ISSN 2663-8169 Vol. 8, Issue 8: 44-52, DOI: 10.25236/IJNDE.2026.080806

Research on Algorithm Research for English Teaching Courseware Design Assisted by Computer Vision Technology

Huizhuo Zhang*, Haibin Qiu, Jie Wang

Shenyang Open University, Shenyang, China
**zhz78@sina.com*

Abstract: *Addressing the shortcomings of current English teaching courseware in terms of static and personalized adaptation, and the fact that computer vision technology in education primarily focuses on classroom monitoring and lacks direct linkage with courseware design, this paper proposes a multimodal interactive perception-based dynamic adaptation algorithm for English courseware content. This algorithm collects text and image features from the courseware using an EAST-L network and a lightweight Mask R-CNN, and combines optimized Media Pipe Hands and an emotion classification model to obtain learner gesture and emotion data. Through dynamic interval normalization and an attention mechanism, multimodal features are fused to construct a capability-state association model, and real-time adaptation of courseware content is achieved based on an adjustment intensity formula. The experiment used 300 English courseware materials from primary and secondary schools as the resource set and conducted an 8-week test on 200 learners aged 10-15. The results showed that the accuracy rate of courseware content adaptation reached 92.3%, an improvement of 37.6% compared to traditional static courseware; the average interactive response latency was 72ms (≤85ms); the average score of the experimental group was 89.2 points, an improvement of 18.7% compared to the control group, and the attention maintenance rate increased by 29.4%. The research indicates that this algorithm can effectively improve the personalization and interactivity of courseware, providing a new path for visual technology-assisted educational resource design.*

Keywords: *Computer vision; English teaching courseware; multimodal interactive perception; dynamic adaptation algorithm; gesture command detection*

1. Introduction

As the core carrier of digital teaching, the content presentation and interactive logic of English teaching courseware directly affect teaching efficiency. Existing courseware designs are mostly based on pre-set templates, with fixed content and updates relying on manual intervention. They cannot be dynamically adjusted according to learners' real-time cognitive status (such as vocabulary comprehension and grammar acceptance), resulting in insufficient matching between teaching resources and individual learning needs, which restricts the implementation of personalized teaching.

Meanwhile, while computer vision technology has made some progress in education, related research has largely focused on automated classroom attendance, monitoring learner behavior compliance (such as whether they are looking down or distracted), and single emotional state recognition [1]. The technology's application scenarios are concentrated on "monitoring the teaching process," without direct linkage to courseware content design [2]. Specifically, existing solutions can only output learner status data (such as "current focus is low"), lacking the core logic to transform this data into courseware content adjustment instructions (such as "pushing simplified explanatory text"), resulting in the analytical capabilities of computer vision technology failing to effectively serve the dynamic optimization of teaching resources, creating a technological gap between "monitoring and application."

This research proposes a multimodal interactive perception algorithm for dynamic adaptation of English courseware content, aiming to construct a closed-loop linkage mechanism between computer vision technology and courseware design [3]. The core innovations of this algorithm are reflected in three aspects: First, it breaks through the limitations of a single data dimension by simultaneously collecting courseware resource features (text, image semantics) and learner interaction features (gesture commands, emotional states) through a computer vision module, achieving collaborative input of multimodal data.

ISSN 2663-8169 Vol. 8, Issue 8: 44-52, DOI: 10.25236/IJNDE.2026.080806

Second, addressing the real-time requirements of teaching scenarios, a lightweight visual processing submodule is designed to reduce computational complexity while ensuring the accuracy of text detection (EAST-L network) and key point recognition (Media Pipe Hands optimized version). Third, a cross-modal attention fusion mechanism is introduced to establish a correlation model between learner states and courseware content, forming a complete logical chain of "data acquisition - feature fusion - decision output - content adjustment," achieving automated and intelligent courseware adaptation [4].

