

AI-powered Personalized Online Drawing Assistant with Hand Skeleton Analysis

Chiao-Wen Kao,¹ Wen-Hsuan Wu,¹ and Chi-Sheng Huang^{2*}

¹Department of Applied Artificial Intelligence, Ming Chuan University,
5 De Ming Rd., Gui Shan District, Taoyuan County 333024, Taiwan

²Department of Computer Science and Information Engineering, National Taichung University of Science and
Technology, No. 129, Section 3, Sanmin Road, North District, Taichung City 404336, Taiwan

(Received August 5, 2024; accepted February 12, 2025)

Keywords: assistive technology, MediaPipe, deep learning models, real-time feedback system

In this paper, we present a computer-vision-based system designed to enhance online drawing instruction by offering prompt feedback and evaluating learners' hand movements and pen handling. The system utilizes only webcam data, avoiding the need for expensive external devices. It consists of three main components: hand skeleton analysis, drawing tool isolation, and movement similarity assessment. MediaPipe is employed for hand skeleton analysis, interpreting the pen-holding postures of both the instructor and the learner. A dual-phase strategy is employed to isolate drawing tools. You Only Look Once (YOLO) v5 performs precise object detection to locate the drawing tool within the image, and U-Net is then used to separate the tool from the surrounding area in the identified region. Movement similarity is assessed using dynamic time warping and Procrustes-aligned mean per-joint position error, comparing the hand movements of the instructor and learner. This drawing assistant system is designed to effectively replicate the instructor's role in online environments, focusing on improving learning outcomes and experiences through interactivity and personalization. Experimental evaluation confirms the system's effectiveness and usability. Results demonstrate that the system provides accurate and timely feedback and assessment for online drawing instruction, thereby enhancing learners' drawing skills and learning experiences.

1. Introduction

The internet age has transformed many aspects of society, such as converting paper documents into electronic files and moving traditional business transactions onto e-commerce platforms. Online learning platforms have also become prevalent in the education sector, enabling remote distance education for university courses. Students can access learning materials from any location with internet connectivity, without being constrained by time and space. This change not only expands educational opportunities but also promotes the dissemination and sharing of knowledge. Particularly during the COVID-19 pandemic, online

*Corresponding author: e-mail: vcshuang@nutc.edu.tw
<https://doi.org/10.18494/SAM5281>

asynchronous courses have replaced traditional offline teaching, offering convenience and flexibility to learners.

Online asynchronous courses provide diverse learning resources that cater to various individual needs, creating an inclusive and accessible learning environment. This mode of learning supports people who face challenges in traditional classrooms, such as those with disabilities, impairments, social anxiety, and other difficulties, thereby promoting fairness and equality in education. However, online courses also pose significant challenges in the field of art education, which typically depends on in-person demonstrations and face-to-face guidance from instructors.⁽¹⁾ For skills like brushwork, teachers need to show and explain techniques in real time, while observing and adjusting to each student's progress and needs. This hands-on instruction is essential for effective learning, especially for beginners who require personalized instruction and real-time feedback.⁽²⁾ However, it is difficult to capture and assess students' creative processes and progress accurately in an online environment, where the visual and audio quality may be limited or distorted.

The field of arts has experienced a growing interest in the use of AI for various purposes, such as art analysis and generative AI. Art analysis applies AI to extract and comprehend different features of artworks, such as the identity of the artist, the mode of expression, and the historical context.^(3–7) Generative AI has been a focus of attention, especially for creating AI Art, which employs AI models to emulate the styles of renowned artists or produce original artistic works.^(8,9) A notable example of this innovation is OpenAI's DALL-E,⁽¹⁰⁾ a sophisticated neural network that can create visual representations from textual inputs across a variety of themes.

There have been some attempts to apply AI methods in the field of art education. For example, Singh *et al.* proposed an AI system that aims to make the painting process more understandable and engaging for human users by providing feedback and guidance.⁽¹¹⁾ Lüthi *et al.* integrated optical flow sensors with stylus devices to enhance the interaction and precision of digital drawing.⁽¹²⁾ He *et al.* simulated the traditional art of calligraphy by using projectors and a single camera to capture and display the strokes on a virtual canvas.⁽¹³⁾ Huang *et al.* designed an innovative portrait drawing interface that helps users create detailed and recognizable face sketches through a two-stage approach that combines sketching and shading.⁽¹⁴⁾ Križnar *et al.* presented a specific application that enables real-time collaboration between teachers and students in remote drawing by using hand tracking technology.⁽¹⁵⁾ The application utilizes Google's MediaPipe library for hand pose estimation and Node.js for server communication.⁽¹⁵⁾

Traditional drawing instruction, which involves free-hand drawing with physical pens and brushes, is a vital component of art education.⁽¹⁶⁾ It fosters the development of fundamental drawing skills, such as brushwork and pen techniques, which are essential for artistic expression and communication. Traditional drawing also reveals the learners' learning process and creative development more clearly than digital methods, as it captures the nuances and details of the hand movements and strokes. However, most AI applications in the arts have mainly focused on digital drawing, aiming to teach AI to generate digital artwork, while neglecting the importance of integrating traditional drawing instruction.

Hence, in this study, we address the current challenges in online drawing instruction by incorporating computer vision technologies that can operate with webcam data only, without the need for any additional devices. The proposed approach relies on four key technologies: hand skeleton analysis, object detection, image segmentation, and movement similarity assessment. Hand skeleton analysis uses skeletal models to interpret and track the users' hand movements, allowing them to interact with the drawing interface accurately using only a webcam. Object detection identifies the type and category of the pen that the users are holding, which helps to reduce the computational load for image segmentation. Image segmentation can accurately separate the drawing tool from the object detection bounding box. By capturing this feature, we calculate the exact position of the drawing tool and combine it with the joint points obtained from hand skeleton analysis to further determine motion similarity. Lastly, the proposed assessment method of motion similarity is used to compare the degree of alignment between the instructors' and learners' hand movements, providing objective and comprehensive evaluation metrics for the drawing performance. The proposed assessment method integrates the dynamic time warping (DTW) and Procrustes-aligned mean per-joint position error (PA-MPJPE). DTW is a widely used method for comparing similarity in time-series data, which can effectively handle discrepancies in sequence length or speed. By computing the distance between sequences of hand movements, DTW quantifies their similarity and provides an objective comparison metric. PA-MPJPE, on the other hand, is a method for evaluating the precision of hand movements by aligning joint positions of two actions and calculating the average bias, which accurately reflects their differences.

