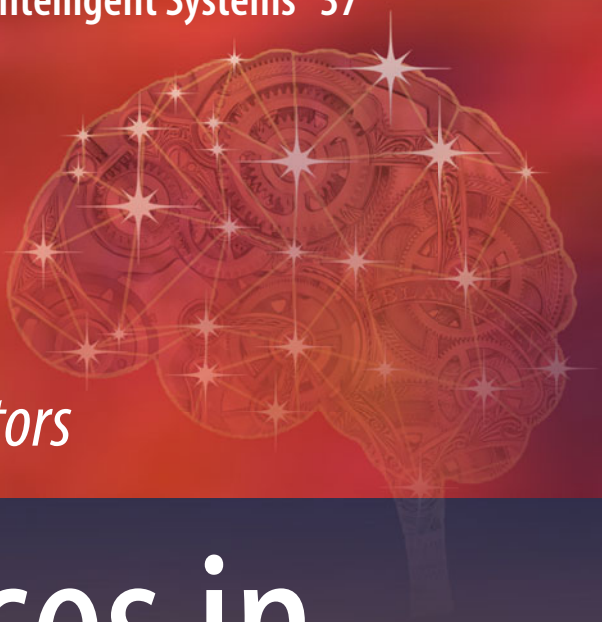
Learning and Analytics in Intelligent Systems 37



Jianbo Su
Xiuquan Qiao *Editors*

Advances in Haptics and Virtual Reality

Proceedings of 2023 International Conference on Haptics and Virtual Reality

Springer

Learning and Analytics in Intelligent Systems

Volume 37

The main aim of the series is to make available a publication of books in hard copy form and soft copy form on all aspects of learning, analytics and advanced intelligent systems and related technologies. The mentioned disciplines are strongly related and complement one another significantly. Thus, the series encourages cross-fertilization highlighting research and knowledge of common interest. The series allows a unified/integrated approach to themes and topics in these scientific disciplines which will result in significant cross-fertilization and research dissemination. To maximize dissemination of research results and knowledge in these disciplines, the series publishes edited books, monographs, handbooks, textbooks and conference proceedings.

Indexed by EI Compendex.

Jianbo Su · Xiuquan Qiao
Editors

Advances in Haptics and Virtual Reality

Proceedings of 2023 International Conference on Haptics and Virtual Reality

Springer

Editors
Jianbo Su
Department of Automation
Shanghai Jiao Tong University
Shanghai, China

Xiuquan Qiao
Beijing University of Posts
and Telecommunications
Beijing, China

ISSN 2662-3447 ISSN 2662-3455 (electronic)
Learning and Analytics in Intelligent Systems
ISBN 978-3-031-56520-5 ISBN 978-3-031-56521-2 (eBook)
https://doi.org/10.1007/978-3-031-56521-2

Contents

Research on Surface Defect Detection of Microchannel Aluminum Flat Tubes Based on Improved Faster RCNN

Wei Zhang, Ziman Guo, Junpeng Xu, Xinjian Li, and Xiangbo Zhu

Abstract To address the challenges of low detection rates for surface defects in microchannel aluminum flat tubes and poor detection effectiveness for small target defects using existing deep learning methods, a Faster RCNN based detection method is proposed. To overcome the limitations of multi-scale recognition in Faster RCNN, the defects were addressed by replacing the feature extraction network with ResNet152 and integrating FPN. These enhancements significantly improved the detection of intricate defects. The experimental results reveal that the improved model achieves an impressive mean average precision of 96.1% in detecting surface defects of microchannel aluminum flat tubes, which is 6.4% higher than the original Faster RCNN, and the detection accuracy of scuffing, scratches, and dirty spot defects is improved by 7.0, 8.9, and 4.1%, respectively. Moreover, the proposed model surpasses alternative target detection algorithms, maintains a low false detection rate, and significantly improves the detection of small target defects on microchannel aluminum flat tubes.

Keywords Microchannel aluminum flat tube · Surface defect detection · Faster RCNN · Residual network · Feature pyramid network

1 Introduction

Microchannel aluminum flat tubes are a unique form of flat tubular aluminum, characterized by their thin-walled porous structure, which is achieved through a hot extrusion process, surface zinc spraying and other operations, which is widely used in air conditioning and other systems. The heat exchange efficiency of microchannel aluminum flat tube is higher than the traditional copper tube fin type [1], reducing the refrigerant charge, and the all-welded performance attenuation is small, and the demand for microchannel aluminum flat tube will increase in many domestic

W. Zhang (✉) · Z. Guo · J. Xu · X. Li · X. Zhu
School of Mechanical and Electrical Engineering, Henan Institute of Science and Technology, Xinxiang 453003, China
e-mail: 13598710802@139.com

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_1

fields [2]. Due to the influence of many factors from raw materials and industrial control systems and workers' operation level, the surface of microchannel aluminum flat tube is prone to scuffing, scratches, dirty spots and other problems, and these defects will seriously affect the performance of the aluminum flat tube [3]. The micro-channel aluminum flat tube of the air conditioning heat exchanger contains refrigerant medium, and surface defects are prone to oxidative leakage due to pitting corrosion action, and thus the entire heat exchanger will be replaced. Therefore, effective detection of microchannel aluminum flat tube defects to improve the yield is necessary.

Currently for microchannel aluminum flat tube defects to be able to obtain the type, size and location of defects, the detection of its detection difficulty lies in the detection of small target defects [4]. Traditional image vision methods mainly use traditional machine learning or joint image processing methods means of detection [5]. Feature extraction based methods mainly use feature information in the image and classify these features using classifiers to achieve the purpose of defect detection. Shipway et al. [6] for the detection of defects in titanium plates by means of traditional image processing techniques combined with random forest algorithms. In addition to the manual extraction of features from traditional images, traditional industrial defect detection techniques such as magnetic particle [7], penetration, eddy current, ultrasonic [8] and magneto-optical imaging are commonly used in the identification of defects in microchannel aluminum flat tubes in order to detect and identify cracks, white dots, defects, and internal problems in the samples. These methods often can only detect one or several types of defects, so these traditional methods are gradually replaced when facing multiple types of defects [9].

The deep learning-based approach solves the problem that traditional machine vision requires different image processing algorithms to categorize different tasks [10, 11], but the approach still has a lot of room for improvement. Neuhauser et al. [12] used neural networks to distinguish aluminum defects to meet production requirements. Target detection is divided into one stage and two stage algorithms. YOLO and SSD as single stage algorithms are based on direct detection of whole image. Yin et al. [13] used Yolov3 to detect defects in sewage pipes and obtained 85.37% mAP. But the emphasis on accuracy in the field of microchannel aluminum flat pipe defect detection cannot meet the industrial needs. The two-stage target detection algorithm first generates a large number of candidate frames that may contain the target, and then classifies the target. He et al. [14] achieved 82.3% mAP in strip steel surface defects using Faster RCNN method with ResNet50 as backbone.

Based on the above research, at the present stage of defect detection of microchannel aluminum flat tubes, a suitable model method for defect detection of aluminum flat tubes is proposed. In order to solve the problem of small samples and small targets, we propose a target defect detection method based on Faster RCNN, introduce the feature pyramid network (FPN) to transform the features of different layers into scales, and replace the VGG16 network with a deeper Resnet152 residual network for feature extraction. Finally, experimental validation was performed using the microchannel aluminum flat tube defect dataset.

2 Algorithm Design and Improvement

With the rapid development of CV, Ren et al. [15] proposed the Fast RCNN for object detection in 2015. Faster RCNN is used to extract image features and then multiple regions are selected from these feature maps, and pooling and fully connected operations are performed in these regions for determining whether they contain the target or not and refining the bounding box locations. The region recommendation network is responsible for generating prediction frames, which are mainly used to produce recommendation regions, which may contain target features. Finally, the region of interest containing all the information is fused with pooling layer with fully connected operations to perform softmax classification and bounding box regression tasks.

2.1 RPN Basic Structure and ROI Align

Region Proposal Network (RPN) as a deep learning network, is used to generate candidate regions in an image that may contain objects and provide effective inputs for subsequent target classification and localization tasks, and is an important part of target detection algorithms, which can significantly reduce the computational volume compared with the traditional sliding window method.

RPN performs 3×3 convolutional sliding on the shared feature maps to obtain anchor boxes at each pixel point corresponding to the center point on the original image. In this anchor boxes three different area ratios are used, 128^2, 256^2 and 512^2, and nine anchor boxes with different aspect ratios are generated. The anchor frames are used at each grid point on the feature layer to predict the presence or absence of an object in the target frame. Candidate frames are mapped in a shared feature map, the feature matrix undergoes pooling layer and fully connected layer operations, the feature matrix is changed in length and width and evaluated, and scaled to 7×7 size to complete the bounding box regression and classification operations.

This time, ROI Align is used instead of the original ROI Pooling. The original ROI Pooling operates in the candidate box using the insert neighboring pixel value method, where the mapped floating-point type coordinates are rounded down and quantized, but this produces a loss of information resulting in region mismatches. For subtle defects in microchannel aluminum flat tubes, ROI Align does not perform the quantization step of the regression box, and uses single and bilinear interpolation to obtain pixels for a more accurate regression box of the defect target.The advantage of RoI Align over RoI Pooling is that it is more accurate and can retain more spatial information, especially when dealing with small RoIs.

2.2 Region Proposal Network with FPN Convergence

The RPN in Faster RCNN extracts from high-level features, which can make small target defect detection ineffective. The semantic features of the higher layers can easily and comprehensively describe the object, but the deeper the layers are, the more likely it is to lose some of the detailed feature information of the defective image of the microchannel aluminum flat tube. Introducing RPN at the end of the feature extraction network usually uses image pyramid techniques to enhance images with multi-scale variations, but this increases the amount of computation and does not effectively fuse high-level semantic and low-level detail information.

For the small target defect problem, the traditional Faster RCNN model is optimized by introducing the feature pyramid network (FPN) [16], FPN has a unique feature pyramid structure, which performs scale transformation and information fusion on different layers of features, enabling the high-level network to access the information of the lower layers. These improvements effectively enhance the accuracy of detecting surface defects, especially small defects, on microchannel aluminum flat tubes. The feature pyramid network utilizes a unique feature pyramid design to more efficiently handle multi-scale problems in target detection. FPN constructs a feature pyramid through the multi-scale, hierarchical structure of deep convolutional networks, which first extracts features from the bottom-up, then uses up-sampling from the top-down to obtain low-level features, and finally fuses the feature maps of the first two parts together through horizontal connectivity.

This time, we use feature pyramid fused with region recommendation model to solve the problem of unbalanced aspect ratio and subtle defects in microchannel aluminum flat tube images. After feature extraction, the C1, C2 and C3 feature maps are horizontally connected with the P1, P2 and P3 feature maps for fusion to form a new feature map. In order to better fuse the shallow features, a 3×3 convolutional layer is introduced, which then performs the classification and regression tasks via RPN. Compared with the original Faster RCNN model, the network structure after adding FPN effectively improves the network's ability to cope with the generalization of small target defects in microchannel aluminum flat tubes, and significantly improves the accuracy of surface defect detection in microchannel aluminum flat tubes. As shown in Fig. 1 is the region recommendation network that incorporates the feature pyramid.

2.3 Improved Feature Extraction Network

VGG16 is Faster CNN's original feature extraction network, the multilayer convolution of VGG16 gives it a good nonlinear mapping capability, which makes the network fit well. The VGG16 network extracts features from the input image and achieves a strong nonlinear mapping capability through multilayer convolution. When operations such as convolution and pooling are performed, information in

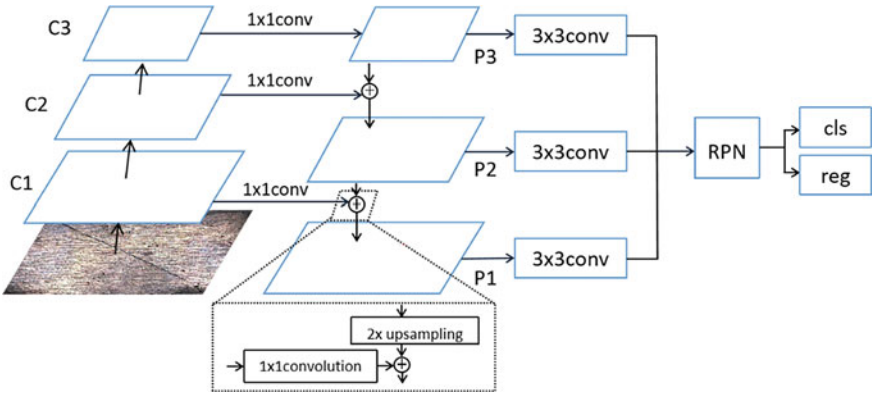


Fig. 1 Region proposal network with fused feature pyramid network

the microchannel aluminum flat tube image data is inevitably lost, and the network may also degrade as the number of network layers gradually deepens.

For the problems that arise and then improve the feature extraction network, the deep residual network Resnet is added to realize the image feature extraction. Resnet passes jumps by means of short connections, allowing inputs to be passed directly to the next layer, avoiding information loss and better mitigating the problem of vanishing gradients. We achieved significant performance improvement by optimizing the network structure to make it more adaptive to the target. In this work, we use the deeper Resnet152 network instead of the VGG16 network for feature extraction. Resnet152 is a further improvement on the Resnet series of models, which is considerably deeper, with a 152-layer network structure, including 151 layers of convolutional layers and 1 layer of fully connected layers. The model uses multiple residual blocks, which contain convolutional and normalization layers, and employs jump-joining, which passes the previous features directly to the later layers, effectively mitigating the problem of vanishing gradients, and the network is able to train more deeply while gaining some robustness.

2.4 Improved Faster RCNN Target Detection Algorithm Flow

This optimization for the Faster RCNN network is mainly in two aspects, firstly, the deep Resnet152 residual network is used instead of the VGG16 network, and in order to solve the problems of easy loss of small targets and unbalanced aspect ratios of multi-category defective targets, the feature pyramid network structure is added, and the multi-scale feature maps are extracted and fused to cleverly integrate the features of the low-level and high-level semantics, which enhances the detection accuracy, and effectively improves the network's generalization performance for

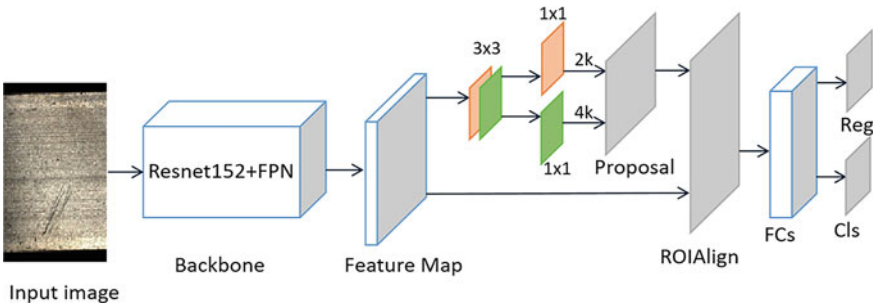


Fig. 2 Improved faster RCNN target detection algorithm

multi-category and small target defects. The flow of the improved Faster RCNN algorithm is shown in Fig. 2.

3 Acquisition and Analysis of Data

In this paper, the microchannel aluminum flat tube defect data set from the actual production of defective samples, using Hikvision MV-CE200-10GC industrial camera shooting collection, the light source using a spherical light source for dimming. The total number of defects collected is 2114, of which 781 are scuffing defects, 703 are scratching defects, and 630 are dirty spot defects. The excess black background during shooting is eliminated by cropping to reduce the size of the data image occupying space bytes for better training of the dataset. The defect dataset comes from the defect samples in the factory, and the amount of data is not a lot, which cannot reflect the optimization and training effect of the network. This time in the original defect data set on the data enhancement, mainly through the microchannel aluminum flat tube image horizontal and vertical rotation, to increase the noise for blurring, and compression processing image size, data enhancement after a total of 8432 images, and then labeling work.

This time, LabelImg is used to label the defective data, and the data is randomly divided according to the ratio of 80% of the training set and 20% of the validation set, and the defective labels are shown in Fig. 3, and the division of defects data is shown in Table 1. The aspect ratio of each type of defects is not consistent, and there are a variety of shapes of the same defects, such as scuffing have a large area but fuzzy, while the area of small target dirty spots is very small, while the reflection of the microchannel aluminum flat tube metal increases the difficulty of shooting and detecting defects, and the border labeling situation and the size of the labeling when manually annotated also affects the accuracy of the detection algorithm.

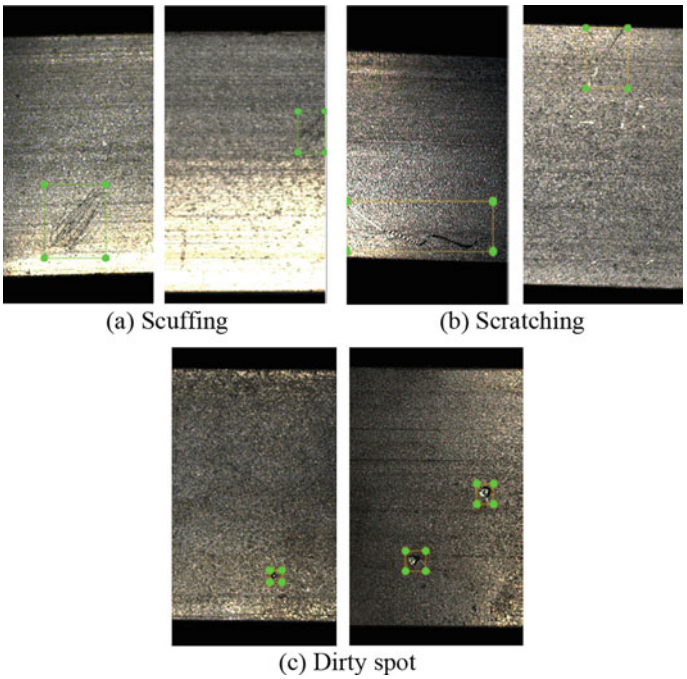


Fig. 3 Defective label image

Table 1 Defect classification data

Data set	Training set	Test set	Total
Scuffing	2500	624	3124
Scratching	2250	562	2812
Dirty spot	2016	504	2520

4 Experimental Results and Analysis

4.1 Hardware and Software Platform Setup

The CPU configuration in the experiment is Intel Xeon Platinum 8358@2.60 GHz, memory is 80G, Nvidia RTX 3090 (24 GB), and the driver is Cuda11.3. Running on Linux operating system, the software is configured with ubuntu20.04, and the deep learning training framework uses Pytorch 1.11.0 version, while using Python version 3.8, image annotation tools using LabelImg tool.

4.2 Assessment of Indicators

Common evaluation metrics such as average precision (AP), which is the average of each category of picture accuracy, and mean average precision (mAP), which is the average of the APs of all defect types, were used in this evaluation. In the evaluation process, manually labeled bounding boxes and detection markers are compared to derive the results and an evaluation value is derived. In this evaluation, intersection over union >0.5 is used as the criterion for correctly detecting defective targets. Other important evaluation metrics include Precision, Recall, and F1 score, which can be used to comprehensively evaluate the performance of the model.

4.3 Comparative Experimental Results Analysis

In this experiment, a training batch size of 4 was used, and a total of 30 rounds of training were performed, using SGD as the optimizer, with the initial learning rate set to 0.02, and weight decay of 0.0001 was introduced.

The results of this detection using the improved Faster RCNN target detection algorithm are shown in Table 2. As seen in the table, the original Faster RCNN is less effective in detecting each class, the mAP value of the network improves by 2.9% after using the residual network Resnet50 to replace the VGG network for feature extraction, the mAP value of the network improves by 2.0% by using the deeper Resnet152 network instead of the Resnet50 network, and the fusion of the feature pyramid network the mAP value is further improved by 1.5%. This improved model has an overall improvement of 6.4% over the original model, including 4.1% for dirty spot defects, 8.9% for scratch defects, and 7.0% for scuffing defects, which proves that the detection model has a strong generalization for the detection of defects in microchannel aluminum flat tubes. The total training loss curve is shown in Fig. 4, Fig. 4 horizontal coordinate is the number of iterations, and the loss is generally decreasing, and the training starts to converge and stabilize at the 18th epoch.

From Table 3, improved Faster RCNN performs better than the popular target detection algorithms today, with accuracy 11.6% higher compared to YOLOv3, and 7.2 and 6.4% higher than SSD300 and SSD512 networks, respectively. The F1 Score as an important index of target detection, is 96.9% for the improved network model,

Table 2 Comparison of evaluation metrics of improved Faster RCNN for defects

Detection network	Dirty spot AP/%	Scratching AP/%	Scuffing AP/%	mAP/%
Faster RCNN + VGG16	89.2	88.5	90.7	89.7
Faster RCNN + Resnet50	90.8	93.4	93.6	92.6
Faster RCNN + Resnet152	92.1	96.4	95.3	94.6
Faster RCNN + Resnet152 + FPN	93.3	97.4	97.7	96.1

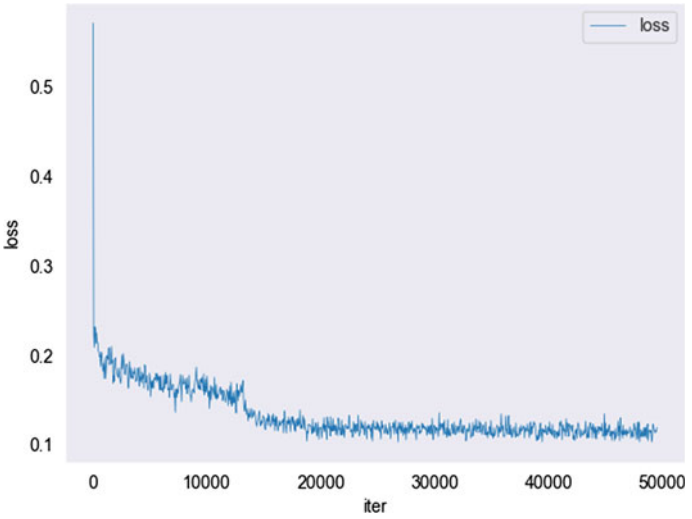


Fig. 4 Total training loss curve

Table 3 Comparison of detection performance of different algorithms for defects

Algorithm	Precision/%	Recall/%	F1 Score/%
SSD300	88.9	94.2	91.5
SSD512	89.7	95.4	92.5
YOLOv3	84.5	93.2	88.6
Improved Faster RCNN	96.1	97.7	96.9

which is 8.3% higher compared to YOLOv3, and 5.4 and 4.4% higher than SSD300 and SSD512 networks, respectively, which proves that the improved network model has strong stability. The Faster RCNN has high accuracy compared to other algorithms, but for the YOLO algorithm compared to its detection time is long and still needs to be improved.

5 Conclusion

In this work, an improved Faster RCNN algorithm is proposed for the problem of detecting surface defects on aluminum flat tubes with microchannels, and for the problems of inconspicuous defects on small targets and uneven aspect ratios of multiple defects. In the improved model, a deeper residual network is used to solve the gradient vanishing. In addition, the feature pyramid network is integrated to make the network more effective in detecting the detail information. The network model of this design has strong stability and generalization, which effectively solves the

problem of detecting multiple defects in microchannel aluminum flat tubes, dramatically improves the defect detection accuracy, and proposes an effective method for machine vision to solve the problem of small targets and multi-category defects.

References

1. Chen, R., Quan, Z., Zhao, Y., et al.: Simulation study on heat transfer performance of adsorption bed based on parallel flow aluminum flat tube. Renew Energy **38**(03), 312–318 (2020)
2. Zeng, C.: Market prospect and production technology of microchannel aluminum flat tube for heat exchanger. Nonferrous Metal Process. **49**(05), 6–8 (2020)
3. Wang, X., Sun, K, Yan, C., et al.: Research on heat transfer performance of aluminum flat tube for air conditioning. Light Indus Stand Qual **2022**(2), 105–109 (2022)
4. Zhou, P., Zhou, G.B., Li, Y.M., et al.: A hybrid data-driven method for wire rope surface defect detection. IEEE Sens. J. **20**(15), 8297–8306 (2020)
5. Jin, Y., Zhang, T., Yang, Y., et al.: A review of product defect detection methods based on deep learning[J]. J Dalian Univ Nationalities **22**(5), 420–427 (2020)
6. Shipway, NJ, Barden, TJ, Huthwaite, et al.: Automated defect detection for fluorescent penetrant inspection using random forest. NDT & E Int. **101**, 113–123 (2019)
7. Ma, G, Jia, H, Lu, C., et al.: Application of magnetic particle testing and penetration testing in nondestructive testing of structural components of construction machinery. Nondestr. Test. **41**(2), 68–70 (2019)
8. Sheng-nan, X.U.E., Qi-chi, L.E., Yong-hui, J.I.A., et al.: Ultra sonic flaw detection of discontinuous defects in magnesium alloy materials. China Foundry **16**(4), 256–261 (2019)
9. Zhang, X., Huang, D.: Deep learning based surface defect detection of aluminum. J. East China Normal Univ. (Nat. Sci. Ed.) **2020**(6), 105–114 (2020)
10. Luo, D., Cai, Y., Yang, Z., et al.: A review of deep learning methods for industrial defect detection. Sci. China: Inform. Sci. (052–006) (2022)
11. Yang, C., Zhang, X.: A review of material defect detection applications based on machine vision and deep learning. Mater. Herald **36**(16), 20070136–20070139 (2022)
12. Neuhauser, F.M., Bachmann, G., Hora, P.: Surface defect classification and detection on extruded aluminum profiles using convolutional neural networks. Int.J. Mater. Form. **13**(4), 591–603 (2020)
13. Yin, X., Chen, Y., Bouferguene, A., et al.: A deep learning-based framework for an automated defect detection system for sewer pipes. Autom. Constr. **109**, 102967 (2020)
14. He, Y., Song, K., Meng, Q., Yan, Y.: An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Trans. Instrum. Meas. **69**(4), 1493–1504 (2020)
15. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks (2015). ArXiv preprintarXiv:1506.01497
16. Lin, T.Y., Dollar, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)

VIOFormer: Advancing Monocular Visual-Inertial Odometry Through Transformer-Based Fusion

Jie Zhao, Yuanwei Zhu, Yakun Huang, Xiuquan Qiao, Meng Guo, Hongshun He, and Yang Li

Abstract Visual-Inertial Odometry (VIO) algorithms are pivotal for localization and navigation, with deep learning enhancing their potential. However, long-distance pose estimation still poses challenges. While most VIO algorithms use Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), the promise of Transformer models remains largely unexplored. We present VIOFormer, a new monocular visual-inertial pose estimation technique harnessing Transformers. By leveraging their attention mechanism and long-range dependencies, VIOFormer addresses the cumulative error issue linked with IMU and augments pose accuracy. Its architecture combines the preprocessing of adjacent frame sequences with IMU data, an encoder, and a decoder, allowing it to discern features from both visual and IMU data for precise pose estimation. Testing on the KITTI dataset confirms VIOFormer's superiority over both traditional and contemporary deep learning-based VIO approaches.