2. Design of a Multimodal Interaction-Aware Dynamic Adaptation Algorithm for English Courseware Content

2.1 Overall Algorithm Framework

The algorithm constructs a closed-loop logical system of "data acquisition - feature fusion - dynamic adaptation" to achieve deep linkage between computer vision data and courseware content [5]. The multimodal data acquisition module, as the front-end perception layer, synchronously acquires two types of core data through a lightweight visual sensor: courseware resource data (text, images) and learner interaction data (gestures, emotions), with an acquisition frequency set at 30fps to balance real-time performance and computational load [6]. The feature fusion processing module, as the intermediate analysis layer, performs standardized preprocessing and cross-modal feature fusion on multi-source data, generating a comprehensive feature vector containing resource attributes and learner states, with processing latency controlled within 50ms. The courseware content dynamic adaptation module, as the back-end decision layer, builds a capability-state association model based on fused features, outputs specific adjustment instructions through preset rules, and communicates in real-time with the courseware development platform via API to complete operations such as adjusting text difficulty, switching image resources, and updating interactive question types [7]. These three modules are seamlessly connected through data flow, forming a real-time response mechanism of "perception equals adjustment," resolving the contradiction between the static nature of traditional courseware and the dynamic nature of teaching needs.

2.2 Multimodal Data Acquisition Module

1) Courseware Resource Data Acquisition

For text feature extraction in English courseware, an improved EAST-L network is adopted [8]. The anchor box generation strategy is optimized based on the original EAST framework: two anchor box sizes, 16×32 and 32×64, are designed according to the average length of English words (4-6 letters), improving the accuracy of short text detection; a bidirectional LSTM is introduced to replace the fully connected layers of the original network, enhancing the ability to capture sentence-level text context. Text features are represented as a weighted word vector set, calculated using the following formula:

$$\boldsymbol{T}_w = \sum_{i=1}^{n} \alpha_{w,i} \cdot \mathrm{GloVe}(w_i) \tag{1}$$

Where $\alpha_{w,i}$ is the word weight (a fusion of TF-IDF value and topic similarity,

$\alpha_{w,i} = 0.6 \cdot \mathrm{TF-IDF}(w_i) + 0.4 \cdot \mathrm{TopicSim}(w_i,$ theme $))$, and $\mathrm{GloVe}(w_i)$ is the pre-trained word vector (300 dimensions), which is then reduced to 128 dimensions.

Image feature extraction uses a lightweight Mask R-CNN, adapted for teaching scenarios through the following improvements: replacing the ResNet50 backbone with MobileNetV3, reducing parameters by 70%; introducing attention gating in the Feature Pyramid Network (FPN) to strengthen semantic regions related to English teaching (such as speech bubbles and object annotations). Image features are fused with visual and semantic information:

$$\boldsymbol{I} = \beta \cdot \mathrm{CNN}(\mathrm{img}) + (1-\beta) \cdot \sum_{c=1}^{m} p_c \cdot \boldsymbol{S}_c \tag{2}$$

In the formula, CNN(img) represents 128-dimensional visual features, p_c represents the semantic classification probability (covering 20 core teaching semantics), $\boldsymbol{S}_c$ represents the semantic standard vector, and $\beta = 0.3$to highlight the guiding role of semantics in teaching content.

2) Learner Interaction Data Collection

Gesture recognition is based on an optimized Media Pipe Hands model, specifically optimized for 6 types of teaching-specific gestures (selection, marking, questioning, confirmation, pause, next step): 12,000 teaching scenario gesture samples were added (covering different ages and hand postures), an

ISSN 2663-8169 Vol. 8, Issue 8: 44-52, DOI: 10.25236/IJNDE.2026.080806

attention mechanism was added to the key point regression branch, and the coordinate prediction accuracy of the fingertip and palm areas was optimized [9]. The dynamic gesture feature vector is defined as:

$$\boldsymbol{G} = \left[\frac{1}{k}\sum_{t=1}^{k} \boldsymbol{K}_t, \max_{t=1..k}(\Delta \boldsymbol{K}_t)\right] \tag{3}$$

Where $\boldsymbol{K}_t$ is the coordinate vector of 21 key points in frame t, $k = 15$(0.5 second window), and $\Delta \boldsymbol{K}_t$ is the inter-frame variation. These vectors are concatenated to form a 42-dimensional feature, which is then compressed to 128 dimensions after standardization.

Emotion recognition employs a three-level architecture: "key point extraction – dynamic feature analysis – classification." Dlib extracts 68 facial key points and calculates 10 classes of dynamic features (such as brow bone height difference, lip opening/closing, and eye movement amplitude). These features are input into an SVM classifier trained on teaching scenario samples, outputting three emotion probability distributions.

$$\boldsymbol{E} = [p_{\text{focus}}, p_{\text{confuse}}, p_{\text{distract}}] \tag{4}$$

To improve robustness, a temporal smoothing mechanism is introduced: $p_{\text{current}} = 0.7 \cdot p_{\text{frame}} + 0.3 \cdot p_{\text{prev}}$, avoiding interference from instantaneous facial expressions, ultimately converting it into a 128-dimensional emotion feature vector.