The main contributions of this paper are as follows:

- A computer-vision-based system that uses webcam data to provide real-time feedback and assessment for online drawing instruction is proposed. This approach reduces the cost and equipment requirements, making online drawing instruction more accessible and feasible.
- Four core techniques are integrated: hand skeleton analysis, object detection, image segmentation, and motion similarity assessment. These techniques address the issues of hand posture, the location and orientation of the drawing tool, and the degree of motion similarity, respectively.
- The motion similarity assessment method integrates DTW and PA-MPJPE to evaluate hand movement similarity. This method provides a comprehensive analysis, from time-series data to detailed joint position comparisons, offering robust tools for pattern recognition in hand movement research.

The structure of this paper is outlined in the following manner. In Sect. 2, details of the methods proposed in this research are described. In Sect. 3, numerical analysis and a comparison of performance metrics are presented. In Sect. 4, the discussion and conclusions are provided.

2. Materials and Methods

In this section, we describe the proposed approach for offering real-time feedback and evaluations during online drawing tutorials using only webcam data. The proposed system consists of four fundamental methodologies: hand skeleton analysis, object detection, image

segmentation, and motion similarity assessment. These techniques are designed to tackle the challenges of interpreting and evaluating the similarity of hand movements and brush strokes between instructors and learners. The following subsections provide a detailed explanation of each technique and present the experimental results.

2.1 Data collection

In this research, we focus on the interactions between hands and drawing tools, which are essential for online drawing instruction. Unlike existing hand models that mainly track hand movements, this approach also needs to detect the drawing tools that the hands are holding. Although existing large datasets, such as EgoHands,⁽¹⁷⁾ contain extensive hand-related data, as shown in Fig. 1, these datasets do not meet the specific needs for “pen-holding” data required for this research. Additionally, publicly available datasets such as MS-COCO do not include categories for items such as pens, as illustrated in Fig. 2.⁽¹⁸⁾ Owing to the lack of appropriate datasets for this task, relevant photographs should be collected.

To compile the needed imagery, a single camera setup was used to photograph various hand positions involved in drawing tool usage from multiple viewpoints. The webcam is approximately 50–80 cm away from the user’s dominant hand, as shown in Fig. 3. Hands and drawing tools need to be captured as clearly as possible to avoid obstruction. The initial collection featured over 200 snapshots of right-handed individuals holding drawing tools. Data augmentation



Fig. 1. (Color online) Example of Egohands.⁽¹⁷⁾



Fig. 2. Categories in MS-COCO dataset.⁽¹⁸⁾



Fig. 3. (Color online) Example of collected datasets.

techniques were then employed in the preprocessing phase to enrich the dataset. Leveraging the hand's symmetry, image flipping was performed to construct a set of corresponding left-hand posture images. These procedures amassed an enhanced collection of images illustrating drawing tool grips and postures.

2.2 Procedure

In this section, we outline the combined approach adopted in this research to improve online drawing tutorials by employing sophisticated computer vision technologies. The method integrates MediaPipe Hands for the accurate tracking of the hand's detailed structure, You Only Look Once (YOLO) v5 for the rapid identification of drawing tools, and U-Net segmentation for the precise delineation of key features in the drawing. This combination allows for the comprehensive and efficient tracking and recognition of hand movements, tools, and drawing elements. Additionally, the DTW and PA-MPJPE methods are used to assess the similarity in hand poses and hand movement amplitudes between the student and the instructor. Figure 4 shows the flowchart of the proposed system, with each part explained in detail in the subsequent sections.

2.2.1 Hand skeleton analysis

Next, we introduce the first component of the proposed system, which is hand skeleton analysis using MediaPipe technology. MediaPipe Hands is a state-of-the-art hand tracking and landmark localization framework developed by Google,⁽¹⁹⁾ which can process video input and recognize key landmarks on the hand, such as joints and fingertips. It pinpoints 21 significant points on the hand, as shown in Fig. 5, in terms of x - and y -coordinates within the image frame and provides a z -coordinate for the depth of each point with respect to the wrist landmark.



Fig. 4. (Color online) Flow chart of the personalized online drawing assistant.

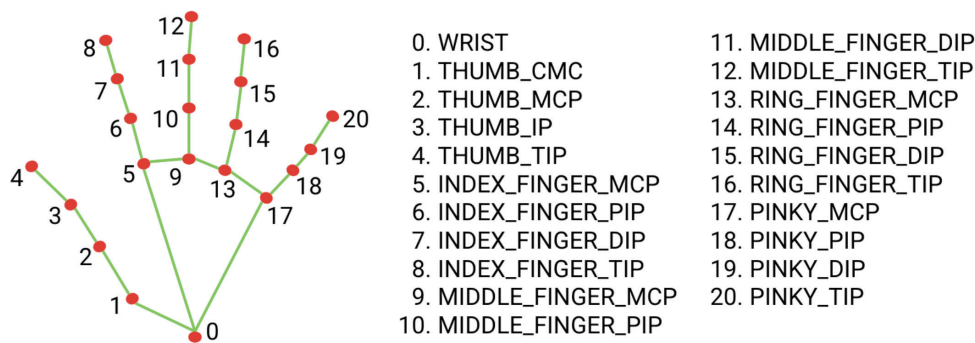


Fig. 5. (Color online) Hand landmarks topology defined by MediaPipe.⁽¹⁹⁾

This hand landmark model has been successfully utilized for a number of applications, including the recognition of sign language,^(20,21) the reconstruction of 3D hand models,⁽²²⁾ and gesture-based navigation in rehabilitation systems with the aid of 2.5D hand poses.⁽²³⁾ These studies demonstrate that it can effectively enable the system to capture and interpret intricate hand movements and gestures with high precision and accuracy. The system leverages this capability to provide immediate and personalized feedback on the learners' pen-holding postures and hand positioning, which are essential skills for drawing. Upon determining the positions of the finger joints, we will utilize the selected hand landmarks and drawing tool positions to compute the intervening angles. Consequently, the procedure to extract the drawing tool will be described in detail.

