J Inf Process Syst, Vol.20, No.3, pp.328~336, June 2024
https://doi.org/10.3745/JIPS.02.0214
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Design of HCI System of Museum Guide Robot Based on Visual Communication Skill

Qingqing Liang[1,2,*]

Abstract

Visual communication is widely used and enhanced in modern society, where there is an increasing demand for spirituality. Museum robots are one of many service robots that can replace humans to provide services such as display, interpretation and dialogue. For the improvement of museum guide robots, the paper proposes a human-robot interaction system based on visual communication skills. The system is based on a deep neural mesh structure and utilizes theoretical analysis of computer vision to introduce a Tiny+CBAM mesh structure in the gesture recognition component. This combines basic gestures and gesture states to design and evaluate gesture actions. The test results indicated that the improved Tiny+CBAM mesh structure could enhance the mean average precision value by 13.56% while maintaining a loss of less than 3 frames per second during static basic gesture recognition. After testing the system's dynamic gesture performance, it was found to be over 95% accurate for all items except double click. Additionally, it was 100% accurate for the action displayed on the current page.

Keywords

Guided Robot, HCI, Neural Network, Visual Communication

1. Introduction

The rapid advancement of wireless network technology, communication, and computing has greatly contributed to the overall progress of society [1,2]. Robot skills are often seen as a representative characteristic of scientific and technological progress, encompassing communication and artificial intelligence abilities. They are a product with a high level of technical proficiency [3,4]. In light of the service realms and objects of the robot system, intelligent robots can generally be divided into three categories: industrial robots, life service robots, and specialized robots in other fields. Among them, service robots primarily serve human beings and provide convenience for human economy and life [5]. Various application scenarios have various requirements for robots, and there are also many engineering designs for robots [6,7]. From an appliance scenario perspective, museum guide robots are a type of service robot that can replace manual labor to provide multiple services. They not only provide a good user experience but also save manpower. Therefore, this paper focuses on the increasing spiritual needs of the Chinese people and explores the use of museums as a scenario for studying the human-computer interaction (HCI) system. The goal is to improve the service function, attractiveness, and service level of museums by enhancing the museum guide function.

Manuscript received September 28, 2022; first revision April 19, 2023; second revision May 30, 2023; accepted June 6, 2023.

* **Corresponding Author:** Qingqing Liang (liangqingqing6888@163.com)
[1] School of Humanities, Art and Design, Guangxi University of Science and Technology, Liuzhou, China (liangqingqing6888@163.com)
[2] School of Art and Design, Guilin University of Electronic Technology, Guilin, China

1.1 Related Work

Due to the crucial role of vision in human-machine interaction, computer vision-based user interfaces have a significant impact on the development of perceptual interfaces. This has garnered attention from numerous experts and scholars. Fang et al. [8] developed a computer vision-based technique to detect seat belt usage among workers at heights. The technique used two convolutional neural network models to determine whether a worker is wearing a seat belt. The test results indicated that the accuracy rate of the convolutional neural network model was 80%, the recall rate was 98%, and the accuracy rate of the regional convolutional neural network was 99% and the recall rate was 95%. Brkic et al. [9] proposed a computer vision-based de-identification pipeline to study neural art algorithms. The algorithms presented pedestrian images in various styles using the responses of deep neural networks. The appearance of segmented pedestrians was changed for identification. The technique successfully completed the identification of non-living organisms, soft organisms, etc., in experiments. Garcia-Pulido et al. [10] developed an automatic expert system based on computer vision to address the flawed navigation ability of unmanned aerial vehicles. The system aided in safe landings by locating drones and platforms.

As a representative characteristic of scientific and technological progress integrating artificial intelligence, news, and communication, robots have attracted the attention of many scholars at home and abroad. Shuai and Chen [11] combined motion planning with neural networks to enable robots to recognize human interference, sensor errors, and part wear during motion execution. This was a step towards creating autonomous service robots that could tolerate unexpected environmental changes. Wang et al. [12] designed a new iterative learning control technique for perspective dynamic systems for the uncertainty in path tracking of portable service robots. Experiments indicated that the framework ran efficiently and met the trajectory accuracy requirements of portable service robot path tracking. Sawadwuthikul et al. [13] proposed a framework for communicating visual targets to robots through interactive two-way communication. The test results indicated that the proposed framework could help reduce the number of required bidirectional interactions and increase the robustness of the predictive model.