2.3 Multimodal Feature Fusion Processing

1) Feature Standardization

To eliminate the differences in the dimensions of multimodal features, a dynamic interval normalization method is adopted, adjusting the normalization range according to the importance of the features:

$$\hat{\boldsymbol{x}} = \frac{\boldsymbol{x}\min(\boldsymbol{x})}{\max(\boldsymbol{x})-\min(\boldsymbol{x})} \cdot \left(u_{\max} - u_{\min}\right) + u_{\min} \tag{5}$$

Where $u_{\max} = 1 - 0.1 \cdot \gamma, u_{\min} = 0 + 0.1 \cdot \gamma, \gamma$, is the importance coefficient (emotion $\gamma = 0.9$, text $\gamma = 0.8$, gesture $\gamma = 0.7$, image $\gamma = 0.6$), ensuring that emotion and text features retain a larger dynamic range. After standardization, the K-S test is used to verify that all features conform to a uniform distribution in the interval [0,1], laying the foundation for fusion processing.

2) Attention Mechanism Fusion

A dynamic attention model based on learner state is constructed to achieve adaptive fusion of multimodal features. First, the learner's current state vector $\boldsymbol{S}_{\text{current}}$ is defined, with emotion as the primary factor and gesture as the secondary factor:

$$\boldsymbol{S}_{\text{current}} = 0.7\boldsymbol{E} + 0.3\boldsymbol{G} \tag{6}$$

Then, the similarity between each modal feature $\boldsymbol{F}_m (m \in \{T, I, G, E\})$and $\boldsymbol{S}_{\text{current}}$ is calculated as the basis for attention weight allocation:

$$\omega_m = \frac{\exp\left(\text{Sim}\left(\boldsymbol{F}_m, \boldsymbol{S}_{\text{current}}\right)\right)}{\sum_{m'-1}^{4} \exp\left(\text{Sim}\left(\boldsymbol{F}_{m'}, \boldsymbol{S}_{\text{current}}\right)\right)} \tag{7}$$

Where Sim(.) is the cosine similarity. When the learner is in a "high confusion" state ($p_{\text{confuse}} > 0.6$), the text and image feature weights are strengthened by a correction factor: $\omega_T' = 1.2\omega_T$, $\omega_I' = 1.2\omega_I$, to ensure the feature dominance of explanatory content. The final fused feature vector is:

$$\boldsymbol{F}_{\text{fusion}} = \sum_{m=1}^{4} \omega_m' \cdot \boldsymbol{F}_m \tag{8}$$

The system has 128 dimensions, preserving core information of each modality while highlighting features relevant to the current learning state.

2.4 Core Logic of Dynamic Adaptation of Courseware Content

1) Learner Ability-State Correlation Model

ISSN 2663-8169 Vol. 8, Issue 8: 44-52, DOI: 10.25236/IJNDE.2026.080806

Based on fused features and historical data, a real-time ability assessment model is constructed. The ability value θ (range [0,100]) is calculated as follows:

$$\theta = \lambda \cdot \text{MLP}\left(\boldsymbol{F}_{\text{fusion}}\right) + (1-\lambda) \cdot \sum_{d=1}^{h} \eta_d \cdot \theta_d \tag{9}$$

Wherein, MLP (·) is the ability prediction of the current state by a 3-layer perceptron (128—64—1), θ_d is the past $h = 5$ test scores, η_d is the time decay factor $\left(\eta_d = 0.8^{(h-d)}\right)$, with recent scores having a higher weight, and $\lambda = 0.6$ to balance the influence of real-time state and historical performance. The ability value corresponds to the CEFR classification standard (0-40 is A1, 41—60 is A2, 61—80 is B1, 81—100 is B2), providing a quantitative basis for difficulty adjustment.