2.2.2 Drawing tool detection

As mentioned above, in this study, we aim to explore the similarity of learners' imitation of instructors' drawing processes, focusing on pen-holding posture, drawing tool location, and the angle between them and the drawing paper. To measure these factors accurately, it is essential to not only locate the hand structure but also detect the drawing tool and determine its position range and orientation. This constitutes the second important task of the proposed system, after the first task of extracting the key hand landmarks from the image. However, performing semantic segmentation on the entire image to find the drawing tool is not feasible owing to several drawbacks. Semantic segmentation relies on pixel-level classification, which consumes numerous computing resources. Additionally, the drawing tool can also be easily occluded by the thumb and produce discontinuous segmentation, affecting the accuracy of the subsequent feature extraction and action evaluation.

To perform semantic segmentation on the drawing tool region more accurately and efficiently, we adopted a dual-phase strategy inspired by Sánchez-Brizuela *et al.*,⁽²⁴⁾ namely, drawing tool segmentation leveraging object detection. Object detection is a process that not only identifies which objects are present in an image, but also determines their location and size. Therefore, the results of object detection are usually represented in the form of bounding boxes, like using a rectangular box to mark the range of the object. However, bounding boxes cannot fully reflect the shape and details of the object, especially when the object has curved or irregular contours. Therefore, to further understand the fine features of the objects in the image, we need to perform image segmentation, which assigns each pixel in the image to a certain object or background class, thereby obtaining the precise outline and more detailed semantic information of the object.

Image segmentation is a more advanced task than object detection, as it requires a deeper understanding and analysis of the objects in the image. Therefore, the quality of image segmentation largely depends on the accuracy of object detection. If object detection can accurately identify and locate various objects in the image, then image segmentation can build on the object detection results and use more fine-grained neural network models to better segment and label the edges and internal details of the object. Therefore, under the premise of image segmentation, object detection is a very important and critical step, and only under the condition of relatively accurate object detection can better image segmentation results be produced.

Hence, initially, YOLO v5 is employed as the object detection component to accurately locate the drawing tool in the image.⁽²⁵⁾ At this stage, we retrain the pretrained model of YOLO v5 with our own collected dataset, and the relevant model parameters and outcomes will be discussed in the following section. Once the drawing tool is detected, the system isolates the region of interest (ROI) where the drawing tools are located using a bounding box for further processing. The process of bounding the drawing tool area is illustrated in Fig. 6.

Next, U-Net,⁽²⁶⁾ an advanced convolutional neural network model, is utilized for pixel-level semantic segmentation on the selected ROI; the process of isolating the drawing tool is illustrated in Fig. 7. With its encoder-decoder structure featuring skip connections, U-Net stands out for requiring minimal training images to generate precise segmentation masks. It excels in

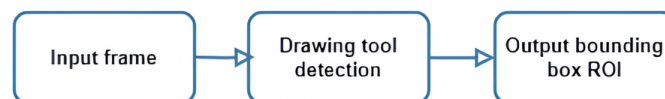


Fig. 6. (Color online) Process of bounding the drawing tool area.



Fig. 7. (Color online) Process of isolating the drawing tool.

delineating accurate masks that distinguish the drawing instrument and its associated attributes from the backdrop, elevating feature extraction precision and diminishing the impact of nonrelevant areas in further processing stages. Implementing U-Net on a small-scale ROI also reduces computational cost and lessens the complications of occlusions created by the thumb.

Furthermore, the proposed system calculates the vectors directly from the segmented pen region and provides a precise measure of the pen's tilt and rotation relative to a predefined reference frame, such as the hand's wrist and thumb orientation. By analyzing the angles between the pen's directional vector and the reference vectors, the system offers a detailed evaluation of learners' pen manipulation skills, providing insights into how effectively they control the pen's position and orientation during various drawing tasks.

After extracting the key hand landmarks, the drawing tool landmarks, and the angles between the orientation vectors, the system proceeds to the final task: evaluating the similarity of learner imitation. This evaluation involves calculating the motion similarity between the learner and the instructor, based on factors such as hand pose and the angle between the drawing tool and the paper. The resulting similarity scores serve as indicators of learning effectiveness. The detailed methodology and results of this evaluation will be presented in the subsequent section.

2.2.3 Evaluation of the learner's imitation similarity

In this section, we introduce a sophisticated methodology for computing quantitative measures based on fundamental pen features such as position, angle, and contour, extracted through techniques detailed in the preceding section. These features serve as the foundation for a comprehensive assessment of the learner's skill. To evaluate the similarity to the instructor's movements, we employ a multifaceted comparison between the learner and the instructor, focusing on three critical aspects: hand posture, pen direction, and drawing speed. This comparative analysis utilizes advanced similarity measures, including DTW and PA-MPJPE, ensuring a robust and nuanced evaluation. Central to our methodology is the introduction of a comprehensive similarity score, designed as a holistic indicator of learning effectiveness. This score is underpinned by three carefully defined indicators: pen-holding posture similarity (PHP_s), wrist rotation angle similarity (WRA_s), and relative angle between the pen and the fingers (RA_{pf}).

