational systems. In *Proceedings of the Artificial Intelligence in Education: 21st International Conference, AIED 2020, July 6–10, 2020, Part II* 21 (Ifrane, Morocco, 2020), Springer, pp. 3–9.

[ASS02] Avcibas I., Sankur B., Sayood K.: Statistical evaluation of image quality measures. *Journal of Electronic Imaging 11*, 2 (2002), 206–223.

[BDCM20] Botarleanu R.-M., Dascalu M., Crossley S. A., McNamara D. S.: Sequence-to-sequence models for automated text simplification. In *Proceedings of the Artificial Intelligence in Education: 21st International Conference, AIED 2020, July 6–10, 2020, Part II* 21, (Ifrane, Morocco, 2020), Springer, pp. 31–36.

[Cad20] Cader A.: The potential for the use of deep neural networks in e-learning student evaluation with new data augmentation method. In *Proceedings of the Artificial Intelligence in Education: 21st International Conference, AIED 2020, July 6–10, 2020, Proceedings, Part II* 21 (Ifrane, Morocco, 2020), Springer, pp. 37–42.

[CCH16] Chen C. H., Chou Y. Y., Huang C. Y.: An augmented-reality-based concept map to support mobile learning for science. *The Asia-Pacific Education Researcher 25*, 4 (2016), 567–578.

[CEF20a] Chanaa A., El Faddouli N.-E.: Bert and prerequisite based ontology for predicting learner's confusion in MOOCs discussion forums. In *Proceedings of the Artificial Intelligence in Education: 21st International Conference, AIED 2020, July 6–10, 2020, Part II* 21, (Ifrane, Morocco, 2020), Springer, pp. 54–58.

[CEF20b] Chanaa A., El Faddouli N.-E.: Predicting learners need for recommendation using dynamic graph-based knowledge tracing. In *Proceedings of the Artificial Intelligence in Education: 21st International Conference, AIED 2020, July 6–10, 2020, Part II* 21, (Ifrane, Morocco, 2020), Springer, pp. 49–53.

[CLL*15] Cortes C., Lawarence N., Lee D., Sugiyama M., Garnett R.: Advances in neural information processing systems 28. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems* (2015).

[CLP*20] Chen P., Lu Y., Peng Y., Liu J., Xu Q.: Identification of students' need deficiency through a dialogue system. In *Proceedings of the Artificial Intelligence in Education: 21st International Conference, AIED 2020, July 6–10, 2020, Part II* 21 (Ifrane, Morocco, 2020), Springer, pp. 59–63.

[CWB*17] Choi H., Wang Z., Brooks C., Collins-Thompson K., Reed B. G., Fitch D.: Social work in the classroom? A tool to evaluate topical relevance in student writing. In *Workshop and Tutorials Chairs* (2017), ERIC, pp. 386.

[CYH14] Chiang T. H., Yang S. J., Hwang G.-J.: An augmented reality-based mobile learning system to improve students' learning achievements and motivations in natural science inquiry activities. *Journal of Educational Technology & Society 17*, 4 (2014), 352–365.

[DD14] Dunleavy M., Dede C.: Augmented reality teaching and learning. *Handbook of Research on Educational Communications and Technology* (2014), New York, Springer, pp. 735–745.

[DLSC20] Draxler F., Labrie A., Schmidt A., Chuang L. L.: Augmented reality to enable users in learning case grammar from their real-world interactions. In *Proceedings of the 2020 CHI*

Conference on Human Factors in Computing Systems (2020), pp. 1–12.

[FB05] Frankland P. W., Bontempi B.: The organization of recent and remote memories. *Nature Reviews Neuroscience 6*, 2 (2005), 119–130.

[FYZL21] Feng W., Yin F., Zhang X.-Y., Liu C.-L.: Semantic-aware video text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 1695–1705.

[GLY*19] Guo X., Li S., Yu J., Zhang J., Ma J., Ma L., Liu W., Ling H.: PFLD: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859* (2019).

[GMB*20] Gliser I., Mills C., Bosch N., Smith S., Smilek D., Wammes J. D.: The sound of inattention: Predicting mind wandering with automatically derived features of instructor speech. In *Proceedings of the Artificial Intelligence in Education: 21st International Conference, AIED 2020, July 6–10, 2020, Part I* 21, (Ifrane, Morocco, 2020), Springer, pp. 204–215.

[GSV19] George N., Sijimol P., Varghese S. M.: Grading descriptive answer scripts using deep learning. *International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8*, 5 (2019).

[HMMH21] Hadi Mogavi R., Ma X., Hui P.: Characterizing student engagement moods for dropout prediction in question pool websites. *Proceedings of the ACM on Human-Computer Interaction5, CSCW1* (2021), 1–22.

[Hof96] Hoffman B.: anaging the information evolution: Planning the integration f school technology. *Nassp Bulletin 80*, 582 (1996), 89–98.

[HS97] Hochreiter S., Schmidhuber J.: Long short-term memory. *Neural Computation 9*, 8 (1997), 1735–1780.

[Kel87] Keller J. M.: Development and use of the arcs model of instructional design. *Journal of Instructional Development 10*, 3 (1987), 2–10.