In summary, the exchange of news between humans and computers has been studied in the fields of neuroscience, news science, intelligence science, and psychology, and has yielded positive results. However, there is a lack of research on human spiritual needs. It is important to note that spiritual needs evolve with social progress. For example, modern people's concept of going to museums and their requirements for museums are various from those in the past. In addition, science and skill are constantly improving, and the museum itself needs to adapt to the times and change. Therefore, based on the visual communication skill, the study has carried out a new design of the HCI system of the museum guide robot.

2. Construction of a Museum Guide Robot HCI System based on Visual Communication Skill

2.1 Improved Gesture Recognition based on Deep Neural Network

The HCI system of the museum guide robot has multiple functions, including basic news management, voice interaction management, gesture recognition, and path navigation. The study primarily emphasizes

gesture recognition experiments and analysis. Gestures are considered as a natural and intuitive way of HCI, and gesture-based HCI skill has important study and appliance value [14,15]. The deep neural network model has a strong ability to learn nonlinearly, making it a popular choice for gesture recognition. YOLOv4-Tiny is a target detection grid constructed from convolutional neural lattices, which is fast and lightweight, and it is very suitable for gesture recognition. Therefore, the paper expects to carry out a study on the construction of a human-robot interaction system for museum tour guide robots from the YOLOv4-Tiny network. The structure of the YOLOv4 small object detection network is shown in Fig. 1.

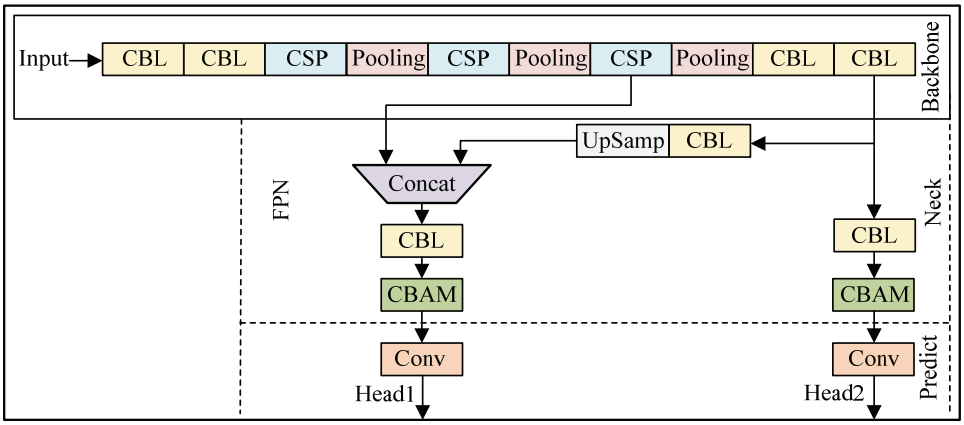

Fig. 1. YOLOv4-Tiny structure diagram of target detection network.

The Conv-Batch normalization-Leaky ReLU (CBL) layer, CSP layer and pooling layer constitute the backbone feature network. The CBL layer consists of convolution operations, batch normalization and activation functions. The expression of convolution is shown in Formula (1):

$$f(x) = \sum_{i,j}^{m} \theta_{i,j} x_{i,j} + \varepsilon \tag{1}$$

In Formula (1), the two position parameters of the convolution kernel are expressed as *i* and *j*. The data in the original image of the location is represented as $\theta_{i,j}$. The data *j* in the convolution kernel of the location is represented as $x_{i,j}$. The weight of this position is expressed as ε. The convolution Kernel size is denoted as m. The pooling layer has features that combine similar semantics, which can reduce the number of network parameters and the dimension of the feature map. The calculation is shown in Formula (2):

$$x_j = \beta_j down\left(x_j^{-1}\right) + b_j \tag{2}$$

In Formula (2), x_j and x_j^{-1} represent the *j* feature mapping in the pooling layer and convolution layer. β_j and b_j represent the weight and offset of the *j* feature mapping in the pooling layer. $down$ represents the pooling function. The role of the fully connected layer is to convert the two-dimensional feature map into a one-dimensional feature vector. To improve the feature extraction capability while increasing the

detection accuracy of the model on the basis of ensuring a certain recognition speed, the attention mechanism module is introduced in the feature fusion network to improve the YOLOv4-Tiny network. In Fig. 2, the attention mechanism module is positioned after the two effective feature layers to enhance the model's ability to represent the features.

