Received: 13 June 2023 | Accepted: 7 September 2023

DOI: 10.1002/cav.2218

SPECIAL ISSUE PAPER

WILEY

GestureTeach: A gesture guided online teaching interactive model

Hongjun Liu[1] | **Chao Yao**[1] | **Yalan Zhang**[1] | **Xiaojuan Ban**[1,2]

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, Institute of Artificial Intelligence, School of Computer and Communication Engineering, School of Intelligence Science and Technology, Key Laboratory of Intelligent Bionic Unmanned Systems, Ministry of Education, University of Science and Technology Beijing, Beijing, China

[2]Institute for Materials Intelligent Technology, Liaoning Academy of Materials, Shenyang, China

Correspondence
Chao Yao and Xiaojuan Ban, Beijing Advanced Innovation Center for Materials Genome Engineering, Institute of Artificial Intelligence, School of Computer and Communication Engineering, School of Intelligence Science and Technology, Key Laboratory of Intelligent Bionic Unmanned Systems, Ministry of Education, University of Science and Technology Beijing, Beijing, 100083, China.
Email: yaochao@ustb.edu.cn and banxj@ustb.edu.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2022ZD0118001; National Natural Science Foundation of China, Grant/Award Numbers: 61972028, 62332017, U22A2022; Guangdong Basic and Applied Basic Research Foundation

Abstract

Online education has become more popular and effective due to the availability of high-speed internet and technological innovations, which allow people from different locations to access educational resources and opportunities. However, online classes often face challenges such as limited interactivity and display options, which can affect the quality and effectiveness of the online learning experience. In this article, we propose GestureTeach, a new pedagogical paradigm that enables free handwriting interaction and animation generation for online teaching. GestureTeach uses gestures as a natural and intuitive way of interaction, which enhances the teacher's intention expression and the student's engagement. GestureTeach also generates animations from handwritten sketches, which improves the display effects of the interaction and the student's knowledge comprehension. We conducted a two-stage study with 15 teachers and 90 students to evaluate the effectiveness of GestureTeach in facilitating classroom interaction. The results show that GestureTeach is preferred by both teachers and students over traditional online teaching methods and has the potential to transform the online teaching landscape by providing a seamless and interactive experience.

KEYWORDS

animation generation, gesture recognition, online education

1 | INTRODUCTION

In recent years, the importance of online education has become increasingly prominent, leading to a surge in online teaching, with 90% of it being a simple classroom relocation.[1] However, 80% of college and university teachers have never participated in online teaching through live broadcasts, leading to challenges in adapting to this new format.[2]

One significant advantage of offline teaching is the ability to present information on a blackboard through writing demonstrations using chalk with high display efficiency, a clear display process, and low hardware dependence. However, these advantages are difficult to replicate in online teaching scenarios through existing platforms and technologies. To tackle the challenge of the limitations and drawbacks of existing online teaching methods, we propose GestureTeach, a novel pedagogical paradigm that enables high freedom handwriting and vivid animation generation. As depicted in Figure 1, teachers make use of predesigned gestures captured by regular cameras to facilitate the process of handwriting, such as forming the word "apple" or creating arbitrary sketches. Subsequently, the corresponding animations are intelligently generated, providing a user-friendly teaching method for both teachers and students.

To enable highly freedom handwriting, we have designed a rapid multi-stage pipeline that enhances the fluency of the handwriting process and allows for the creation of more user-friendly interactive gestures. To further enhance the aesthetics of handwritings, GestureTeach includes an automatic text animation that enhances the readability of the text by converting handwritten fonts to print fonts. Moreover, GestureTeach generates high-quality image animations through a quick sketch composed of hand tips trajectories, combining a diffusion process[3] with a data-driven conditioning mechanism, which generates coherent and meaningful animation frames. By leveraging the power of cutting-edge handwriting recognition and animation generation techniques, our proposed method may represent a significant advance in the field of online teaching and has the potential to revolutionize the way teachers and students interact in virtual classrooms.

Finally, we conducted a comprehensive longitudinal study to evaluate the effectiveness and usability of GestureTeach for online teaching tasks. The study consisted of several phases, including an initial interview, a training session to introduce GestureTeach and its corresponding gestures, and a comparison with two widely-used online teaching approaches with Tencent-Meeting and PPT. After the initial evaluation, participants were involved in a daily practice of GestureTeach

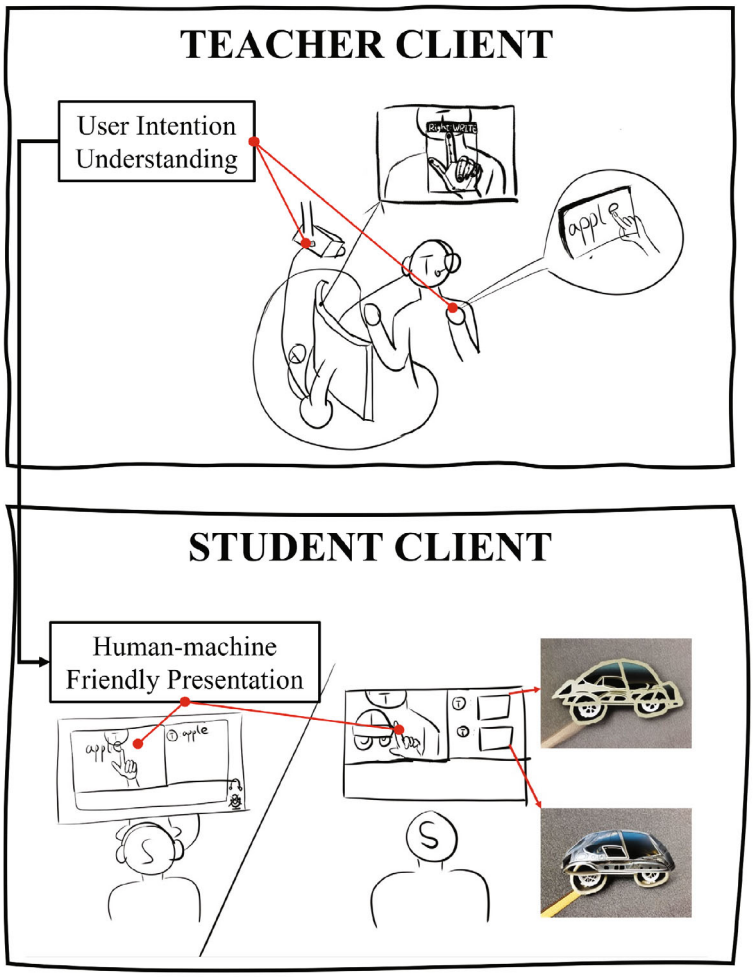

FIGURE 1 The prototype of GestureTeach.