Keywords Monocular Visual-inertial odometry · Transformer · Pose estimation

1 Introduction

Visual Odometry (VO), essential for visual Simultaneous Localization and Mapping (SLAM) [1], plays a critical role in Augmented Reality (AR), autonomous driving, and robotics. However, it struggles with scale determination and swift movements. Visual-Inertial Odometry (VIO) addresses these by combining Inertial Measurement Unit (IMU) data to estimate scale and reduce motion-related errors. Though advanced VIO methods like MSCKF [2] and VINS-Mono [3] are effective, their intricate calibration, initialization, and high computational demands hinder their adoption in

J. Zhao · Y. Zhu · Y. Huang · X. Qiao (✉)
Beijing University of Posts and Telecommunications, Beijing 100876, China
e-mail: qiaoxq@bupt.edu.cn

M. Guo · H. He · Y. Li
China Mobile Communications Research Institute, Beijing 100053, China

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_2

resource-limited devices. Moreover, while proficient in long-range pose estimation, they can err significantly over short distances.

Deep learning techniques are emerging in the visual realm. Models leveraging potent CNN architectures for feature extraction and RNNs for time-based processing are being devised for VO and VIO, often outperforming traditional methods. Key deep-learning VO strategies like DeepVO [4], FlowOdometry [5], and FlowNet [6] utilize CNNs for feature discernment between consecutive frames. Other notable techniques are VINet [7] and GANVO [8], introducing deep learning VIO and self-regulated learning via Generative Adversarial Networks (GAN), respectively. Expanding on GANVO, SelfVIO [9] integrates IMU data, delivering near-top-tier estimation outcomes. Furthermore, the Transformer's attention mechanism has enhanced robust feature extraction and improved sequence processing in visual-IMU fusion techniques with CNN and RNN. For example, Informer [10] demonstrates the Transformer's prowess in temporal forecasting by refining its attention process and introducing a generative decoder to enhance prediction efficiency.

Despite these successes, applying Transformer to VIO confronts the challenge of designing a heterogeneous modal input accommodating Transformer network structure, and solving large pose estimation errors in long-distance scenarios with deep learning methods. Another challenge arises from the high-frequency nature (typically over 100 Hz [11]) and the accumulation of errors in IMU data. Also, IMU acceleration, symbolizing displacement, allows algebraic calculations, IMU angular velocity, denoting rotation, requires computation with rotation matrices or quaternions, hence, conflicting with the algebraic procedures employed for IMU acceleration input. Thus, directly incorporating raw IMU angular velocity into the VIO network escalates the model training difficulty, as the model must learn features of angular velocity data and simulate matrix or quaternion operations.

To address these challenges, we propose VIOFormer, a novel Transformer-based architecture for enabling VIO estimation, yielding smaller errors across both short and long distances compared to benchmarks. VIOFormer comprises four key components: preprocessing of adjacent frame sequences, preprocessing of IMU data, an encoder, and a decoder. VIOFormer accommodates decoder input start and end tokens while learning different modal features of visual and IMU data for more precise pose estimation. To address the first challenge, we design a VIO network based on the Transformer to concurrently learn the temporal features of both visual and IMU inputs. This network amalgamates these features, extracting deep feature association information between the modalities, thereby addressing traditional deep learning VIO methods' shortcomings in learning inherent associations of different modal temporal features. To mitigate the heightened training difficulty induced by IMU input, a conversion method for the rotation magnitude into an input form suitable for algebraic computations has been designed before feeding it into the VIO model for training. This approach ensures that the input IMU can intuitively represent motion changes, supplying accurate features for short-distance prediction that are subsequently input into the decoder. The key contributions are outlined as follows:

- Introducing a Transformer-based VIO network that enhances long-distance displacement and angle estimation by learning temporal and intermodal features from visual and IMU inputs.
- Proposing a method converting rotation magnitude into an algebra-friendly form for training, paired with a multi-input encoder-decoder structure addressing IMU data-related training difficulties.
- Demonstrating advantages of the proposed VIOFormer, with significant reductions in Root Mean Squared Error (RMSE) performance.

2 VIOFormer Design

2.1 Overview of VIOFormer Architecture

Figure 1 presents the structure of a transformer-based VIO network, comprising four key components: preprocessing modules for adjacent frame sequences and IMU data, as well as an encoder and a decoder. The frame sequence processing module leverages optical flow to source input from adjacent frames for the VIO model, which outputs pose from multiple inputs. Specifically, we use the FlowNet backbone network [6] for processing the VIO model's input sequence of adjacent frames to extract the inter-frame motion relationship. Meanwhile, raw IMU data, comprising acceleration and angular velocity, is handled by the IMU input preprocessing module. To facilitate training, the module converts rotation magnitude, inherently unsuitable for algebraic operations, into a compatible form before its incorporation into the VIO model. The core of the VIO model comprises a transformer-based encoder and decoder. The encoder processes output from adjacent frame sequence preprocessing, while the decoder integrates output from both the encoder and the IMU data preprocessing to compute the final predicted pose. VIOFormer captures dynamic inter-frame relationships from the camera sequence and motion information shifts from IMU data. The output pose, representing accumulated motion changes, leads the model to estimate these changes. Given its ability to intuitively represent motion changes, IMU data aids in short-distance prediction. Conversely, image data provides stable features for long-distance pose estimation.

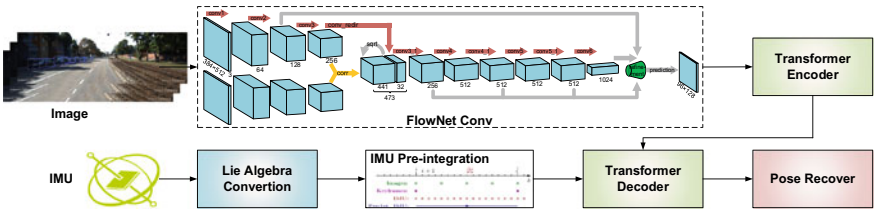


Fig. 1 Network architecture of VIOFormer

2.2 Methods of VIOFormer

Adjacent Frame Sequences Preprocessing. As shown in Fig. 1, we use FlowNet backbone network [6] to ingest images from the current and preceding time steps, employ an array of convolutional and pooling layers for feature extraction, and condense the images. Included in an intermediate layer is a feature-matching layer designed to correlate features of the current and preceding frames. This layer incorporates a pyramid structure to create feature maps at various scales. Ultimately, the extracted features are used as the input for the Encoder module of VIOFormer. The structural design of FlowNet provides an efficient model for the motion between successive frames.

Preprocessing of IMU Data Input. This component primarily aims at enabling the algebraic representation of rotational variables. Several prominent rigid body rotation representations include rotation matrices, axis-angle, Euler angles, and quaternions. (1) Using rotation matrices for learning requires the extraction of matrix operation features and the maintenance of orthogonal matrices for the pose's rotational matrix. This can lead to complexities during model training and convergence. (2) Representing rotations with quaternions imposes quaternion operation rules and unit quaternion constraints, potentially leading to model training and convergence difficulties. (3) Euler angle representation faces the gimbal lock problem [12], and the sequential nature of Euler angle rotations complicates deep learning model training. (4) The axis-angle representation applies minimal restrictions on rotations, merely requiring the product of a unit vector and an angle. As such, it stands out as the most suitable method for developing VIO models based on deep learning. More precisely, from the perspectives of Lie groups and Lie algebras, the rotation of a rigid body can be expressed using the Special Orthogonal group $SO(3)$ as follows:

$$SO(3) = \left\{ R \in \mathbb{R}^{3\times 3} | RR^T = I, det(R) = 1 \right\}. \tag{1}$$

To distinguish between the rotation matrix and the real number space, we denote the rotation matrix as R and the real number space as $\mathbb{R}$. To capture the local properties of the Lie group, the Lie algebra $\mathfrak{so}(3)$ is used, defined as follows:

$$\mathfrak{so}(3) = \phi \in \mathbb{R}^3, \Phi = \phi^{\wedge} \in \mathbb{R}^{3\times 3}, \tag{2}$$

where $\wedge$ represents an antisymmetric matrix. Importantly, employing Lie algebra facilitates handling the differentiation issue of Lie groups. Additionally, the exponential mapping process starting from Lie algebra is as follows:

$$R = cos\theta I + (1 - cos\theta)nn^T + sin\theta n^{\wedge}. \tag{3}$$

In conclusion, the axis-angle representation of rotation vectors is chosen as it puts less constraint on the deep learning model training and complies with the vector space depicted by Lie algebra in motion variation. As the VIOFormer must simultaneously

estimate the orientation and pose, it is vital to transform $\mathfrak{so}(3)$ into $\mathfrak{se}(3)$ to accurately represent the rigid body's translational and rotational movements.

$$\mathfrak{se}(3) = \left\{ \begin{array}{c} \xi = \begin{bmatrix} \rho \\ \phi \end{bmatrix} \in \mathbb{R}^6, \rho \in \mathbb{R}^3, \phi \in \mathfrak{so}(3) \\ \xi^{\wedge} = \begin{bmatrix} \phi^{\wedge} & \rho \\ 0^T & 0 \end{bmatrix} \in \mathbb{R}^{4\times 4} \end{array} \right\}. \tag{4}$$

Given this, the special Euclidean group $SE(3)$ can be represented as follows:

$$SE(3) = \left\{ \begin{array}{c} T = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4} \\ R \in SO(3),\ t \in \mathbb{R}^3 \end{array} \right\}. \tag{5}$$

Utilizing the following transformations:

$$\left\{ \begin{array}{c} \theta = arccos \frac{\mathrm{tr}(R)-1}{2} \\ Ra = a \\ \phi = \theta a \\ J = \frac{\sin\theta}{\theta} I + \left(1 - \frac{\sin\theta}{\theta}\right) aa^T + \frac{1-\cos\theta}{\theta} a^{\wedge} \\ t = J\rho \end{array} \right., \tag{6}$$

we can obtain the Lie algebra $\mathfrak{se}(3)$, effectively achieving the algebraic representation of motion.

Encoder and Decoder. As Fig. 1 illustrates, the encoder uses a Transformer architecture tailored for visual domains, including these components: (1) **Input Embedding.** The optical flow feature vectors are produced by the FlowNet network processing the input image data. (2) **Multi-Head Self-Attention.** The input embedding undergoes mapping to a k-dimensional space, where multi-head self-attention is computed to capture local relationships and global context within the image. (3) **Feed-Forward Network.** The feature vectors are processed by a two-layer feed-forward neural network following the multi-head self-attention, transforming them into a d-dimensional vector. (4) **Normalization Layer.** Post feed-forward network, feature vectors undergo normalization in a normalization layer, which could either be Batch or Layer Normalization. (5) **Residual Connection.** To avoid vanishing and exploding gradient issues, the output of the feed-forward network is combined with the input embedding, forming a residual connection.

The Decoder's design slightly deviates from a conventional Transformer. In the translation domain, < BOF > < EOF > typically marks sentence boundaries and are fed into the Decoder for prediction. Despite being encoded as special values for calculations, their inclusion does not yield the anticipated effect in this model. As depicted in Fig. 1, VIOFormer, borrowing from the Informer time series prediction

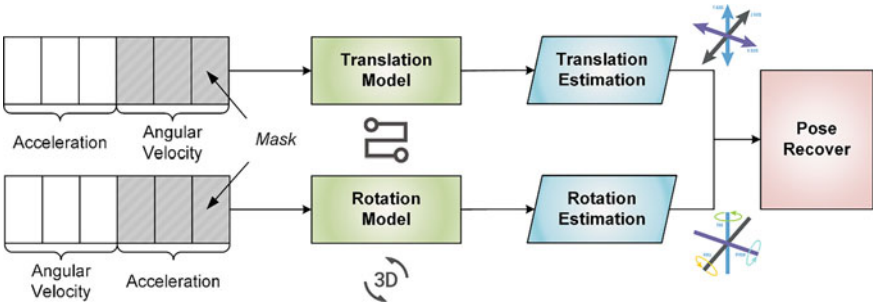


Fig. 2 Mask mechanism during the training process

model [10], appends an output length T_{output} marker sequence P at the start of the input sequence to indicate the output sequence's start position. This marker sequence can be any vector. Informer models often use an all-zero vector. The inclusion of marker sequences in the Decoder lets the model effectively use already predicted output sequences for subsequent predictions, enhancing time series prediction accuracy and stability. Concurrently, marker sequences help the model capture sequence time information and trends, enabling superior time series prediction.

Training and Estimation. VIOFormer's training process encounters a challenge in balancing optimal loss between displacement and angle estimation. To resolve this, a masking mechanism is introduced, separately masking displacement and angle using masks, as depicted in Fig. 2. Two models are trained: one for displacement estimation, and another for angle estimation. Specifically, when training the displacement estimation model, the angle information in the input data is masked, and vice versa for the angle estimation model. The masking-enhanced trained models deliver improved estimation results. During prediction, the displacement model estimates displacements, while the angle model estimates angles, with the results from both models merged in the end. Notably, the Decoder's output is pose change, not real-time pose results. Therefore, all pose changes need to be accumulated according to Eq. (8) to obtain the final pose.

3 Evaluation

3.1 Dataset and Preprocessing

The KITTI Odometry dataset [13] provides continuous image sequences and corresponding IMU data, allowing us to implement and evaluate VIO algorithms. This dataset enables quantitative evaluation of VIO algorithms, including accuracy of pose estimation, consistency of trajectory estimation, and runtime speed. Additionally, the dataset's ground-truth data supports the validation and comparison of algorithm

results. We utilized the sequence data with IDs from 0 to 10 from the KITTI Odometry task to train and evaluate the VIOFormer. Each sequence contains image and IMU data, as well as ground-truth poses. Due to sequence 3's absence, VIOFormer was trained on data from sequences 0 to 8 (excluding sequence 3) and tested on sequences 9 and 10, allowing a fair comparison with existing deep learning-based VO and VIO methods.

The KITTI dataset's image resolution is 1226 × 370, but the preprocessing FlowNet network requires a resolution of 384 × 512. Thus, we widened KITTI images to 384 and centered cropped them to achieve a resolution of 512 × 384 for VIOFormer's encoder input. Further, the KITTI dataset's IMU data length is inconsistent with the ground-truth length, and ground-truth poses lack timestamps. We addressed this by using IMU data's instantaneous motion characteristics and changes in ground-truth poses to align them. After processing IMU data and corresponding ground-truth poses at each KITTI dataset timestep, preintegration was performed based on the IMU data. We then expanded $\mathfrak{so}(3)$ to $\mathfrak{se}(3)$ using rotation's algebraic representation to depict rigid body displacement and rotation. Given that the KITTI dataset's ground-truth poses are provided as $T = R|t$ matrices, we calculated the pose change between adjacent frames

$$T_1 = \Delta T \cdot T_0. \tag{7}$$

Acknowledging the orthogonality of the rotation matrix R, the subsequent deduction ensues directly:

$$T_0^{-1} = \begin{bmatrix} R_T & -R^T t \\ 0^T & 1 \end{bmatrix}. \tag{8}$$

3.2 Performance on Estimation Accuracy

We compared VIOFormer with three prominent VIO methods: (1) **SelfVIO** [9]: it uses deep learning to estimate a camera and robot's relative pose and motion trajectory by integrating visual and inertial sensor data. Unlike traditional VIO techniques, SelfVIO obviates the need for precise sensor calibration or manually designed feature extraction procedures, allowing optimized feature extraction and data association to be learned autonomously via self-supervised learning. Currently, SelfVIO is considered the most accurate self-supervised VIO method. (2) **RNN-based VIO** [14]: it employs both supervised and unsupervised learning, using RNNs for monocular video VO and depth estimation. It combines deep learning and RNNs to learn camera motion and scene structure information from monocular videos. This algorithm can be trained with either supervised or self-supervised learning. For our comparison, we

Table 1 Performance on estimation accuracy

	Seq_09		Seq_10		Mean	
	t_{rel} (%)	$r_{rel}(°)$	t_{rel} (%)	$r_{rel}(°)$	t_{rel} (%)	$r_{rel}(°)$
SelfVIO	1.95	1.15	**1.81**	1.30	**1.88**	1.12
RNN-based VIO	9.30	3.5	7.21	3.9	8.25	3.7
VINS-Mono	41.47	2.41	20.35	2.73	30.91	2.57
VIOFormer	**1.844**	**0.533**	2.338	**0.742**	2.091	**0.638**

consider its supervised learning model. (3) **VINS-Mono** [3]: a non-deep learning-based algorithm used for VIO that employs a monocular camera and an IMU. This method demonstrates excellent accuracy and robustness in practical applications, marking it as a significant methodology in visual-inertial navigation.

SelfVIO, RNN-based VIO, and VIOFormer were trained on the same dataset and tested on the KITTI Odometry Seq_9 and Seq_10. Table 1 presents the comparison results of the test set. Here, t_{rel} (%) represents the percentage error in average displacement between 100–800 m, while $r_{rel}(°)$ denotes the rotation error (°/100m) within the same range. The key findings include: (1) VIOFormer exhibits a significant advantage in $r_{rel}(°)$ and matches the performance of the best competitor in t_{rel} (%). (2) On the Seq_10 dataset, SelfVIO outperforms others in t_{rel} (%), resulting in superior overall average accuracy. However, its performance on other test sequences is subpar. (3) Compared to RNN-based VIO, VIOFormer shows lower t_{rel} (%) and $r_{rel}(°)$ errors, suggesting that Transformer-based VIO yields superior results under supervised learning compared to RNN-based methods. (4) Compared with VINS-Mono, VIOFormer shows lower errors. Specifically, it decreases t_{rel} (%) and $r_{rel}(°)$ errors by 95.55% and 85.96%, respectively, on Seq_09, and by 77.97% and 72.82%, respectively, on Seq_10 compared to VINS-Mono. The average error performance in t_{rel} (%) and $r_{rel}(°)$ is reduced by 93.24% and 75.20% across both test sequences.

3.3 Trajectory Evaluation

We compared the performance of VIOFormer on trajectory plots with advanced algorithms, as shown in Fig. 3. The compared algorithms include: (1) **SfMLearner** [15], which learns camera motion and scene depth from monocular video sequences without traditional handcrafted feature extraction and matching. This was the first unsupervised VIO model. Furthermore, SC-SfMLearner [16] represents an improvement over SfMLearner, ensuring scale consistency. (2) **DF-VO** [17] combines the robustness of traditional VO with the benefits of deep learning feature extraction, providing an accurate and robust solution for real-time visual localization. (3) **ORB-SLAM2** [18], a well-known SL AM algorithm based on ORB feature points. For the comparison in the experiment, we used the version of this algorithm that includes loop detection. The ground truth trajectory is denoted as GT.

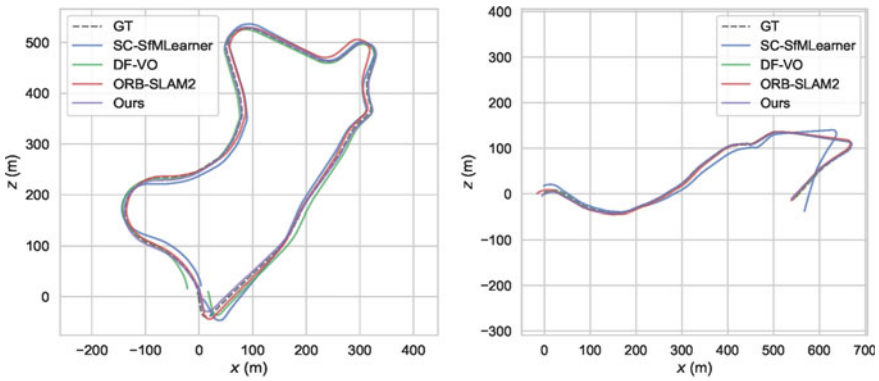


Fig. 3 Trajectory comparison

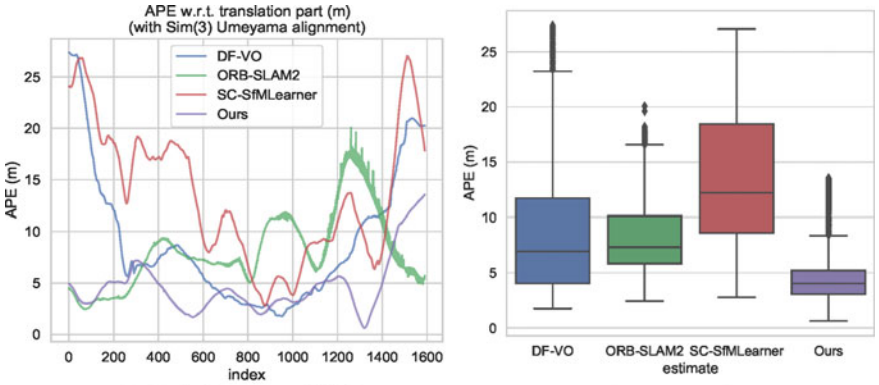
The results indicate that: (1) On the Seq_09 dataset spanning 1705.05m, VIOFormer and ORB-SLAM2 exhibit similar performances, while other methods show larger discrepancies at the start and end of the trajectory. Since the trajectory of sequence 9 forms a closed loop, and ORB-SLAM2 is the only method among these with loop detection functionality, it aligns best with the GT trajectory on this test dataset. (2) On the Seq_10 dataset spanning 919.52m, SC-SfMLearner demonstrates a substantial deviation from the GT trajectory at the beginning and end of the trajectory. In contrast, the VIOFormer-based method shows better alignment with the GT trajectory, confirming the robust generalization performance of VIOFormer across different scenes and datasets.

3.4 Analysis of APE and RPE Results

APE Distribution. Figures 4 and 5 illustrate the Absolute Pose Error (APE) performance of various pose estimation methods on Seq_09 and Seq_10, respectively. Figure 4 reveals that VIOFormer outperforms other deep-learning baselines with the lowest APE values. Figure 4a highlights VIOFormer and ORB-SLAM2's superior pose estimation, especially within the 0–200 and 1200–1600 frame intervals. Notably, past the 1400th frame, ORB-SLAM2's APE value drops rapidly due to its unique loop closure detection. The study's focus is on leveraging transformer structures in VIO models to enhance pose estimation. Future endeavors will aim to enhance accuracy by introducing loop closure detection into deep-learning VIO techniques. Figure 4b confirms VIOFormer's advantage, underscoring the transformative impact of the Transformer on VIO pose estimation accuracy. The results on Seq_10 depicted in Fig. 5 demonstrate outcomes closely aligned with those from Seq_09.

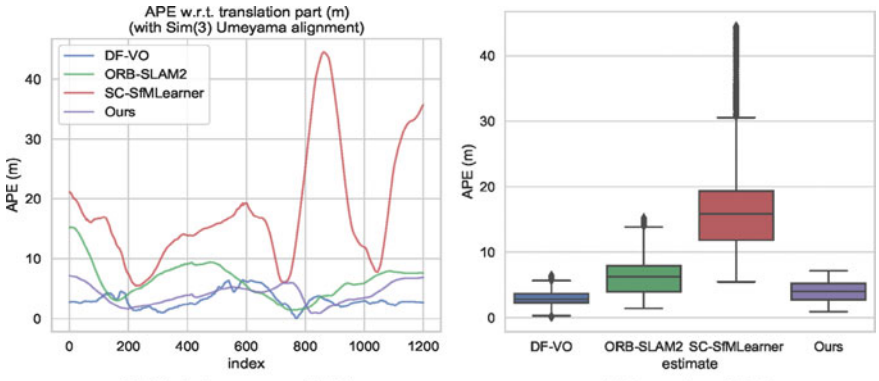
It's noteworthy that within the frame range of 800–1000 as shown in Fig. 5a, SC-SfMLearner exhibits an error exceeding 40 m, while other methodologies, particularly our proposed approach, manifest optimal performance. Observing the outcomes from Fig. 5b, SC-SfMLearner displays the most suboptimal performance across the entire sequence, while our methodology showcases superior efficacy, thereby validating the effectiveness of the Transformer-based VIO fusion scheme proposed in this paper.

Tables 2, 3 comprehensively analyze APE values from various methods on the test set. Evaluating Table 2, VIOFormer consistently outperforms, achieving a 40.3% improvement in RMSE (5.331) over ORB-SLAM2's 8.922. It also records reductions of 43.1% and 44.7% in Mean and Median metrics, respectively, compared to ORB-SLAM2, demonstrating overall stability. For Seq_10, while ORB-SLAM2



(a) Variation curve of APE (b) Box plot of APE

Fig. 4 APE on Seq_09



(a) Variation curve of APE (b) Box plot of APE

Fig. 5 APE on Seq_10

marginally outpaces VIOFormer in certain metrics, the differences are negligible, and VIOFormer remains consistent in error bounds. In essence, VIOFormer demonstrates robust performance across Seq_09 and remains competitive in Seq_10, affirming its stability and precision.

RPE Distribution. Figure 6 explores the Relative Pose Error (RPE) distribution of VIOFormer and other baselines. As RPE quantifies the error in pose increment between estimated and ground-truth values, it mirrors the precision of the estimation at a specific moment. As both VIOFormer and the baselines display minimal RPE values, the boxplot depiction fails to be instructive. Hence, a bar chart is utilized in Fig. 6 for better comparison of the APE of different methods.

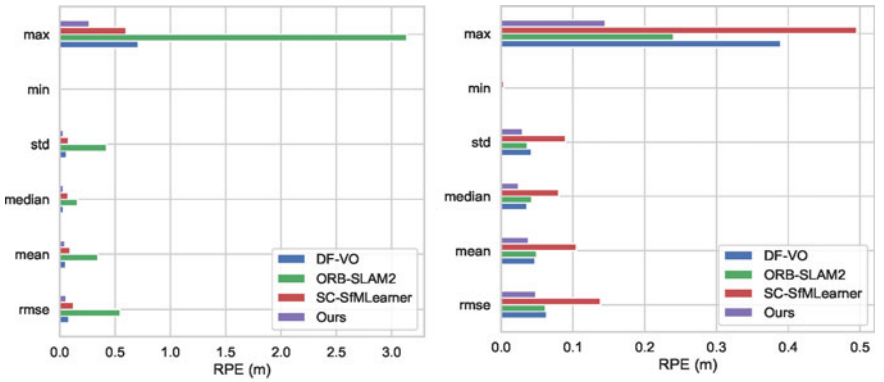
The results in Tables 4 and 5 show that: (1) On Seq_09, our method outperforms in RMSE, mean, and median. VIOFormer acquires an RMSE of 0.059, significantly below DF-VO (0.083), ORB-SLAM2 (0.546), and SC-SfMLearner (0.125). Additionally, VIOFormer shows remarkable performance in mean and median errors, recording 0.047 and 0.032, respectively, significantly lower than the other three methods. (2) On Seq_10, our method again excels in RMSE, mean, and median. VIOFormer achieves an RMSE of 0.048, undercutting DF-VO (0.064), ORB-SLAM2 (0.062), and SC-SfMLearner (0.138). Regarding mean and median errors, VIOFormer outstrips other methods with values of 0.038 and 0.024, respectively. (3) Our method manifests lower standard deviation values, indicative of a more stable error distribution. On Seq_09, VIOFormer's standard deviation is 0.035, undercutting DF-VO (0.062), ORB-SLAM2 (0.424), and SC-SfMLearner (0.080). On Seq_10, VIOFormer's standard deviation is 0.030, less than DF-VO (0.042), ORB-SLAM2 (0.037), and SC-SfMLearner (0.090). (4) Our method also exhibits the smallest maximum error on both datasets, implying that VIOFormer can maintain a low error level even under extreme conditions.