2.2 Dynamic Gesture Recognition based on Static Basic Gestures

After obtaining the static basic gesture, it is necessary to use dynamic gesture recognition to track and judge the gesture action, which can be divided into five parts, as shown in Fig. 2.

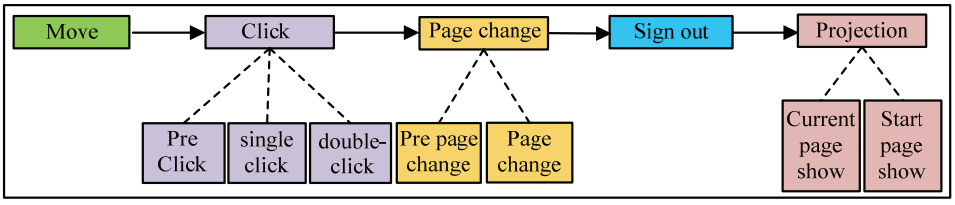

Fig. 2. Steps of dynamic gesture recognition.

In Fig. 2, the dynamic gesture recognition contains five parts: move, click, page change, exit, and indicate. The first time a gesture appears in the screen is recorded as follows. The movement S_f of the hand is recorded as S_m. The action is recorded as maintained S_k. The translation of the hand is recorded as S_{hm}. The rotation of both hands is recorded as S_r. When a single-finger gesture appears for the first time P_w, the current action is defined as establishing a moving area, denoted as M_b. The finger-tip position coordinate of single gesture P_w is set as (x_c, y_c). The width and height of the selection box are set as W_{wd} and h_{ht}. The width of the moving area is W_m. The height is H_m. The width and height of the screen area are W_{sc} and H_{sc}. The width and height of the camera capture area are W_{ca} and H_{ca}. Then the size of the moving area is shown in Formula (3):

$$\begin{cases} (W_m, H_m) = \left(4W_{wd}, \frac{H_{sc}}{m_l}\right), D_m \le D_{ca} \\ (W_m, H_m) = (W_{ca}, H_{ca}), D_m > D_{ca} \end{cases}. \tag{3}$$

In Formula (3), the size and position of the moving area and the camera area covered are D_m and D_{ca}. After setting the moving area, the mapping relationship between the position of the fingertip of the single-finger gesture and the position of the screen area is established. Assuming that the coordinate position of the upper left corner of the screen area is $(0,0)$, the coordinate of the fingertip position is (x_{sc}, y_{sc}), and the coordinate of the upper left corner of the moving area is (x_m, y_m), as shown in Formula (4):

$$\begin{cases} (x_m, y_m) = (0,0), D_m \le D_{ca} \\ (x_m, y_m) = \left(x_c - \frac{x_{sc}}{m_l}, y_c - \frac{y_{sc}}{m_l}\right), D_m > D_{ca} \end{cases}. \tag{4}$$

After the moving area is set, the target is tracked. The specific gesture action commands based on the duration of gesture retention are divided. The action of controlling page switching is called a page change action. The pre-page change action is expressed as M_{ps}. The page change action is expressed as M_{sw}, as

shown in Formula (5):

$$\begin{cases} M_{ps} = P_f + S_f \\ M_{sw} = M_{ps} + S_{hm} \\ A_{ang}(A, A_1) < 30° \,\&\, D_{dist}(A, A_1) > 0.6W_{wid}(A) \end{cases}. \tag{5}$$

When the target shows a five-finger gesture after being tracked, it is a pre-exit action, denoted as M_{pe}. When the fist gesture appears again, it is an exit action, denoted as M_e, as shown in Formula (6):

$$M_e = M_{pe} + P_s, 10U(B, B_1) > 0.1 \,\&\, T_{k2} > 0.5\, s. \tag{6}$$

In Formula (6), T_{k2} is the duration of the five-finger gesture. After target tracking, if both targets are five-finger gestures, they are recorded as P_{f1} and P_{f2}. The pre-screening action is shown in Formula (7):

$$\begin{cases} M_{pp} = P_{f1} + P_{f2} + S_c \\ A_{ang}(E, F) < 30° \,\&\, D_{dist}(E, F) > W_{wid}(E) \end{cases}. \tag{7}$$