and provided feedback to the experimenter. The results of the study showed GestureTeach was highly effective and provided an easy and convenient way for teachers to teach online, which they had never imagined before. Participants provided very positive feedback, indicating that GestureTeach could potentially revolutionize online teaching practices.

To summarize, this article makes the following contributions:

1. We propose a novel pedagogical paradigm and the corresponding technology that offers free handwriting interaction and two kinds of animation generation methods without the hardware facilities, making it a versatile solution for various online teaching scenarios.
2. We design a rapid multi-stage pipeline eliminating the requirement to locate the hand during the hand-tracking process, resulting in fewer redundant calculations and enabling highly freeform handwriting.
3. We introduce a novel gesture-recognition-based animation generation technique that leverages hand tips trajectories to enable teachers to freely handwrite and generate animation during online teaching.

2 | RELATED WORKS

In this section, some existing methods were reviewed with gestures for human–computer interaction and summarize some devices and tools that have been used to support online teaching.

2.1 | Gesture recognition

Most approaches recognize the gestures by applying machine-learning techniques to a set of relevant features extracted from the depth data. In Reference 4, silhouette and cell occupancy features are used to build a shape descriptor that is then fed to a classifier based on action graphs. However, machine-learning based methods are not robust for the hand recognition in the wild. With the rapid growth of computer vision and deep learning, the CNN model has become a popular choice for features extraction and classification tasks. For example, in Reference 5 a robust method was proposed for hand gesture recognition based on a modified convolutional neural network and prepossessed edge data. As for Reference 6, deep learning to the problem of hand gesture recognition was applied for 24 hand gestures, in which an end-to-end convolutional neural network is proposed to recognize American sign language. However, end-to-end networks are not enough for more elaborate tasks.

To solve above problems inherent to end-to-end recognition methods, there are researchers[7] first presenting an approach that uses a multi-camera system to train fine-grained detectors for keypoints that are prone to occlusion, such as the joints of a hand. This method was widely used to train a hand keypoint detector for single images and the resulting keypoint detector runs in realtime on RGB images. In Reference 8, a new method was proposed for 3D hand pose estimation from a monocular image through a novel 2.5D pose representation. However, these models generally perform better but are larger and computationally demanding simultaneously, thus, are not lightweight enough to run real-time on commodity mobile devices.

Models based on stream employing gesture estimation on the input video further analyze just the sequences of skeletal data[9-11] are one of the reasonable solutions. In Reference 9, a novel method using a nonparametric representation was proposed, which is referred to as Part Affinity Fields (PAFs), to learn to associate body parts with individuals in the image. Especially for AR/VR applications, a real-time on-device hand tracking solution was presented that predicts a hand skeleton of a human from a single RGB camera.[10]

2.2 | Animation generation

Text and image animation generation have been active research areas in recent years. In the field of text animation, various techniques have been proposed to generate dynamic text effects, such as text scrolling, typing, and transitioning.[12,13] He et al.[14] proposed a method for text style transfer using disentangled representation learning. The goal was to transfer the style of a given source text into a target text while preserving its content. Gong et al.[15] proposed a method for text animation generation using deep reinforcement learning with motion graphics templates. Huang et al.[16] introduce a method that

decomposes Chinese characters into radicals and generates new characters by combining the radicals with a target style. However, there have been limited studies on techniques for generating animations from handwritten inputs.

On the other hand, image animation generation has also attracted significant attention, with the development of deep learning-based methods that can produce realistic and visually appealing animations from still images or sketches. Luo et al.[17] proposes a method for editing anime hairstyles based on user sketches and inpainting. Ho et al.[18] introduce a series of diffusion models, which are trained on a hierarchy of images with increasing resolutions. By iteratively applying these models in a cascaded fashion, the generated images achieve high fidelity and realism. Zhang et al.[19] proposes a method for adding conditional control to image diffusion models, allowing for more fine-grained control over the image generation process. However, most existing works in the field of image animation generation have focused on generating animations from still sketches or images, overlooking the potential for interaction with users and their high degree of freedom in expressing their intentions.

3 | PROPOSED METHOD

As shown in Figure 2, GestureTeach commences by downsampling the input frame to reduce the calculation scale and processing time. The downscaled input frame is then fed through the palm detection block (PDB) to detect the teacher's palm and the landmark regression block (LRB) to recognize their gestures. Once the hand is accurately tracked, the palm region PR_{M-1} is used directly in the current frame PR_M, resulting in fewer redundant calculations from the PDB.

With the aid of the system's interactive suggestive cues, teachers can effortlessly facilitate the handwriting process, with the results saved as points. Following the completion of the handwriting process, teachers utilize gestures classified by the gesture recognition module (GRM) to indicate different animation types. Text-type sketches are subsequently inputted into the character recognition block (CRB), which outputs printed-font text. While in the intelligent generation of animations, GestureTeach employs an encoder-decoder architecture, comprising the stable diffusion encoder (SDE) and stable diffusion decoder (SDD). The image-type sketches are inputted into the shared-weight SDEs and upsampled, serving as control conditions before being concatenated with the diffusion process features. The generated animations are presented to students in a user-friendly manner via the animation generation module (AGM).