Table 2 APE distribution on Seq_09

	RMSE	Mean	Median	Std.	Min	Max
DF-VO	11.023	8.995	6.936	6.371	1.760	27.381
ORB-SLAM2	8.922	8.131	7.304	3.673	2.438	20.061
SC-SfMLearner	15.019	13.520	12.216	6.540	2.786	27.056
Ours	5.331	4.628	4.037	2.644	0.625	13.547

Table 3 APE distribution on Seq_10

	RMSE	Mean	Median	Std.	Min	Max
DF-VO	3.372	3.082	2.823	1.369	0.078	6.463
ORB-SLAM2	6.969	6.317	6.236	2.942	1.409	15.243
SC-SfMLearner	20.192	17.778	15.872	9.573	5.476	44.506
Ours	4.391	4.074	4.016	1.637	0.918	7.153



(a) RPE on Seq_09 (b) RPE on Seq_10

Fig. 6 RPE performance

Table 4 RPE distribution on Seq_09

	RMSE	Mean	Median	Std.	Min	Max
DF-VO	0.083	0.055	0.034	0.062	0.002	0.711
ORB-SLAM2	0.546	0.343	0.162	0.424	0.005	3.138
SC-SfMLearner	0.125	0.095	0.075	0.080	0.003	0.601
Ours	0.059	0.047	0.032	0.035	0.002	0.266

Table 5 RPE distribution on Seq_10

	RMSE	Mean	Median	Std.	Min	Max
DF-VO	0.064	0.047	0.036	0.042	0.002	0.389
ORB-SLAM2	0.062	0.050	0.043	0.037	0.001	0.240
SC-SfMLearner	0.138	0.105	0.080	0.090	0.004	0.496
Ours	0.048	0.038	0.024	0.030	0.001	0.145

4 Discussion and Conclusion

This paper presents VIOFormer, a Transformer-based architecture aimed at enhancing VIO estimation. By harnessing the temporal processing prowess of Transformers, VIOFormer addresses the scale determination and motion-related errors inherent in swift movements, typically seen in conventional VIO methods. The architecture, including a multi-input encoder-decoder structure and an innovative method for rotation magnitude conversion, demonstrates significant reductions in RMSE performance compared to benchmark methods. Its design facilitates precise pose estimation by concurrently learning temporal features from visual and IMU inputs.

Despite its merits, VIOFormer has limitations. The real-world validation of the architecture remains to be rigorously tested outside controlled experimental setups. Also, the computational demands of the Transformer-based architecture could pose challenges for real-time applications on resource-constrained platforms. Future work should focus on extensive real-world testing, further algorithm optimization to reduce computational demands, and exploring alternative deep learning architectures for potentially better performance. Through continuous refinement, VIOFormer could play a pivotal role in advancing VIO estimation, impacting real-world applications like autonomous driving, robotics, and augmented reality.

Acknowledgements This research was funded in part by the National Key R&D Program of China under Grant 2022YFF0904304.

References

1. Fuentes-Pacheco, J., Ruiz-Ascencio, J., Rendón-Mancha, J.M.: Visual simultaneous localization and mapping: a survey. Artif. Intell. Rev. **43**, 55–81 (2015)
2. Anastasios, I, Mourikis., Stergios I Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In Proceedings 2007 IEEE International Co on robotics and automation, pages 3565–3572. IEEE (2007)
3. Qin, T., Li, P., Shen, S.: Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Trans. Rob. **34**(4), 1004–1020 (2018)
4. Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni.: Deepvo: To- wards end-to-end visual odometry with deep recurrent convolutional neural networks. In 2017 IEEE International Conference on Robotics And Automation (ICRA), pages 2043–2050. IEEE (2017)
5. Peter Muller, Andreas Savakis.: Flowdometry: An optical flow and deep learning-based approach to visual odometry. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 624–631. IEEE (2017)
6. Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox.: Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 2758–2766 (2015)
7. Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, Niki Trigoni.: Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 31 (2017)
8. Yasin Almalioglu, Muhamad Risqi U Saputra, Pedro PB De Gusmao, Andrew Markham, Niki Trigoni. Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks. In 2019 International conference on robotics and automation (ICRA), pages 5474–5480. IEEE (2019)
9. Almalioglu, Y., Turan, M., Muhamad Risqi, U., Saputra, P.P.B., de Gusmão, A., Markham, and Niki Trigoni,: Selfvio: Self-supervised deep monocular visual–inertial odometry and depth estimation. Neural Netw. **150**, 119–136 (2022)
10. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In AAAI Conference on Artificial Intelligence **35**, 11106–11115 (2021)
11. Zichao Zhang, Davide Scaramuzza.: A tutorial on quantitative trajectory evaluation for visual-inertial odometry. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 7244–7251. IEEE (2018)

12. Ameesh Makadia, Kostas Daniilidis.: Direct 3d-rotation estimation from spherical images via a generalized shift theorem. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., volume 2, pages II–217. IEEE (2003)
13. Geiger, A., Lenz, P., et al.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
14. Rui Wang, Stephen M Pizer, Jan-Michael Frahm. Recurrent neural network for unsupervised learning of monocular video visual odometry and depth. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5555–5564 (2019)
15. Tinghui Zhou, Matthew Brown, Noah Snavely, David G Lowe.: Unsupervised learning of depth and ego-motion from video. In Proceed- ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1851–1858 (2017)
16. Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. Advances in Neural Information Processing Systems, 32 (2019)
17. Huangying Zhan, Chamara Saroj Weerasekera, Jia-Wang Bian, Ravi Garg, and Ian Reid. Df-vo: What should be learnt for visual odometry? arXiv preprint arXiv:2103.00933. (2021)
18. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Trans. Rob. **33**(5), 1255–1262 (2017)

Incorporating Feature Signal Transmission with Block-Based Haptic Data Reduction for Time-Delayed Teleoperation

Hongjun Wu, Xiao Xu, Zhi Jin, and Fanle Meng

Abstract This paper presents an innovative feature signal transmission approach incorporating block-based haptic data reduction to address time-delayed teleoperation. Numerous data reduction techniques rely on perceptual deadband (DB). In the preceding block-based approaches, the whole block within the DB is discarded. However, disregarding all signals within the DB loses too much information and hinders effective haptic signal tracking, as these signals contain valuable information for signal reconstruction. Consequently, we propose a feature signal transmission approach based on the block algorithm that aggregates samples as a unit, enabling high-quality haptic data reduction. In our proposed approach, we employ max-pooling to extract feature signals from the signals within the DB. These feature signals are then transmitted by adjusting the content of the transmission block. This methodology enables the transmission of more useful information without introducing additional delay, aside from the inherent algorithmic delay. Experimental results demonstrate the superiority of our approach over other state-of-the-art (SOTA) methods on various assessment measures under distinct channel delays.

Keywords Feature signal transmission · Haptic data reduction · Block algorithm

H. Wu · Z. Jin (✉)
School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-Sen University, Shenzhen 518107, Guangdong, China
e-mail: jinzh26@mail.sysu.edu.cn

X. Xu
Media Technology and Munich Institute of Robotics and Machine Intelligence, Technical University of Munich, 80333 Munich, Germany

F. Meng
Information Science Academy, China Electronics Technology Group Corporation, Beijing, China

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_3

1 Introduction

Haptic teleoperation systems have gained significant traction in various domains, including remote surgery [1], underwater exploration [2], and extraterrestrial missions [3]. Figure 1 illustrates a typical example of a time-delayed haptic teleoperation system with the block-based haptic data reduction scheme. In this system, the primary device transmits motion signals to control the secondary device, while the haptic signals (e.g., force and velocity samples) collected by the secondary device are sent back to the primary device as feedback. To ensure system stability in the presence of time delays, a stability control module, such as time-domain passivity approaches, is employed in both devices [4].

Efficient haptic signal communication is crucial for a haptic teleoperation system. Typically, the haptic signal sampling rate is set to a high frequency, such as 1 kHz, to ensure system stability. However, this high sampling rate causes issues in terms of heavy transmission burden and packet header redundancy. The conventional method widely employed to mitigate these issues is based on perceptual deadband (DB) [5, 6]. In this approach, only signals falling outside the DB range are transmitted, and a zero-order-hold (ZOH) method is used at the receiver for signal reconstruction. The sample-based approach significantly reduces the packet rate during transmission. However, it requires precise adjustment of the deadband parameter (DBP) to avoid a severe decline in the quality of reconstructed signals. Recently, more methods have been proposed for haptic data reduction. Xu et al. [7] proposed a time-based update trigger scheme to constrain communication interruption within the predefined threshold. Ming et al. [8] proposed to control the packet rate by developing a peak-suppressing adaptive DB based on the transmission history. Wang et al. [9] proposed a novel quality assessment approach for predicting subjective experience to guide the development of new haptic data reduction methods. Later, a novel block-based approach for haptic data reduction has been introduced [10]. This approach addresses the challenges of haptic signal transmission by aggregating samples into blocks and transmitting them as single units, respectively. Furthermore, the block-based approach incorporates the perceptual DB scheme to achieve even better data reduction. However, the issue still exists in the approaches incorporating the DB scheme, where a low packet rate can lead to a decrease in signal transmission quality, and it is hard to achieve a good balance between signal quality and packet rate by adjusting the DBP accurately.

Fig. 1 Overview of a typical time-delayed haptic teleoperation system with the block-based haptic data reduction scheme

To overcome this issue, it is crucial to extract feature signals from the signals within the DB, as they contain valuable information for signal restoration. In this paper, we propose a feature signal transmission approach integrated with haptic data reduction, building upon the block-based approach. Our approach consists of two key components: feature signal extraction and signal adjustment within the block. Specifically, max-pooling is employed to extract feature signals. Additionally, as some signals within the DB are determined to be transmitted because of the block algorithm, we can leverage the feature signals to replace these signals within the blocks to optimize the transmission content. This approach enhances the quality of reconstructed signals while maintaining a low packet rate.

To validate the effectiveness of our proposed approach, we conduct experiments to evaluate the haptic quality. The results demonstrate that our approach surpasses other state-of-the-art (SOTA) methods across various assessment measures. In general, the contributions of this work can be summarized as follows:

- We propose a novel scheme for DB feature signal extraction, which enables the collection of important signals in one coding block within the DB. This extraction process enhances the quality of reconstructed signals while only introducing a limited number of additional transmission packets.
- We leverage the signals within the DB to adjust the signals within the block that require transmission. This adjustment improves the restoration of the signal, contributing to stable teleoperation.
- Through comprehensive experiments, we compare our proposed approach with several SOTA methods. The results demonstrate that our approach outperforms these approaches in terms of various haptic quality assessment measures.

2 Technical Background

2.1 Sample-Based Haptic Packet Rate Reduction

The sample-based approach is commonly employed to reduce the packet rate in haptic teleoperation systems. In this approach, each sample is evaluated according to the perceptual DB range to determine whether it needs to be transmitted. This decision is guided by Weber's Law [11], which posits that there exists a DB threshold that exhibits a linear correlation with the physical stimulus. According to this law, changes in stimuli within the DB threshold are not perceptible to humans. Specifically, the size of the DB range can be calculated as follows:

$$dI = K * I \tag{1}$$

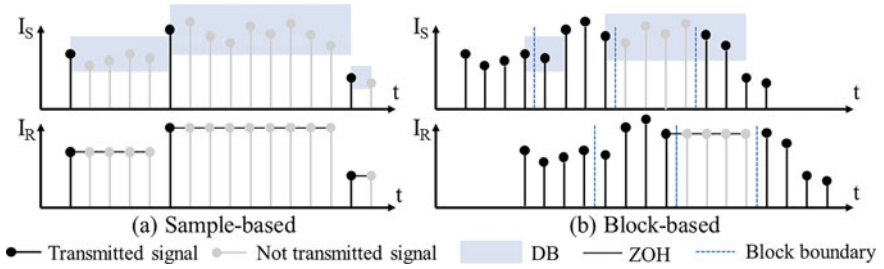


Fig. 2 Comparison between the sample-based and block-based approaches. I_S stands for the original signal collected at the sender, and I_R represents the reconstructed signal at the receiver

where K represents DBP, I is the physical stimulus, and the threshold of DB is represented as $I \pm dI$. For the sample-based haptic packet rate reduction approach, only the signals that fall outside the DB are transmitted. The process of a typical sample-based approach is demonstrated in detail in Fig. 2a. When a signal is not transmitted at a specific timestamp, the receiver reconstructs the corresponding signal using the ZOH method. This approach effectively reduces the burden of signal transmission while introducing a certain degree of signal loss.

However, the sample-based approach has notable drawbacks. Firstly, the packet rate is not stable because it is unpredictable if the next signal needs to be transmitted. This introduces variability in the packet rate, which can impact the overall system performance. Secondly, it is noticeable that a larger DBP can effectively reduce the packet rate. However, a larger DBP also leads to a decrease in signal transmission fidelity and results in excessive information loss. Achieving a balance between packet rate reduction and signal quality is crucial in the sample-based approach.

2.2 Block-Based Haptic Packet Rate Reduction

To address the objectives of reducing the packet rate, enhancing signal quality at the receiver, and maintaining a stable transmission packet rate, the block-based haptic packet rate reduction approach [10] is proposed. In this approach, a block is formed by aggregating n continuous samples of signals, where n is a predefined hyperparameter, and the block is considered the smallest unit for transmission. For a given block length of n, the packet rate decreases by a factor of $1/n$, and the rate is stable and accurately controllable. However, this introduces an additional delay of $n-1$ ms due to the need to form a complete block. Therefore, it is key to balance the packet rate reduction and the quality deterioration. In scenarios with large communication delay, the block coding delay can be ignored, while in a system with a relatively small communication delay, a smaller block length is suggested.

To incorporate the benefits of the sample-based approach, the block-based approach utilizes DB during transmission. Let I denote the magnitude of the signal,

K represents the DBP, i indicates the start position index of the current block, and n denotes the length of the block. If all the signals within the current block satisfy the following condition, the block is considered within the DB and is not transmitted:

$$I_k - I_{i-1} \leq K * I_{i-1} \quad \forall k \in [i, i+n) \tag{2}$$

The implementation details of the block-based approach are depicted in Fig. 2b. While incorporating the DB can further reduce the payload of packet transmission, the setting of the DBP may not be optimal for all situations. Additionally, discarding all samples within the DB can result in a decrease in signal fidelity. Building upon the block-based approach, this work proposes a feature signal transmission approach. This approach leverages representative signals within the DB to assist the receiver in achieving better signal reconstruction.

3 Method

In this work, we introduce a feature signal transmission approach that encompasses two primary schemes: feature signal extraction within the DB and signal adjustment within the block. This section provides a detailed explanation of our proposed approach.

3.1 Feature Signal Extraction Within the DB

The feature signal extraction scheme of our proposed approach is depicted in Fig. 3. Max-pooling is adopted to extract feature signals from the block. With a large step size (st) of max-pooling, more loss of signal information occurs, rendering the feature signals less effective in tracking signal changes. Through comparative analysis, st is set as 2 for better signal transmission quality. When a block at the sender is determined not to be transmitted, all signals within that block are temporarily stored. As the stored signals accumulate, they are compressed into a block using max-pooling once the number of stored signals reaches $st * n$ (where n represents the block length). This block formed by the feature signals is then transmitted to the receiver.

For instance, as exemplified in Fig. 3, the blocks colored green and purple are within the DB. The signals within the green block are stored temporarily after formation, and ZOH is used to generate the second block at the receiver. Eventually, when the stored signals can be compressed into a block through max-pooling (e.g., the number of stored signals equals $2 * 4$), the extracted feature signals are aggregated and transmitted to the receiver. Considering that the timeliness of signal transmission is crucial if a block out of the DB is formed at the sender, it is transmitted immediately, and the stored signals are relinquished due to expiration.

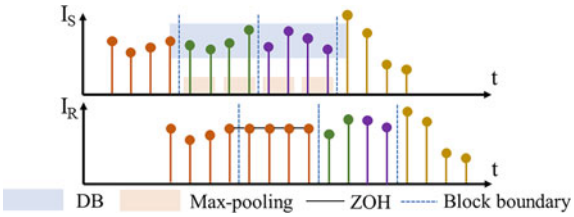


Fig. 3 A signal extraction scheme application instance. I_S stands for the original signal collected at the sender, and I_R represents the reconstructed signal at the receiver. The signals in different blocks with a length of 4 are marked with different colors

Regarding the extraction of feature signals in Fig. 3 at the sender, the first and the second signals undergo max-pooling, and the first signal is selected as the first feature signal, given its larger value. Subsequently, for the third and fourth signals at the sender, the signal with a larger value (i.e. the fourth signal) is picked as the second feature signal. A similar operation is conducted on the purple block to extract the third and fourth feature signals. In this manner, four feature signals are extracted and unified into a block ready for transmission, constituting the third block at the receiver.

The proposed signal extraction scheme offers several advantages, making it a robust and efficient method for improving the quality of transmitted signals. When determining the transmission signal, our scheme takes into account not only the immediate preceding signal but also considers the former or future signal within the block. This consideration allows for a more comprehensive analysis of signal patterns and ensures that important signals are not mistakenly discarded. Unlike the sample-based approach, which relies solely on the preceding signal for decision-making and requires careful selection of the DBP to avoid discarding too many useful signals, our scheme exhibits improved robustness and reliability.

Moreover, the proposed scheme demonstrates a clear advantage when the changes between signals at the sender are not readily apparent, resulting in a large number of signals falling within the DB. In such scenarios, the sample-based approach's application of the ZOH method over a wide range may lead to excessive smoothing signals. Conversely, the signal extraction scheme in our approach excels at restoring signals in these situations, resulting in enhanced signal reconstruction performance.

Overall, the proposed signal extraction scheme addresses the limitations of the previous approaches and improves the signal quality effectively.

3.2 Signal Adjustment Within the Block

In order not to increase the delay caused by the block algorithm, it is imperative to transmit the non-DB block as soon as sufficient samples have been accumulated, even if some signals have been stored. It should be acknowledged that the stored

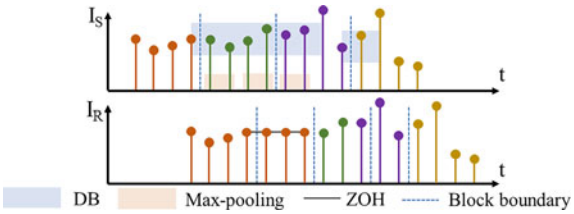


Fig. 4 A signal adjustment scheme application instance. I_S stands for the original signal collected at the sender, and I_R represents the reconstructed signal at the receiver. The signals in different blocks with a length of 4 are marked with different colors

signal plays a constructive role in signal reconstruction at the receiver. Furthermore, within the block earmarked for transmission, there exist signals within the DB that can be amalgamated with the stored signals and compressed into feature signals. Hence, apart from appending additional signals for packet transmission, we propose the signal adjustment scheme to refine the signals in the block slated for transmission.

Figure 4 provides a visual demonstration of the signal adjustment scheme. In this scheme, each sample within the block is evaluated whether it falls within the DB when collected. If a sample is outside the DB, the action taken depends on the number of stored signals in our scheme. If the number of stored signals is equal to or exceeds the block length n, all stored signals are compressed using max-pooling and transmitted as a packet to the receiver. Then the remaining samples within the block are aggregated and transmitted as another separate packet. For example, in Fig. 4, when the third sample of the purple block is collected in the sender, the green and previous purple signals are compressed and transmitted as one packet, while the remaining samples in the block are transmitted as another packet. On the other hand, if the number of stored signals is less than n, all the samples within the block are transmitted together as in the traditional block-based approach (e.g., the yellow block in Fig. 4).

The advantage of the signal adjustment scheme lies in its flexibility, enabling more efficient feature signal extraction. By adjusting the content of the block with feature signals, more significant information is transmitted to the receiver while no extra delay is introduced. This ensures the stability of the teleoperation system.

4 Experiments

To demonstrate the superiority of the proposed feature signal transmission approach, comprehensive comparative experiments are conducted. In the following experiments, we simulate a teleoperation system using two Geomagic Touch devices in Matlab, with system parameters set according to the references [12–14]. To ensure

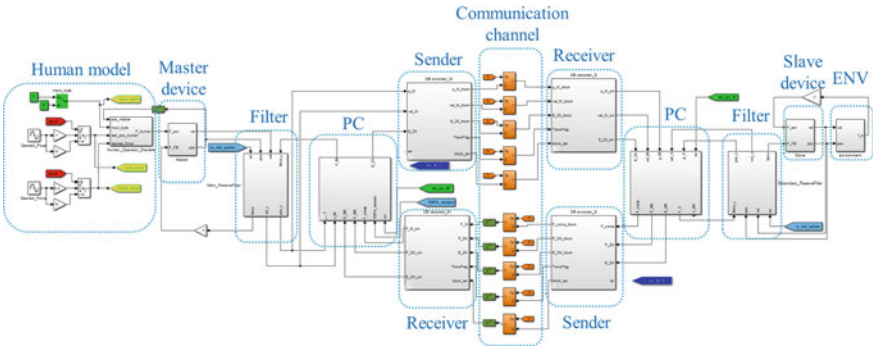


Fig. 5 Overview of the simulation teleoperation system in the experiment. PC represents the passivity controller. ENV represents the environment model. Different haptic data reduction approaches are employed in the senders and receivers

reproducibility, we apply a segment of a sinusoidal signal as the input to the teleoperation system. The overview of the simulation teleoperation system is presented in Fig. 5.

The force signals from the sender of the slave device and receiver of the master device are utilized to measure the transmission quality and efficiency of various methods. To assess the haptic quality, we employ the following objective measures: Haptics Structural Similarity (HSSIM) [15], Haptic Perceptually Weighted PSNR (HPW-PSNR) [16], Mean Squared Error (MSE), and Perceptual MSE (P-MSE) [17]. Higher values of HSSIM and HPW-PSNR indicate better results, while lower values of MSE and P-MSE indicate better performance. To verify the effectiveness of our approach in different scenes, the experiments are conducted in channel delays of 10, 50, and 100 ms, respectively.

We compare the following approaches for haptic quality evaluation. The first two approaches represent the traditional sample-based approach with DB. The third one is the conventional block-based approach, and the fourth one is our proposed approach.

- The DB sample-based approach [6].
- The DB sample-based approach with regular flag transmission with the interval of 40 ms [7].
- Block-based approach with fused DB [10]. The DBP is set as 0.02.
- Block-based approach with feature signal transmission. The DBP is set as 0.02.

The results of the quantitative experiments are presented in Fig. 6. It is evident that the block-based approach with feature signal transmission outperforms other methods across various packet rates. In the case of a 10 ms delay, our approach demonstrates superior performance in terms of HSSIM, HPW-PSNR, and MSE. However, the improvement in P-MSE is not as obvious. As the delay increases, the improvement becomes much more significant across the measures including HSSIM,

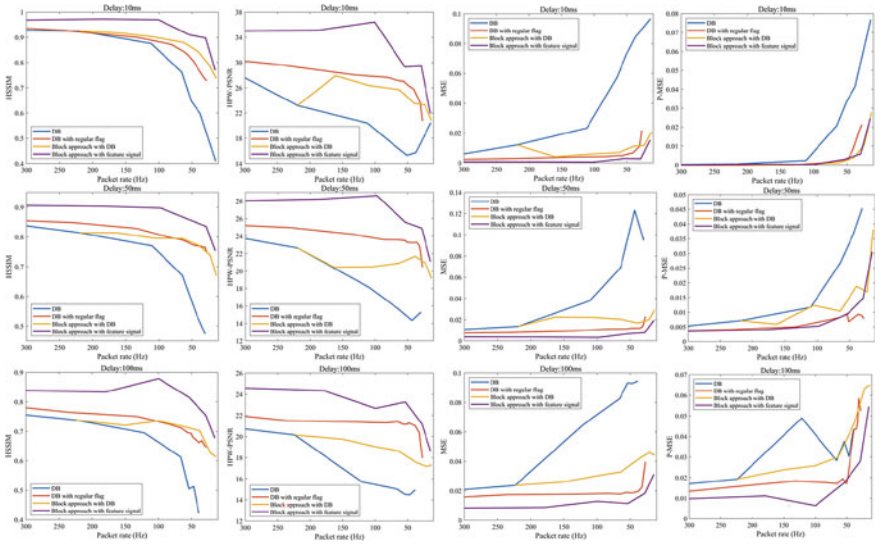


Fig. 6 Transmission quality evaluation between different haptic data reduction approaches

MSE, and P-MSE. This observation is expected since the block-based approach introduces an additional delay due to block forming. The impact of this additional delay is more noticeable in systems with lower communication delays. Therefore, our proposed approach exhibits substantial improvements over other methods, particularly in systems with higher latency. Furthermore, it is noticeable that as the packet rate declines, since our approach extracts feature signals and retains important information about signal changes, the signal quality of our approach maintains a good and stable level in a wide range of packet rates, especially in terms of HSSIM.

Upon the application of our proposed feature signal transmission approach, the workload of the signal transmission network is reduced, and the risk of packet delay and losses is obviously decreased, in this way the teleoperated robot system stability is guaranteed. Furthermore, as shown by various measures mentioned above, the data transmission fidelity is improved significantly, which leads to the more precise control of teleoperated robot motion, and the user experience is increased.

5 Conclusion

In this paper, we present a feature signal transmission approach that incorporates block-based haptic data reduction for time-delayed teleoperation. We recognize that the signals within the DB still contain valuable information for signal reconstruction. To improve transmission quality while maintaining a low packet rate, we extract feature signals from the DB using max-pooling. These feature signals are then utilized

to adjust the content within the block that needs to be transmitted. Experimental results demonstrate that our approach outperforms other SOTA methods in terms of various quality assessment measures and exhibits greater robustness of declining packet rates in scenarios with different communication delays. Moving forward, our future work will focus on exploring more effective methods for extracting feature signals and developing a flexible scheme to adjust the content of the block based on the extracted feature signals and block length.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant No. 62071500, Supported by Shenzhen Fundamental Research Program (Grant No. JCYJ20230807111107015), and the Sino-Germen Mobility Programme M-0421.