PHP_s is calculated on the basis of the angles of the finger joints, as these angles reflect the posture and motion of the fingers; the visualization result is presented in Fig. 8. The posture and motion of the fingers are essential elements that affect drawing skills, so they need to be checked and evaluated for their accuracy and smoothness. Several important finger joints were selected, such as the first and second joints of the index finger, the base and the first and second joints of the thumb, and the wrist landmark. The cosine theorem was used to calculate the angles between these joints, as shown in Eq. (1). Assuming there are three points a , b , and c , where b is the center point and a and c are two adjacent points, the finger joint angle can be calculated as

$$\cos(\theta) = \frac{\overline{AB} \cdot \overline{BC}}{\|\overline{AB}\| \|\overline{BC}\|}, \quad (1)$$

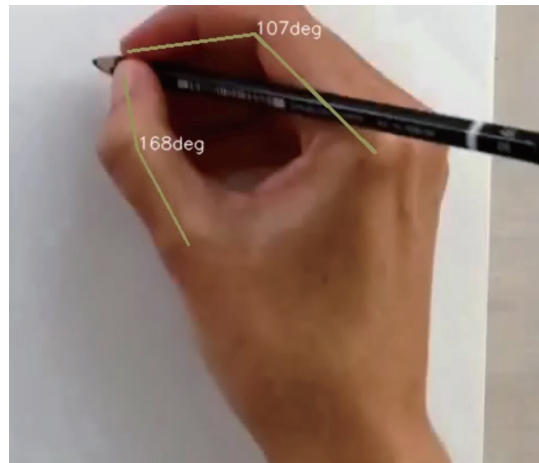


Fig. 8. (Color online) Visualization of angles of finger joints.

where \overline{AB} is the vector from joint A to joint B and \overline{BC} is the vector from joint B to joint C .

After obtaining the finger joint angles in each frame, PA-MPJPE was used to compare the similarity degrees of the overall pen-holding postures of the learners and instructors, which is a method of calculating the Euclidean distance between two poses. The formula of PHP_s is

$$PHP_s = \frac{1}{N} \sum_{i=1}^N \|t_i - s_i\|, \quad (2)$$

where t_i and s_i represent the angles of corresponding joints from the teacher and student, respectively, and N denotes the total number of joints. This function computes the average Euclidean distance between corresponding angles of instructor and student hands, reflecting the degree of posture similarity crucial for evaluating instructional feedback efficacy.

Aside from the finger joint angles, WRA_s and RA_{pf} were also considered. WRA_s is calculated on the basis of the wrist rotation angle, which is the angle between the vector from the wrist to the thumb base and the reference vector; the pen and finger similarity is calculated on the basis of the vector of the pen, as these are the key factors that affect the drawing quality. The visualization of the DTW results of RA_{pf} is shown in Fig. 9. Since drawing is a continuous action, DTW can be used to compare changes in the learner and instructor during the drawing process to evaluate the similarity degree of their motion continuity and consistency.

DTW is a method to measure the similarity of time series, which can adapt to the variations of motion speed and duration. The calculation formula of DTW distance is

$$DTW(i, j) = d(i, j) + \min[DTW(i-1, j), DTW(i, j-1), DTW(i-1, j-1)], \quad (3)$$

where $d(i, j)$ represents the Euclidean distance between points i and j in the sequences.

By utilizing these three indicators, we can generate an array of numerical data. During this process, the instructor's demonstration video is initially recognized and recorded. Subsequently,

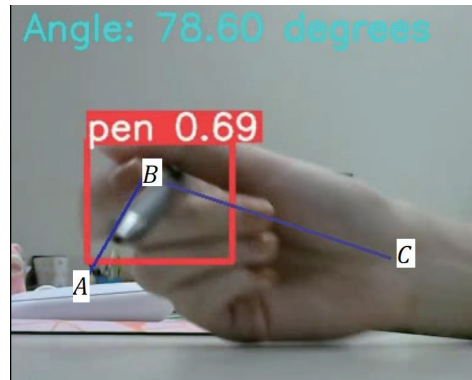


Fig. 9. (Color online) Visualization of DTW results of RA_{pf} .

students' videos are compared against the stored array, and the system annotates the playback with painting instruction cues, assisting students in retrospectively reviewing and improving their actions. In the following section, we will demonstrate the potential for using this system in painting instruction through data visualization techniques such as DTW path diagrams and scatter plots. Thus, we can quantify the performance of the system in replicating an online guided painting scenario.

3. Results and Discussion

The efficacy of the developed system was evaluated through comprehensive experiments focusing on both hand and pen movements during online drawing instruction sessions. In this section, we present the results obtained from applying the proposed evaluation metrics.

3.1 Experimental environment

3.1.1 Experimental platform

The experiments were conducted on a personal computer equipped with an Intel Xeon E5-1630 v4 CPU, 16 GB of RAM, and an NVIDIA GeForce RTX 1080 GPU, running Windows 10. The data collection setup environment illustration is presented in Fig. 10. The development environment was Python 3.8. A Logitech C270 HD Webcam, capturing images at a resolution of 1280×720 pixels and a frame rate of 30 fps, was used to record the students' process of imitating drawing. YOLO v5, trained using PyTorch 2.2.2 and CUDA 11.8, with weights from the official YOLOv5 GitHub repository, was employed for object detection. MediaPipe Hands, a cross-platform framework, was utilized for hand pose estimation. The DTW algorithm was implemented using the fastdtw library to approximate the optimal DTW distance efficiently. Evaluation metrics were calculated using NumPy libraries, and results were visualized with Matplotlib.

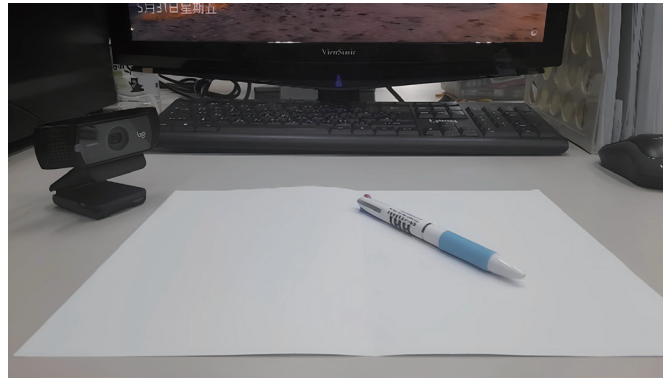


Fig. 10. (Color online) Experimental setup used in this paper.