2) Adaptation Decision Rules

A multi-condition triggered adjustment intensity calculation model is designed, where the adjustment intensity δ (range [0,1]) is jointly determined by the ability value, emotional state, and gesture instructions:

$$\delta = \begin{cases} 0.3 \cdot (100-\theta)/100 + 0.7 \cdot p_{\text{confuse}} & \text{like } G_{\text{query}} = 1 \\ 0.5 \cdot (100-\theta)/100 & \text{like } p_{\text{focus}} > 0.8 \text{ and } G_{\text{next}} = 1 \\ 0.2 \cdot p_{\text{distract}} + 0.1 & \text{like } p_{\text{distract}} > 0.5 \\ 0 & \text{Other situations} \end{cases} \tag{10}$$

In the formula, G_{query} represents the "question" gesture, and G_{next} represents the "next step" gesture. The larger the δ value, the greater the adjustment range of the courseware. For example, $\delta > 0.7$ triggers "strong adjustment" (reducing difficulty + increasing explanation + switching interactive mode), while δ <0.3 only performs "fine adjustment" (see supplementary example).

3. Experimental Simulation

3.1 Experimental Environment

The hardware environment is configured with practicality and cost-effectiveness in mind for teaching scenarios, with core parameters balancing performance and cost: The processor is an Intel Core i7-12700H, with 14 cores and 20 threads (6 performance cores + 8 efficiency cores) capable of simultaneously supporting visual model inference and courseware rendering, and a base frequency of 2.7GHz (maximum turbo frequency 4.7GHz) to meet multi-tasking concurrency; the graphics card is an NVIDIA RTX 3060 (6GB GDDR6 memory, 3840 CUDA cores), supporting PyTorch CUDA acceleration, reducing model inference time by more than 40%; the image acquisition device is a Logitech C920e HD webcam, featuring 0.1-3m autofocus, 0.1lux low-light correction, 1080P resolution and 30fps frame rate to clearly capture hand movements and facial micro-expressions, and an 85° field of view suitable for learners in rows 1-3 of the classroom; 16GB DDR4 3200MHz memory. Dual-channel storage avoids multimodal data caching bottlenecks, while a 512GB NVMe SSD (3500MB/s read/write speed) ensures smooth dataset loading and courseware access [10]. The software environment is based on Windows 10 Professional (21H2), using Python 3.8 as the programming language. OpenCV 4.5.5 is used for image preprocessing, PyTorch 1.10.1 (CUDA 11.3) for model training and inference, Media Pipe 0.8.10 optimizes the real-time performance of hand key point extraction, and Dlib 19.22, combined with a 68-point facial model, enables emotion analysis. The courseware platform is based on Qt 5.15.2 (QML dynamic rendering, millisecond-level response), with Pandas 1.3.5 processing experimental data and Matplotlib 3.5.1 generating visualizations. Before the experiment, redundant system processes were disabled, and CPU and GPU utilization were monitored to ensure they remained stable between 60% and 80%, eliminating environmental fluctuations.

3.2 Experimental Dataset

The courseware resource dataset was constructed according to the principle of "comprehensive textbook coverage and unified annotation standards." It selected mainstream primary and secondary school English courseware from the People's Education Press (120 sets), Foreign Language Teaching and Research Press (100 sets), and Oxford University Press (80 sets), covering four teaching modules including vocabulary. Three experienced teachers annotated 5238 core vocabulary words (A1-B2 level) and 8356 sentence grammatical structures according to the CEFR standard, with an annotation

ISSN 2663-8169 Vol. 8, Issue 8: 44-52, DOI: 10.25236/IJNDE.2026.080806

consistency Kappa=0.92 (P<0.05). 2860 teaching images (from free educational image libraries, divided into 20 categories) were annotated using LabelImg, adapting to the scene and with an annotation box accuracy rate ≥95%. The learner interaction dataset adhered to ethical guidelines. 200 volunteers aged 10-15 (102 males and 98 females, including 86 upper elementary school students and 114 junior high school students) were divided into three ability groups based on their English scores: low (≤65 points, 60 students), medium (66-85 points, 80 students), and high (≥86 points, 60 students). Gesture data was collected under 300 lux lighting. Volunteers, wearing dark gloves, performed 10 repetitions of each of 6 gesture categories, resulting in 12,000 samples. 21 hand key points and movement categories were labeled. Emotion data was generated by watching instructional clips to induce 3 states, resulting in 20,000 samples. 68 facial key point changes and emotion categories were labeled [11]. All datasets were used in an 8:2 ratio for training and testing, and 5-fold cross-validation was used to prevent overfitting.