References

1. El Saddik, A.: The potential of haptics technologies. IEEE Instrum. Meas. Mag. **10**, 10–17 (2007)
2. Sheridan, T.B.: Space teleoperation through time delay: review and prognosis. IEEE Trans. Robot. **9**, 592–606 (1993)
3. Artigas, J., et al. Kontur-2: force-feedback teleoperation from the international space station. IEEE ICRA, 1166–1173 (2016).
4. Ryu, J.H., et al.: Time domain passivity control with reference energy following. IEEE Trans. Control Syst. Technol. **13**, 737–742 (2005)
5. Steinbach, E., et al.: Haptic codecs for the tactile internet. Proc. IEEE **107**, 447–470 (2018)
6. Hinterseer, P., et al.: Perception-based data reduction and transmission of haptic data in telepresence and teleaction systems. IEEE Trans. Signal Process. **56**, 588–597 (2008)
7. Xu, X., et al.: Integrating haptic data reduction with energy reflection-based passivity control for time-delayed teleoperation. In; IEEE Haptics Symposium, pp. 109–114 (2020)
8. Gui, M., et al.: Adaptive packet rate control for the mitigation of bursty haptic traffic in teleoperation systems. In: IEEE Haptics Symposium, pp. 134–139 (2020)
9. Wang, Z., et al.: Towards subjective experience prediction for time-delayed teleoperation with haptic data reduction. In: IEEE RO-MAN, pp. 129–134 (2022)
10. Gui, M., et al.: Block-based novel haptic data reduction for time-delayed teleoperation. IEEE IROS, 6183–6188 (2022)
11. Fechner, G.T.: Elemente der psychophysik, 1–2 (1860)
12. Ajoudani, A., et al.: Reduced-complexity representation of the human arm active endpoint stiffness for supervisory control of remote manipulation. IJRR **37**, 155–167 (2018)
13. Diolaiti, N., et al.: Stability of haptic rendering: Discretization, quantization, time delay, and coulomb effects. IEEE Trans. Robot. **22**, 256–268 (2006)
14. Ryu, J.H., et al.: A passive bilateral control scheme for a teleoperator with time-varying communication delay. Mechatronics **20**, 812–823 (2010)
15. Hassen, R., et al.: HSSIM: an objective haptic quality assessment measure for force-feedback signals. QoMEX, 1–6 (2018)
16. Sakr, N., et al.: A perceptual quality metric for haptic signals. In: IEEE HAVE, pp. 27–32 (2007)
17. Chaudhari, R., et al.: Towards an objective quality evaluation framework for haptic data reduction. In: IEEE WHC, pp. 539–544 (2011)

Sailboat Simulation System Based on Natural Interaction and Eye-Tracking

Xuqi Pan, Weicheng Hu, Xinkai Lv, Cui Xie, Junyu Dong, and Xiaofeng Chang

Abstract In an effort to promote and popularize sailing, we have developed a natural interaction-based sailing simulation system that realistically simulates the experience of maneuvering a J/80 sailboat. This system encompasses a multi-sail four degrees of freedom (4DOF) maneuvering motion model to represent the actual movement of the sailboat and a simulated wind and wave navigation environment. Users can steer the sailboat by controlling the rudder, raising and lowering sails under varying wind and sea conditions. Our system introduces an innovative natural interaction method based on data gloves, which includes static gestures interaction, virtual hand grasping, pulling, and other actions for steering the virtual sailboat. This significantly enhances user immersion and participation, making the handling of sailboat more natural. Furthermore, the system integrates eye-tracking technology to provide prompts based on the user's gaze point, assisting them in better steering the sailboat and effectively reducing the learning curve. User study indicates that these interaction methods offer a deeper sense of immersion with a more natural and efficient interaction compared to VR controller. Therefore, our system provides users with a good sailboat handling experience, allowing them to experience the charm of sailing sports even without a real sailboat and venue.

Keywords Virtual reality · Sailboat simulation · Gesture-based interaction · Eye-tracking

X. Pan · W. Hu · X. Lv · C. Xie (✉) · J. Dong · X. Chang (✉)
Ocean University of China, Qingdao 266100, China
e-mail: spring@ouc.edu.cn

X. Chang
e-mail: 2007050@ouc.edu.cn

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_4

1 Introduction

Sailing is a globally popular sport, but due to natural constraints, many inland enthusiasts struggle to access environments suitable for sailing. Researchers have developed sailing simulators to enable people to experience sailing, but their simulation results are not satisfactory. One key element of simulating sailing is to establish a model for sailing maneuvers. Current simulators typically only consider the propulsive force generated by a single sail, failing to accurately simulate multi-sail navigation [1]. Human–computer interaction is also crucial in sailing simulators. Existing simulators primarily use PCs or a limited number of VR immersive devices to reconstruct specific boat types [2], allowing users to move around and interact with the boat to replicate sailing scenarios or complete specific sailing tasks [3]. However, the interaction modes of these systems are usually keyboard-mouse or VR controllers, which do not take into account the natural actions of humans in real situations and lack realism and immersion.

The contributions of our sailing simulator are summarized as follows: (1) A multi-sail 4DOF maneuvering motion model is proposed; (2) The natural virtual hand interactions and gesture-based interactions are integrated into the system; (3) The eye-tracking technology is used to comprehend users' visual intentions by real-time informational prompts. The following sections of the paper will introduce "Related Works, Sailing System, User Study, Conclusion and Future Work".

2 Related Works

Research on sailing simulators primarily focuses on two aspects: modeling of sailing dynamics and interaction in sailing maneuvers.

2.1 Modeling of Sailing Dynamics

There are three representative methods for modeling sailing dynamics. The first type is the Abkowitz holistic model [4], which establishes a function between the state variables of ship motion and the parameters of propulsion system. This involves numerous parameters, and some of which lack physical significance. The second type is the matrix vector model [5], which describes the forces and moments acting on the ship in the form of matrix vectors, analyzing the stability and passivity of ships. The MMG model [6] separately models the forces and moments acting on the hull, propeller, rudder, and their interactions. All parameters in this model have physical significance. This paper is based on MMG model for modeling.

2.2 *Interaction in Sailing Maneuvers*

Currently, common interaction methods in simulators mainly rely on VR controllers, leading to unnatural interactions and poor user experience. Natural interaction can significantly enhance user immersion and task completion rates [7]. Moreover, designing interaction gestures based on real-life situations can improve interaction quality and user experience [8, 9]. Technologies like Leap Motion [10], which are based on visual gesture recognition, have limitations due to their restricted detection range. Continuous hand placement in front of these head-mounted devices for recognition causes user arm fatigue. Additionally, interaction techniques based on eye-tracking not only align with human interaction habits but can also be combined with hand gestures to simplify user operations and enhance interaction performance [11].

3 Sailing System

First, we established a sailing maneuver motion model of sailboat J/80 to simulate the maritime sailing scenarios, along with an interaction interface (see Fig. 1a). The system allows users to interact with the sailboat through ropes and the rudder, inputting control strategies into the maneuver motion model. It then calculates the motion state of the sailboat in real time and updates it in the scene, enabling users to learn and practice sailboat operations in this system.

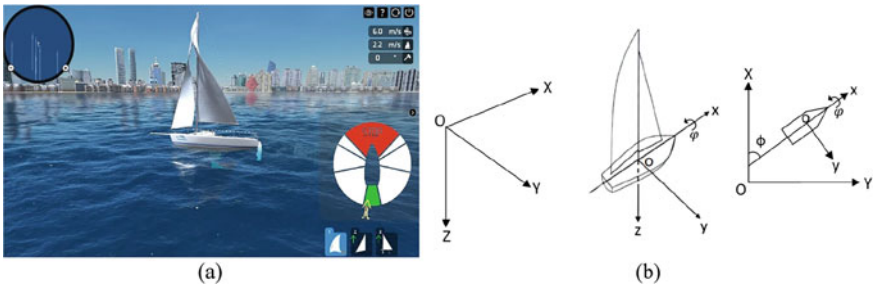


Fig. 1 **a** Sailing manipulation simulator. **b** Inertial coordinate system (O-XYZ) and body attached coordinate system (o-xyz), both of which are right hand coordinate systems

3.1 4DOF Maneuvering Motion Model Integrating Multiple Sails

Assuming the sailboat is a rigid body, with the coordinate system (see Fig. 1b), we establish a 4DOF maneuvering motion equation for the integration of multiple sails, as seen in Eq. (1). Here, m represents the mass of the hull; I_{XX} and I_{ZZ} are the moments of inertia of the hull mass around the x and z axes, respectively; u and v are the longitudinal and lateral velocities of the hull; $\dot{u}$ and $\dot{v}$ are the longitudinal and lateral accelerations of the hull; $\dot{p}$ and $\dot{r}$ are the roll and yaw angular accelerations of the hull; X, Y, K, N are the different forces experienced by the sailboat in forward, drifting, rolling, and yawing motions. The subscript H refers to the hull, K refers to the keel, R refers to the rudder, $Wave$ refers to the wave, s refers to the mainsail of the J/80 sailboat, hs refers to the jib, and sp refers to the genoa. X_s, X_{hs}, X_{sp}, Y_s, Y_{hs}, Y_{sp}, K_s, K_{hs}, K_{sp}, N_s, N_{hs}, N_{sp} are used to describe the impact of multiple sails on the motion of the sailboat under the influence of wind.

$$\begin{cases} m(\dot{u} - vr) = X_H + X_R + X_K + X_{Wave} + X_s + X_{hs} + X_{sp} \\ m(\dot{v} + ur) = Y_H + Y_R + Y_K + Y_{Wave} + Y_s + Y_{hs} + Y_{sp} \\ I_{XX}\dot{p} = K_H + K_R + K_K + K_{Wave} + K_s + K_{hs} + K_{sp} \\ I_{ZZ}\dot{r} = N_H + N_R + N_K + N_{Wave} + N_s + N_{hs} + N_{sp} \end{cases} \tag{1}$$

The force on the sail as seen in Eq. (2). L_{s_i} and D_{s_i} represent the lift and drag forces respectively, exerted by aerodynamic forces on the sail, as well as the sail area (S_{s_i}), lm_{s_i} represents the projected distance on the G_x axis from the mast to the resultant force point of airflow on the sail, while lom_{s_i} is the horizontal distance from the mast to the coordinate origin. α_{as_i} is the relative wind angle, α_{s_i} is the sail's angle of attack, and δ_{s_i} is the sail angle. si includes s, hs, sp, which refer to mainsail, jib, and genoa.

$$\begin{cases} X_{s_i} = L_{s_i}\sin\alpha_{s_i} - D_{s_i}\cos\alpha_{s_i} \\ Y_{s_i} = -L_{s_i}\cos\alpha_{s_i} - D_{s_i}\sin\alpha_{s_i} \\ K_{s_i} = \left(L_{s_i}\cos\alpha_{ar} + D_{s_i}\sin\alpha_{ar}\right) \cdot z_{s_i} \\ N_{s_i} = \begin{bmatrix} \left(-L_{s_i}\cos\alpha_{ar} - D_{s_i}\sin\alpha_{ar}\right) \cdot \left(lom_{s_i} - lm_{s_i} \cdot \cos\delta_{s_i}\right) - \\ \left(L_{s_i}\sin\alpha_{s_i} - D_{s_i}\cos\alpha_{s_i}\right) \cdot lm_{s_i} \cdot \sin\delta_{s_i} \end{bmatrix} \end{cases} \tag{2}$$

3.2 Natural Interaction Based on Data Gloves

Our system implements natural virtual hand interactions and global gesture interactions based on static gesture recognition in the virtual environment. We have chosen the Noitom Hi5 2.0 data glove as the interaction device. This data glove uses inertial

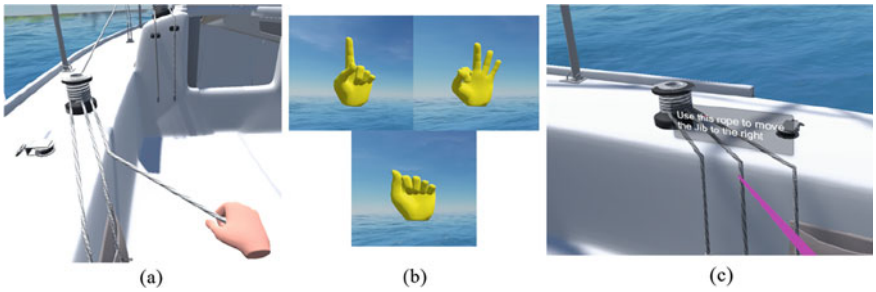


Fig. 2 Interaction design: **a** Grasping the rope. **b** IndexFinger, OK and fist gesture. **c** Eye-tracking real-time information prompts

sensors to collect hand information, displaying a virtual hand model and its motion changes within the scene.

Virtual Hand Grasping and Pulling. In the virtual J/80 sailboat, there are nine ropes for different sailing maneuvers. The cabin houses three ropes for the mainsail's lateral maneuvering and elevation changes. On the left side of the boat, three ropes control the jib's elevation, leftward movement, and the genoa's leftward movement. The right side contains three ropes for raising or lowering the genoa, moving the jib right, and moving the genoa right. As illustrated in Fig. 2a, users interact with these sails by approaching the corresponding rope in the virtual environment, grasping and pulling the virtual rope with hands clad in data gloves to raise, lower or swing sails for various maneuvers. This method provides a more natural and intuitive maneuvering experience as real-life sailing.

Global Static Gesture Recognition Interaction. Our system offers a training mode where users can pause, continue, or reset the training for adjustments. We implemented three gesture types for interaction: IndexFinger, OK, and Fist gestures, illustrated in Fig. 2b. Unlike standard VR applications that rely on UI panel clicks via VR controllers, sailing demands rapid response to environmental changes. We thus opted for static gesture recognition for real-time interaction needs. Gesture recognition occurs when the hand model's collision bodies interact. The IndexFinger gesture pauses the sailboat for adjustments, the OK gesture resumes movement, and the Fist gesture restarts the task. Data gloves used for gesture interaction mitigate the hardware limitations of gesture recognition, reduce arm fatigue from extended arm raising. The system processes gestures via gloves' sensor data, independent of hand placement, with user actions inputted into the motion model for computation.

3.3 Real-Time Information Prompts Combined with Eye Tracking

In virtual environments, users operate three sails using nine distinct ropes, each with a specific function. Remembering the purpose of each rope is challenging, with the potential for confusion and mistakes. A UI widget for hints near each rope could lead to information overload and visual clutter, negatively impacting the experience. To simplify the learning process, we employ an eye-tracking module in the HTC Vive Pro to track users' gaze points and directions. Rope selection hinges on where the user looks, showing real-time function information for the focused rope only. This information vanishes when the gaze shifts, reducing interference and improving interaction efficiency, as depicted in Fig. 2c.

4 User Study

We invited 11 users (10 males, 1 female, around 23 years old) to test our system, including 2 students majoring in sailing and 9 students from the computer science department. They evaluated both the natural interaction enabled by data gloves and the real-time information prompts facilitated by eye-tracking. The experiment devices included a computer, an HTC Vive Pro VR headset, two Hi5 2.0 data gloves.

4.1 Virtual Hand and Controller Interaction Evaluation

Users operated the sailboat using both controllers and data gloves, executing ten actions, including raising the genoa, lowering the genoa, raising the jib, moving the mainsail left, moving the mainsail right, moving the jib right, moving the genoa left, pausing, resuming, and resetting. The time spent by users to complete these tasks was recorded for a comparative analysis between the two methods, as depicted in Fig. 3a and b. The less time it takes to complete the task, the better the interaction method.

4.2 Eye-Tracking Real-Time Information Prompts Evaluation

The evaluation assessed user performance using our system, comparing scenarios with and without real-time information prompts from eye-tracking. Users randomly selected and utilized nine ropes for operational tasks, testing their understanding of each rope's role. We recorded their selection accuracy. After all users completed

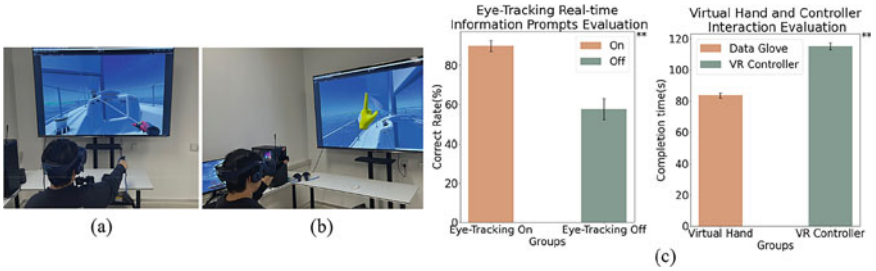


Fig. 3 **a** Tasks performed using VR controllers for interaction. **b** Tasks performed using data gloves for interaction. **c** Evaluation results bar chart

tasks, we aggregated the data to analyze choice accuracy, conducting a comparative analysis. We hypothesize that higher task completion accuracy indicates a more effective interaction method.

4.3 Results and Analysis

The experiment (refer to Fig. 3c) employed ANOVA ($\alpha = 0.05$) to assess significant differences in user performance across two interaction methods. The analysis identified significant differences in both methods' impact on performance ($p < 0.05$).

In the Virtual Hand and Controller Interaction Evaluation, users averaged 83.5 s to complete tasks using data gloves, against 115 s with VR controllers ($p < 0.05$), indicating that virtual hand interaction significantly reduces task completion time compared to VR controllers, enhancing sailboat maneuvering efficiency.

During the Eye-Tracking Real-time Information Prompts Evaluation, task completion accuracy with eye-tracking prompts was 90%, versus 58% without ($p < 0.05$). This highlights that our eye-tracking based real-time prompts aid in quicker and more accurate task completion, effectively lowering the learning curve.

We complemented objective metrics with NASA-TLX assessments to understand user subjective preferences. Results indicated a majority favoring virtual hand interaction for its enhanced performance. However, some users found data gloves small and lacking in durability. To address this, we're considering acquiring various glove sizes for better user fit. Most users also preferred performing tasks with eye-tracking real-time prompts, citing reduced mental demand and frustration. Conversely, a minority found these prompts abrupt and immersion-breaking, recommending an eye-tracking toggle option for user customization.

5 Conclusion and Future Work

In this paper, we designed and developed a 4DOF multi-sail sailing simulation system and proposed two interaction methods that enhance the sailing manipulation experience. Users wearing data gloves can interact naturally with ropes and the ship's rudder, combining conventional sailing techniques. Using global hand gesture recognition reduces fatigue and expands the range of gesture recognition. The use of an eye-tracking-based real-time information prompt system reduces the learning curve and enhances interaction performance and experience.

In the future, we aim to enhance hand–eye coordination to improve interaction efficiency. We also plan to integrate open-source voice recognition plugin to implement voice command control for sailing, improve interaction efficiency and training quality.

Acknowledgements This work was supported in part by the Fundamental Research Funds for the Central Universities (No.202264002).

References

1. Setiawan, J.D., Chrismianto, D., Ariyanto, M., Sportyawan, C.W., Widyantara, R.D., Alimi, S.: Development of dynamic model of autonomous sailboat for simulation and control. In: 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), pp. 52–57 (2020).
2. Barreau, J.B., Nouviale, F., Gaugne, R., Bernard, Y., Llinares, S., Gouranton, V.: An immersive virtual sailing on the 18th-century ship Le Boullongne. Presence: Teleoperators Virtual Environ. **24**(3), 201–219 (2015).
3. Novitzky, M., Semmens, R., Franck, N.H., Chewar, C.M., Korpela, C.: Virtual reality for immersive human machine teaming with vehicles. In: Virtual, Augmented and Mixed Reality. HCII 2020, Lecture Notes in Computer Science, vol. 12190, pp. 575–590 (2020)
4. Lewis, E.V.: Principles of naval architecture second revision. SNAME, Jersey City, New York **2**, 152–157 (1988)
5. Fossen, T.I.: Handbook of marine craft hydrodynamics and motion control. John Wiley & Sons, Manhattan (2011)
6. Yasukawa, H., Yoshimura, Y.: Introduction of MMG standard method for ship maneuvering predictions. J. Mar. Sci. Technol. **20**(1), 37–52 (2015)
7. Ricca, A., Chellali, A., Otrnane, S.: The influence of hand visualization in tool-based motor-skills training, a longitudinal study. In: IEEE Virtual Reality and 3D User Interfaces (VR), pp. 103–112. Lisboa, Portugal (2021)
8. Blaga, A.D., Frutos-Pascual, M., Creed, C., Williams, I.: Freehand grasping: an analysis of grasping for docking tasks in virtual reality. In: IEEE Virtual Reality and 3D User Interfaces (VR). pp. 749–758 (2021)

9. Vuletic, T., Duffy, A., Hay, L., McTeague, C.: Systematic literature review of hand gestures used in human computer interaction interfaces. Int. J. Hum. Comput. Stud. **129**, 74–94 (2019)
10. Alimanova, M. et al.: Gamification of hand rehabilitation process using virtual reality tools: using leap motion for hand rehabilitation. In: First IEEE international conference on robotic computing (IRC), pp. 336–339. Taichung, Taiwan (2017)
11. Wang, Z., Wang, H., Yu, H., Lu, F.: Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system. IEEE Trans. Hum. Mach. Syst. **51**(5), 524–534 (2021)

Virtual Re-Creation in Augmented Reality for Artistic Expression and Exhibition

Jeffrey Price, Bryon Caldwell, Brandon Coffey, Hamida Khatri, Scott Krabbenhoft, and Justin Shaw

Abstract The convergence of fine art and technological innovation has led to a paradigm shift in creative expression, challenging the conventional constraints of traditional art and redefining the role of the artist. Through a collaborative effort with the New Mexico Museum of Art and Vladem Contemporary, this study examines the Grand Opening exhibition showcasing the transformative nature of Judy Chicago's "Kitty City" cat sculptures and exploring the complexities involved in animating them inside an augmented reality environment. While the project magnifies the polygonal relationship between art interpretation and ownership, it emphasizes the need to strike a balance between preserving the original essence of the artwork and introducing technological ingenuity. The paper sheds light on the multilayered process of how artistic authenticity is transformed into a digital domain with the interplay between the medium and the message. This involved a comprehensive examination of the process of 3D scanning cat sculptures and the use of the software Maya and Unity for the purposes of animation, rigging, and texture development. Furthermore, transitioning the animated cat rigs to the Meta Quest Pro AR/VR headset (HMD) included a series of procedural stages, including boundary delineation inside the art gallery, calibration, spatial anchor deployment, and control design. The entire process was centered on guaranteeing an accurate representation of creative production via the use of precise lighting and shadow methods, providing a full understanding of the integration of fine art and modern technology.

Keywords Virtual reality · Augmented reality · Fine art · Emerging technology · Representation · Expression · Animation · Ownership · Sculpture

J. Price · B. Caldwell · B. Coffey · H. Khatri · S. Krabbenhoft (✉)
University of Texas at Dallas, Richardson, TX 75080, USA
e-mail: Scott.Krabbenhoft@UTDallas.edu

J. Shaw (✉)
Oklahoma City University, Oklahoma City, OK 73106, USA
e-mail: Justin.Shaw@okcu.edu

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_5

1 Introduction

The convergence of artistic mastery with technical proficiency introduces a distinct facet to the realm of art, which disrupts traditional limitations and reevaluates the artist's role. This study explores the intricate process of portraying Judy Chicago's three-dimensional "Kitty City" cat sculptures inside a virtual reality (VR) environment. This endeavor emphasizes the delicate balance between the interpretation and ownership of the artist's work. The key aspect is in achieving a harmonious equilibrium between maintaining authenticity to the original artwork and using technological advancements without compromising the artist's intended message. Through the lens of augmented reality (AR), we sought to breathe life into Judy Chicago's cat sculptures, presenting them as lively entities in a real-world setting. This project demanded a sensitive approach to avoid overshadowing the artist's original message with technological dexterity. The journey involved layers of separation from the original artwork, the challenges of transitioning from a fine artist to a technical realm, and the dynamics of changing roles within the production team. In this paper, we reflect upon the complexities encountered during this project, examining challenges around creative authenticity, the potential pitfalls of misinterpretation, and the dynamic relationship between medium and message in the digital age.

2 Artistic Expression Using Augmented Reality

2.1 Judy Chicago's Feline Sculptures in Augmented Reality

For the Grand Opening exhibition on September 23, 2023, in New Mexico, our team joined hands with Mark White, Director of the New Mexico Museum of Art, and Vladem Contemporary. This collaboration centered around the acclaimed artist Judy Chicago and her efforts to bring to life four cat sculptures from her "Kitty City: A Feline Book of Hours" through AR technology [1] (See Fig. 1).

Tracing back to the 1970s, Judy Chicago delved deep into the world of felines, a journey that birthed a collection of intricate watercolor paintings over the years. The intricacy of these artworks arose from the challenges of capturing the fleeting stillness of cats. These paintings, later featured in a 2005 publication, found their muse in the medieval "book of hours"—sacred scripts from the Middle Ages. The devotional nature of Judy Chicago's artwork mirrors her profound connection with the cats she shared her life with, alongside her spouse, Donald Woodman. Every piece speaks of the distinct character of the felines, reinforcing Judy Chicago's reverence for life in all its forms. This sentiment is beautifully encapsulated in her reference to John Gardner's quote, "Always be kind to animals. Morning, noon, and night. For animals have feelings too. Moreover, they bite" [1]. In her writings, she emphasizes, "Through my art, I've persistently aimed to champion the voiceless, especially those from other species" [1]. Contrary to many women of her era, her life revolved predominantly

Fig. 1 Judy chicago's ceramic cat sculptures: (Left to Right) Veronica, Inka, Milagro, Poppy

around her art, leaving little room for traditional caregiving roles. Her bond with Sebastian, her first cat, was transformative. His medical needs introduced her to the profound realization that "true love demands unwavering care for the cherished" [1]. Integrating Judy Chicago's artistic representation of cats into AR avatars offers viewers a palpable presence of these feline entities in the real-world environment, a way to memorialize them. Beyond the tangible beauty of her sculptures, the AR adaptation immerses the audience in a lively experience, blurring the lines between memories and present interactions.

2.2 AR Immersive Experiences

The captivating allure of AR lies in its ability to blend digital and real-world elements seamlessly. In a project like this, where cats are presented in an AR setting, they appear almost tangible, reinforcing the belief that they coexist in our environment. Animated movements mirroring the natural behavior of cats amplify this illusion of realism [2].

2.3 Data Collection of Observed Behavior/Capturing Visitor Reactions

During the museum's inauguration, over 500 visitors were in attendance. Of these, approximately 100 chose to put on the Meta Quest Pro AR/VR headset (HMD), immersing themselves in the virtual world of animated cats.

2.4 Goals of the Project Scope

In the original scope the virtual cats were to serve as interactive guides in the museum setup, directing visitors to specific artworks within the galleries [3]. This functionality of the system was successfully conducted in the research laboratory; nevertheless, its performance was hindered by the final gallery (architecture and lighting) configuration, resulting in its failure to operate as intended, therefore realigning our project scope.

3 Modeling and Animating the Cats

3.1 Previsualization

The project began with a meticulous preparation and research phase. A series of 3D scans of Judy Chicago's ceramic cat models were collected from the museum and her studio in New Mexico, serving as the foundation for the subsequent work (see Fig. 2). To understand the artist's materials and art style, a material analysis of the ceramics was conducted, focusing on material and texture properties. In addition, an in-depth study of feline anatomy was conducted to adjust the model's proportions according to the specifications of Judy Chicago's sculptures.

3.2 Models and Textures

Our first attempt at modeling the scanned cats, was to use a photogrammetry software called "Reality Capture" to convert the photographs to a point cloud [4]. This did not give the quality of mesh we were intending; therefore, the decision was made to have each sculpture be made digitally. We chose to sculpt the cat representations using ZBrush software, which is an advanced digital sculpting tool (see Fig. 2) [5].

In the next stage, the 3D scanned models were moved into Substance Painter for detailed texturing [6]. Each texture was designed to mirror the unique materials and aesthetic of Judy Chicago's ceramic cats (see Figs. 3 and 4).