3.1 | Gesture recognition module

The proposed GRM utilizes a handwriting-smooth satisfied pipeline consisting of two main blocks working together: a PDB that processes on a full input image to locate palms and a LRB that processes on the cropped hand region box provided

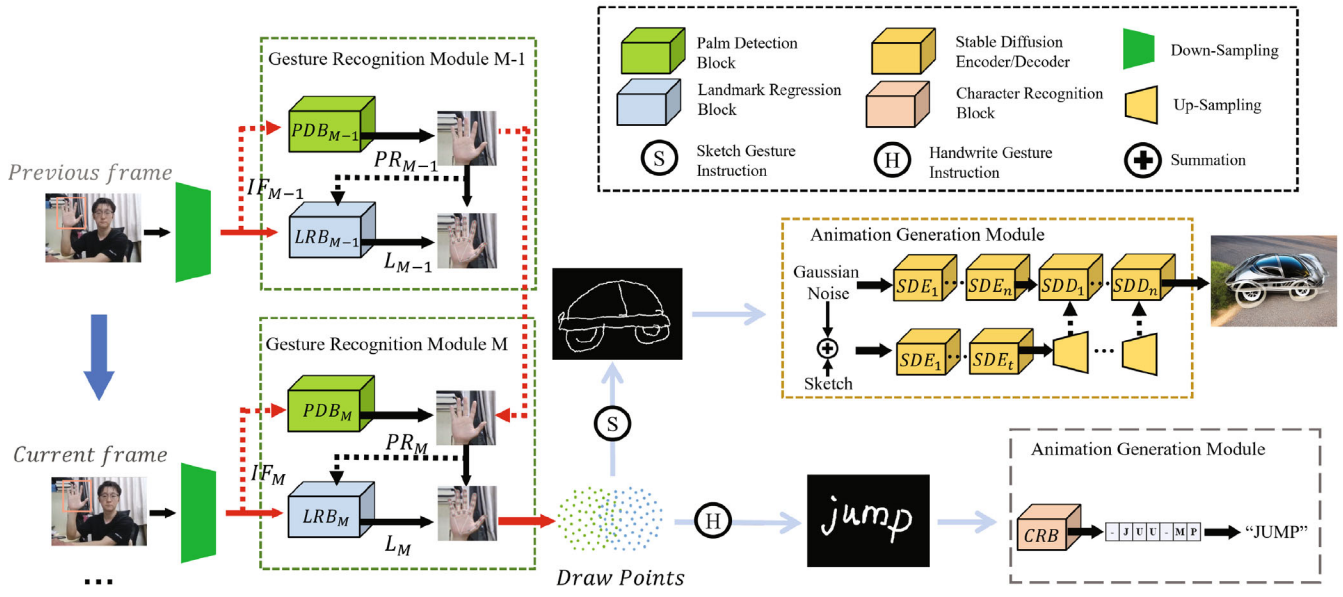

FIGURE 2 GestureTeach system overview.

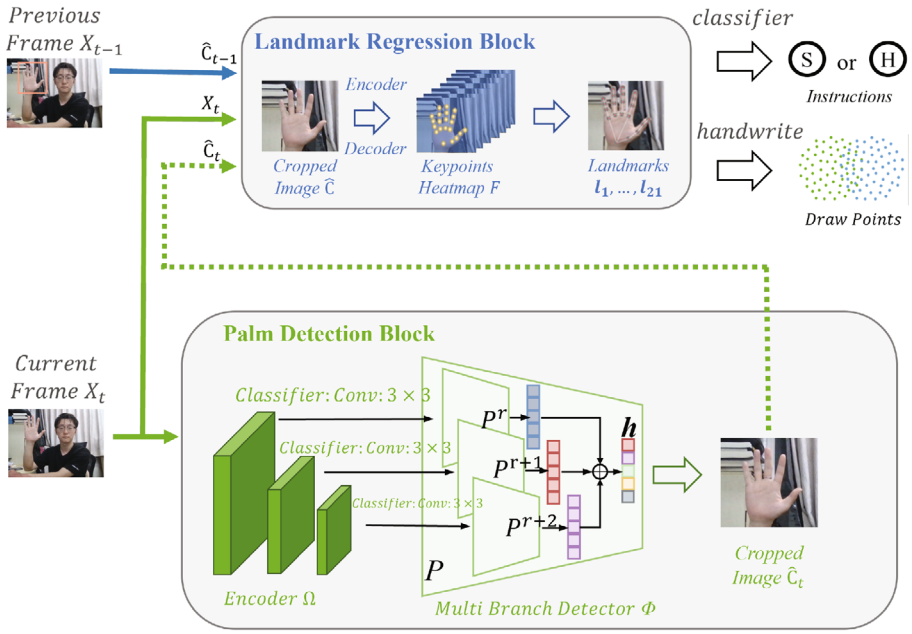

FIGURE 3 Architecture of gesture recognition module.

by the PDB. As shown in Figure 3, for the need of smooth and real-time handwriting, GestureTeach directly feed the palm region detected in the previous frame X_{t-1} into the LRB. When the existence $\hat{E}$ calculated later is false, indicating the non-existence of hands, the PDB will recalculate the palm position based on the image input of the current frame, which simultaneously means that the hand tracking fails. The idea behind this pipeline is to reduce the time consumption for hand region localization, thus saving computational resources.

Concretely, the PDB first employs an encoder Ω to process the full input image and produces a set of different feature maps p of size $(128, 32, 32)$, $(256, 16, 16)$, and $(256, 8, 8)$. The outputs p are then up-sampled and concatenated, passed through an encoder architecture. Prior to the execution of landmark regressor, output tensor f is first transformed from the center position, rotation angle, and scaling factor of the palm region into the coordinates of the four corners of the hand region. This is done to calculate the affine transformation matrix that will transform the hand region into a square output of the specified resolution.

Then, the LRB takes a cropped image as input and processed through an encoder-decoder architecture in sequence to extract multi-level features, getting 21 landmark heatmaps, which are defined as probabilistic images with the value of each pixel in the range from 0 to 1, and corresponding landmark coordinates $l_1, \ldots, l_{21}$.