3.1.2 Custom dataset building

Building a custom dataset was necessary for this project, as there was no existing dataset that matched the requirements. Therefore, a diverse and representative dataset of 200 images of different types of pens and hand poses was collected from various online sources, such as drawing tutorials and videos. Data augmentation techniques, such as rotation, flipping, and scaling, were applied to increase the size and variety of the dataset. Table 1 shows the detailed outcomes of the data augmentation.

LabelMe was used to manually annotate each image with a bounding box and a label of the drawing tool object.⁽²⁷⁾ The bounding box coordinates and the label were stored in an XML file for each image. The collected dataset was divided into 80% training, 10% validation, and 10% test sets, and saved in txt format for YOLO v5 training. The COCO format is a JSON file that contains the image file name, the image size, and the bounding box information for each image in the dataset.



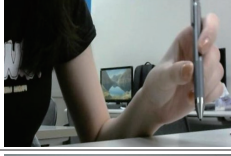

3.2 Drawing tool detection performance

In this section, we will demonstrate the annotation precision of the drawing tool through a two-step process. Initially, we perform object detection across the entire frame and mark the ROI. Then, we conduct semantic segmentation within the ROI to pinpoint the drawing tool's exact location more precisely.

3.2.1 Performance of drawing tool detection

To detect the drawing tool in each frame, we retrained the YOLO v5 model on a custom dataset of 500 images that contained different types of pens and hand poses, with annotations of the bounding boxes and labels of the drawing tool objects. We evaluated the model on a test set of 50 images and computed the following metrics: mean average precision (mAP) at IoU thresholds of 0.9 and 0.50, recall, and precision. These metrics are commonly used to measure the quality of object detection models. mAP is the average of the maximum precision values at

Table 1
(Color online) Data augmentation.

Techniques	Amount before augmentation	Amount after augmentation	Example
Original	50	50	
Vertical flipping	20	40	
Horizontal flipping	100	200	
Random rotation 15 degrees	70	210	

different recall levels, and it reflects how well the model can detect all the relevant objects in an image. IoU is the intersection over union of the predicted and ground truth bounding boxes, and it indicates how accurate the model is in locating the objects. Recall is the ratio of true positives to the sum of true positives and false negatives, and it shows how many of the actual objects the model can find. Precision is the ratio of true positives to the sum of true positives and false positives, and it measures how many of the detected objects are relevant.

The performance indicators are presented in Table 2. These metrics demonstrate high precision and recall rates, highlighting the robust performance of the system in detecting the drawing tool. After achieving reliable performance in object detection, the next step is to evaluate the efficacy of image segmentation, which refines the boundaries and contours of the detected ROI of the drawing tool.

The results are visualized as shown in Fig. 11. The images with and without the hand skeleton points are shown, and the score in the upper left corner represents the confidence value. This value is based on the model's certainty of the presence of the "drawing tool" in the image. The clearer and more intact the object is in the image, the higher the confidence value will be. As shown in Fig. 11(b), the pen is clearly obscured by the hand, which results in a lower confidence value, but this is not caused by the hand skeleton points.

3.2.2 Performance of drawing tool segmentation

To further refine the drawing tool detection results, U-Net, a deep neural network model image segmentation methodology, is adopted to isolate the drawing tool from the ROI and obtain

Table 2
YOLO v5 object detection results.

	Precision	Recall	mAP_50	mAP_90
Value	0.996	0.995	0.995	0.798



Fig. 11. (Color online) Detection results with and without hand skeleton points.

detailed masks of its shape and orientation. Table 3 shows the performance of image segmentation using U-Net.

Three metrics are used to evaluate the quality of the segmentation masks: mean pixel accuracy (mPA), mean intersection over union (mIoU), and value accuracy. mPA measures the ratio of correctly classified pixels to the total number of pixels in the image. mIoU measures the degree of overlap between the predicted mask and the ground truth mask. Value accuracy measures the ratio of correctly classified pixels to the number of pixels in the pen region. These metrics indicate how well the network can segment the pen from the background and capture its shape and orientation. As shown by the results, U-Net achieves high scores on all three metrics, demonstrating its effectiveness for image segmentation.

The results are visualized as shown in Fig. 12, which displays the static state of the drawing tool when it is placed on the desk and the state of holding the drawing tool in hand. The static state is used in the system to determine whether the learner has started or finished the practice.

3.3 Results of evaluating the similarity of learner's imitation

In this section, we present the evaluation metrics used to measure the similarity of the learner's imitation and the results of applying these metrics to a dataset of instructional sessions. These evaluation metrics are based on the PA-MPJPE and the DTW distance between the hand and pen movements of the instructor and learner. We also provide visualizations of the DTW paths to illustrate the temporal alignment and synchronization of the movements.

DTW is used to compare hand posture changes in hand action sequences, such as angle changes between the hand and the horizontal vector. Figure 13 shows the changes in DTW

Table 3
mIoU of U-Net segmentation.

	mIoU	mPA	Accuracy
Value	92.8%	96.43%	99.62%

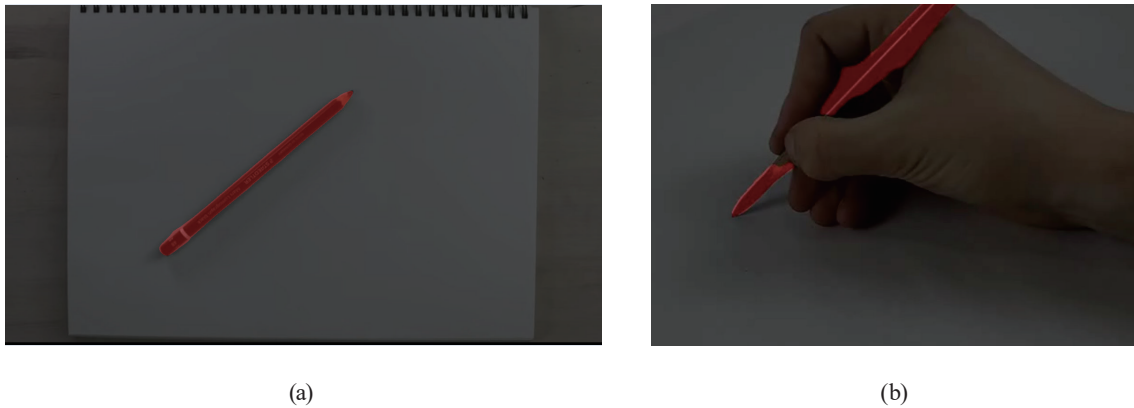


Fig. 12. (Color online) Two states of drawing tool. (a) Placed on the desk. (b) Holding the drawing tool.