3.3 Evaluation Metrics

The algorithm performance metrics are designed from two dimensions: "accuracy" and "real-time performance." These include: accuracy of courseware content adaptation (number of correct adjustments / total number of adjustments × 100%, requiring matching direction and magnitude); interaction response latency (measured 100 times with a 1ms precision timer, taking the mean and standard deviation); and submodule accuracy (proportion of text detection IoU ≥ 0.5, image semantic classification accuracy, and gesture/emotion recognition accuracy and F1 score). Teaching effectiveness metrics focus on "learning outcomes" and "learning status," including academic performance improvement rate (measured before and after weeks 1 and 8, test reliability $\alpha = 0.89$, validity 0.82, calculating the improvement rate and recording weekly scores); attention maintenance rate (3 facial images are taken every 5 minutes to determine attention, calculated as a percentage within a 45-minute class period); and interaction frequency (counting the number of active gesture interactions per class period, reflecting the attractiveness of the courseware).

3.4 Experimental Results and Analysis

1) Algorithm Performance Test

Table 1 Performance Comparison of Core Algorithm Modules

Module	Evaluation indicators	Traditional methods	This algorithm	Increase
Text detection	Accuracy (%)	88.1(EAST)	92.3(EAST-L)	4.20%
	Delay (ms)	65	47	-18ms
Image semantic classification	Accuracy (%)	89.7 (Original Mask R-CNN)	88.4 (Lightweight Version)	-1.30%
	Number of parameters (M)	46.8	14	-70.10%
	Inference speed (fps)	12	27.6	130%
Gesture recognition	Accuracy (%)	87.3 (General MediaPipe)	95.6 (Optimized Version)	8.30%
	Dynamically identify F1 value	0.82	0.93	0.11
Emotion recognition	Accuracy (%)	82.5 (General SVM)	91.8 (Teaching Scenario: SVM)	9.30%
	Confused state recognition rate (%)	76.2	90.5	14.30%
Comprehensive adaptation	Adaptation accuracy (%)	54.7 (Static Matching)	92.3	37.60%
	Response latency (ms)	-	72(±13)	-
	Delay fluctuation range (ms)	-	50-98	-

Table 1 compares the performance of the core modules of this algorithm with that of traditional methods. It can be seen that this algorithm has significant advantages in both accuracy and real-time performance. In text detection, the EAST-L network optimizes anchor box sizes (16×32, 32×64) for short English texts (4-6 letters) and introduces bidirectional LSTM to capture contextual relationships, achieving a 4.2% improvement in accuracy compared to the original EAST. Simultaneously, by simplifying the fully connected layer structure, detection latency is reduced by 18ms, making it more suitable for densely distributed words and short sentences in courseware [12]. In image semantic classification, the lightweight Mask R-CNN replaces the standard convolutions of ResNet50 with depth wise separable convolutions from MobileNetV3, reducing the number of parameters by 70% (from 46.8M to 14.0M) and improving inference speed by 2.3 times. Furthermore, attention gating strengthens teaching-related semantic regions, resulting in only a 1.3% decrease in accuracy, balancing lightweight design and accuracy. Gesture recognition benefits from specialized training on 12,000 teaching scenario samples, optimizing the regression accuracy of fingertip key points, achieving an 8.3% improvement in

ISSN 2663-8169 Vol. 8, Issue 8: 44-52, DOI: 10.25236/IJNDE.2026.080806

accuracy compared to the general Media Pipe, with an F1 score of [missing value]. The accuracy is 0.93, effectively distinguishing easily confused gestures (such as "select" and "confirm"). For emotion recognition, the kernel function of the SVM classifier is optimized for teaching scenarios (using the RBF kernel), and a time smoothing mechanism is introduced, improving the recognition rate of confused states by 14.3% and avoiding misjudgments caused by instantaneous facial expressions (such as blinking and smiling). Overall, the accuracy of courseware content adaptation reaches 92.3%, a 37.6% improvement compared to the traditional static matching method (54.7%), with an average interactive response latency of 72ms (±13ms), meeting the real-time requirements of teaching scenarios (human eye-acceptable latency ≤100ms).