Mathematically, the cropped image is analyzed by an encoder-decoder network Θ, generating a set of confidence maps $F_{j,k}$ for each landmark j and each hand k. Let $x_{j,k} \in R^2$ be the groundtruth position of hand landmark j for hand k in the image. The value at location $p \in R^2$ in $F_{j,k}$ is defined as,

$$F_{j,k}(p) = \exp\left(-\frac{||p - x_{j,k}||_2^2}{\sigma}\right), \tag{1}$$

where σ controls the spread of the peak. The groundtruth heatmap predicted by the block is an aggregation of the individual confidence maps via a max operator,

$$F_{j,k}(p) = \max_k F_{j,k}(p). \tag{2}$$

The base module of the PDB and LRB is organized in a residual way, concretely, for the lth layer of palm detection network, the input is $X_{2D}^{l-1} \in R^{D \times W \times H}$ and the process can be formalized as:

$$X_{2D}^{l} = \text{Conv}(X_{2D}^{l-1}) + X_{2D}^{l-1}. \tag{3}$$

Given 2*D* gesture key point coordinates as described above, GestureTeach will further implement several interactive functions based on the gesture classifier, comprising of an input layer, three hidden layers with 32, 32, and 16 units respectively, and an output layer with classification units. For the Handwriting gesture, GestureTeach will record the trajectory of the index finger tip, as shown in Figure 5. Please note that caution should be taken. For the text animation generation gesture, GestureTeach will perform handwritten character recognition based on the recorded trajectory, and then generate vivid animations based on the recognized characters. For image animation generation, GestureTeach incorporates hand-drawn sketches as control conditions into the generation model, achieving semantic-based animation generation based on the sketches.

3.2 | Animation generation module

In the context of online teaching, the interaction between teachers and students mainly comes from the blackboard, which highlights the importance of the AGM. According to the control gesture, AGM takes handwriting trajectories, or sketches in other words, as input, and outputs text animation or image animation as shown in Figure 2.

For text animation generation, the proposed CRB comprises of two main parts: a text box orientation classifier, and a pre-trained text recognizer. The text box orientation classifier takes the original binary image calculated from handwriting trajectories as input, and uses MobileNetV3 as the backbone network to determine the orientation of the text boxes. The pre-trained text recognizer takes the horizontally aligned rectangular text box obtained through transformation as input, and uses a deep bidirectional LSTM network to obtain predictions for each character.

For image animation generation, we present a novel architecture to sketching and 2D animation generation by capturing the trajectory of the user's index finger. The AGM uses stable diffusion techniques[20] to generate high-quality animations from the captured sketches. The stable diffusion algorithm combines a diffusion process with a data-driven conditioning mechanism, which generates coherent and meaningful animation frames.

To enhance the quality of the generated animations, we introduce a stable diffusion block based on the stable diffusion architecture, incorporating the handpainting sketch as an additional input into the stable diffusion model and putting corresponding learned features into SDD. This enables the user to control various aspects of the generated animations, such as the contour, direction, and style. Specifically, SDE comprising of a normalization layer, a self-attention layer, another normalization layer and a multilayer perceptron. Mathematically, SDE can be interpreted as an equally weighted sequence of denoising autoencoders $\epsilon_\theta(x_t, t); t = 1 \ \ldots \ T$, which are trained to predict a denoised variant of their input x_t, where x_t is a noisy version of the input x. The corresponding objective can be simplified as follows:

$$L_{DM} = E_{x,\epsilon\sim N(0,1),t}[||\epsilon - \epsilon_\theta(x_t, t)||_2^2]. \tag{4}$$

In practice, the GRM captures and stores the unprocessed trajectory points in an array during handwriting, which are then used as control inputs for AGM to generate high-quality and visually appealing animation frames. To clarify, "unprocessed points" refers to the trajectory points that have not yet been processed into a graphical representation, such as a printed font or an animation frame. With this interactive animation feature, users can effectively convey complex ideas in an intuitive and captivating manner. This allows for a more engaging and effective online teaching experience.

4 | EXPERIMENTS

4.1 | Gesture recognition experiment

The experiments were conducted on a laptop equipped with an Intel i7-7700k@4.20GHz CPU, a GeForce GTX 1080Ti GPU, and 16GB of RAM. The GRM was trained on the SCUT-Ego-Gesture dataset,[21] which contains 59,111 RGB images of 16 different egocentric hand gestures. We implemented the GRM using PyTorch and integrated it into a Python script with Python library OpenCV.[22] To simulate the online teaching scenario and not just test simple gesture recognition on individual images, we recorded a live video annotated containing 6190 images with the built-in camera of the laptop.

In the annotation video the GRM successfully recalled 6014/6190 recognitions (recall rate: 97.15%). For CPU latency, the average time each component took to process a single frame (Resolution: 640 × 480) was 27.12 MS for the combination of the PDB and the LRB, 3.76 MS for the gesture classifier, and 0.04 MS for the trajectory points conservation. For GPU

latency, the average time of tracking is 12.27 MS. Most of the false negatives came from that the hand is too parallel to the camera, making the fingertip invisible or being excessively dark that could be cut after segmentation. Most of the false negatives came from that the hand is too parallel to the camera, making the fingertip invisible or being excessively dark that could be cut after segmentation.

As shown in Table 1, we evaluated the recognition accuracy and computational efficiency of our gesture recognition compared to other algorithms.[23-25] The classification accuracy was measured by the ratio between the number of correctly predicted hand gestures and the total number of validation dataset.

To evaluate the usability of our predesigned gestures and better understand the teaching process of teachers, we conducted a formative study and identified common issues such as initial difficulties in understanding some hand gestures and the need for feedback on gesture success. We followed the design principles of implicit metaphor[26] and intuitiveness[27] and proposed three gesture instruction mapping principles: intuitive and natural, appropriate difficulty of gestures, and less cross-functional. We designed gestures that are consistent with users' background culture, understanding intention, and life experience, so that the meaning of gestures corresponds to the subjective understanding of users. When implementing gestures, using tangible or symbolic metaphors that are consistent with users' life experience is more intuitive. We also considered user fatigue and proposed ten simple but well-designed gestures as shown in Figure 4.