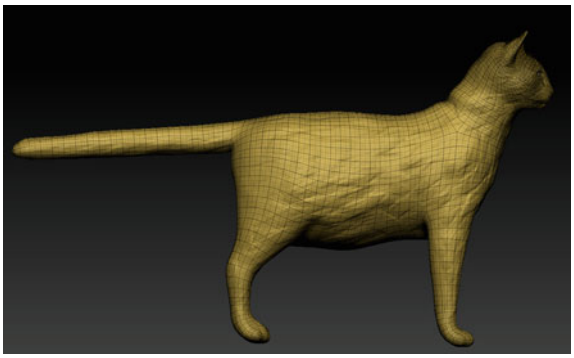


Fig. 2 Judy Chicago's Zbrush model

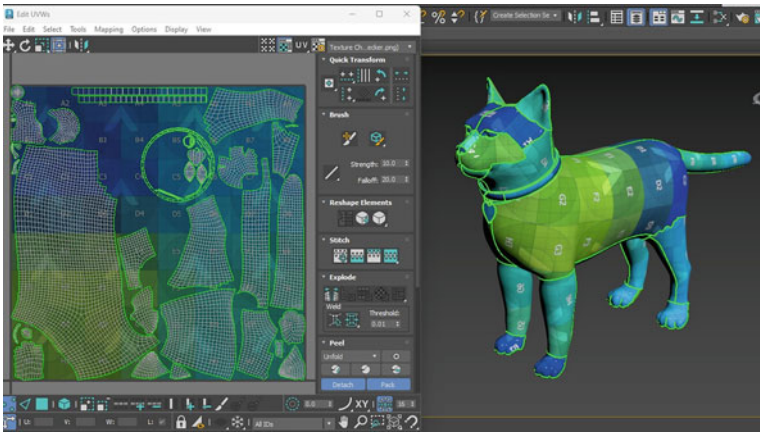


Fig. 3 UV mapping layout for texturing

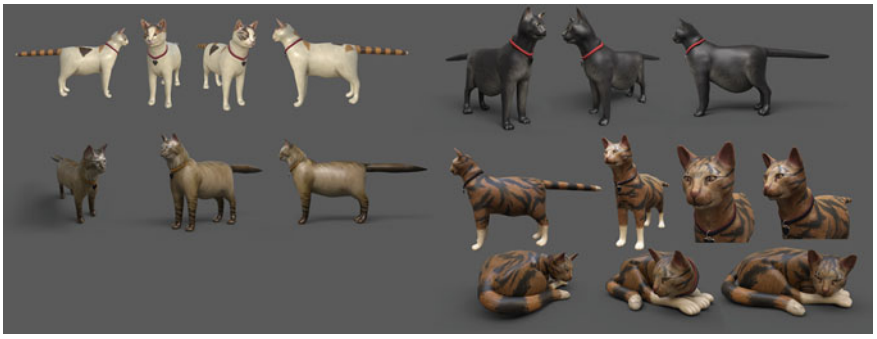


Fig. 4 All four 3D cat models with final textures

3.3 Rigging and Animation

The development of the animation began with a search across online vendors to find pre-made rigs suitable for adaptation and retargeting to our cat 3D models. Realizing retargeting would not work, the decision was made to engage a specialized rigger to create custom rigs for the cats. The goal for the specialized rig would be to construct a skeletal framework and controlling mechanisms, dictating the model's diverse movements and transformations (see Fig. 5).

One significant modification involved relocating the "Root Control" from the hips to the spine's midpoint, enhancing the flexibility and controllability of the animations (see Fig. 6). With the completion of the cat animation rig, the focus shifted to character animation, encompassing six basic actions: running, walking, sitting idly, sitting, and licking, transitioning from walking to sitting, and sitting to stretching. To infuse the animations with authenticity and capture genuine cat-like movements, comprehensive research was conducted by studying hours of live-action footage of real house cats. Select video clips were chosen to serve as motion reference, ensuring the accuracy of the cats' body mechanics throughout the animation process. In addition to Judy Chicago's insightful reviews and critiques of the work in progress, her book "Kitty City" provided valuable insights into specific poses and personality descriptions, enriching the depth and believability of the cat characters' movements [1]. As each action was animated and finalized, the Maya animation data was exported and adapted for the other three cats: Milagro, Veronica, and Inka. Poppy, the fourth cat, had a sleeping position originally from the artist so we decided not to have it moving with the rig (see Fig. 1).

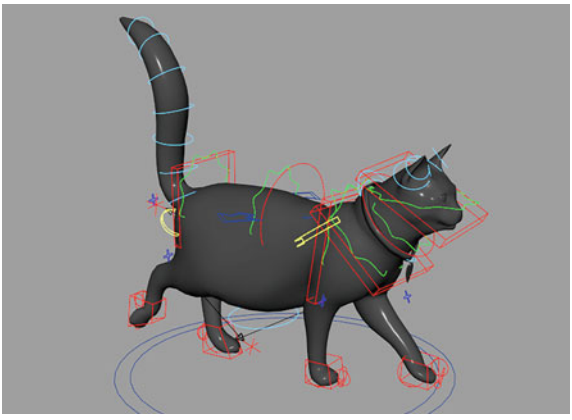


Fig. 5 Generic cat rig used and modified for each cat

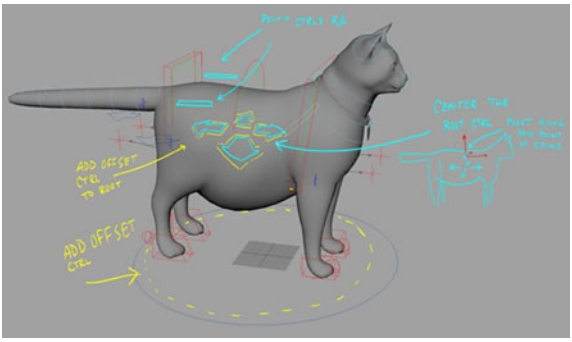


Fig. 6 Cat rig controls, position, and movement

3.4 Maya to Unity

Once the first few animations were completed, they were exported as.FBX files and imported into Unity. Unity's simplification of rigs cannot handle these components separately, and this led to broken animations. (see Fig. 5). To fix this issue, a second duplicate joint structure with the same bone weights was created and then positionally constrained to the matching joints from the source rig. This new rig followed the animation's motion, with minor offsets due to missing the vertex scaling, but it worked in Unity without major scaling issues. Once imported, however, one issue persisted in relation to the animation, as the skin of the cats was deforming unnaturally. The identification of the issue was promptly ascertained to be the skin weights, as the rig was only tested in Maya and each vertex on the mesh was affected by up to 10 joints. Under default settings, Unity only supports vertex transformations from either 2 or 4 joints, which led to this "artifacting," a glitch characterized by faulty texture (see Fig. 7). By changing the project settings under graphics to allow unlimited joints to affect each vertex, and then setting each cat animation to allow up to 8 joints provided an optimal solution to the issue while minimizing performance degradation.

Subsequently, it was discovered that an additional segment of the animation was missing and exclusively controlled by the rig. The Unity software lacks support for the importation of "blendshape" keyframes, necessitating the implementation of a customized solution (see Fig. 8).

A collection of joints without any connection to the rig was generated to transmit this data to Unity. These joints included positional offsets equivalent to the blend shape value derived from the rig. In Unity, this value is read at each frame update and then used to modify the blend shape values of the skeletal mesh. Once a pipeline to import the animations into Unity was implemented, a state machine using Unity's "Mecanim" (a system that employs a visual layout similar to the flow-chart to represent the state of the machine to enable to control and sequence the animation clips used for the character/object) was created with all states (character engaged in a particular kind of action in a given time) that would be animated for each of the cats (see Fig. 9). From a scripting standpoint, primarily Boolean values were used

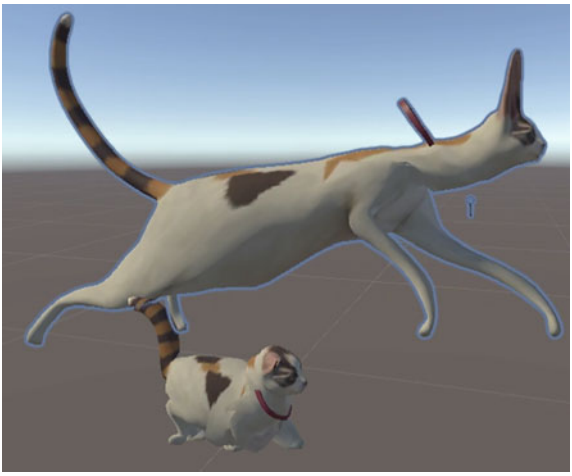


Fig. 7 Generic cat rig used and modified for each cat

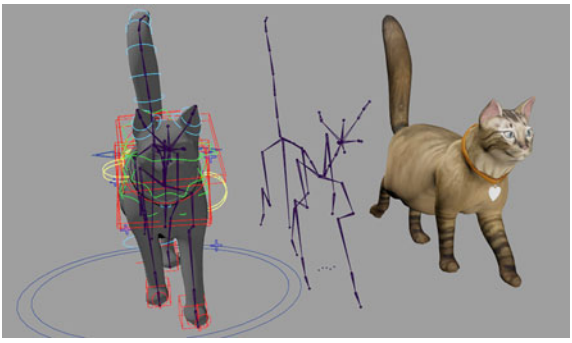


Fig. 8 Left: Source cat rig that was animated; Middle: Duplicate rig structure without scale, blendShape values separate at the bottom; Right: Final mesh output compatible with unity

as parameters for the state machine, such as "IsSitting" or "IsStretching." Then transitions between the states were set up with varying time intervals to "lerp" (Linear Interpolation: used for creating simple movement and animation over a fixed time interval) between the animations for the most realistic result. Some animations, such as "WalkToSit" required unique animations for the transition to look more realistic, rather than using the "lerp" functionally. A basic test case that moved the cats forward while walking and stopped them when sitting was set up, which revealed additional specific cases that needed custom logic.

During the "WalkToSit" transition, it was observed that the cat abruptly stopped as soon as the animation started, which was incongruous, given that the cat proceeded to take an additional step subsequent to the initiation of the animation. In order to achieve the desired animation effect, it was determined that the cat should go forward and then stop. To do this, "Mecanim" system was applied, whereby an animation trigger was implemented to execute custom movement manipulation logic. This logic allowed the continuation of the cat's forward motion even after it had come to a stop. Moreover, throughout the "Stretch" animation, the cat proceeded by taking a

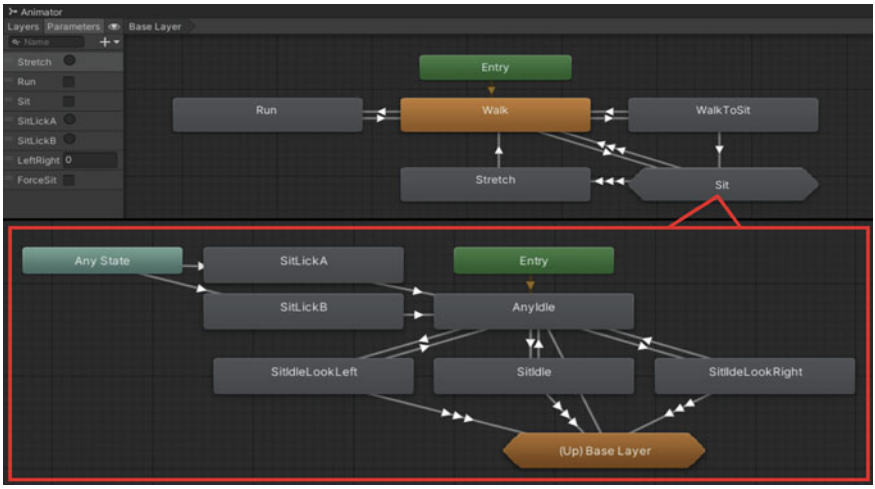


Fig. 9 Unity Mecanim setup for cats, minus personality-based animations specific to any cat

single step forward, thereafter, stopped momentarily, and subsequently resumed its forward locomotion [7]. The process required the use of three triggers: one to initiate the customized movement, another to halt the movement, and a final trigger to resume movement till the cat reached the desired walking motion. To further enhance the coherence of the transitions, the custom movement logic used a snap "lerp" function to mitigate abruptness in the initiation and cessation of movement.

3.5 Virtual Walking Cats on a NavMesh

The cats in Unity were configured as "NavMeshAgents," which are components that enable mobile characters or objects to navigate the virtual environment. This navigation is facilitated by a "NavMesh," which determines the character's path and assesses the walkability of specific areas within the game [8]. The agents have the capability to be given 3D coordinates in space as their designated destinations, and thereafter determine the optimal route to traverse towards these spots, while actively circumventing any impediments encountered along the way [9]. Subsequently, by using the previously established floor plan during the previsualization phase, cats were able to effortlessly navigate a digital reproduction of the exhibition area. Nevertheless, using a completely random wandering strategy was suboptimal due to the cats' inclination to attempt exiting the area, often resulting in their entrapment near the room's walls and corners. The current approach lacked optimality and did not provide a streamlined method for obtaining a random location within the confines of the "NavMesh." Consequently, an alternative strategy was implemented whereby a set of predetermined places were designated as destinations for each of the cats to

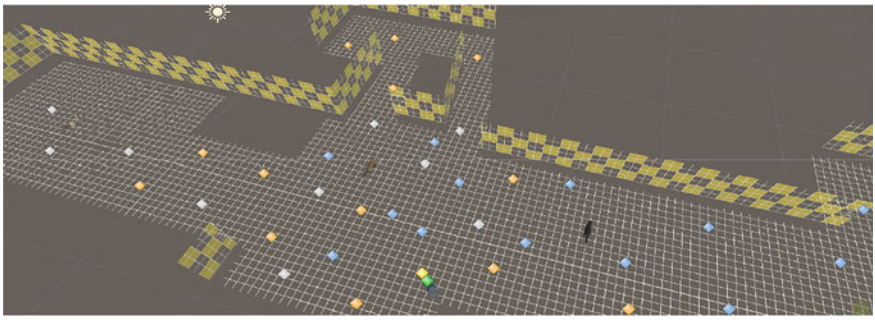


Fig. 10 Floor plan layout with colored points that match each cat's assigned destinations

navigate towards. The allocation of specific spots enabled the cats to remain confined inside designated zones, giving the developers a heightened level of control over the cats' perception (see Fig. 10). In order to determine their behaviors, each cat was allocated a set of potential actions, including walking, sitting, stretching, and so on [2]. The cats were designed with a programming mechanism that mandated them to choose a random action at every given moment, followed by a period of waiting characterized by a certain range of predetermined values assigned to each action. When the cat is strolling, it will have a tendency to wait until it arrives at the designated location before making the decision to sit down. In order to determine the frequency of certain actions performed by the cats, we devised a weighted algorithm to guide the selection of random activities.

To enhance the individuality and authenticity of each cat, distinct weights were assigned to each, drawing inspiration from the insights provided by Judy Chicago and her book [1]. This observation resulted in the cat's exhibiting behavior that more closely resembled their natural tendencies [10]. Once the appropriate plans were made for the cats, they freely explored the designated area and independently engaged in various activities, exercising their autonomy to make choices and pursue their own preferences. Additionally, a system for unique actions for specific cats was also added [1]. As an example, Inka, a cat of considerable large size, was supplied with a food receptacle to which she often meandered, consuming a little portion before succumbing to slumber (see Fig. 11). Following a period of rest, she assumed an upright position, proceeded to cleanse her extremities by means of licking, engaged in a brief period of observation, executed a stretching maneuver, and then embarked along a meandering trajectory in a different orientation. The primary objective of this system of animation was to enhance the cats' realism and facilitate their engagement in cat-like behaviors, simulating a sense of autonomy [10].

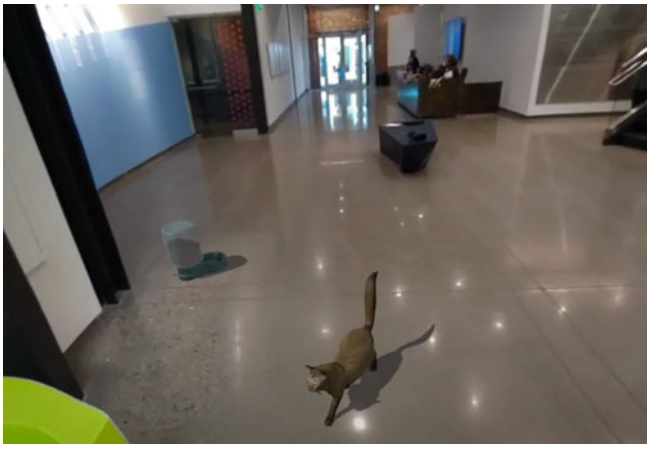


Fig. 11 Cats as seen through AR with meta quest Pro AR/VR headset

4 Augmented Reality Integration

4.1 Choosing an HMD: Meta Quest Pro Versus Microsoft Hololens 2

During the preliminary phase of the project, a deliberation arose on the selection of hardware to be used, namely between the Meta Quest Pro and the Hololens 2. The Hololens 2 emerged as a promising first contender because to its proximity to genuine AR [11]. This device effectively projects visual elements onto a transparent surface positioned in front of the user's eyes, enabling them to see the physical surroundings independently from the augmented visual content. In contrast, the Meta Quest Pro is a comprehensive VR headset that utilizes its front-facing cameras to replicate the user's visual perception and then generates them as the backdrop for a VR encounter. This phenomenon results in visual distortion when the camera feeds overlap. Furthermore, the quality of the cameras and their renderings have a crucial role in determining the visual output, typically exhibiting graininess. Upon first use of the Hololens 2, some concerns promptly surfaced. Specifically, the presence of a limited visual rendering area inside the glass frame resulted in a significant portion of the user's field of vision being unable to see virtual items. While not inherently problematic, it is suboptimal for creating an immersive experience. Additionally, the user retains the ability to visually perceive the physical headgear. Notwithstanding these shortcomings, the Hololens 2 affords users the ability to retain their whole peripheral vision while also providing an unobstructed and lucid perspective of the physical environment. In contrast, Meta Quest Pro does not provide a comprehensive field of view, although it does provide the most optimal integration of virtual objects with real-world visuals. This is precisely why it was selected for the present undertaking. When comparing

the two headsets, it was evident that the Meta Quest Pro offered a higher level of immersion and realism in its visual experience. This is achieved by replicating the user's actual world, although at the expense of reduced peripheral vision.

With a headset selected, it was time to decide on a development game engine. The decision to use Unity was made promptly due to its superior compatibility with the mobile headset and the developers' existing familiarity with the platform. In order to configure a Unity project for use with the Meta Quest Pro for AR or VR, the Oculus Integration SDK and OpenXR had to be imported as plugins and set up in the scene using the given "OVRCameraRig" prefab, which is a replacement to Unity's main camera providing the key interface to the VR hardware of Meta Quest Pro. On this prefab, the settings on the "OVRManager" were modified to require AR Passthrough (allowing camera overlay of the real world to be projected into the user's VR view) and an "OVRPassthroughLayer" was added to the camera rig. Building this as an APK and using the Oculus Developer Hub to export the file to the headset, the application became operational and had a comprehensive AR display. This ensured that any virtual objects placed in the scene would show up in AR relative to where the headset was positioned when the application was launched. Subsequently, by using the Meta Quest Developer Hub, applications were efficiently deployed and effectively controlled on interconnected headsets.

4.2 Meta Quest Pro Boundary

Upon activation of the passthrough, the Meta Quest Pro displayed a comprehensive wireframe representation of the gallery's floor layout inside the scene, therefore revealing the initial problem associated with the headset. In the process of configuring the headset, it is necessary to define a "boundary," which refers to the designated region demarcated as a safe zone for engaging in gaming and other VR activities. When wearing the headset, if the user approaches or exceeds the border, or ventures beyond it, any open application is deactivated and prompts the user to either return to the boundary or establish a new one. The efficacy of this method is very notable for a majority of applications. However, when confronted with the task of delineating a boundary on the size of the gallery's floor plan, it encountered a limitation in terms of its maximum distance capability. The prescribed boundary is established at a distance of 15 m from the location of the headset and is not adjustable. The gallery in which the experience took place spanned a minimum length of 50 m. Participants were required to traverse the entire gallery, making it clear that the boundary was not an option for this large of a full-scale experience.

The boundary system may be deactivated by using a headset with developer mode enabled, which is a feasible alternative due to the user's constant visual access to the surrounding environment inside the program. Nevertheless, this state of disablement is not without its own array of challenges. The boundary feature plays a crucial role in establishing the minimum degree of immersion, and when deactivated, the Meta Quest Pro loses its ability to precisely detect the vertical distance to the floor.

Additionally, the absence of the border prevents the proper functioning of another experimental feature known as "Room Setup." The Room Setup feature enables users to accurately depict the layout of a room by drawing up to 20 walls, regardless of size. This functionality is particularly valuable for developers seeking to create an AR experience, as the application can accurately identify the precise location of the walls, resulting in a seamless integration with the physical space. The inclusion of this element would have been ideal for the overall experience, provided that it functioned properly when the border was deactivated, allowing the experience to include the whole extent of the space. Furthermore, it is important to acknowledge that there exists the possibility of delineating numerous borders, allowing users to traverse from one boundary to another. However, it is challenging to effectively communicate this transition to the user without interruption, since the headset often encounters difficulties in identifying the subsequent boundary while simultaneously tracking the present one.

4.3 Spatial Anchors and Calibration

With the boundary disabled, a new method to both ground the floor and know where the walls are positioned was necessary. Meta Quest Pro has a solution for retaining information about where a virtual object is in the real world through a feature called "Spatial Anchors" (see Fig. 12). This allows users to place down a point in space, save its position in the real world to the headset, and then recall that position at any given time, having that point reappear at the same location relative to the real world. Using spatial anchors, a user could place down a cat at a specific point, and no matter if the headset was turned off or put away, the same cat could reappear at that same point when the application is opened again. A system was set up for a developer to match the floor plan of the digital room to the actual floor plan so anchors could be placed on that virtual floor plan and then saved as spatial anchors [12]. Then, when the application is opened, the saved spatial anchors can be loaded, and the application could use the positions of the anchors to place the virtual floor plan in the position that would match the actual floor plan. This system worked seamlessly and without error, as it updated live if the headset started to drift or temporarily lost tracking. Additionally, using Meta's Platform for Developers, one headset could be used to set up the floor plan and can then save those anchors to the cloud and remotely share that data to additional headsets. This reduced the total amount of work to a considerable extent to set up the experience.

4.4 Custom Developer Controls

In order to restrict the placement of the floor plan and the management or debugging of the program to certain users, a collection of customized developer controls was

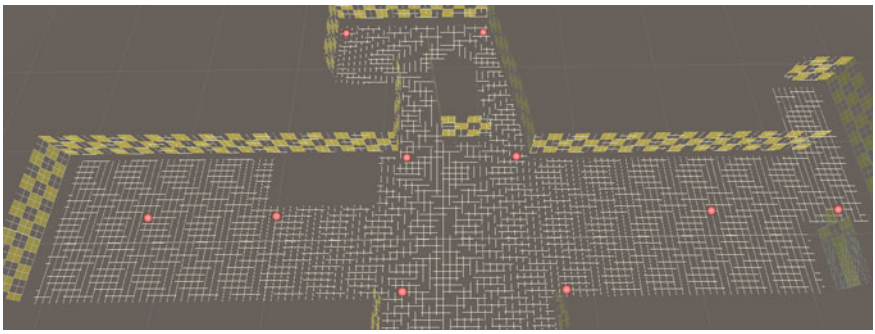


Fig. 12 Floor plan layout with red points marking the stored spatial anchor positions

developed (see Fig. 13). Given that the end users lack the necessary controllers, the design was implemented in such a way that the controls of the experience can only be accessible by a user possessing the controller and pushing a designated button. This measure aims to mitigate the risk of inadvertent access to the controls by end users, hence minimizing the potential for unintended modifications or confusion. The controller had a personalized array of choices that could be navigated and chosen by means of the "A" button. The closure of the menu may be achieved by pressing the "B" button. Maintaining a simplified menu was deemed crucial in order to accommodate museum staff who may possess little technical expertise but may still need the ability to troubleshoot any issues. A method was devised to facilitate the placement of the floor plan, enabling the user to see a wireframe grid representation of the floor plan and align it with the actual floor plan via the use of AR, so permitting simultaneous observation of both (see Fig. 14). This was developed with controls that have been designed to be as intuitive as feasible. Subsequently, additional functionalities were included to enable the preservation or removal of pre-existing anchors from the floor plan, as well as the adjustment of the floor plan's position and rotation.

Fig. 13 Developer tool menu as well as lower floor layout seen from the upper floor

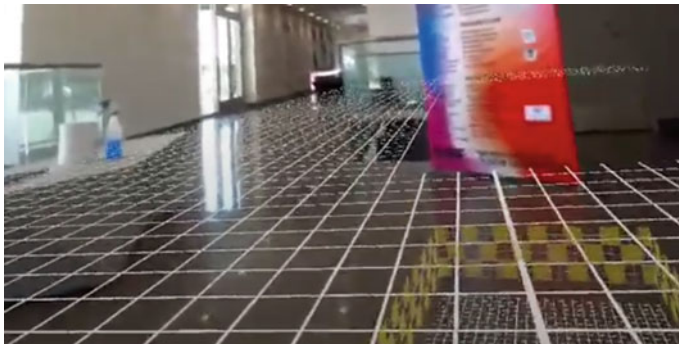


Fig. 14 Placing the upper floor to match the room layout as closely as possible

4.5 Lighting, Shadows, and Realism

When developing a project in Unity, users are presented with a selection of render pipelines to pick from. The Universal Render Pipeline (URP) is well-suited for this particular project because of its specific design aimed at facilitating the development of mobile apps while maintaining a high level of realism in the rendered output. Utilizing the available models and textures, the task of generating a material for the purpose of rendering cats was accomplished by establishing linkages between the distinct files. In Unity, channel packing is used for grayscale texture maps, such as metallicness and smoothness, which can be packed in photo editing software applications. Once configured, the cats simulate with a level of realism commensurate with the accuracy of the lighting setup. Initially, a rudimentary directed light source was used due to its ability to provide optimal rapid outcomes. The cats were visually appealing when seen against the default Unity skybox. As the project approached its final stages, a high dynamic range image (HDRI) depicting a room that closely resembled the physical environment in which the event would take place was included as the skybox. Additional lights could then be created to replicate the actual lighting of the environment if desired, which made the cats feel much more dynamic and realistically placed in the environment.

One further concern with the visual representation of cats using AR was the potential for them to have a floating or partly submerged look when seen from certain angles. In order to address this concern, it was possible to generate shadows on an imperceptible surface positioned underneath the cats. This approach served a double purpose: enhancing the authenticity of the cats' appearance and facilitating the viewers' cognitive connection between the shadow plane and the physical ground. However, accomplishing this task was not a straightforward endeavor due to the fact that the shadows are aligned with the transparency level of the rendered mesh, resulting in the shadow itself being imperceptible. One possible solution to address this issue involved implementing a customized shader that samples

the "MainLightRealtimeShadow" function and then applied it to the opacity (see Fig. 15).

By enabling alpha clipping, just the shadow was shown. The enhancement of rendering realism in AR for cats was achieved by implementing two modifications. Firstly, the opacity of the objects was reduced to provide better integration with the actual ground, resulting in a more seamless visual experience. Secondly, the shader was extended to accommodate more light sources, so further enhancing the authenticity of the rendered cats in the AR environment.

After the implementation of appropriate lighting for the cats, a last concern about the level of realism emerged, namely, the cats' visibility through walls and other objects. This phenomenon resulted in peculiar outcomes when the cats manifested as ethereal specters, apparently levitating without adherence to the spatial constraints of their surroundings (see Fig. 16).