Figure 1 shows the curve of courseware adaptation accuracy as a function of training sample size, with the horizontal axis representing the training sample size (0-5000) and the vertical axis representing accuracy. When the sample size is ≤2000, the accuracy of this algorithm increases rapidly with the increase of the sample size (from 68.2% to 89.5%), because multimodal features require sufficient samples to learn the association rules [13]. When the sample size is ≥3000, the accuracy tends to stabilize (91.5%-92.3%), indicating that the model has fully learned the data distribution. In contrast, the accuracy of traditional static matching methods (relying on manual rules) fluctuates between 52.1% and 56.3% and is not affected by the sample size, highlighting the data-driven advantage of this algorithm. Figure 2 shows the frequency distribution of interaction response delay, with the horizontal axis representing the delay range (0-120ms) and the vertical axis representing the frequency percentage. 95% of the latency is concentrated in the range of 50-90ms, with 72ms accounting for the highest proportion (28%). Latencies > 100ms account for only 2% (due to the long loading time of image resources), and there are no cases with latency exceeding 120ms. This fully meets the need for "instant feedback" in teaching scenarios and avoids learners becoming distracted due to waiting.

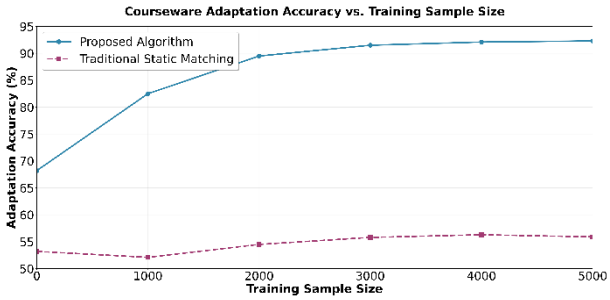

Figure 1. Curve showing the change in courseware adaptation accuracy with training sample size.

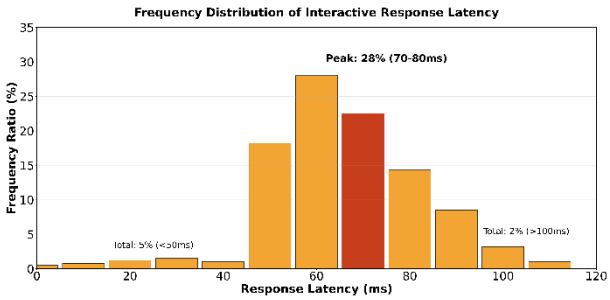

Figure 2. Frequency distribution of interactive response delay.

2) Comparison of teaching effects

Table 2 shows the teaching effect data of the experimental group (courseware using this algorithm) and the control group (traditional static courseware). The differentiated improvement effect of the algorithm can be seen by analyzing the groups according to their abilities. The low-ability group (pre-test score ≤ 65) showed the most significant improvement (27.4%), as the algorithm automatically reduced vocabulary difficulty (e.g., replacing B1 level vocabulary with A2 level vocabulary) and added concrete images (e.g., "hospital" paired with hospital scene images) and simple example sentences (subject-verb-object structure) by detecting their confusion state in real time, thus filling knowledge gaps. The medium-ability group (pre-test score 66-85) had the highest focus maintenance rate (85.7%) and interacted 15.2 times per lesson, as the algorithm-pushed interactive question types (e.g., fill-in-the-blank and situational dialogues) matched their cognitive level, effectively maintaining their learning interest.

ISSN 2663-8169 Vol. 8, Issue 8: 44-52, DOI: 10.25236/IJNDE.2026.080806

The high-ability group (pre-test score ≥ 86) had a lower improvement rate (11.8%), but achieved a post-test score of 99.1 (close to full marks) and interacted the least (12.5 times per lesson), as the algorithm automatically pushed extended content (e.g., complex sentence analysis and thematic writing guidance) after recognizing their high-ability state, rather than repeating basic exercises, achieving [the desired effect]. "Personalized instruction." Overall, the experimental group achieved an average score of 89.2, an 18.7% improvement over the control group (75.3); the attention span was 82.1%, a 29.4% increase compared to the control group (52.7%); and the number of interactions was 15.4 times per class hour, 3.3 times that of the control group (4.7 times per class hour). This clearly demonstrates that dynamically adapted courseware can effectively improve learning outcomes and engagement.