In the light of the relative angle size and the length of time in the show judgment state, it is segmented into the current page indicate and the start page show, as shown in Formula (8):

$$\begin{cases} M_{cp} = M_{pp} + S_c \\ A_{ang}(E, F) < 30° \,\&\, D_{dist}(E, F) > W_{wid}(E) \\ M_{sp} = M_{pp} + S_r \\ A_{ang}(E, F) > 60° \,\&\, T_{k3} > 0.5\, s \end{cases}. \tag{8}$$

In Formula (8), when the relative angle is less than 30°, and the relative distance keeps getting smaller until it is smaller than the width of the marquee, the current page is shown. When the relative angle keeps increasing until it is greater than 60°, and the projection judgment state time is greater than 0.5 seconds, it is the start page show. The study uses mean average precision (mAP) and a performance metric to validate the validity and superiority of the study method. The mAP is a crucial metric in target detection. It is calculated as a weighted average of the correct detection rates for all categories. Frames per second (FPS) is the number of images that can be processed per second or the time it takes to process an image, and is used to evaluate the efficiency of an algorithm's operation.

3. Performance Analysis of HCI System of Museum Guide Robot

The migration learning idea is used, and the YOLOv4-Tiny. The conv.29 weights pre-trained on the COCO dataset from the YOLO website are used as the initial weights for training. The batch size is set to 64, the sub-vision is 8, the momentum parameter is 0.9, the weight decay is 0.0005, the maximum number of training iterations is 5,000 and the initial learning rate is 0.001. The experimental dataset is obtained from a large number of videos of gesture news collected using a crowd-sourcing approach. Dividing the 2,750 data sets into 2,305 training sets and 445 test sets. Based on the network structure of the first to eighth convolutional layers, the loss curve of the 5,000 iteration training and validation set is shown in Fig. 3.

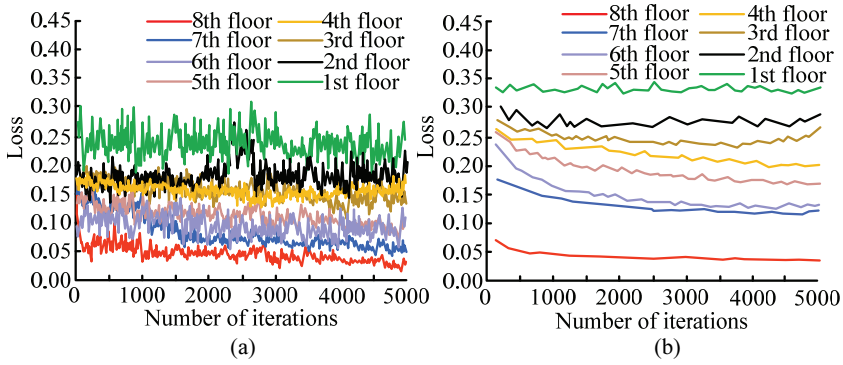

Fig. 3. Fully convolutional network (FCN) corresponding loss of various convolution layers: (a) training set and (b) validation set.

As the number of convolutional layers decreases, the convergence speed becomes slower. When there are only 3 convolutional layers left, the model cannot converge. When combined with 8 to 4 convolutional layers, with the reduction of the number of convolutional layers, the running speed becomes faster, and the space required for the model becomes smaller, so the study chooses 4 convolutional layers. In the study, gesture recognition is used to design the news interaction system, and the graphics are designed accordingly. In gesture recognition, the domain average technique is used to denoise the gesture news. Therefore, in the test, the gesture news is denoised first. The effect is analyzed to determine whether the gesture recognition news in the system can be used, as shown in Fig. 4.

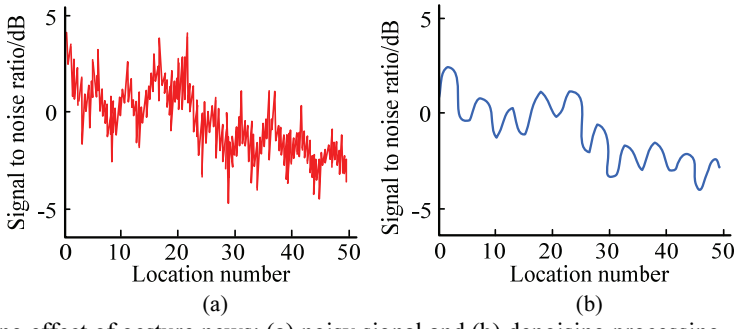

Fig. 4. Denoising effect of gesture news: (a) noisy signal and (b) denoising processing.