[KK92] Kalman B. L., Kwasny S. C.: Why tanh: Choosing a sigmoidal function. In *Proceedings of the IJCNN International Joint Conference on Neural Networks* (1992), IEEE, vol. *4*, pp. 578–581.

[LCS*20] Liu Y., Chen H., Shen C., He T., Jin L., Wang L.: ABCNet: Real-time scene text spotting with adaptive Bezier-curve network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9809–9818.

[LKWW16] Limsukhawat S., Kaewyoun S., Wongwatkit C., Wongta J.: A development of augmented reality-supported mobile game application based on jolly phonics approach to enhancing English phonics learning performance of ESL learners. In *Proceedings of the 24th International Conference on Computers in Education. India: Asia-Pacific Society for Computers in Education* (2016).

[LLCH20] Lu S.-J., Liu Y.-C., Chen P.-J., Hsieh M.-R.: Evaluation of AR embedded physical puzzle game on students' learning achievement and motivation on elementary natural science. *Interactive Learning Environments 28*, 4 (2020), 451–463.

[LXQ22] Li J., Xu S., Qin X.: A hierarchical model for learning to understand head gesture videos. *Pattern Recognition 121* (2022), 108256.

[MBGM19] Makransky G., Borre-Gude S., Mayer R. E.: Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. *Journal of Computer Assisted Learning 35*, 6 (2019), 691–707.

[MH20] Mitrović A., Holland J.: Effect of non-mandatory use of an intelligent tutoring system on students' learning. In *Proceedings of the Artificial Intelligence in Education: 21st International Conference, AIED 2020, July 6–10, 2020, Part I* 21 (Ifrane, Morocco, 2020), Springer, pp. 386–397.

[MJXQ21] Ma J., Jiang X., Xu S., Qin X.: Hierarchical temporal multi-instance learning for video-based student learning engagement assessment. In *IJCAI* (2021), pp. 2782–2789.

[ML18] Makransky G., Lilleholt L.: A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development 66*, 5 (2018), 1141–1164.

[NNT*21] Nguyen N., Nguyen T., Tran V., Tran M.-T., Ngo T. D., Nguyen T. H., Hoai M.: Dictionary-guided scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 7383–7392.

[Oso19] Osokin D.: Real-time 2D multi-person pose estimation on CPU: Lightweight openpose. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2019* (2019), pp. 744–748.

[P*91] Pintrich P. R., Smith D. A. F., Duncan T., Mckeachie W. J.: A manual for the use of the motivated strategies for learning questionnaire (MSLQ). *Ann Arbor. Michigan 48109* (1991), 1259.

[PPŘ*17] Pelánek R., Papoušek J., Řihák J., Stanislav V., Nižnan J.: Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction 27* (2017), 89–118.

[Rya82] Ryan R. M.: Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology 43*, 3 (1982), 450.

[Tİ15] Tokel S. T., İsler V.: Acceptance of virtual worlds as learning space. *Innovations in Education and Teaching International 52*, 3 (2015), 254–264.

[TMBF17] Tobar-Muñoz H., Baldiris S., Fabregat R.: Augmented reality game-based learning: Enriching students' experience during reading comprehension activities. *Journal of Educational Computing Research 55*, 7 (2017), 901–936.

[VNL*17] Vazquez C. D., Nyati A. A., Luh A., Fu M., Aikawa T., Maes P.: Serendipitous language learning in mixed reality. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017), pp. 2172–2179.

[WB96] Weiser M., Brown J. S.: Designing calm technology. *PowerGrid Journal 1*, 1 (1996), 75–85.

[WWSS18] Wang L., Wang Y., Shan S., Su F.: Scene text detection and tracking in video with background cues. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval* (2018), pp. 160–168.

[WYG*19] Wang K., Yang J., Guo D., Zhang K., Peng X., Qiao Y.: Bootstrap model ensemble and rank loss for engagement intensity regression. In *Proceedings of the 2019 International Conference on Multimodal Interaction* (2019), pp. 551–556.

[WYWH20] Wu J., Yang B., Wang Y., Hattori G.: Advanced multi-instance learning method with multi-features engineering and conservative optimization for engagement intensity prediction. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (2020), pp. 777–783.

[WZW*19] Wu J., Zhou Z., Wang Y., Li Y., Xu X., Uchida Y.: Multi-feature and multi-instance learning with anti-overfitting strategy for engagement intensity prediction. In *Proceedings of the 2019 International Conference on Multimodal Interaction* (2019), pp. 582–588.

[YCLC19] Yang T.-Y., Chen Y.-T., Lin Y.-Y., Chuang Y.-Y.: FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 1087–1096.

[YHP*21] Yu H., Huang Y., Pi L., Zhang C., Li X., Wang L.: End-to-end video text detection with online tracking. *Pattern Recognition 113* (2021), 107791.

[ZCL*21] Zhu Y., Chen J., Liang L., Kuang Z., Jin L., Zhang W.: Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 3123–3131.

[ZL22] Zhang C., Liu X.: Dense embeddings preserving the semantic relationships in wordnet. In *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN)* (2022), IEEE, pp. 01–08.

[ZSHC14] Zhang J., Sung Y.-T., Hou H.-T., Chang K.-E.: The development and evaluation of an augmented reality-based armillary sphere for astronomical observation instruction. *Computers & Education 73* (2014), 178–188.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information

Video S1

Video S2