The occurrence of this visual anomaly hindered the ability of users to completely immerse themselves in the virtual environment, hence impeding their capacity to

Fig. 15 Cats as seen with shadows

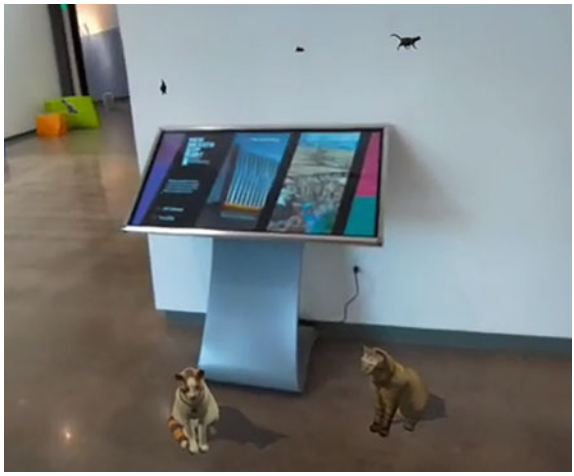


Fig. 16 Cats with the final lighting setup, users can see cats through the walls

actively interact with simulated cat objects. The cats can also be seen through any person walking around the space. One potential approach to address this issue was using the floorplan configuration to render the walls and thereafter utilizing them as a mask to disregard any items situated behind them. This method proved effective for the walls, if it aligned precisely with the dimensions of the room. However, it was not infallible, since there were other items inside the room that also obstructed the cats' transparency. Regrettably, the Meta Quest Pro does not provide a viable resolution for this issue. If the headset has the capability to generate a depth channel via the use of technologies like LiDAR, it would be feasible to compare this channel with a depth channel generated from Unity. Consequently, the system could choose to show the visual representation that is closer in proximity. The proposed technique had promise in effectively masking people, even the user's own hands. However, its implementation proved challenging due to the inherent complexity associated with developing a multi-camera system. In any case, the issue at hand remained unsolved within the context of this project, since there was no need to rectify it. In fact, it was deemed more advantageous to retain the ability to discern the whereabouts of the cats even when they were obscured by other objects. This was done to prevent users from erroneously perceiving a lack of visual stimuli when the cats vanished behind walls.

5 Conclusion

This paper delves into the intricate journey of marrying traditional art with cutting-edge technology, focusing on the digital revival of Judy Chicago's cat sculptures within an augmented reality sphere. Through a detailed process encompassing 3D scanning, advanced digital modeling, and the incorporation of AR tools, we navigated both artistic and technological challenges. The utilization of game design software played a pivotal role in bringing the "Kitty City" cat sculptures to life with fidelity and dynamism.

References

1. Chicago, J.: Kitty city: a Feline book of hours. Harper Design (2005)
2. Mertens, C.: Human-cat interactions in the home setting. Anthrozoös **4**(4), 214–231 (1991)
3. Bachiller, C., Monzo, J. M.: Augmented and virtual reality to enhance the didactical experience of technological heritage Museums. In: Special Issues Virtual Reality Technology and Applications. Appl. Sci. **13**(6) (2023)
4. Vajak, D., Livada, Č.: Combining photogrammetry, 3D modeling and real time information gathering for highly immersive VR experience. In: 2017 Zooming Innovation in Consumer Electronics International Conference (ZINC). Novi Sad, Serbia (2017)
5. Spencer, C.: ZBrush Character Creation. Wiley, Advanced Digital Sculpting (2011)

6. Potter, R., Rönnlund, R., Wallensten, J.: An evaluation of substance painter and mari as visualisation methods using the Piraeus lion and its runic inscriptions as a case study. Herit. Sci. **11**, 226 (2023)
7. Çimen, G.: Animation models for interactive AR characters. In: ETH Zurich (2019)
8. Unity Documentation. NavMesh Surface. https://docs.unity3d.com/560/Documentation/Manual/class-NavMeshSurface.html. Accessed 18 January 2024
9. Shrestha, S., Mukhiya, S., Bhandari, B., Mishra, P. K., Mohd, A. W.: Indoor navigation using augmented reality. NeuroQuantology **20**(20), 2977–2983 (2022)
10. Bernstein, P.L., Mickie, S.: A game of cat and house: spatial patterns and behavior of 14 domestic cats (Felis Catus) in the home. Anthrozoös. **9**(1), 25–39 (1996)
11. Hammady, R., Ma, M., Strathearn, C.: User experience design for mixed reality: a case study of HoloLens in museum. Int. J. Technol. Mark. **13**(3–4), 354–375 (2020)
12. Ong, S.: Using spatial mapping. In: Beginning Windows Mixed Reality Programming. Apress (2017)

Design and Experimental Study of Automatic Phase Adjustment System for Combined Filter Rods Based on Visual Detection

Changfeng Qin, Liang Han, Yingze Lin, Yangzhen Gao, Fei Lu, and Shuaishuai Fan

Abstract In order to improve the product quality of combined filter rods and reduce the phase offset problem of combined filter rods. This paper takes the combined filter rod machine as the research object, and designs a combined filter rods phase visual detection and automatic adjustment system based on visual detection technology. The system includes three modules: filter rod phase detection, camera calibration and automatic phase adjustment. The edge detection operator is used to determine the edge position of the two ends of the filter rod, and the gradient calculation is used to obtain the mutation region of the combined filter rod image, so as to calculate the phase length of the two ends of the combined filter rod and achieve automatic adjustment according to the process requirements of the combined filter rod and the actual phase deviation. According to the experimental results, after the automatic adjustment system is turned on in the combined filter rod machine, the failure rate of the products is reduced by about 37% over the same period of time, and the average value of phase deviation is reduced by about 33% over the same period of time, which achieves the expected goal of automatic adjustment of phase of combined filter rods, and effectively improves the product quality of combined filter rods.

Keywords Combined filter rods · Phase offset · Visual detection · Autonomous adjustment

1 Introduction

In recent years, with the diversified development of filter rod products, grooved filter rods, empty tube filter rods, spice filter rods [1] and so on have become the main products in the market. The application of these special filter rods on cigarettes is mainly

C. Qin · L. Han (✉) · Y. Lin · Y. Gao · S. Fan
School of Mechanical Engineering, Southeast University, Nanjing, China
e-mail: melhan@seu.edu.cn

C. Qin · F. Lu
Nantong Cigarette Filter Co., Ltd., Nantong, China

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_6

in the form of combined filter rods, thus forming a variety of combined filter rod product styles commonly found in the market. However, in the process of producing combined filter rods on the combined filter rod machine, two or more different sizes of rods are cut into segments and then transferred to the conveyor mechanism through chain transmission and dial header. Between each two sections of the material rod section according to the specifications of the demand for alternating placement, transported to the smoking gun rolled into shape, after gluing adhesive, iron heating, knife cutting to form an equal length of the combined filter rods. However, due to mechanical gain, servo control zero position repeatability, knife box cutter position deviation and other problems, will lead to the production of combined filter rods appear part of the quality problems, such as the appearance of phase offset, the gap error, the alignment error, the length of the rods is wrong, the number of segments is wrong and so on, which is not in line with the production process of the defects [2]. These flaws can result in compromising the user vaping experience. Therefore, it has become an important research direction for the combined filter rod production equipment to increase the phase detection and phase automatic adjustment function, so as to ensure the qualification rate and reliability of the production of combined filter rods.

2 Related Works

In order to improve the production qualification rate of combined filter rods and reduce the labor rate of human detection. At present, scholars at home and abroad have developed and researched different ways of detecting algorithms for various defects produced by combined filter rods [3]. Yin Xiangyun [4] used the machine vision technology, to develop a measuring system to measure the geometrical parameters of the equipment to measure the cross-section of filter rods to meet the requirements of speed and accuracy. Guo Yasheng [5] proposed a machine vision based dimensional detection method for cylindrical rollers, using images of cylindrical rollers to achieve simultaneous detection of cylindrical roller diameter and length. It can be seen that machine vision technology plays an important role in dimensional measurement applications [6], but at present the application of visual detection and dimensional measurement to the defect detection and closed-loop control of combined filter rods is still a technology gap. Therefore, the main work and innovation points of this paper include: a. Applying machine vision to phase detection of combined filter rods, completing the design and implementation of algorithms for phase visual detection; b. Researching and applying algorithms for closed-loop control of visual detection and phase adjustment, realizing real-time automatic control of combined filter rods' phases, and advancing the application of intelligent manufacturing in the process of combined filter rods' production.

3 Combined Filter Rods Phase Visual Detection Subsystem

The combined filter rods phase visual detection system is programmed and developed by Visual Studio 2019 software through C++ language, and based on servo motion control to achieve phase autonomous adjustment, the system flow block diagram is shown in Fig. 1.

3.1 Combined Filter Rod Image Preprocessing

The combined filter rod video stream image captured by the camera is a continuous sequence of RGB colour images as shown in Fig. 2.

Grayscale Processing. As the camera acquires images, it is easily affected by the light source causing low contrast in the acquired images. The contrast is stretched in order to improve the dynamic range of grey level during image processing [7]. Assuming that the original image pixel coordinates are (x, y), the corresponding grey value is $f(x, y)$, the range of grey scale values is $[a, b]$; The grey scale value of the image after linear transformation of grey scale is $g(x, y)$, the range of variation of grey values after the transformation is $[c, d]$. Then the mathematical expression for the grey scale linear transformation is shown in Eq. (1).

$$g(x,y) = \begin{cases} c & c \le f(x,y) < a \\ \dfrac{d-c}{b-a}[f(x,y)-a]+c & a \le f(x,y) < b \\ d & b \le f(x,y) \le M \end{cases} \tag{1}$$

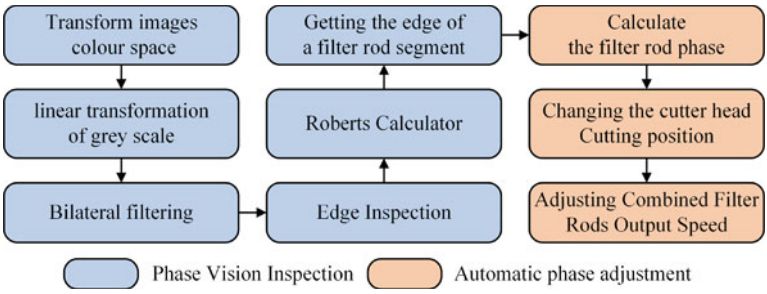


Fig. 1 Flow chart of phase detection and automatic adjustment

Fig. 2 Original diagram of combined filter rod

where M is the maximum grey value. Based on the grey scale characteristics of the image itself, the values of a, b, c, and d are adjusted to get a better contrast in the image. After several comparisons it was found that setting a = 100, b = 227.5, c = 0, d = 255.

Image Filtering. Image filtering refers to the method of removing unimportant content in an image and making the content of interest appear more clearly, such as removing noise from an image and extracting certain information from an image. The bilateral filtering algorithm used in this paper can achieve the removal of noise and smoothing of local edges on the basis of retaining regional information. Bilateral filtering plays a role in smoothing the high frequency fluctuation signals, while retaining the signal fluctuations of large variations, and thus achieve the role of retaining the edge information in the image [8]. The mathematical expression of the bilateral filtering principle is shown in Eq. (2).

$$g(i,\ j) = \frac{\sum_{k,l} f(k,l)\omega(i,\ j,\ k,\ l)}{\sum_{k,l} \omega(i,\ j,\ k,\ l)} \tag{2}$$

where $\omega(i,\ j,\ k,\ l)$ is the weighting factor. The bilateralFilter function in OpenCV is used to realize the bilateral filter function, and the parameters are set to d = 7, sigmaColor = 10, sigmaSpace = 20, respectively.

Image Edge Detection. Since each segment of the combined filter rod is made of a different material, there is a difference in the grey value of each segment when illuminated by a white light source, so the boundary line of the combined segment is located based on the grey mutation at the junction of the combined segment. The purpose of image edge detection is to better separate the combined filter *rod* from the surrounding background, so as to find out the image edge of the combined filter *rod* and measure the size of each segment of the base *rod* in the combined filter *rod*. The common edge detection algorithms are Canny operator [9], Sobel operator [10], Robert operator, Prewitt operator, etc. Sobel operator is an edge detection operator that finds the edge of the image by discrete differentiation method. This operator method is computationally simple, fast, better noise immunity and has good handling of grey scale gradients, so Sobel operator method is used to detect the edges of the image.

Modulus of the gradient vector of the image f(x, y) at position (x, y), ∇f is represented as shown in Eq. (3).

$$\nabla f = mag(\nabla \mathbf{f}) = \left[G_x^2 + G_y^2\right]^{1/2} \approx |G_x| + \left|G_y\right| \tag{3}$$

where G_x and G_y represent the first order partial derivatives in the image $f(x, y)$, for the x and y directions respectively. Subsequently for the Sobel edge detection operator processed image to do image binarization, set the grey scale threshold T to judge the combined filter rod edge points. In order to more accurately locate the combined filter rod edge features, the gradient in the diagonal direction, i.e., the Roberts operator [11], also needs to be computed.

3.2 Combined Filter Rod Phase Detection Principle

In order to accurately detect the phase size of the combined filter rod, the method used in this paper is to calibrate the camera with a standard-size rectangular calibration block with a length of 15 mm and a vertical side angle of 0°, which has a size error of no more than 1 μm.

When the angle of the standard block is detected, the data is valid, then the data is recorded as *Test_distance*, expressed as the number of pixels occupied by the calibration block. The resolution of the CMOS camera used in this paper is 3072 × 2048, through Eq. (4).

$$\frac{\text{Test\_distance}}{3072} = \frac{15}{\text{Window\_size}} \tag{4}$$

where *Window_size* is the actual lateral size in the spatial coordinate system corresponding to the image captured by the camera, unit: mm; *Test_distance* is the number of pixels occupied by the calibration block in the image obtained by the camera.

The lateral real size *Window_size* of the camera's field of view can be calculated. finally in detecting the combined filter rods according to Eq. (5).

$$\frac{\text{Window\_size}}{3072} = \frac{\text{Real\_Length}}{\text{Distance}} \tag{5}$$

where *Real_Length* is the actual size of the combined filter rod, unit: mm, and *Distance* is the number of pixels occupied by each segment of the combined filter rod. The phase detection schematic of the combined filter rod is shown in Fig. 3. 1063 means that the pixel size of the whole shooting frame occupied by the whole rod is 1063 pixels. $L_{all} = 120.01$ mm, $m_0 = 19.64$ mm, $m_1 = 20.21$ mm, $w = 40.08$ mm, $n_1 = 20.43$ mm, $n_0 = 19.64$ mm. Assuming that the permissible error in the length of each section of the filter rod is ± b. The system measures the length of each section of the filter rod as m_0, m_1, w, n_1 and n_0.

4 Experimental Study

Based on the combined filter rods phase visual detection system introduced in the previous chapter, this chapter will carry out the program design, experimental platform hardware construction, and visual parameter debugging of the combined filter

Fig. 3 Schematic diagram of phase detection of combined filter rod

rod phase automatic adjustment system. The experimental results are analyzed to verify the feasibility and reliability of the combined filter rod phase automatic detection and adjustment system.

4.1 Experimental Platform Construction

Set the allowable error of each section of the combined filter rod length for ± b, the system measures the length of the two ends of the filter rod as m_0 (left end) and n_0 (right end), the specific workflow of the phase automatic adjustment system is shown in Fig. 4.

Combined with the principle of combined filter rod phase detection, the combined filter rod phase automatic detection hardware platform is shown in Fig. 5. The hardware platform consists of cigarettes separation wheel, detection wheel, rejection wheel, transmission wheel, industrial camera and phase adjustment motor and other components. The combined filter rods are passed from the cigarettes separation wheel to the detection wheel, and the industrial camera is installed right above the detection wheel and shoots and detects the combined filter rods one by one. The unqualified combined filter rods are rejected on the reject wheel by controlling the solenoid valve to generate the reject air pressure; the qualified filter rods are transferred to the conveyor belt through the transfer wheel.

The test equipment selected for this test is a conventional combined filter rod forming machine to verify the combined filter rod phase automatic adjustment system

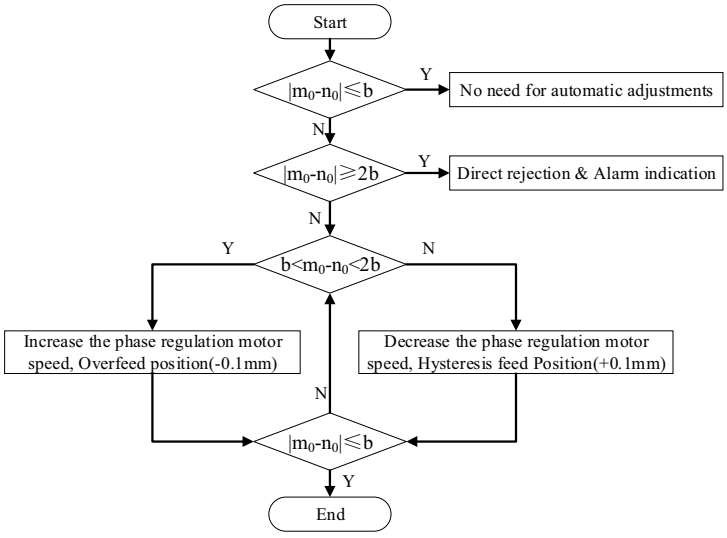


Fig. 4 Flow chart of automatic adjustment

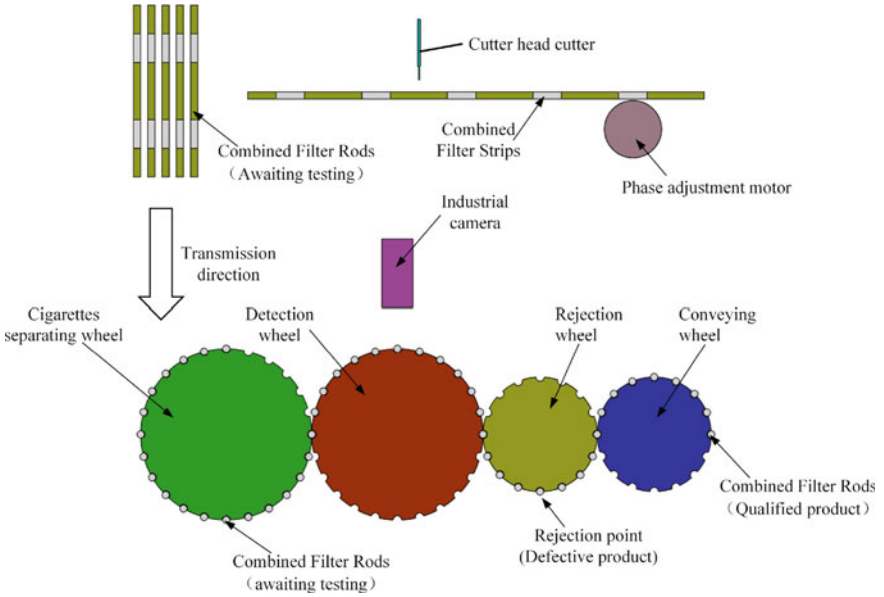


Fig. 5 Experimental platform construction diagram

based on visual detection. The specification of the combined filter rod for this test is DC16.70 mm × 120 mm × 3000 Pa (circumference, length and suction resistance), and the speed of the production vehicle is 120 m/min.

According to the product quality requirements, the length of the combined filter rod machine at both ends is set to be 20.00 mm in the parameter page of the detection software, and the permissible error of phase deviation is ± 1.0 mm. The allowable error of phase deviation is ± 1.0 mm, the phase deviation exceeding this range will be rejected, and the phase will be automatically adjusted when the average value of phase deviation of 10 consecutive combined filter rods is more than 0.5 mm. The system parameters set the camera parameter exposure to 2500, and the accuracy is calibrated to 0.1129. In order to fully verify the accuracy of the combined filter rod phase automatic adjustment, this test respectively closes the automatic adjustment and opens the automatic adjustment system, and compares the unqualified number of detected branches with the average value of phase deviation under the condition of the same detection branches, so as to fully verify the application effect of the combined filter rod phase automatic adjustment system.

4.2 Experimental Data Analysis

As can be seen from Table 1, when the phase auto-adjustment system is turned on, when the phase deviation of the combined filter rod is detected to be beyond the

Table. 1 Comparison of unqualified rate of combined filter rods before and after operating

Operating conditions	Phase offsetting mean/mm	Defective product number /unit	Total number/unit	Defective rate (%)
OFF	0.67	645	121,867	0.53
OFF	0.69	1369	287,307	0.48
ON	0.40	606	195,521	0.31
ON	0.52	759	230,082	0.33

control requirements, the phase adjustment motor of the system carries out closed-loop adjustment until the phase meets the requirements. Under the same test conditions, after the equipment opens the phase self-adjustment function, the value of phase deviation of combined filter rods and the number of unqualified branches are significantly reduced, which verifies the validity of the combined filter rods phase self-adjustment method based on visual detection, in which the average value of phase deviation decreases by 33% from 0.68 to 0.46 mm and the rate of defective products decreases by 37% from 0.51 to 0.32%.

5 Conclusion

The combined filter rod phase automatic adjustment system realizes real-time detection of combined filter rod phase by adopting visual detection technology, and sends the detection results to the control system of the combined filter rod production equipment in real time to control the phase adjustment mechanism in real time to adjust the phase of the composite filter rod, so as to realize real-time closed-loop control of the visual detection of combined filter rod phase and phase adjustment, and to improve the precision of the phase, which has not been found to be similarly studied and applied in the combined filter rod production and manufacturing industry. Similar research applications have not been found in the combined filter rod manufacturing industry. The experimental results show that when the combined filter rod production equipment is in operation, the phase automatic adjustment system can effectively reduce the phase shift of the combined filter rod and lower the rate of defective products caused by the phase shift.

References

1. Ai, Y.: Study on classification of special filter rods for tobacco. Sci. Technol. & Innov. **16**, 101–102 (2020). (in Chinese)
2. Shang, Y.: Study on the correlation between filter rod storage environmental conditions and key indexes of filter rod. Light. Ind. Sci. Technol. **10**, 79–82 (2021). (in Chinese)
3. Guanhua Wang, Guoxian Xu, Zhongxiang Chen.: The detection system of midline glue based on computer vision. Plant Maint. Eng. 17:79-80. (in Chinese) (2021)
4. Yin, X., et al.: Automatic measurement system of section geometry parameters for composite filter rods based on machine vision. IEEE Trans. Instrum. Meas.Instrum. Meas. **69**, 5037–5050 (2020)
5. Guo, Y., Zhang, S., Zhang, A.: Cylindrical roller size detection method based on machine vision. Mod. Manuf. Eng. **04**, 109–113 (2021). (in Chinese)
6. Liu, Z., Yang, G., Tang, W.: Research on Multi-type workpiece measurement system based on machine vision. Mach. Tools & Hydraul. **50**(04), 6–12 (2022). (in Chinese)
7. Gonzalez, R.C., et al.: Digital image processing, second edition. Publ. House Electron. Industry. (2007)
8. Zhen Feng, Yanning Guo, Yueyong Lv.: Open CV4 Quick Start. Posts & Telecom Press.. (in Chinese) (2021).
9. Li, J., et al.: Improved canny algorithm for image edge enhancement. J. Geomat.S Sci. Technol. **38**(04), 398–403 (2021). (in Chinese)
10. Li, Y., et al.: Study on edge detection algorithm of urine test strip combined with Sobel operator and morphology. J. Biomed. Eng. Research. **38**(01), 43–47 (2019). (in Chinese)
11. Lin, T., Han, Y., Fan, S.: Analysis of Roberts edge detection method. Arch. Eng. Technol. Design. **36**, 2380–2382 (2017). (in Chinese)

Multi-Sensor SLAM Assisted by 2D LiDAR Line Features

Zhanhong Shi, Ping Wang, Wanquan Liu, and Chenqiang Gao

Abstract In the domain of indoor localization, visual SLAM has gained prominence as a popular approach. However, visual information is frequently limited by feature degradation, resulting in diminished accuracy or location completely lost. To address this challenge, IMU and expensive 3D LiDAR are conventionally employed as solutions. Considering the low cost, we propose a multi-sensor fusion system with a camera, IMU, and 2D LiDAR. Firstly, we extract straight-line features from 2D LiDAR by Random Sample Consensus (RANSAC). Secondly, in scenarios where visual features degrade, the real-time pose will be performed based on the 2D LiDAR and IMU. Thirdly, in instances where visual features remain adequate, we further enhance the accuracy of pose estimation by the 2D LiDAR straight-line features. The experimental results demonstrate that our method has reliable accuracy and robustness in situations where visual features are degraded. Furthermore, our proposed method outperforms both the visual method and the visual-inertial method in terms of accuracy, particularly within indoor environments such as in rooms and corridors.

Keywords Pose estimation · Multi-sensor fusion · 2D LiDAR · Straight-Line features

1 Introduction

Traditional simultaneous localization and mapping (SLAM) algorithms based on a single-sensor have matured, but they always have limitations. When facing some special environments, a single-sensor system is likely to experience a decrease in the accuracy of localization or even get lost. Therefore, a multi-sensor system is an inevitable development trend in the future [1].

In indoor environments, rooms and corridors are the most common scenes, both of which have structural features, such as walls. Cameras often face problems such as lack of texture information when encountering white walls, which can lead to

Z. Shi · P. Wang · W. Liu (✉) · C. Gao
Shenzhen Campus of Sun Yat-Sen University, Shenzhen, Guangdong 518107, P.R. China
e-mail: liuwq63@mail.sysu.edu.cn

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_7

unreliable visual feature points or even unable to extract feature points. Some methods try to use a multi-camera setup to expand the perspective views to obtain more visual features [2, 3]. In addition, the fusion of the camera and IMU is also a common solution [4, 5], which can effectively improve the robustness during rapid turns or spins. However, when the camera fails for a long time, it is difficult to track the pose by IMU alone. Therefore, the most robust solution is combining LiDAR with the camera. Some loosely coupled methods simply combine the pose of the camera and LiDAR [6, 7]. Although they reduce the complexity of algorithm design, they are not as accurate as tightly coupled methods which have unified the information of different sensors for joint optimization [8–10].

Currently, mature camera, IMU, and LiDAR fusion systems mostly use 3D LiDAR. The more lines a 3D LiDAR has, the higher the price, which can reach hundreds of thousands of yuan, while a 2D LiDAR costs only tens of thousands of yuan. In indoor scenes, robots mostly move on a 2D plane, and based on visual information, 2D LiDAR is sufficient to assist in localization. Furthermore, 2D LiDAR has a small amount of point cloud data and consumes less computational resources to process it. Therefore, compared to 3D LiDAR, low-cost 2D LiDAR fusion with a camera and IMU is a suitable solution in indoor localization. Different from the loosely coupled method [11], we will conduct pose estimation with a tightly coupled approach by joint nonlinear optimization based on 2D LiDAR, camera, and IMU for the ORB-SLAM3 system [12]. Compared to other tightly coupled methods of directly using 2D LiDAR point cloud [13, 14], we combine visual point features with 2D LiDAR straight-line features which have advantages over point cloud in a structured environment. Visual features tend to degenerate in a structured environment, such as white walls, so the line features from the 2D LiDAR have complementarity with visual feature points.

The main contributions of the proposed method are as follows.