4.2 | Animation generation experiment

For text animation generation, we train a word recognition model using a PaddlePaddle public handwritten English character dataset, as it contains a large number of air-written English character, totally involves 3400 recordings covering 26 English characters. To train a single digit classifier, MNIST dataset was used. MNIST dataset is composed of a good number of handwritten digit images from "0" to "9" of size 28×28. It contains 60,000 samples for training and 10,000 samples for validation. We optimized the network for 20 epochs using Adam optimizer with a learning rate of 0.001 and set the batch size to 64. For training the word recognition model, we selected a PaddlePaddle public handwritten English character dataset as it includes a significant number of air-written English characters, covering 24 characters across 3400 recordings.

We noted the difference between blackboard writing and the GestureTeach method, which incorporates pen lifting and dropping movements to improve the writing experience. Therefore, we design two hand gestures, namely, the pen-down

TABLE 1 Deep learning components comparison in text animation generation module.

Methods	Tracking speed CPU (ms)	Tracking speed GPU (ms)	Recognition accuracy CPU %
25	39.43	30.57	93.32
23	44.92	35.64	87.66
24	37.77	27.64	95.45
Ours	39.43	30.57	93.32

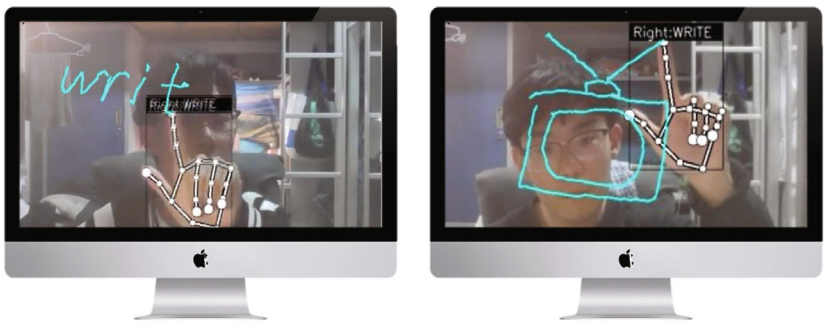

FIGURE 4 Handwriting gesture recognition.

and pen-up gestures, based on the principles of real-world handwriting. As shown in Figure 5, previous works on trajectory capture and recognition[23,25] did not consider the control of pen lifting and dropping gestures, resulting in unnecessary trajectories that could interfere with the machine and lead to inaccurate handwriting recognition, text animation and user perception of the written content contrast to ours.

We assessed the performance of the text animation generation model using character error rate (CER) and word error rate (WER), which corresponds to normalized Levenshtein distances between the predicted and ground truth character sequences. Models[28-30] were trained and evaluated on the corresponding Aachen splits of the IAM dataset. Our approach is compared to state-of-art methods with varying characteristics, as displayed in Table 2. Our model is able to compete with the approaches.

For image animation generation, we utilized stable diffusion 1.5 as the base model and obtained 500K scribble-image pairs from the internet. We trained a modified SDE and Decoder architecture with a combination of adversarial and mean squared error loss. To increase the diversity of the training data, we employed data augmentation techniques such as random rotation, scaling, and flipping, and implemented an early stopping strategy to prevent overfitting. We used a pre-trained model as the starting checkpoint and trained the model for 150 GPU-hours using NVIDIA 3090*Ti*. We assessed the overall image animation generation capability of GestureTeach compared to the method proposed by Isola et al.[31] for each sketch, as illustrated in Figure 6. The time taken by both methods is approximately 3 to 5 s, however, the images generated by the GestureTeach are more realistic.

FIGURE 5 Different handwriting of GestureTeach (down) and others (up) and corresponding text animation generation results.[23]

TABLE 2 Deep learning components comparison in text animation generation module.

Methods	CER %	WER %	Speed (s)
28	3.2	10.5	3.7
29	3.59	9.44	2.58
30	3.03	8.66	3.3
Ours	3.23	8.79	2.70

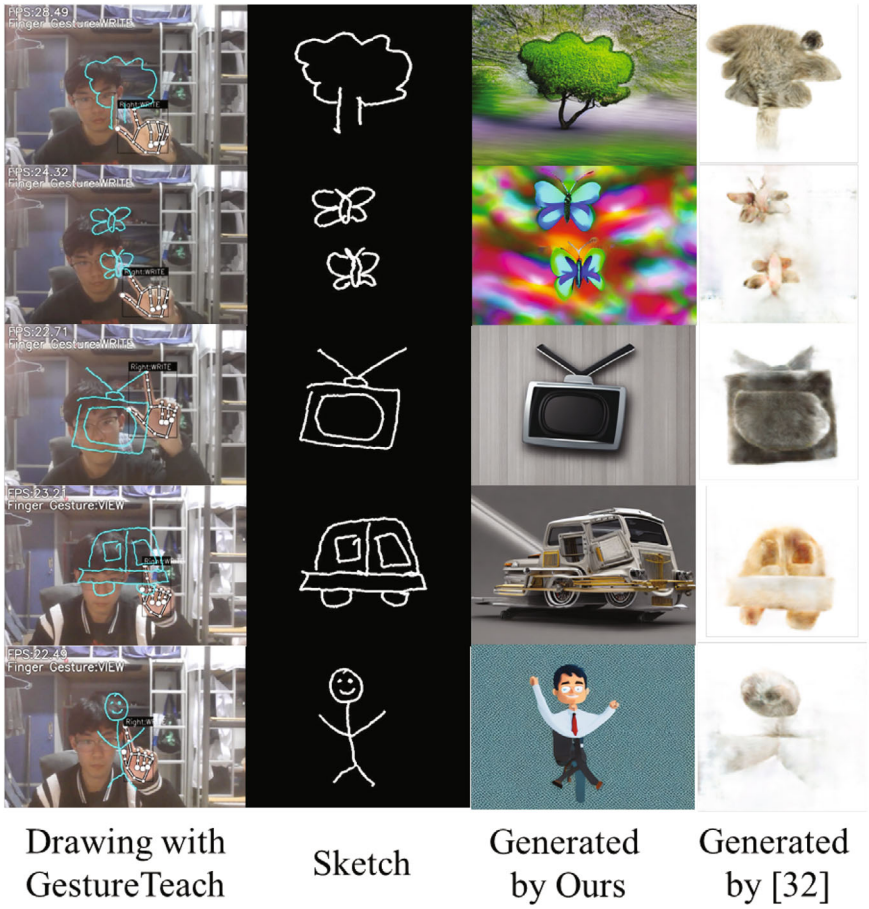

FIGURE 6 Generated image animation.