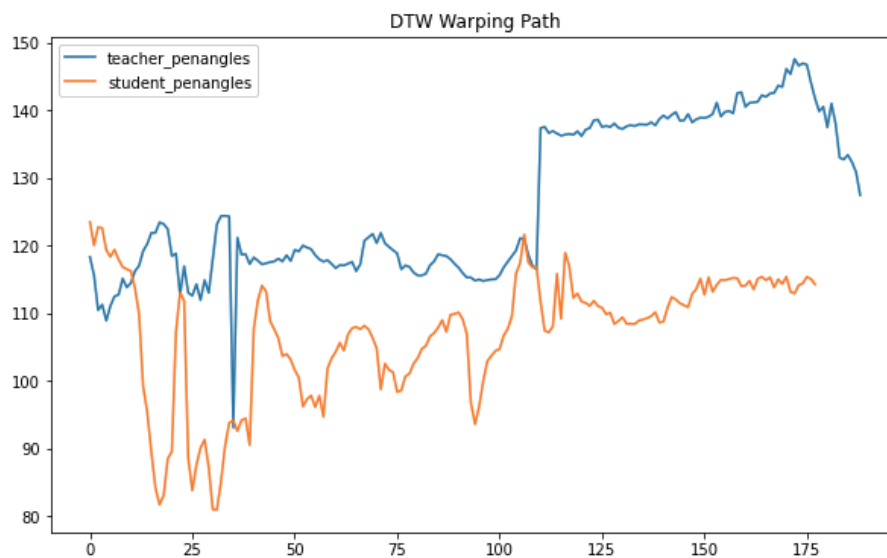


Fig. 13. (Color online) DTW shows the difference between the instructor and the learner.

values of the student, reflecting the consistency and changes in their hand postures when imitating the teacher's demonstration.

Furthermore, the DTW warping path provides valuable insights into the temporal alignment and synchronization of movements between the instructor and the learner. RA_{pf} and WRA_s are shown in Figs. 14 and 15, respectively. The DTW warping path visually depicts how the learner's hand movements deviate from the ideal instructor-guided path over time. A clearer understanding of the DTW warping path reveals instances where the student's movements may diverge or lag the expected pattern, showing potential errors or inconsistencies in their execution of drawing

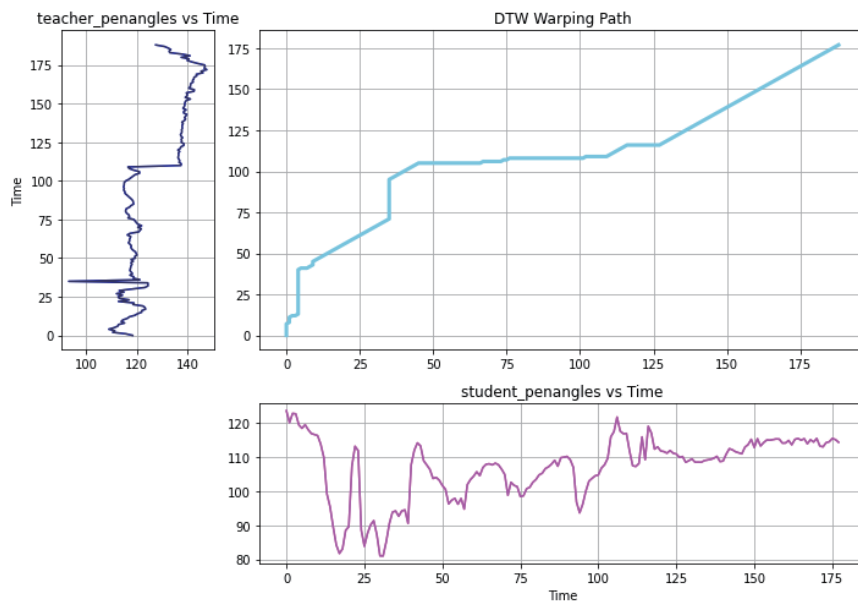


Fig. 14. (Color online) The DTW path reveals insights into the changes in angle between the pen and the hand during drawing (RA_{pf}).

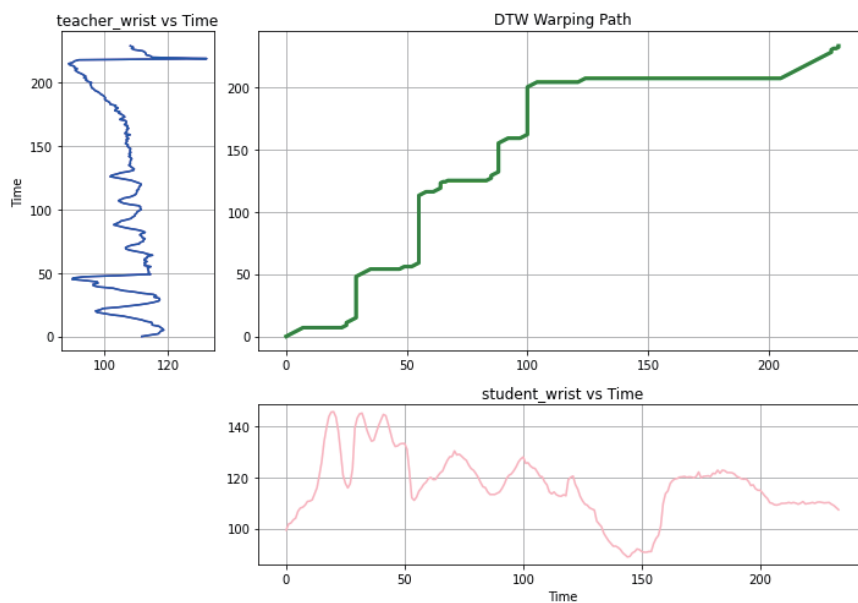


Fig. 15. (Color online) Wrist rotation angle similarity (WRA_s).

techniques. This visual representation aids in pinpointing specific segments or time intervals during which corrective feedback or additional guidance may be necessary to improve the student's performance.