- A multi-sensor fusion algorithm that combines low-cost 2D LiDAR with camera and IMU is proposed. It enables real-time pose estimation through alternative sensors in the event of features missing, which enhances the system's robustness.
- Our method significantly reduces the accumulation of lateral errors caused by the robot's long-term forward movement in structured environments through the 2D LiDAR straight-line features.
- Our method combines visual point features and 2D LiDAR straight-line features. This fusion enhances pose accuracy in both textured environments, suitable for camera-based methods, and structured environments with less texture conducive to 2D LiDAR.

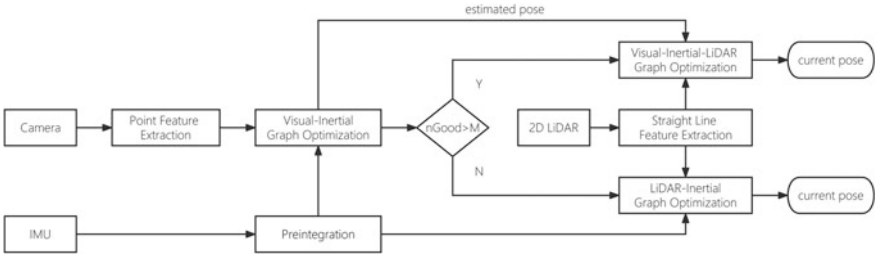


Fig. 1 The Framework of our system

2 The Proposed Method

2.1 System Overview

As shown in Fig. 1, firstly, one obtains the camera estimated pose by visual-inertial odometry and the number of visual feature points (nGood) that can be successfully matched with the visual-inertial odometry. Secondly, when visual features degrade, LiDAR-inertial odometry can combine 2D LiDAR and IMU to track the current pose. Thirdly, in the case of sufficient visual features, the system will further perform tightly coupled nonlinear optimization of 2D LiDAR straight-line features and the camera estimated pose to obtain a more accurate current pose.

2.2 Residuals

The straight-line feature on the 2D LiDAR plane is defined by slope and intercept (k, b). For each line, we extract minP representative points, where minP represents the minimum number of points to ensure that at least two points can be successfully matched as much as possible. $p_{i,q}$ represents the q-th representative point of the i-th frame. Then, we match the nearest point $p_{j,nearest[q]}$ in previous j-th frame corresponding to $p_{i,q}$ by the kd-tree algorithm, and compare the lines which they belong to and round them out if they differ greatly. According to the 2D LiDAR pose $T_{L_i} = \left(R_{L_i}, t_{L_i}\right)$ of the i-th frame, points are converted to the world coordinate system. Similarly, we can obtain (k_q, b_q) of the line corresponding to $p_{j,nearest[q]}$ in the world coordinate system. We calculate the residual between the i-th frame and the previous j-th frame by the distance from point $p_{i,q}$ to line (k_q, b_q) on the 2D LiDAR plane as shown in Eq. (1), where $n = (k_q, -1, 0)$.

$$r_{L_{i,j}} = \left[n(R_{L_i} p_{i,q} + t_{L_i}) + b_q\right] \Big/ \sqrt{k_q^2 + 1} \tag{1}$$

Through bundle adjustment (BA) optimization, the estimated pose T_{O_i} can be obtained by ORB-SLAM3. T_{O_i} is a transformation matrix that represents the camera pose of the *i*-th frame. Therefore, by the transformation matrix T_C^L between the camera and 2D LiDAR, we can convert T_{O_i} to the 2D LiDAR coordinate system. The residual between the camera estimated pose and the 2D LiDAR pose from the previous *j*-th frame to the *i*-th frame can be expressed as Eq. (2), where $\log(\bullet)$ represents a logarithmic map that converts the transformation matrix into Lie algebraic form, returning a six-dimensional vector.

$$r_{O_{i,j}} = \log\left(\left(\left(T_{O_j}T_C^L\right)^{-1}\left(T_{O_i}T_C^L\right)\right)^{-1}\left(T_{L_j}^{-1}T_{L_i}\right)\right) = \log\left(\left(T_C^L\right)^{-1}T_{O_i}^{-1}T_{O_j}T_C^L T_{L_j}^{-1}T_{L_i}\right) \quad (2)$$

2.3 Nonlinear Optimization Model

We count the number of visual feature points (nGood) with errors less than a certain threshold. If nGood is less than parameter M, the pose of visual-inertial odometry will be discarded, and LiDAR-inertial odometry is used to get the pose estimation. We use adjacent frames to estimate the LiDAR-inertial odometry. The LiDAR-inertial model is shown in Eq. (3), where the set L represents all representative points extracted from line features of the *i*-th frame. $r_{I_{i,i-1}}$ is IMU residual including the rotation, speed, and position errors of IMU from the *i-1*-th frame to the *i*-th frame [12].

$$T_{L_i}^* = \arg\min_{T_{L_i}}\left(\sum_{q\in L}\left\|r_{L_{i,i-1}}\right\|^2 + \left\|r_{I_{i,i-1}}\right\|^2\right) \quad (3)$$

If nGood is greater than M, the estimated pose T_{O_i} from visual-inertial odometry should be reliable but maybe not sufficently accurate, and in this case, it will be further optimized with residual $r_{L_{i,j}}$ from the 2D LiDAR. In the case of no degradation of visual features, the pose of the keyframe is optimized by the global optimization in ORB-SLAM3 and has high credibility, so we will optimize the current pose based on the last keyframe. The joint optimization model for camera, IMU, and 2D LiDAR is shown in Eq. (4), where the *j*-th frame is the last keyframe of the *i*-th frame.

$$T_{L_i}^* = \arg\min_{T_{L_i}}\left(\sum_{q\in L}\left\|r_{L_{i,j}}\right\|^2 + \left\|r_{O_{i,j}}\right\|^2\right) \quad (4)$$

3 Experiment

All experiments are based on the Intel Realsense D455 camera and the Hokuyo UST-10LX LiDAR. The IMU information can be obtained from the D455 camera. We mainly compare and analyze our method with the visual method and visual-inertial method in ORB-SLAM3, and conduct comparative experiments in a long corridor, room 1, and room 2 to verify the accuracy and robustness of our method. Due to the page limitation, we only included representative figures and tables.

3.1 Trajectories Comparison

In Room 1, as shown in Fig. 2, there is no visual texture feature in the red area where the robot will face a white wall. The trajectories obtained by different methods: (a) The localization is directly lost in the red area; (b) The pose estimation is performed based on IMU in the red area, resulting in significant errors and ultimately getting lost; (c) Our method maintains accurate real-time pose estimation.

In the long corridor, visual texture features are few. When the robot moves along the corridor, the 2D LiDAR point cloud will be similar. Therefore, the corridor is not conducive to using a camera or 2D LiDAR alone. The results are shown in Fig. 3: (a) Once a turn is made, the positioning is lost and recalculated from the origin; (b) The first turn in the red circle can successfully maintain positioning, but the second turn in the green circle still gets lost; (c) We specifically did not allow the robot to move a complete circle, verifying that at both turns, one can successfully track real-time pose without loop detection.

Experimental results show that our algorithm has good robustness, and can continue to maintain real-time pose estimation by 2D LiDAR and IMU when visual features degrade.

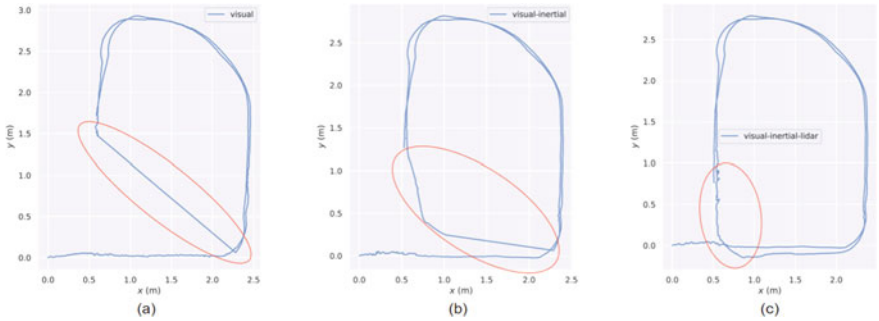


Fig. 2 Trajectory in room 1 with fewer texture features. **a** ORB-SLAM3(visual); **b** ORB-SLAM3(visual-inertial); **c** Our method

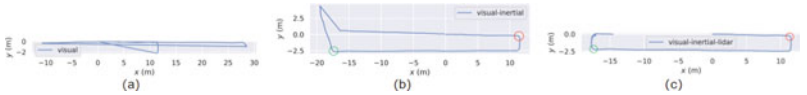


Fig. 3 Trajectory in the long corridor. **a** ORB-SLAM3(visual); **b** ORB-SLAM3(visual-inertial); **c** Our method

3.2 Pose Estimation Accuracy Comparison

As shown in Fig. 3, we made two marker points, the red and green ones. The robot will stay for a short time at marker points to record the pose. Due to similar environments, 2D LiDAR may generate significant errors in the x-axis direction. As shown in Table 1, the error on the x-axis is not much different from the visual-inertial method because we only extracted straight-line features that are easy to match. Through the constraints of the straight-line features, our method's error on the y-axis has reduced by 3.5 cm at the red marker, while at the green marker, the error is reduced by up to 40 cm.

In Room 2 with rich features, the robot smoothly circled three times and we calculated the absolute trajectory error (ATE) and relative pose error (RPE) [15]. As shown in Table 2, the visual method's accuracy is the lowest. Our method has improved accuracy compared to the visual-inertial method, with approximately 0.5% improvement in ATE and 0.3% improvement in RPE.

Our method significantly reduced the cumulative lateral error of the robot's long-term forward movement in a structured environment through 2D LiDAR line features. In addition, in scenes with more visual features that are completely favorable to ORB-SLAM3, our system can also perceive some structured information, further improving localization accuracy.

Table 1 Comparison of errors(m) between our method and ORB-SLAM3 on marker points

	ORB-SLAM3(visual-inertial)	Our method
Red marker point(x)	0.383430	0.391267
Red marker point(y)	0.200967	**0.165035**
Green marker point(x)	0.732422	0.744195
Green marker point(y)	0.706182	**0.303863**

Table 2 Comparison of errors(m) between our method and ORB-SLAM3 in room 2

	ORB-SLAM3(visual)	ORB-SLAM3(visual-inertial)	Our method
Mean ATE	0.204398	0.106928	**0.106704**
Rmse ATE	0.297788	0.112492	**0.111892**
Mean RPE	0.023392	0.023114	**0.020923**
Rmse RPE	0.053995	0.052558	**0.052401**

4 Conclusion

We propose a multi-sensor fusion system based on ORB-SLAM3, which is tightly coupled with a camera, IMU, and 2D LiDAR. Our method addresses the challenge of difficulty in relying on cameras for pose estimation in indoor environments with few visual features by a low-cost 2D LiDAR instead of an expensive 3D LiDAR. It optimizes visual point features and 2D LiDAR straight-line features to achieve enhanced pose accuracy. The experimental results confirm that our method ensures robust positioning, regardless of whether visual features degrade or not. Moreover, it significantly reduces lateral errors in structured environments, and outperforms the state-of-the-art ORB-SLAM3 in texture environments. Future research will explore the fusion of 3D LiDAR, camera, and IMU, which aims to verify that our proposed method in this paper can offer reliable accuracy with a reduced system cost relative to 3D LiDAR.

References

1. Chen, W., Zhou, C., Shang, G., Wang, X., Li, Z., Xu, C., Hu, K.: SLAM overview: from single sensor to heterogeneous fusion. Remote Sensing **14**(23), 6033 (2022)
2. Urban, S., Hinz, S.: Multicol-slam-a modular real-time multi-camera slam system. arXiv preprint arXiv:1610.07336 (2016)
3. Won, C., Seok, H., Cui, Z., Pollefeys, M., Lim, J.: Omni SLAM: Omnidirectional localization and dense mapping for wide-baseline multi-camera systems. In: IEEE International Conference on Robotics and Automation (ICRA) 2020, pp. 559–566. IEEE (2020)
4. Qin, T., Pan, J., Cao, S., Shen, S.: A general optimization-based framework for local odometry estimation with multiple sensors. arXiv preprint arXiv:1901.03638 (2019)
5. Von Stumberg, L., Cremers, D.: Dm-vio: Delayed marginalization visual-inertial odometry. IEEE Robotics and Automation Letters **7**(2), 1408–1415 (2022)
6. Zhang, J., Singh, S.: Laser-visual-inertial odometry and mapping with high robustness and low drift. Journal of Field Robotics **35**(8), 1242–1264 (2018)
7. Zhao, S., Zhang, H., Wang, P., Nogueira, L., Scherer, S.: Super odometry: IMU-centric LiDAR-visual-inertial estimator for challenging environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2021, pp. 8729–8736. IEEE (2021)
8. Shan, T., Englot, B., Ratti, C., Rus, D.: Lvi-sam: Tightly-coupled lidar-visual-inertial odometry via smoothing and mapping. In: IEEE International Conference on Robotics and Automation (ICRA) 2021, pp 5692–5698. IEEE (2021)
9. Lin, J., Zhang, F.: R3LIVE: A Robust, Real-time, RGB-colored, LiDAR-Inertial-Visual tightly-coupled state Estimation and mapping package. In: International Conference on Robotics and Automation (ICRA) 2022, pp. 10672–10678. IEEE (2022)
10. Shao, W., Vijayarangan, S., Li, C., Kantor, G.: Stereo visual inertial lidar simultaneous localization and mapping. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2019, pp. 370–377. IEEE (2019)
11. Chan, S. H., Wu, P. T., Fu, L. C.: Robust 2D indoor localization through laser SLAM and visual SLAM fusion. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC) 2018, pp. 1263–1268. IEEE (2018)
12. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. IEEE Trans. Rob. **37**(6), 1874–1890 (2021)

13. Mu, L., Yao, P., Zheng, Y., Chen, K., Wang, F., Qi, N.: Research on SLAM algorithm of mobile robot based on the fusion of 2D LiDAR and depth camera. IEEE Access **8**, 157628–157642 (2020)
14. Peng, G., Zhou, Y., Hu, L., Xiao, L., Sun, Z., Wu, Z., Zhu, X.: VILO SLAM: Tightly Coupled Binocular Vision-Inertial SLAM Combined with LiDAR. Sensors **23**(10), 4588 (2023)
15. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of RGB-D SLAM systems. In: IEEE/RSJ international conference on intelligent robots and systems 2012, pp. 573–580. IEEE (2012)

Transforming Healthcare with Immersive Visualization: An Analysis of Virtual and Holographic Health Information Platforms

Z. YongQi, S. Chan-Bormei, and H. Miri

Abstract Advancements in technology have opened new possibilities for revolutionizing healthcare systems. One such emerging concept is the use of virtual and holographic health information platforms that aim to provide interactive and personalized medical information to users. This paper highlights the need for information visualization and 3D representation. It proceeds to provide background knowledge on information visualization and historical developments in the 3D visualization technology. Additional domain knowledge concerning holography, holographic computing, and mixed reality are then introduced, followed by highlighting some of their common applications and use-cases. The discussion then focuses on the importance of virtual and holographic visualization in medicine, detailing current research areas and applications in digital holography and its role in medical genetics and genomics in particular. The principles and concepts underlying virtual and holographic health information systems, as well as their potential healthcare implications, are subsequently analyzed. The paper concludes by examining some of the notable mixed reality applications and systems that aid doctors in visualizing diagnostic and genetic data, as well as assist in enhancing patient education and communication. This study serves as a valuable resource for researchers, developers, and healthcare professionals exploring virtual and holographic technologies for healthcare improvement.

Keywords Virtual reality · Augmented reality · Mixed reality · Holography · Information visualization

Z. YongQi · S. Chan-Bormei · H. Miri (✉)
AI & Computer Engineering, CMKL University, Bangkok, Thailand
e-mail: miri@cmkl.ac.th

Z. YongQi
e-mail: yzhou@cmkl.ac.th

S. Chan-Bormei
e-mail: csuy@cmkl.ac.th

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_8

1 Introduction

The development of advanced technologies and innovative solutions has opened up exciting new possibilities for revolutionizing healthcare systems. One such emerging concept is the use of *virtual and holographic health information platforms* that aim to provide *interactive and personalized medical information* to users. This paper begins by highlighting the need for information visualization and 3D representation. It then proceeds to provide background knowledge on information visualization and historical developments in the 3D visualization technology. Additional domain knowledge concerning holography, holographic computing, and mixed reality are then presented, followed by highlighting some of their common applications and use-cases. After setting the scene and defining the context, the need for and importance of virtual and holographic visualization in *medicine* is discussed. Subsequently, some of the current research areas and applications of digital holography and holographic technology are explored, alongside the importance and role of virtual and holographic visualization in *genetics and genomics*. An analysis of the key principles and concepts underlying virtual and holographic health information systems is presented, as well as their potential implications for healthcare is pointed out. The paper concludes by examining the most notable existing mixed-reality applications and systems that help doctors *visualize diagnostic and genetic data*, and assist in *patient education and communication,* followed by discussing some implications and providing conclusions.

2 Background

There has been a remarkable transformation in mankind's means of communication throughout history. From the early days, where communication was conveyed through cave paintings, to the development of intricate human languages, we have continually evolved our methods of expression. However, in the modern era, with the rise of the digital age and the exponential growth of data-driven decisions, there is a need for a medium to effectively represent, analyze, and interpret this vast amount of data. For a long time, 2D visualization has served this purpose. However, given that the human brain is inherently a three-dimensional processor and intuitively understands and navigates the world in 3D, there is a growing need for a new form of *visualization* that offers more depth and allows for a more intuitive, interactive, and comprehensive representation of data: *3D visualization.*

2.1 The Need for Information Visualization and 3D Representation

"*A picture is worth a thousand words.*" This adage is reflective of the numerous studies conducted on visual perception [1] which demonstrates that vision is our most prominent sensory modality. In the face of information overload, data often lacks clear visualization. Information visualization seeks to enhance understanding by enabling knowledge acquisition through sight, facilitating personal cognition, and supporting long-term memory retention or interactive analysis for problem-solving. The representation of information plays a crucial role in shaping comprehension, influencing individual behavior and reactions. The increasing demand for innovative visual representation has led to the widespread adoption of Stereoscopic 3-Dimensional (S3D) technology. Rooted in the quest for a more *spatial and immersive viewing experience*, S3D enhances understanding by accurately depicting depth and spatial relationships between objects. Unlike traditional 2D methods, this technology mirrors real-life perception, aligning with our natural 3D environment and the capabilities of binocular vision. The transition to 3D is driven by a desire for a more engaging and relatable experiences.

The extensive exploration of 3D technology across various domains has revealed manifold advantages over conventional 2D representation methods. For example, Amin et al. [2] investigated the impact of S3D technology on human behavior during learning and long-term memory, and found that individuals processed S3D content using widespread cortical networks, compared to 2D content. This property makes S3D technology highly promising for applications in education and advertising. Similarly, in industrial settings, 3D technology can effectively address the limitations of traditional 2D visualization. For instance, multi-dimensional data that is difficult to display in tables and 2D diagrams can be effectively represented in various forms using 3D technology. Additionally, 3D technology enables users to view the spatial structure of objects and to observe both the overall structure and local details, making it useful for analyzing hierarchical and complex objects. In the field of simulation, 3D systems provide a realistic representation of physical interactions, enabling a more immersive and user-friendly experience. In architecture and engineering design, architects and engineers use 3D models to explore the design of a building before it is constructed. This enables them to pinpoint potential issues and implement necessary adjustments prior to commencing construction, thereby streamlining the design process. Coupled with virtual reality and holographic technologies, such endeavors enable designers and developers to also test certain psychological aspects and features of future environments [3, 4]. Furthermore, it allows product managers and clients to get a better understanding of the design and make informed decisions, improving communication by creating interactive prototypes. Moreover, 3D representation offers unique advantages in data storage [5], spatial structure display [6], innovative human–computer interaction [7], and artistic expression [8].

2.2 The History and Development of 3D Visualization Technology

The history and development of 3D visualization technology has undergone significant advancements since its inception in the Renaissance era. The first breakthrough in 3D visualization can be traced back to the fourteenth century when Italian architect and designer Filippo Brunelleschi painted the world's first painting using the concept of linear perspective, as seen in Fig. 1. Brunelleschi's discovery paved the way for the use of perspective drawing to create the illusion of depth on 2D surfaces. Static cues such as interposition, shading, projection, size reduction in the distance, and texture gradient were also developed and widely used to enhance the sense of 3D in paintings. These mathematical and cognitive-based principles allowed for the imagination and perception of stereoscopic space from 2D paintings.

The advent of computer graphics in the mid-twentieth century marked the beginning of a new era in 3D visualization. The term Computer Graphics (CG) was coined by William Fetter, a graphic designer for Boeing, in 1960 [9]. One of the earliest and most innovative computer graphic design systems, Sketchpad, was developed by Ivan Sutherland in 1963 and inspired by the TX-2 computer [10]. In the late 1960s, IBM released the first commercial graphic computer, the IBM 2250 [11]. This period also witnessed the emergence of early video games and animations.

With the advent of powerful computers and advanced graphics hardware, modern rendering software became increasingly sophisticated, capable of producing photo-realistic images. The concept of Computer-Aided Design (CAD) revolutionized industrial design and architecture with the development of powerful and widely used software such as AutoCAD, SolidWorks, 3DMax, among others. Additionally, Computer-Generated Imagery (CGI) technology became mainstream in gaming,

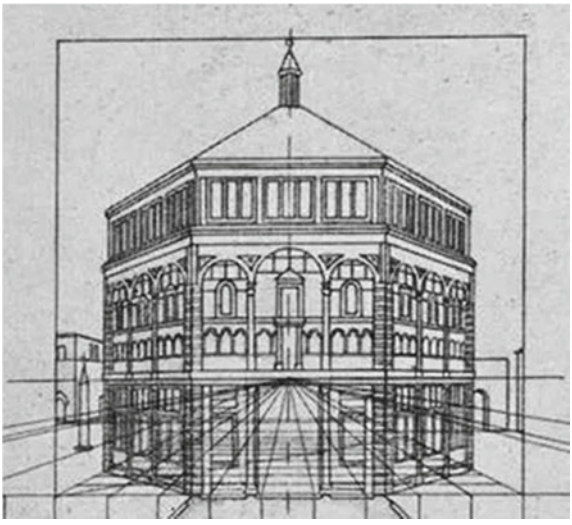


Fig. 1 Diagram demonstrating Filippo Brunelleschi's perspective technique from a lost painting of the Battistero di San Giovanni

multimedia, film, and animation. In the early 1990s, the first CGI TV series, La Vie Des Betes, was released in France. In 1995, Pixar released its first commercial animated film, Toy Story, which was a huge success both commercially and in the field of CG. The same decade saw the rise of 3D games including racing games, first-person shooter games, and fighting games such as Virtual Fighter, Tekken, and Super Mario, which captivated audiences with their interfaces and gaming experiences.

The field of 3D visualization has progressed considerably, with the introduction of Head-Mounted Display (HMD) equipment, like VR headsets, adding a new dimension. Entertainment VR devices often work seamlessly with established game engines, enabling the easy deployment of Unity 3D-developed first-person games on VR headsets like Oculus Rift and HTC Vive. Additionally, the integration of HMD interfaces into rendering and modeling software, such as Autodesk 3ds Max, has enabled architects, 3D artists, and designers to experience and interact with their visualized spaces in a more immersive way. In addition to VR technology, the products of AR and MR have also attracted significant investment and been introduced into the mass market. These new technologies provide a unique medium for exploring imaginary spaces and have the potential to greatly influence the development of new ideas. Despite these advancements, the field of 3D visualization is still in its early stages of development and has yet to reach its full potential in Thailand. One of the latest and most exciting developments in this field is non-glasses holographic visualization, which introduces a novel and natural way for displaying 3D visualizations and aims to provide a more immersive and interactive experience for users. Innovative companies like Zebra Imaging and Musion 3D are exploring holographic technology, suggesting the potential for holograms to become the primary visualization tool, akin to the ubiquity of PCs and 3D visualization software.

3 What is Holography?

A hologram is a 3D image produced from a pattern of interference created by a split coherent beam of radiation, such as a laser. It carries intensity, color, depth, and directional information of the original scene, and can reconstruct the corresponding true-light wavefronts. The technique of holography is used to perform imaging of multi-dimensional objects and makes use of hardware, software, and more exotic types of programming to create a hologram image. Simply put, it is a technique used to create 3D images for computing purposes. The theory of holography was developed by Hungarian physicist Dennis Gabor in 1947 to increase the resolving power of electron microscopes, and the invention of the laser allowed its prosperous development in various fields in the 1960s [12].

To create a hologram, a high-powered laser is split into two beams: an object beam (which bathes a 3D object) and a reference beam (which illuminates the imaging medium from a different angle). The hologram provides the brain with two images of an object from slightly different angles. When these images are displayed to each eye separately, the brain believes that it is seeing a 3D object. In reality, a hologram

acts like a complex lens through which each of the infinitesimally-small points on the object that reflected light onto the film (by interfering with the reference beam) are neatly focused to their respective positions in 3D space. Two researchers at the University of Michigan, Emmett Leith and Juris Upatnieks, attracted the attention of the popular press for their work on off-axis holography [13] which suggested the transition from describing holography as a visualization method to a means of transmitting information.

Advancement in the theory of information and computing opened the door for a new era in the field of holography. Soon after the off-axis solution to the twin image problem was proposed, it became apparent that either generating or displaying the hologram could be performed digitally rather than optically. Digital Holography (DH) which emerged in the early 1990s, revolutionized analogue processing methods and became an appealing research field. DH technology generally consists of two different stages: the hologram acquisition process, which converts the optical wavefront of an object sense into a digital hologram, and the hologram display and reconstruction technology, which involves displaying the digital hologram as a visually observable 3D optical image through high-resolution devices or other mediums [14].

Among these aspects, Computer Generated Holography (CGH) is particularly important and relevant to our work. CGH consists of three basic elements: the light source, the hologram, and the image. If any two of these elements are pre-determined, the third can be computed. For example, if we have a parallel beam of light of a certain wavelength and a "double-slit" system (a simple hologram), we can calculate the diffraction pattern. Likewise, knowing the diffraction pattern and the details of the double-slit system, we can calculate the wavelength of the light. Therefore, we can design any pattern we want to see. CGH is used to make holographic optical elements (HOE) for scanning, splitting, focusing, and, in general, controlling laser light in many optical devices, such as a common CD player [15].

4 Mixed Reality and Holographic Equipment and Manufacturers

In recent years, the VR/AR/MR market has seen a surge in the number of vendors and companies offering a diverse range of holographic devices and services catering to consumers, businesses, and developers. Industry leaders in gaming, computing, and Internet technologies such as Microsoft, Google, AMD, NVIDIA, and Samsung, have invested heavily in this technology, giving them a significant edge in the market. However, several start-ups, including Oculus VR, Next/Now, and Magic Leap, have also captured public attention, with some beginning with successful crowdfunding rounds. Below are some of the most popular companies in the VR/AR/MR space and the products they offer.