4.3 | Usability experiment

4.3.1 | Participants and apparatus

We recruited 15 teachers (8 female, 7 male) and 90 students from our university for the evaluation study. Teachers' ages ranged from 31 to 44. All of them were native speakers of Chinese and had basic knowledge of gesture-based interaction method. Six teachers were computer teachers, four were math teachers, five were English teachers. All teachers were right-handed. The study was conducted with laptops and Windows 10.

4.3.2 | Design and procedure

We first gave a brief interview, asking the participants' demographic information and their prior experience with online teaching. An experimenter then introduced the concept of GestureTeach and the tasks in this study. Then teachers were shown the functions of GestureTeach and corresponding gestures, which they also try to learn to master in the meantime. This learning process lasted about an hour until all the teachers thought they had mastered all the functions. Teachers were also asked to handwrite using other two widely-used online teaching approaches with Tencent-Meeting and PPT. After the whole first day evaluation, teachers were asked to fill out a questionnaire with three items on 7-point Likert scale about quality of the manuscript in three approaches. What's next, participants were asked to use GestureTeach to pretend to teach online without students 20 min per day for 5 days. Finally, they went through a true online class with students. We set the daily practicing time for 20 min to keep a balance between the floor effect and ceiling effect.[32] They were asked to note down their daily learning activities and send them to the experimenter.

Teachers were informed that on day 7, a final testing session would assess their learning outcomes. After the testing session on day 7, teachers were asked to give scores for their agreements with six metrics to evaluate their experience on 7-point Likert scale. What's more, students participated in the test session were also asked to fill out a questionnaire with four items on a 7-point Likert scale. All of the issues asked about subjective feelings about GestureTeach.

4.3.3 | Subject measurement

From our observations, teachers can be proficient with the system after around ten minutes of practicing and successfully complete online courses after a five-day practicing stage. The results of the teachers' questionnaire showed in Figure 7 indicated that teachers generally recognize the performance of handwriting in Gesture-Teach contrast to TencentMeeting and whiteboard pencil solution. Figure 8 showed that teachers appreciated the interaction techniques in online teaching and felt that the interaction techniques were fun to use and easy to learn. As for the students' questionnaire, all students gave very positive ratings towards the idea of learning online through GestureTeach. Although not even asked by the experimenter, two teachers expressed their strong interesting continuing to use GestureTeach after the study.

After sorting out the above questionnaire, we tested the significance of the questionnaire data to verify that the experimental results were not obtained by accidental error. We perform one-way ANOVA analysis on the survey data as shown

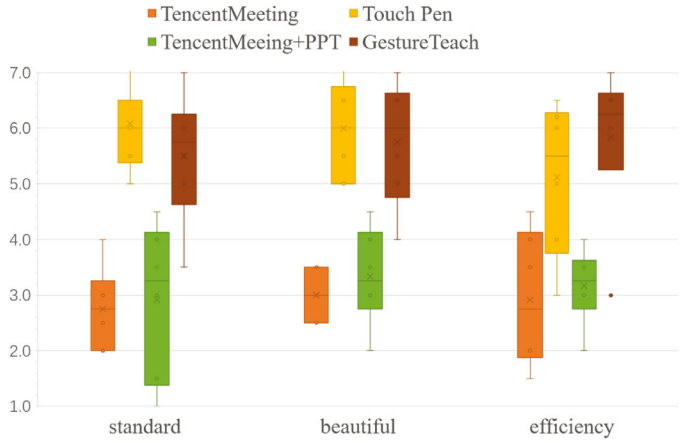

FIGURE 7 The subjective ratings on handwriting in GestureTeach.

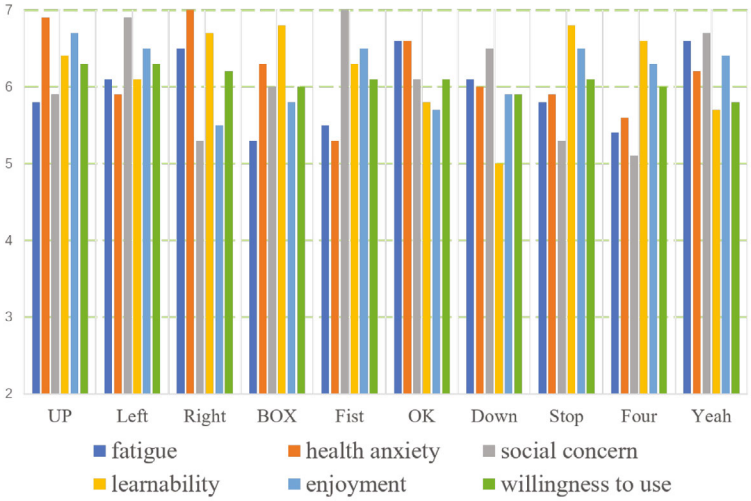

FIGURE 8 The subjective ratings of ten gesture interaction techniques in GestureTeach.