PA-MPJPE was used to evaluate the difference between the learner's hand posture and the teacher's ideal posture. Figure 16 shows the average angular displacement of different learners at

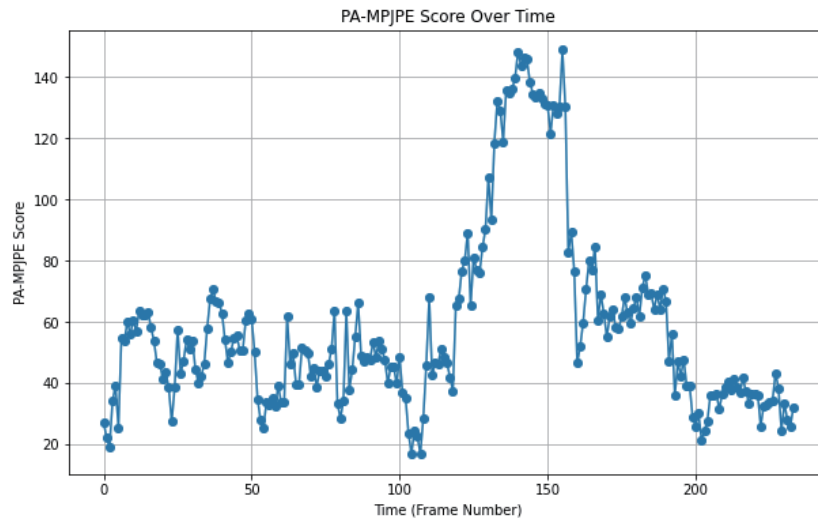


Fig. 16. (Color online) Average angular displacement.



Fig. 17. (Color online) Results of annotated video.

different points in time. The proposed evaluation metrics were applied to a dataset of instructional sessions, and the results demonstrate effectively the overall performance of learners in replicating instructor-guided hand and pen movements.

On the basis of the results obtained from DTW and PA-MPJPE calculations, we incorporated the visual guidance prompts into the videos recorded by the students. If PA-MPJPE, WRA_s , and RA_{pf} show low similarity, the instruction ‘Hand gesture error’, ‘The wrist rotation is too much’, and ‘Improper pen-holding technique’ will be displayed on the frame, respectively. As illustrated in Fig. 17, errors made by the students are displayed in the upper left corner of the screen, enabling students to make corrections on the basis of the prompts.

4. Conclusions

In this paper, we proposed a computer-vision-based system that uses a webcam to implement an AI-powered personalized online drawing assistant, which uses a series of hand skeleton analysis, two-stage drawing tool extraction, and action similarity evaluation techniques to perform three evaluation metrics on the learner's hand movements and pen usage, namely, the pen-holding posture similarity, the wrist rotation angle similarity, and the relative angle between the pen and the fingers. By recording the learner's imitation of the instructor's drawing process, the system generates text feedback on the imitation drawing process, annotates the actions needing modification on the original video, and replays it back to the learner. The learner can clearly see the actions that need correction by comparing their own movements with the instructor's movements.

The primary innovation of this research lies in the integration of MediaPipe, DTW, and PAMJPE for real-time feedback in online art education. While existing studies have explored AI-based applications for freeform creativity in art education, they have not focused on applying these technologies to hand posture comparison and real-time feedback. In contrast, our research combines these technologies to provide the real-time assessment of hand posture and pen manipulation during online drawing instruction, offering a novel approach to evaluating drawing skills. Compared with current studies, our system emphasizes the accuracy of hand posture and pen manipulation, leading to more concrete improvements in learners' drawing techniques and performance.

We selected post-playback feedback in this study to avoid disrupting the learner's workflow and compromising their drawing experience, as constant feedback during the drawing process could interfere with the learning flow. Moreover, drawing is a personalized creative process, unlike mathematics subjects that require every step to be completely correct before proceeding to the next step. Therefore, our study presents feedback in a more neutral manner, akin to a real-world learning environment where teachers do not typically provide quantitative scores for paintings. Instead, teachers usually offer suggestions for improvement. The proposed system is completely simulating the role of the teacher in the online environment, providing learners with software that enhances their drawing skills. We prioritize enhancing learning motivation over providing scores.

In future recommendations, although our system performed well in the experiments, there are still some limitations for improvement. For instance, the quality and positioning of the webcam can affect the accuracy and completeness of hand or pen detection. To address this, we plan to investigate more robust and stable methods for hand and pen detection and tracking to handle more complex and variable drawing scenarios. Additionally, we aim to incorporate more feedback and evaluation metrics, such as pen pressure, speed, and thickness, to provide a more comprehensive assessment of drawing skills and identify potential issues. Finally, we hope to collaborate with professional drawing instructors and students to validate the usability and effectiveness of our system in real teaching environments.

Acknowledgments

This research was supported by the Ministry of Science and Technology (MOST), Taiwan, under Contract No. NSTC 112-2222-E-130-001 and the Ministry of Education (MOE), Taiwan, under Contract No. PEE1134196.