In Fig. 4(a), the gesture signal has obvious burrs at various positions, indicating that there is a connection between the points in the gesture signal at this time, and the noise is obvious. Fig. 4(b) shows the changes in gesture signals after denoising using domain averaging technology. At this point, the outline of the gesture image is clear and the meaning conveyed by the gesture can be well recognized. The above results indicate that domain averaging technology can effectively remove noise from the initial gesture image, making it have clear gestures and distinct features. The static basic gesture recognition results of YOLOv4-Tiny network and improved Tiny+CBAM network are shown in Fig. 5.

In Fig. 5, it can be observed that the mAP of the Tiny+CBAM network has increased by 13.56% compared to YOLOv4-Tiny, and the loss is less than 3 FPS. The operational efficiency of the HCI system based on this network is much higher than the other three. The results of dynamic gesture recognition detection are shown in Table 1.

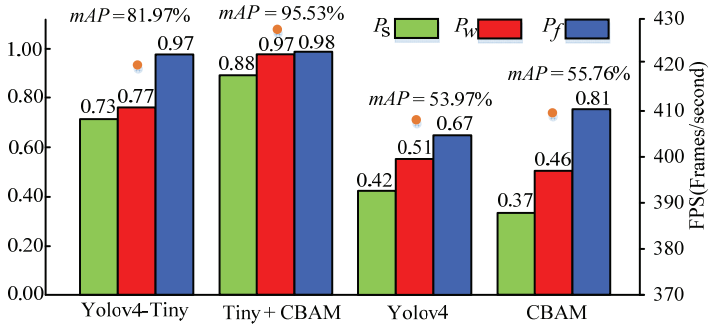

Fig. 5. Four kinds of network mAP test results.

From Table 1, the study has carried out practical appliance detection on the designed actions and possible wrong actions. The total number of actual appliances is 1,600, the total number of false detection is 56, and the total correct rate is 96.5%. Among them, except the number of page changes is 400 times, the rest of the actions are 200 times. The accuracy rates of click, page change, exit, current page display, and start page display are all over 95%, and the accuracy rate of double-click is the lowest, which is 91%. In addition, the current page shows no false detection in the 200 actual appliance times.

Table 1. Statistical table of practical appliance results of dynamic gesture recognition

	Number of samples	Number of false detection	Accuracy (%)
Single click	200	8	96.0
Double-click	200	18	91.0
Page change	400	14	96.5
Sign out	200	6	97.0
Current page show	200	0	100
Start page show	200	2	99.0
Wrong action	200	8	96.0
Total	1,600	56	96.5

4. Conclusion

Perceiving human-computer interface is the progress tendency of HCI in the future. Among them, the skill of human-computer interface based on machine vision has also received increasing attention from researchers. To improve the performance of the museum guide robot, a study based on visual communication skills and the introduction of deep neural networks to the HCI system was proposed, which could result in a new design for the museum guide robot. The static basic gesture recognition utilized the Tiny+CBAM network, which integrates an attention mechanism. The dynamic gesture recognition also employed the basic gesture in combination with the gesture state. The results suggested that selecting four layers from the multilayer convolutional layers was the most suitable option. In the performance test of the YOLOv4-Tiny network and the Tiny+CBAM network, the mAP test results for the two networks were 81.97% and 95.53%, respectively. The mAP of the Tiny+CBAM network was improved by 13.56%. With a total sample of 1,600 actions, the total number of false detection of the

system was only 56, and the correct rate could reach 96.5%. The system improves the recognition accuracy and ensures the ductility of gesture actions while maintaining high speed and light weight. The study is carried out in various experiments indoors. Future research can also consider how to perform accurate gesture segmentation outdoors to further improve the robustness of recognition algorithms and gesture segmentation.

Conflict of Interest

The author declare that they have no competing interests.

Funding

None.