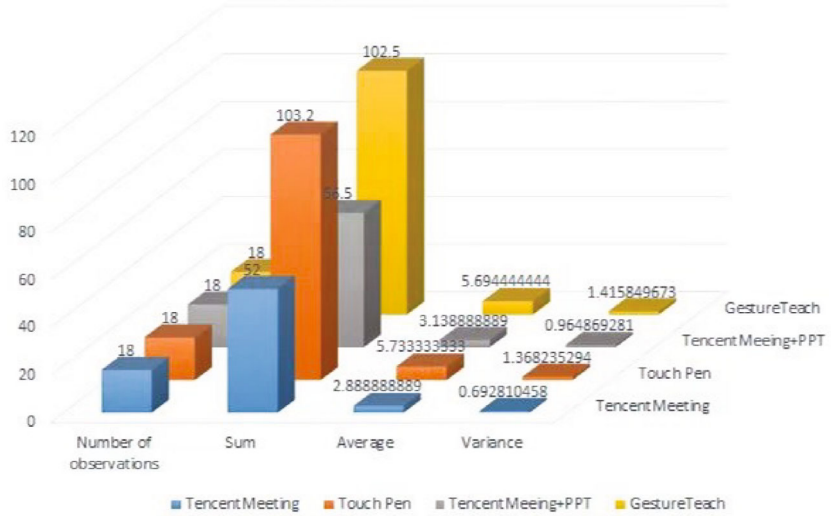

FIGURE 9 Significance test of the questionnaire data evaluated for the four teaching methods.

in Figure 9. That is, we propose the null hypothesis $H0$ that people's evaluation of writing function is independent of different teaching methods, and determine the significance level $\alpha = 0.05$. The calculated p value is $6.54 \times 10^{-15} << .05$, so the null hypothesis is overturned, and a significant correlation between people's evaluation of writing function and different teaching methods is concluded, indicating that our experimental results are not random errors, but real and valid.

4.4 | Discussion

In this work, we presented a novel pedagogical paradigm and the corresponding technology for online teaching. We demonstrated gestures can function as a general interaction channel to successfully complete online course. Table showcases different deep learning components for gesture recognition and animation generation respectively, demonstrating the superiority of our proposed methods. Additionally, figure illustrates the design of pen lifting and dropping movements for text animation generation is useful and efficient.

For the user experiments, we find that existing solutions tend to pose some objective problems and GestureTeach can go beyond them and replace them for online teaching. Due to the limited precision of mouse writing, the use of software electronic white-board can only carry little writing information in a specific area. It takes time to switch back and forth with the courseware and the teacher's picture when using the camera to display paper and pen writing, which affects the classroom fluency and disrupts the teaching rhythm. Figure showcases the significance test of the questionnaire data evaluated for the four teaching methods. The results of median score with three judging criteria indicate that GestureTeach received more positive ratings than TencentMeeting methods and almost equal score than touch pen.

In this study, we have introduced a novel pedagogical paradigm and the accompanying technology for online teaching. We have demonstrated the effectiveness of gestures as a versatile interaction channel for successfully conducting online courses. Table 1 highlights the diverse deep learning components used for gesture recognition and animation generation, showcasing the superiority of our proposed methods. Furthermore, Figure 5 showcases the design of pen lifting and dropping movements for text animation generation, which proves to be useful and efficient.

Through user experiments, we have identified certain objective issues with existing solutions, and our GestureTeach system surpasses them, offering a viable alternative for online teaching. The limited precision of mouse writing and the restricted information capacity of software electronic whiteboards in specific areas are noteworthy challenges. Additionally, the time required for switching between courseware and the teacher's visual representation when using a camera to display paper and pen writing can disrupt the classroom flow and teaching rhythm. To assess the effectiveness of our approach, we conducted a significance test of the questionnaire data, as depicted in Figure 8. The median scores based on three evaluation criteria indicate that GestureTeach received more positive ratings than TencentMeeting methods and achieved comparable scores to touch pen methods.

5 | CONCLUSION

In this article, we introduce GestureTeach, a low-cost and accessible system designed to assist teachers in creating high-quality animations through handwriting. Our proposed interaction system includes free interaction channels and intelligent display options, all tailored to the needs of online teaching. Given user sketches, GestureTeach allows teachers to generate realistic animations through user-friendly predesigned gestures, offering a high degree of freedom and creative interaction intention expression method. In our long-term usability evaluation study, the result proves GestureTeach promises an effective user intention understanding mechanism for teachers and presents the corresponding visual contents for students in a human–machine friendly style, offering a smooth and interactive teaching procedure for both teachers and students. Our work highlights the potential of using gestures as a means of facilitating online teaching, and we hope to inspire other researchers to explore the use of gestures and other interactive methods in assisting teachers in online education.

ACKNOWLEDGMENTS

This article is sponsored by National Key R&D Program of China (2022ZD0118001), National Natural Science Foundation of China under Grant 61972028, 62332017 and U22A2022, and Guangdong Basic and Applied Basic Research Foundation.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Hongjun Liu https://orcid.org/0009-0002-2661-8047