References

- 1 M.-C. Chiu, G.-J. Hwang, L.-H. Hsia, and F.-M. Shyu: *Interact. Learn. Envir.* **32** (2024) 824. <https://doi.org/10.1080/10494820.2022.2100426>
- 2 S. P. Wu and M. A. Rau: *Learning and Instruction* **55** (2018) 93. <https://doi.org/10.1016/j.learninstruc.2017.09.010>
- 3 L. Shamir and J. A. Tarakhovsky: *ACM J. Comput. Cult. Heritage* **5** (2012) 1. <https://doi.org/10.1145/2307723.2307726>
- 4 E. Cetinic and S. Grgic: *Proc. ELMAR-2013* **5** (2013) 19–22. <https://ieeexplore.ieee.org/document/6658309>.
- 5 E. Cetinic, T. Lipic, and S. Grgic: *Expert Syst. Appl.* **114** (2018) 107. <https://doi.org/10.1016/j.eswa.2018.07.026>
- 6 Z. Falomir, L. Museros, I. Sanz, and L. Gonzalez-Abril: *Expert Syst. Appl.* **97** (2018) 83. <https://doi.org/10.1016/j.eswa.2017.11.056>
- 7 Z. Bai, Y. Nakashima, and N. Garcia: *Proc. IEEE/CVF Int. Conf. Computer Vision* (2021) 5422–5432. <https://doi.org/10.48550/arXiv.2109.05743>
- 8 A. Mordvintsev, C. Olah, and M. Tyka: *Google research blog* **20** (2015) 5. <https://research.google/pubs/inceptionism-going-deeper-into-neural-networks/>
- 9 L. A. Gatys, A. S. Ecker, and M. Bethge: *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2016) 2414–2423. <https://doi.org/10.1109/CVPR.2016.265>
- 10 M. D. M. Reddy, M. S. M. Basha, M. M. C. Hari, and M. N. Penchalaiah: *UGC Care Group I J.* **8** (2021) 71. <https://api.semanticscholar.org/CorpusID:261026641>
- 11 J. Singh, C. Smith, J. Echevarria, and L. Zheng: *Proc. Computer Vision – ECCV 2022* **13676** (2022) 685–701. https://doi.org/10.1007/978-3-031-19787-1_39
- 12 G. Lüthi, A. R. Fender, and C. Holz: *Proc. 35th Annu. ACM Symp. User Interface Software and Technology* **57** (2022) 1. <https://doi.org/10.1145/3526113.3545655>
- 13 Z. He, H. Xie, and K. Miyata: *Proc. 2020 Nicograph International (NicoInt)* (2020) 55. <https://doi.org/10.1109/NicoInt50878.2020.00018>
- 14 Z. Huang, Y. Peng, T. Hibino, C. Zhao, H. Xie, T. Fukusato, and K. Miyata: *Comput. Visual Media* **8** (2022) 63. <https://doi.org/10.1007/s41095-021-0227-7>
- 15 V. Križnar, M. Leskovešek, and B. Batagelj: *Proc. 2021 44th Int. Conv. Information, Communication and Electronic Technology (MIPRO)* (2021) 804–809. <https://doi.org/10.23919/MIPRO52101.2021.9596976>
- 16 S.-Y. Chen, P.-H. Lin, and W.-C. Chien: *Front. Psychol.* **13** (2022) 823078. <https://doi.org/10.3389/fpsyg.2022.823078>
- 17 S. Bambach, S. Lee, D. J. Crandall, and C. Yu: *Proc. IEEE/CVF Int. Conf. Computer Vision* (2015) 1949–1957. <https://doi.org/10.1109/ICCV.2015.226>
- 18 T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick: *Proc. Computer Vision–ECCV 2014* **8693** (2014) 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
- 19 F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, M. Grundmann: *arXiv. abs/2006.10214* (2020). <https://doi.org/10.48550/arXiv.2006.10214>
- 20 J. Cheng, F. Wei, Y. Liu, C. Li, Q. Chen, and X. Chen: *Math. Probl. Eng.* **2020** (2020) 8953670. <https://doi.org/10.1155/2020/8953670>
- 21 J. Shin, A. Matsuoka, M. A. M. Hasan, and A. Y. Srizon: *Sensors* **21** (2021) 5856. <https://doi.org/10.3390/s21175856>
- 22 M. Seeber, R. Poranne, M. Pollefeys, and M. R. Oswald: *Proc. 2021 Int. Conf. 3D vision (3DV)* (2021) 22–31. <https://doi.org/10.1109/3DV53792.2021.00013>
- 23 F. Xiao, Z. Zhang, C. Liu, and Y. Wang: *Biomed. Signal Process.* **79** (2023) 104089. <https://doi.org/10.1016/j.bspc.2022.104089>
- 24 G. Sánchez-Brizuela, A. Císnal, E. de la Fuente-López, J.-C. Fraile, and J. Pérez-Turiel: *Virtual Reality* **27** (2023) 3125. <https://doi.org/10.1007/s10055-023-00858-0>

- 25 Ultralytics: <https://doi.org/10.5281/zenodo.7347926> (accessed December 2024).
- 26 O. Ronneberger, P. Fischer, and T. Brox: Proc. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 **9351** (2015) 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
- 27 B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman: Int. J. Comput. Vision **77** (2008) 157. <https://doi.org/10.1007/s11263-007-0090-8>

About the Authors



Chiao-Wen Kao received her B.S. and M.S. degrees from the Computer Communication and Engineering from Ming-Chuan University, Taiwan, and Ph.D. degree from the Department of Computer Science and Information Engineering, National Central University, Taiwan, in 2018. She is currently an assistant professor at the Department of Applied Artificial Intelligence, Ming-Chuan University, Taiwan. Her research interests include pattern recognition, image processing, and interactive media. (chiaowen@mail.mcu.edu.tw)



Wen-Hsuan Wu received her M.S. degree from the Department of Applied Artificial Intelligence, Ming-Chuan University, Taiwan, in 2025. Her research interests include image processing and interactive media. (12756012@me.mcu.edu.tw)



Chi-Sheng Huang received his Ph.D. degree from the Computer Science and Information Engineering from National Central University, Taiwan, in 2018. He then joined the Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, Taiwan, as an assistant professor in 2022. His research interests include data mining, big data analysis, and cloud computation. (vcshuang@nutc.edu.tw)