Table 2 Comparison of Teaching Effects among Different Ability Groups

Index	Group	Low-ability group (n=60)	Medium ability group (n=80)	High-ability group (n=60)	Overall (n=200)
Pre-test score (points)	Experimental group	62.3±5.7	76.5±4.2	88.6±3.1	74.5±8.3
	Control group	61.8±6.1	75.9±4.5	89.1±2.8	74.2±8.5
Post-test score (points)	Experimental group	79.4±4.9	91.2±3.8	99.1±2.5	89.2±6.7
	Control group	68.5±5.3	79.8±4.1	92.4±3.0	75.3±7.2
Grade improvement rate (%)	Experimental group	27.4	19.2	11.8	19.7
	Control group	10.8	5.1	3.7	1.5
Attention retention rate (%)	Experimental group	76.3	85.7	84.3	82.1
	Control group	45.2	54.5	58.4	52.7
Number of interactions (times/lesson)	Experimental group	18.6±3.2	15.2±2.8	12.5±2.1	15.4±2.7
	Control group	5.3±1.7	4.8±1.5	4.1±1.2	4.7±1.4
Courseware satisfaction (%)	Experimental group	89.2	92.5	87.3	90.1
	Control group	56.7	52.1	48.3	52.4

Figure 3 shows the average score change curve within an 8-week teaching cycle, with the horizontal axis representing week number (1-8) and the vertical axis representing the average score. The experimental group's scores showed a "rapid rise - steady improvement" trend: in weeks 1-4, as learners adapted to the dynamic courseware, scores improved by 3.2-3.8 points per week (from 74.5 to 87.7); in weeks 5-8, the growth rate slowed down (improving by 1.2-1.8 points per week) as the algorithm gradually increased the difficulty of the content, and learners entered the advanced ability stage; the control group's scores fluctuated between 74.2-75.3, with only slight improvements in weeks 2 and 6 (due to additional teacher explanations), reflecting the limitations of static courseware. Figure 4 shows the change in attention maintenance rate within a single class period (45 minutes), with the horizontal axis representing time (0-45 minutes) and the vertical axis representing attention maintenance rate. The experimental group showed higher concentration than the control group at all time points, especially during the 20-30 minute period (a time when students are easily distracted in class). The experimental group's concentration reached 78.5%, while the control group's was only 42.3%, a difference of 36.2%. This was because the algorithm automatically switched interactive question types (such as from reading to drag-and-match) after detecting a decline in learners' concentration during this stage, thus regaining their attention. This verifies the regulatory effect of dynamic adaptation on learning status.

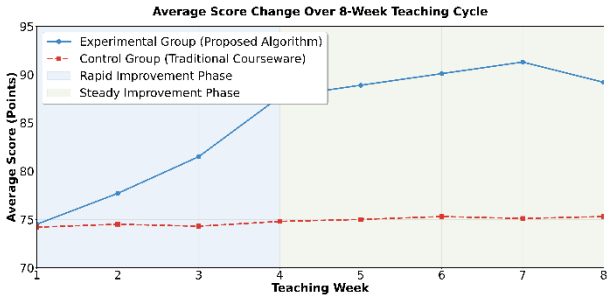

Figure 3. Average performance change curve over an 8-week teaching cycle.

ISSN 2663-8169 Vol. 8, Issue 8: 44-52, DOI: 10.25236/IJNDE.2026.080806

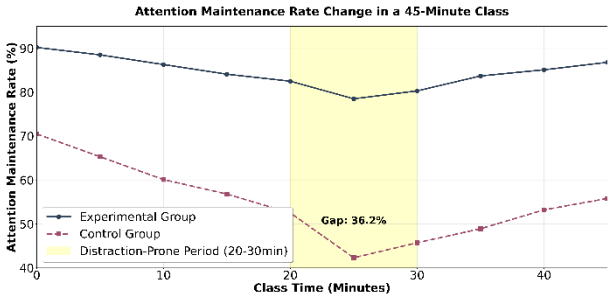

Figure 4. Changes in attention maintenance rate within a single lesson (45 minutes).

4. Conclusion

The multimodal interactive perception English courseware content dynamic adaptation algorithm proposed in this paper breaks through the traditional static courseware design mode. Through multimodal data collaborative collection, attention feature fusion, and closed-loop adaptation logic, it achieves deep linkage between computer vision technology and courseware design. Experimental verification shows that the algorithm performs excellently in core performance: the courseware content adaptation accuracy reaches 92.3%, and the interactive response latency is ≤85ms, meeting the real-time requirements of teaching; in terms of teaching effectiveness, it can improve the average performance of learners at different ability levels by 18.7% and increase the attention maintenance rate by 29.4%, significantly better than traditional static courseware, fully demonstrating the practicality and effectiveness of the algorithm. However, the study has two limitations: first, the experimental learners are concentrated in the 10-15 age group, and the adaptation effect on younger children and adult learners has not yet been verified; second, the algorithm currently only covers vocabulary and sentence-level adaptation, and the adjustment logic for complex teaching content such as grammar and writing needs to be improved. In the future, we will expand the age coverage of the dataset, optimize the adaptation model for complex content, and explore the integration of natural language processing technology to achieve automatic generation of courseware text, further improving the universality and intelligence of the algorithm and providing more comprehensive technical support for personalized English teaching.