References

[1] Y. Lou, J. Wei, and S. Song, "Design and optimization of a joint torque sensor for robot collision detection," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6618-6627, 2019. https://doi.org/10.1109/JSEN.2019.2912810

[2] D. Brscic, T. Ikeda, and T. Kanda, "Do you need help? a robot providing information to people who behave atypically," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 500-506, 2017. https://doi.org/10.1109/TRO.2016.2645206

[3] G. Doisy, J. Meyer, and Y. Edan, "The impact of human–robot interface design on the use of a learning robot system," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 6, pp. 788-795, 2014. https://doi.org/10.1109/THMS.2014.2331618

[4] S. Haghzad Klidbary, S. Bagheri Shouraki, and S. Sheikhpour Kourabbaslou, "Path planning of modular robots on various terrains using Q-learning versus optimization algorithms," *Intelligent Service Robotics*, vol. 10, pp. 121-136, 2017. https://doi.org/10.1007/s11370-017-0217-x

[5] A. Sahai, E. Caspar, A. De Beir, O. Grynszpan, E. Pacherie, and B. Berberian, "Modulations of one's sense of agency during human–machine interactions: a behavioural study using a full humanoid robot," *Quarterly Journal of Experimental Psychology*, vol. 76, no. 3, pp. 606-620, 2023. https://doi.org/10.1177/17470218221095841

[6] D. Ruhlmann, J. P. Fouassier, and F. Wieder, "Relations structure-proprietes dans les photoamorceurs de polymerisation—5. Effet de l'introduction d'un groupement thioether," *European Polymer Journal*, vol. 28, no. 12, pp. 1577-1582, 1992. https://doi.org/10.1016/0014-3057(92)90154-T

[7] A. R. Habib, G. Crossland, H. Patel, E. Wong, K. Kong, H. Gunasekera, et al., "An artificial intelligence computer-vision algorithm to triage otoscopic images from Australian Aboriginal and Torres Strait Islander children," *Otology & Neurotology*, vol. 43, no. 4, pp. 481-488, 2022. https://doi.org/110.1097/MAO.0000000000003484

[8] W. Fang, L. Ding, H. Luo, and P. E. Love, "Falls from heights: a computer vision-based approach for safety harness detection," *Automation in Construction*, vol. 91, pp. 53-61, 2018. https://doi.org/10.1016/j.autcon.2018.02.018

[9] K. Brkic, T. Hrkac, and Z. Kalafatic, "Protecting the privacy of humans in video sequences using a computer vision-based de-identification pipeline," *Expert Systems with Applications*, vol. 87, pp. 41-55, 2017. https://doi.org/10.1016/j.eswa.2017.05.067

[10] J. A. Garcia-Pulido, G. Pajares, S. Dormido, and J. M. de la Cruz, "Recognition of a landing platform for unmanned aerial vehicles by using computer vision-based techniques," *Expert Systems with Applications*, vol. 76, pp. 152-165, 2017. https://doi.org/10.1016/j.eswa.2017.01.017

[11] W. Shuai and X. P. Chen, "KeJia: towards an autonomous service robot with tolerance of unexpected environmental changes," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, no. 3, pp. 307-317, 2019. https://doi.org/10.1631/FITEE.1900096

[12] Y. Wang, F. Zhou, Y. Zhao, M. Li, and L. Yin, "Iterative learning control for path tracking of service robot in perspective dynamic system with uncertainties," *International Journal of Advanced Robotic Systems*, vol. 17, no. 6, article no. 1729881420968528, 2020. https://doi.org/10.1177/1729881420968528

[13] G. Sawadwuthikul, T. Tothong, T. Lodkaew, P. Soisudarat, S. Nutanong, P. Manoonpong, and N. Dilokthanakul, "Visual goal human-robot communication framework with few-shot learning: a case study in robot waiter system," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1883-1891, 2022. https://doi.org/10.1109/TII.2021.3049831

[14] J. L. Xu, C. Riccioli, and D. W. Sun, "Comparison of hyperspectral imaging and computer vision for automatic differentiation of organically and conventionally farmed salmon," *Journal of Food Engineering*, vol. 196, pp. 170-182, 2017. https://doi.org/10.1016/j.jfoodeng.2016.10.021

[15] T. Toulouse, L. Rossi, A. Campana, T. Celik, and M. A. Akhloufi, "Computer vision for wildfire research: an evolving image dataset for processing and analysis," *Fire Safety Journal*, vol. 92, pp. 188-194, 2017. https://doi.org/10.1016/j.firesaf.2017.06.012

Qingqing Lian **https://orcid.org/0000-0003-0502-2009**

She graduated from Guilin University of Electronic Technology, majoring in the School of Art and Design, with a bachelor's degree (Sep 2017–Jul 2020) and a master's degree. She works at the School of Humanities, Arts and Design, Guangxi University of Science and Technology. The research direction is product design. She has published four academic papers, participated in two scientific research projects, and six other academic research achievements and projects.