REFERENCES

1. Huang J. Successes and challenges: online teaching and learning of chemistry in higher education in China in the time of COVID-19. J Chem Educ. 2020;97(9):2810–4.
2. Jin H, Zhang M, He Q, Jun G. Over 200 million students being taught online in China during COVID-19: will online teaching become the routine model in medical education? Asian J Surg. 2021;44(4):672.
3. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach F, Blei D, editors. Proceedings of the 32nd International Conference on Machine Learning. Volume 37. Cambridge, MA: PMLR; 2015. p. 2256–65.
4. Kurakin A, Zhang Z, Liu Z. A real time system for dynamic hand gesture recognition with a depth sensor. 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO). Piscataway, NJ: IEEE; 2012. p. 1975–9.
5. Yingxin X, Jinghua L, Lichun W, Dehui K. A robust hand gesture recognition method via convolutional neural network. 2016 6th International Conference on Digital Home (ICDH). New York: IEEE; 2016. p. 64–7.
6. Oyedotun OK, Khashman A. Deep learning in vision-based static hand gesture recognition. Neural Comput Appl. 2017;28(12):3941–51.
7. Simon T, Joo H, Matthews I, Sheikh Y. Hand keypoint detection in single images using multiview bootstrapping stacked hourglass approach. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2017. p. 1145–53.
8. Iqbal U, Molchanov P, Breuel T, Gall J, Kautz J. Hand pose estimation via latent 2.5D heatmap regression. Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer; 2018. p. 118–34.
9. Cao Z, Simon T, Wei S-E, Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE; 2017. p. 7291–9.
10. Zhang F, Bazarevsky V, Vakunov A, Tkachenka A, Sung G, Chang C-L, et al. Mediapipe hands: on-device real-time hand tracking. arXiv preprint arXiv:2006.10214, 2020.
11. Liu M, Jin S, Yao C, Lin C, Zhao Y. Temporal consistency learning of inter-frames for video super-resolution. IEEE Trans Circuits Syst Video Technol. 2023;33:1507–20.
12. Krishnan P, Kovvuri R, Pang G, Vassilev B, Hassner T. Textstylebrush: transfer of text aesthetics from a single example. IEEE Trans Pattern Anal Mach Intell. 2023;45:9122–34.
13. Yao C, Xiao J, Zhao Y, Ming A. Video streaming adaptation strategy for multiview navigation over dash. IEEE Trans Broadcast. 2019;65(3):521–33.
14. He J, Wang X, Neubig G, Berg-Kirkpatrick T. A probabilistic formulation of unsupervised text style transfer. arXiv preprint arXiv:2002.03912, 2020.
15. Gong H, Bhat S, Wu L, Xiong JJ, Hwu W-m. Reinforcement learning based text style transfer without parallel training corpus. arXiv preprint arXiv:1903.10671, 2019.
16. Huang Y, He M, Jin L, Wang Y. RD-GAN: few/zero-shot Chinese character style transfer via radical decomposition and rendering. In: Vedaldi A, Bischof H, Brox T, Frahm JM, editors. Computer Vision–ECCV 2020. Cham: Springer; 2020. p. 156–72.

17. Luo S, Xie H, Miyata K. Sketch-based anime hairstyle editing with generative inpainting. 2021 Nicograph International (NicoInt). Piscataway, NJ: IEEE; 2021. p. 7–14.
18. Ho J, Saharia C, Chan W, Fleet DJ, Norouzi M, Salimans T. Cascaded diffusion models for high fidelity image generation. J Mach Learn Res. 2022;23(47):1–33.
19. Zhang L, Agrawala M. Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543, 2023.
20. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE; 2022. p. 10684–95.
21. Wu W, Li C, Cheng Z, Zhang X, Jin L. Yolse: egocentric fingertip detection from single RGB images. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Piscataway, NJ: IEEE; 2017. p. 623–30.
22. Bradski G. The OpenCV library. Dr Dobb's J. 2000;25(11):120–5.
23. Lee S-K, Kim J-H. Air-text: air-writing and recognition system. Proceedings of the 29th ACM International Conference on Multimedia. New York: Association for Computing Machinery; 2021. p. 1267–74.
24. Xia Z, Xing J, Li X. Gesture tracking and recognition algorithm for dynamic human motion using multimodal deep learning. Secur Commun Netw. 2022;2022:4387337.
25. Kim U-H, Hwang Y, Lee S-K, Kim J-H. Writing in the air: unconstrained text recognition from finger movement using spatio-temporal convolution. IEEE Trans Artif Intell. 2022;1–13.
26. Ruiz J, Yang L, Lank E. User-defined motion gestures for mobile interaction. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11. New York: Association for Computing Machinery; 2011. p. 197–206.
27. Blackler A, Popovic V, Mahar D. Intuitive interaction applied to interface design. New Design Paradigms: Proceedings of International Design Congress (IDC) 2005. Sippy Downs, QLD: University of the Sunshine Coast; 2005. p. 1–10.
28. Bluche T, Messina R. Gated convolutional recurrent neural networks for multilingual handwriting recognition. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Volume 1. Piscataway, NJ: IEEE; 2017. p. 646–51.
29. Kizilirmak F, Yanikoglu B. CNN-BiLSTM model for English handwriting recognition: comprehensive evaluation on the IAM dataset; 2022.
30. Xiao S, Peng L, Yan R, Wang S. Deep network with pixel-level rectification and robust training for handwriting recognition. SN Comput Sci. 2020;1:1–13.
31. Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE; 2017. p. 1125–34.
32. Yan C, Wang L. Experienced EFL teachers switching to online teaching: a case study from China. System. 2022;105:102717.

AUTHOR BIOGRAPHIES

Hongjun Liu received B.S. degree from University of Science and Technology Beijing (USTB), Beijing, China, in 2022. He is currently pursuing a Ph.D's degree in Computer Science at USTB. His research interests mainly focus on physiological time series analysis.

Chao Yao received the B.S. degree in computer science from Beijing Jiaotong University (BJTU), Beijing, China, in 2009. He received the Ph.D. degree from the Institute of Information Science at BJTU in 2016. From 2014 to 2015, he served as a Visiting Ph.Dad student at Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland. Since July 2016, he served as a Postdoc in Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His current research interests include image and video processing and computer vision.

Yalan Zhang received her Ph.D. degree from the University of Science and Technology Beijing, Beijing, in June 2020. Prior to completing her Ph.D., she was a visiting researcher at the Institute for Data Learning and Applications at Rutgers University, USA, from 2017 to 2018. Her research focuses on intelligent simulation, 3D visualization, and computer vision.

Xiaojuan Ban received the Ph.D. degree from the University of Science and Technology Beijing, Beijing, in 2003. She is currently a Ph.D. Supervisor with the University of Science and Technology Beijing. She is the Managing Director of the Chinese Association for Artificial Intelligence (CAAI). She has received the New Century Excellent Talent of the Ministry of Education. Her current research interests are artificial intelligence, natural human–computer interactions, and 3D visu alization. Co-corresponding author of this paper.

How to cite this article: Liu H, Yao C, Zhang Y, Ban X. GestureTeach: A gesture guided online teaching interactive model. Comput Anim Virtual Worlds. 2023;e2218. https://doi.org/10.1002/cav.2218