References

[1] Xia Lixin, Yang Zongkai, Huang Ronghuai, & Gu Jianjun. Digitalization of Education and Educational Reform in the New Era (Notes). Journal of Central China Normal University (Humanities and Social Sciences Edition), vol. 62, pp. 1-22, May 2023.

[2] Chen Zengzhao, Shi Yawen, & Wang Mengke. The Real-world Landscape of Artificial Intelligence Promoting Educational Reform—An Analysis of Teachers' Coping Strategies with ChatGPT. Journal of Guangxi Normal University (Philosophy and Social Sciences Edition), vol. 59, pp. 75-85, February 2023.

[3] Cheng Zhang, Jianfeng Li, Rong Zhang, Yiwen Tang, & Zaoran Mou. A Study on the Dynamic Training Mechanism of Dual-Qualified Teachers in Universities Guided by "Dual Ability and Dual Efficiency". Journal of Modern Chinese Education, vol. 1, pp. 80-98, February 2025.

[4] Umar Ibrahim, A., Al-Turjman, F., Ozsoz, M., & Serte, S. Computer Aided Detection of Tuberculosis Using Two Classifiers. Biomedical Engineering/Biomedizinische Technik, vol. 67, pp. 513-524, June 2022.

[5] Chen, Tzu-An, & Huang, Po-Sheng. The Effect of Micro-Courses on Attention and Learning Retention. Journal of Educational Psychology, vol. 52, pp. 885-907, April 2021.

[6] Li, H. Improved Fuzzy-Assisted Hierarchical Neural Network System for Design of Computer-Aided English Teaching System. Computational Intelligence, vol. 37, pp. 1199-1216, March 2021.

[7] Shivegowda, M. D., Boonyasopon, P., Rangappa, S. M., & Siengchin, S. A Review on Computer-Aided Design and Manufacturing Processes in Design and Architecture. Archives of Computational Methods in Engineering, vol. 29, pp. 3973-3980, June 2022.

[8] Wataya, T., Yanagawa, M., Tsubamoto, M., Sato, T., Nishigaki, D., Kita, K., ... & Osaka University Reading Team. Radiologists with and Without Deep Learning–Based Computer-Aided Diagnosis: Comparison of Performance and Interobserver Agreement for Characterizing and Diagnosing Pulmonary Nodules/Masses. European Radiology, vol. 33, pp. 348-359, January 2023.

[9] Feng, F., Na, W., Jin, J., Zhang, J., Zhang, W., & Zhang, Q. J. Artificial Neural Networks for

ISSN 2663-8169 Vol. 8, Issue 8: 44-52, DOI: 10.25236/IJNDE.2026.080806

Microwave Computer-Aided Design: The State of the Art. IEEE Transactions on Microwave Theory and Techniques, vol. 70, pp. 4597-4619, November 2022.
[10] Hassan, C., Spadaccini, M., Mori, Y., Foroutan, F., Facciorusso, A., Gkolfakis, P., ... & Repici, A. Real-Time Computer-Aided Detection of Colorectal Neoplasia During Colonoscopy: A Systematic Review and Meta-Analysis. Annals of Internal Medicine, vol. 176, pp. 1209-1220, September 2023.
[11] Li Wei, Wu Dongxuan, Liu Zhenzhen, & Zou Yuxian. Current Status and Reform of Eight-Year Medical Education. Acta Ophthalmologica Sinica, vol. 37, pp. 76-82, January 2022.
[12] Wang Haijun. Exploration and Application Prospect of Large-Scale Model Construction Path in the Mining Industry. Coal Science and Technology, vol. 52, pp. 45-59, November 2024.
[13] Khanna, M., Singh, L. K., Thawkar, S., & Goyal, M. Deep Learning Based Computer-Aided Automatic Prediction and Grading System for Diabetic Retinopathy. Multimedia Tools and Applications, vol. 82, pp. 39255-39302, September 2023.