1. Oculus VR (Meta): Oculus, a well-known VR company, was acquired by Meta in 2014. They offer a range of VR headsets, including the Meta Quest 2, which is currently one of the most popular VR devices on the market. The Quest 2 provides an all-in-one VR experience, allowing users to play games, watch movies, and even work in virtual environments. In addition to the hardware, Oculus also offers a software platform for developers called Meta Quest for Business which enables businesses to create custom VR applications for their needs.
2. HTC: HTC offers the VIVE series headsets with the latest one being VIVE XR Elite which is an all-in-one XR headset that converts into a set of portable immersive glasses, boasting high-performance capabilities. The VIVE XR Elite boasts exceptional graphics and high-resolution passthrough in a compact form factor. HTC also holds a metaverse content ecosystem VIVERSE that aims to connect individuals and communities together and enable the creation and exploration of virtual worlds from any device.
3. Microsoft: Microsoft is a major player in the MR market with its HoloLens 2 headset. The HoloLens 2 is designed for use in enterprise environments, enabling users to view and interact with 3D holograms. To create holographic scenery and mixed-reality experience for HoloLens 2, developers can use popular development platforms such as Unity, Unreal, and Vuforia, and get built-in HoloLens 2 developer support.
4. Magic Leap: Magic Leap is an AR company that recently released the Magic Leap 2 headset. The headset projects virtual images onto the real world, creating an immersive and interactive experience. This series of devices focus on building enterprise AR solutions for manufacturing operations, aiming to benefit every facet of manufacturing.
5. Google: Google's exploration on VR/AR mainly focuses on the software area and affordable lightweight products. Google aims to bridge the physical and virtual worlds without introducing expensive devices. Google Cardboard, for example, provides a simple and fun way to get an immersive experience. Google's typical application includes Google Lens and live view in Google Maps. They also offer ARCore, a software platform that allows developers to create AR applications for Android devices.
6. Varjo: Varjo is a company that offers high-resolution VR and MR headsets for professionals. Their main products are Varjo XR-3, Varjo VR-3, and Varjo Aero headsets, offering industry-leading resolution and display technology. These devices are primarily used in industries such as aerospace, automotive, and architecture. Varjo also has software infrastructure such as Varjo Reality Cloud and Varjo Teleport.

5 What is Holographic Computing?

Holographic Computing is the Computer Science of working with *holograms.* It essentially refers to the use of *holography* and involves the generation and manipulation of *holographic representations of objects or data.*

Holograms are 3D optical objects that represent 3D physical objects. The HoloLens headset, together with the Universal Windows Platform, bring high-definition holograms to life. Here are a few examples of Holographic Computing methods:

- *Holographic Displays*: These display technologies project 3D holographic images, allowing users to interact with virtual objects as if they were present in the real world.
- *Holographic Modeling & Simulation*: This involves creating virtual models of physical objects or systems (i.e., holograms) for simulation and analysis purposes, enabling engineers and researchers to study and optimize designs in a realistic virtual environment.
- *Holographic Data Storage*: Holographic Computing can be used for data storage, utilizing the interference patterns of laser light to store and retrieve large amounts of information in 3D holograms.
- *Holographic Telepresence*: This method enables remote communication and by projecting realistic, life-sized holographic representations of individuals, creating a sense of presence and enhancing the remote interaction experience.

6 Common Applications and Use-Cases

Holographic technology has revolutionized various industries with its ability to create immersive experiences and convey information in an engaging way using holograms. Such innovation caters to two primary user categories: *professionals seeking expert solutions* and the *general public as end-users*. Here are some examples of how holograms have been used.

6.1 Expert-Based Users

6.1.1 Holographic Communication

In April 2017, the two largest operators, Verizon (USA) and Korea Telecom (South Korea), achieved a breakthrough in telecommunications technology by making the first international holographic call using 5G technology [16]. This cutting-edge technology enables the formation of holograms of the interlocutor, which can completely convey the emotions and gestures of the user. During the test, the

hologram of the interlocutor was reflected on the screen of the holographic device, creating a life-like representation of the person being called. It is worth noting that holographic communication technology requires a high bandwidth, making it only possible in 5G networks, which are 10 to 100 times faster than existing networks. This incredible speed and capacity make 5G the ideal platform for this type of innovative communication technology.

6.1.2 Spatial Navigation

Holograms have the potential to help in spatial navigation. In 2017, scientists from the Munich University of Technology developed a method of obtaining 3D holograms using a Wi-Fi router, which could help in creating copies of premises and locating trapped victims [16]. Singh et al. proposed the virtual global landmark in 2021 as a novel AR landmark technique to prevent people's orienting skill loss induced by current navigation support systems, such as Google Maps [17]. Likewise, Liu's team designed and implemented virtual semantic landmarks in MR-based indoor environments and conducted a user study to explore whether such landmarks can assist spatial knowledge acquisition during navigation [18]. Advancements in 3D holograms and virtual landmarks demonstrate hologram technology's potential in enhancing spatial awareness and revolutionizing rescue operations.

6.1.3 Virtual Training

The rise of technology has brought virtual environments into various industries, greatly benefiting design, planning, and training processes. These digital landscapes reduce the necessity for physical prototypes, speed up development timelines, and reduce manufacturing expenses [19]. Notably, they prove to be exceptionally advantageous when it comes to *training* new employees, as they minimize the need for setting up hands-on assembly and enrich the learning experience through simulated practice [20]. In the realm of training, VR and MR also offer a cost-effective alternative to physical training setups, as virtual environments could improve cognitive, perceptual, and motor skill learning [21] with interactivity [22], and stereoscopic images aiding the process [23]. Researchers can use time data to assess learning rates and the number of errors made during assembly tasks in virtual environments [24]. Results vary due to different scenarios and methods used in research, though the consensus is that VR and MR are valuable training tools, improving performance and retention, although it may not fully replace hands-on training.

6.2 General-Based Users

6.2.1 Education

3D hologram technology has been recognized as an effective tool for visualization in education, making learning more interactive and engaging [25]. Korean researchers applied hologram technology in educational materials for high school students, teaching them to produce a device that creates hologram videos, and make and share simple holographic videos based on advanced communication technology [26]. The NYU Medicine School introduced Bio Digital Human, an online 3D interactive medical visualization program. As explained by their paper [27], students can wear 3D glasses to view and explore a life-size 3D virtual human body projected onto a 2D projector screen, magnifying and dissecting organs and anatomical structures.

6.2.2 Games and Entertainment

The use of holograms in games has been gaining traction recently. Siang and Mohamed [28] introduced a method of implementing AR tracking and holographic display in an action card game. Fadzli's team [29] developed an AR battleship board game with a holographic display, adding another layer of excitement to the classic game. The gaming industry has embraced hologram technology in various ways. Some games use holographic displays to create a more immersive gaming experience, such as the AR game Pokemon Go. Additionally, popular mobile games, such as Clash of Clans, have utilized holographic ads to promote their games, further demonstrating the versatility of hologram technology.

6.2.3 Marketing and Advertising

Holographic technology is also being used in marketing and advertising. Companies can create holographic product displays that allow customers to view products from all sides, and these displays can be used to showcase the details and features of a product. Puma, for example, launched a 360-degree model of their new sneakers using holograms in a digital-out-of-home campaign [30]. Holographic animal displays have also been used in awareness campaigns, such as the World Wildlife Fund's campaign, featuring a holographic elephant walking on the streets of London.

6.2.4 Shows and Galas

Holograms have been used to create memorable performances in concerts and other events. For example, the closing of Eric Prydz's EPIC 5.0 show in London featured a hologram of his face accompanied by a laser show. Some K-POP entertainment

contents use hologram technology as a part to arouse spectacles or represent idols as hologram characters [31]. Holograms have also been used in large stage performances and galas to create unique and immersive experiences for audiences.

7 Virtual and Holographic Visualization in Medicine—Importance and Need

In the field of medicine, knowledge and information are often abstract and complex, making it difficult for physicians and patients to effectively communicate and understand medical concepts. However, *medical information visualization* has been developed and researched for a long time to provide practical solutions to this problem. From Vesalius' anatomical drawings to the invention of X-ray and imaging-guided minimally-invasive procedures, researchers have used *visual expression* to enhance understanding of complicated and abstract medical concepts. They have utilized various technologies to display information clearly and vividly, helping doctors diagnose diseases, create treatment plans, and communicate with patients.

Medical systems generate complex data using advanced imaging technologies, such as MRI, UltraSound, and CT scans. Normally, this electronic information is used to display a *flat* image on a computer *screen*. However, with the developments in *holographic technology and 3D visualization*, this data can be used to produce full-color, computer-generated 3D holographic images, which is a game-changer in the medical field. Holographic technology, thus, has significant benefits for the medical industry. It uses digital imaging inputs to provide an extensive visualization of the data for doctors, patients, surgeons, and students. In medical imaging, there is a requirement for imaging internal and external human body parts. Compared to regular 2D images, *holograms* store information about multiple images at different angles in the same visualization. This means the viewer can move around the projections and view them, allowing them to examine different body parts or observe biological structures from different angles and perspectives.

Furthermore, holograms can provide an environment in which the learner is fully engaged, making them not only suitable but in fact invaluable for medical training, disease diagnosis, treatment planning, and medical recording. Additionally, holograms bring a sense of *immersion* and help create immersive and interactive experiences which can, in turn, help medical students as well as doctors and patients gain a better understanding, retain information for longer, and form a longer-term memory of medical concepts and knowledge, thus promoting doctor-patient and doctor-doctor communication. Holograms also make remote therapy more appealing and more memorable. According to the report on Profiles in Innovation of VR and AR by Goldman Sachs, the healthcare industry alone could generate $5.1 billion in revenue from holographic technology, and it is predicted that 3.4 million people will adopt this technology by 2025 [32]. Holographic technology has the potential

to revolutionize medical education, training, diagnosis, and treatment, leading to better healthcare outcomes for patients.

8 Current Research Areas and Applications

After discussing the significance and need for virtual and holographic visualization in medicine, this section delves into the current research areas and applications of digital holography and holographic technology. Here, we explore the developments and applications of these technologies in various fields to demonstrate their applications and potential in revolutionizing medical practices.

8.1 Visualization and Assessment of Abnormalities and Organ Functions

Medical holography can greatly assist in detecting problems in complex organs and examining their abnormalities. An organ hologram can be zoomed, manipulated, and interacted with, in 3D space, which makes the examination and diagnosis processes more effective and precise. It has the potential to significantly disrupt 3D imaging, as it offers better usability and addresses the shortcomings of current 3D solutions Some [33] noteworthy applications include:

(a) In 2016, Pathania et al. utilized holography to accurately assess lymphoma tissue [34]. They devised a system that could detect malignant lymphoma cells labeled with marker-specific microbeads, generating distinctive holographic signatures. This technology is crucial for identifying cancer with greater accuracy and in less time.
(b) Furlong et al. proposed a novel approach to diagnose and treat hearing disorders using a digital holographic system [35]. The system is capable of measuring the structure, shape, and acoustically-induced changes to the membrane of the human middle ear. By utilizing holographic technology, this system can detect complex patterns unfolding across the full surface of the membrane, which provide more information than current admittance or reflectance examination.
(c) Abdelazeem et al. proposed a new optical method in 2020 that uses comparative holographic projection to effectively visualize and assess tumor progression in glioblastoma patients [36]. The method displays a 3D map of the tumor, allowing for fast assessment of the tumor volume after treatment and progression. This method can aid in interpretation and assessment of tumor progression with respect to treatment, improving diagnosis and treatment planning.
(d) Digital holography in the Terahertz spectral region has enabled the reconstruction of medical images with complex structures [37]. It offers superior contrast in imaging soft biomedical tissues compared to X-rays and is free from ionization

damage. Terahertz holographic imaging holds great potential for achieving early detection of cancer [38].

8.2 Surgical Navigation and Intervention

The advancement in holographic technology has the potential to revolutionize surgical interventions. By enhancing precision and optimizing outcomes, holographic imaging has numerous applications in the pre-operative, intra-operative, and post-operative periods. In the pre-operative phase, it can be utilized for surgical planning, while in the intra-operative phase, it can provide real-time assistance to surgeons. In the post-operative phase, holographic imaging can be used for assessment, allowing for a thorough evaluation of the efficacy of the procedure. As such, the integration of holographic imaging promises to enhance patient care and improve the overall quality of surgical interventions.

(a) Computer-assisted intra-operative interventions have long struggled with the challenge of acquiring real-time patient anatomy. Brudfors presented a system that overcomes limitations such as low image quality and radiation exposure, by enabling holography technology [39]. The system employs a tracked conoscopic holography device to acquire the patient's anatomy during surgery.
(b) Simpson et al. introduced an innovative holographic imaging method called the conoprobe [40] which offers a means of documenting tissue resection cavities and evaluating the efficacy of resection on model-based guidance systems. In contrast to traditional MRI scans, their digitalization technology generates real-time data, providing an additional dimension to the non-contact instrumentation framework for soft-tissue deformation compensation in guidance systems.
(c) Müller et al. explored 3D intra-operative fluoroscopy with holographic navigation by head-mounted devices for the use in spine surgery [41]. Their study shows that holographic navigation by use of a head-mounted device achieves accuracy comparable to the gold standard of high-end pose-tracking systems.

8.3 Biomedicine and Physiology

Digital holography is also a powerful technology to improve micro-imaging and quantification. Utilizing advanced imaging tools such as digital holographic microscopy, researchers can obtain detailed 3D quantitative information about specimens, driving progress in the fields of biomedicine and physiology. This technology has the potential to enable unprecedented insights into biological structures and processes at the cellular and sub-cellular levels.

(a) With the significant help of off-axis electron holography of nanometer scale particles, a group of physics, pathology, and biology researchers made a significant progress on accurate identification and quantification of abnormal accumulation of Fe, Cu, and Zn present within amyloid-β plaque cores [42]. These brain metals are critical factors in the assessment of neuro-degenerative diseases.
(b) Digital holography employs a short-coherence source to capture and process data from tissues that are difficult to imagine with traditional techniques. The holographic process uses laser ranging to maintain the integrity of the acquired information from depths much greater than what is achievable through conventional imaging methods. In 2015, researchers utilized digital holography to investigate alterations in adhesion-dependent tissue response in 3D cultures, capitalizing on its capacity to image through dense tissue with improved fidelity [43].
(c) In 2022, Kumar and colleagues published their research on a novel single-shot common-path off-axis digital holographic microscopic system for assessing the morphological and quantitative parameters of human red blood cells [44]. This system is characterized by its simplicity, compactness, affordability, and reduced sensitivity to vibrations, making it a superior option for studying small fluctuations in cell thickness.

8.4 Tele-Monitoring and Medical Education

Holographic technology has immense potential for telemonitoring and medical education. We group these fields together as both are more application-oriented compared to the previous three. Many platforms and tools have already progressed from academic concepts and research to industrial applications and have actually been implemented in real-world scenarios. For tele-monitoring, holographic tools can assist doctors in precisely and continuously monitoring patients' long-term health conditions in real-time. It possesses the capacity to greatly enhance healthcare and empower healthcare providers to deliver care that is more tailored to individuals. As for medical education, holography serves as the technical foundation for constructing virtual training environments or classrooms that offer immersive experiences and physical simulations, increasing learners' engagement, and facilitating better medical education and training.

(a) Salvetti and Bertagni utilized holograms to create medical tests and devices for monitoring various conditions such as infections, cardiac function, and diabetes [45]. This technology is employed for advanced medical simulations and facilitates face-to-face training. Its rapid emergence in healthcare has made the education process more efficient.
(b) The Proximie start-up leverages AR to enable surgeons to virtually scrub in and observe any clinical setting in real-time worldwide. This collaborative visualization tool enables hospitals, surgeons, and medical device companies to capture operating room data, share information from anywhere, and generate

new insights to improve outcomes and productivity. Research conducted on the Proximie platform [46, 47] has demonstrated the feasibility and efficiency of using AR tools to guide surgical tasks and enhance communication between surgical mentors and mentees.

(c) CAE Healthcare is advancing simulation-based training solutions to improve medical education and enhance patient safety using MR technology with Microsoft HoloLens. Their simulation has been proven to lead to better outcomes, faster product adoption, and fewer patient complications. Since 2016, a project between CAE and Leeds Beckett has focused on the impact of simulation-based education on patient outcomes, and it has demonstrated the overall effect of a student's encounter with a patient during a simulated scenario [48].

(d) EchoPixel has developed the True 3D Viewer software which converts 2D images into stereoscopic 3D images. The software allows medical professionals to virtually 'cut' organs, tissues, and other body parts at different angles for diagnostics, surgical planning, and interventional cardiology as well as radiology. With this, image specialists, surgeons, and interventionists can generate multiple cross-sections and identify abnormal tissue growth in any organ [33].

(e) Researchers from the University of Tübingen developed an online holographic application which can be used together with real-world simulations to improve medical training [49]. Their platform enables the trainees to view the scenario from different perspectives and freely explore the environment, helping trainees gain a better understanding of the importance of communication and teamwork.

9 Enhancing Doctor-Patient Communication Through Holography

Considering medical applications requires acknowledging both the perspective of doctors and that of the patients. Research has highlighted the need for improved communication between doctors and patients [50]. Furthermore, evidence suggests that patients have specific preferences and requirements for how information is presented to them [51]. For example, they desire to be more involved and active in the process. Moreover, good communication has been linked to positive patient health outcomes [52], higher patient satisfaction [53], increased patient referrals [54] as well as better compliance with healthcare providers and lower readmission rates [55] plus higher patient satisfaction scores [56]. Effective doctor-patient communication is essential for high-quality healthcare but is not consistently achieved as desired. Currently, doctors use brochures, drawings, and 3D models provided by medical device companies as promotional materials to explain specific surgeries to their patients [57].

However, many medical concepts and mechanisms are complex, making them difficult to explain through 2D imaging, statistical diagrams, or static 3D models. As a result, despite doctors explaining the procedure beforehand and patients signing

an informed consent as required by law, gaps in patients' understanding of the risks, complications, undesired outcomes, and post-operative care often arise [58].

To address the communication challenges faced by doctors and patients, research has been conducted on the development of communication support tools and decision aids that can be used in various clinical settings. Effective communication training tools for medical care providers have also been emphasized [59]. Among the various innovative communication solutions being explored, *holographic tools* such as *immersive environments and holographic models*, offer significant advantages. They can disassemble and display complex biological structures, enhance patient involvement and interest, integrate multi-dimensional information, and provide an innovative interactive experience. The development and implementation of holographic communication and information platforms in healthcare have the potential to greatly enhance healthcare outcomes for individuals and communities worldwide, revolutionizing the way doctors and patients communicate.

10 Genetics and Genomics: Understanding the Presented Information

Genetics is an important branch of biology that involves the study of DNA, genes, genetic diversity, and inheritance in all living organisms. Our health starts with our genes. Many health conditions and genetic diseases are carried in our genes. Most cells in the human body have a complete set of genes. The genome is the complete set of our genes that determines pretty much everything about us. Put simply, a genome encompasses all the genetic material within a living organism. In the case of humans, this genetic blueprint comprises more than 20,000 genes. Each individual possesses a distinct genome, akin to a comprehensive guidebook for the body's functions. This intricate manual furnishes directives that dictate various physical traits—ranging from skin tone and stature to susceptibility to specific ailments. Mutation, which is an alteration in the DNA sequence within our genome, can occur for a variety of reasons, e.g., certain chemicals and environmental factors. We can also inherit gene mutations from our biological parents, which could potentially lead to diseases. Our environment and lifestyle can also contribute to genetic changes, thereby causing health issues. Genetic testing can, therefore, be beneficial in identifying genetic illnesses, inherited conditions and risks, and finding appropriate health solutions. Genetic testing and screening for hereditary mutations, furthermore, help predict future health conditions, enabling us to plan and potentially prevent certain serious conditions.

A medical geneticist is a physician who specializes in the study of genetics and its impact on human health. Geneticists diagnose and treat genetic disorders, and provide counseling to families who have a history of genetic diseases. Therefore, they play a critical role in the detection, diagnosis, and treatment of genetic disorders, and in

helping families understand and manage the risks and impacts of these conditions. The tasks of a medical geneticist may include:

- Performing genetic testing to diagnose and monitor genetic disorders
- Providing genetic counseling to families, including discussing the risks of passing on genetic conditions to future generations
- Developing and managing treatment plans for patients with genetic disorders
- Collaborating with other medical specialists to provide comprehensive care for patients with genetic conditions

It must be noted that *genetics* is the study of individual genes and their role in inheritance, while *genomics* is the study of all genes and their interactions within an organism or a population, i.e., the study of the entire genome, which is the complete set of genetic material of an organism.

11 Conventional Visualization Methods in Genetics

There are a number of conventional methods for *genetic information presentation and analysis.* In 2D approaches, genetic visualizations display data either in sequence coordinate systems or genetic tracks and matrices. Sometimes, it is rather impossible for non-specialists to read or understand this information, which shows that conventional 2D methods of information presentation are not particularly helpful in conveying such information, nor are they conducive to drawing clear spatial conclusions from static graphs and flat images and charts.

As for traditional 3D approaches, some challenges still exist. For instance, legibility can be compromised when presenting track-aligned matrices in the second dimension, necessitating more screen space. Additionally, the vertical alignment in track-based genome browsers becomes challenging for 3D representations of genomes. Furthermore, the exploration of unprecedented levels of detail is uncovering intricate complexities that span multiple scales, states, and time dependencies. Genetic visualization demands new approaches to data access and new layouts for analysis. Presenting dynamic holographic visualizations and virtual educational lessons can significantly assist comprehension and improve decision-making.

12 The Holographic Solution

A holographic solution refers to a technology that utilizes holographic imaging, visualization techniques, and corresponding equipment to generate realistic hologram representations of objects and environments. Within the realms of genomics, genetics, and genetic counseling, holographic solutions offer distinct advantages over traditional 2D and 3D representations.

When patients, and in some cases, their families, seek genetic counseling, they rely on hospitals to provide comprehensive information regarding the risks of genetic disorders and to assist them in making well-informed decisions about their health. Holographic solutions offer medical geneticists a highly interactive and visually immersive method to effectively communicate complex genetic information to patients. For instance, holographic models offer a unique opportunity to visually showcase the intricate structures and functions of genes, as well as the consequences of mutations that affect them.

Additionally, they serve as invaluable tools for illustrating complex inheritance patterns and the specific impacts of genetic mutations, which may prove challenging to grasp through traditional 2D visualization and presentation techniques. Holographic visualization offers a more intuitive approach to grasp genetic information, potentially empowering patients to make well-informed decisions about their health and gain a deeper understanding of the implications of their disorders and test results. Various research prototypes and products are available, either directly in the market or indirectly as previously conducted research within university departments:

- *Genetics Data Visualizer* (MR) In 2017, two UCL researchers [60] developed an MR-based genome browser to provide a novel approach of retrieving and analyzing genomics data. The platform was made to resemble a search engine with the difference being that it only contains genetics-related data, in addition to its integration of MR technology. The user's search inputs are cross-referenced with data from Ensemble REST API, e.g., a DNA sequence, mutation alleles, phenotype traits, etc., and are rendered into an accurate 3D model of the requested genome structure, which can be viewed through the HoloLens headset.
- *CSynth DNA Visualizer* (VR & Web) In 2019, researchers at the MRC Weatherall Institute of Molecular Medicine in Oxford and the Department of Computing at Goldsmiths University developed technology that allows scientists to explore the 3D structure of DNA in VR [61]. They created a 3D genome browser and real-time chromatin modeler to visualize dynamic and interactive models of chromatin capture data. By incorporating the feature to observe and interact with these intricate structures through VR, scientists can engage with the models to enhance their comprehension.
- *Cell Explorer* (VR) An interactive VR application designed for the HTC Vive headset that allows users to grab molecules and DNA inside the nucleus of a cell. The prototype was developed in collaboration with the 3D Visual Aesthetics Laboratory at UNSW Art & Design in Australia.
- *Huawei Genomics Viewer* (VR) A VR application developed by Huawei and designed for the HTC Vive headset that utilizes hand-held controllers to allow users to enter an immersive self-discovery experience into the human body, including basic genomics structures that can be inspected and examined.
- *Cascade RNA Visualizer* (VR & Web) In 2016, to meet the growing demand for cancer genomics data analysis tools, [62] created Cascade, a web-based 3D RNA visualization software. The platform enables request transfer between users and the RNA-seq database. The requested data is displayed on the main webpage in the

form of 3D color-coded hair-ball network diagrams. In addition to its customizable graphics settings (e.g., color selection, font size), Cascade can perform statistical calculation (i.e., gene expression distribution) which are displayed as labeling, aiding users with close examination, and in-depth analysis.

- *Delta.AR Genomics Data Visualizer* (MR and AR) In 2021, Delta-AR was developed as an AR 3D visualization platform which introduces a new approach to display and visualize complex genomic data [63]. It is an accessible web-based genome visualization platform made to reduce the difficulties faced when evaluating complex genomic data. The platform assists scientists and medical professionals by incorporating AR technology in order to display 3D genome structure in real-world environments. Delta.AR consists of 3 main components: the HMD of which HoloLens was selected, a portable 2D display device that complements the HMD, and the software, consisting of two servers - one for omics data and another for AR modeling.
- *HoloAnatomy* (MR) is a holographic anatomy programme developed by Case Western Reserve University for medical students to learn and practice anatomy skills in the dissection laboratory on a holographic body [64]. It has been developed through collaboration between anatomists, artists, programmers, instructional designers and scientists and is to be used in conjunction with HoloLens.

13 Research and Analysis

In this section, we highlight the result of research and analysis of the key principles and concepts underlying MR applications that help doctors visualize diagnostic and genetic data, as well as virtual and holographic systems that assist in patient education and communication. The following fundamental principles and concepts form the bedrock of applications and systems aimed at empowering doctors to visualize diagnostic and genetic data effectively.

- Interactive Interface: A virtual and holographic health information system utilizes MR displays to present medical data and information in a 3D and interactive manner. Users can interact with the holograms and virtual artifacts, manipulate data, and gain a deeper understanding of their health conditions. Such interactive interfaces foster patient engagement and facilitate effective communication between healthcare providers and patients.
- Personalization: A crucial aspect of virtual and holographic health information systems is personalization. By integrating patient-specific data, such as medical history, vital signs, and genetic information, the system can tailor information and recommendations to individuals. Personalized virtual and holographic displays allow patients to visualize their health conditions, treatment options, and progress, enabling them to make informed decisions and actively participate in their care.
- Real-Time Data Integration: virtual and holographic health information systems rely on real-time data integration from various sources, including EHRs, wearable devices, and medical sensors. By collecting and aggregating data, the system can

provide up-to-date information to users, monitor their health parameters, and generate personalized insights. Real-time data integration enhances the accuracy and effectiveness of medical recommendations and interventions.

- Artificial Intelligence: Virtual and holographic health information systems often incorporate AI technologies. AI algorithms analyze vast amounts of medical data to provide predictive analytics, decision support, and personalized recommendations. The inclusion of AI enhances the overall user experience and augments healthcare professionals' capabilities.
- Remote & Collaborative Care: Virtual and holographic health information systems have the potential to enable remote healthcare delivery and facilitate collaboration among healthcare providers. Physicians and interdisciplinary healthcare teams can conduct virtual consultations, remotely examine patients, provide guidance to on-site healthcare professionals, share information and expertise, and collaborate in real-time; all improving patient outcomes and optimizing resource allocation.

14 Implications

The adoption of virtual and holographic health information systems holds several potential implications for healthcare:

- Enhanced Patient Engagement: Their displays as well as interactive interfaces empower patients to actively participate in their healthcare journey, leading to improved treatment adherence and outcomes.
- Improved Communication: They can enhance communication between healthcare providers and patients, promoting shared decision-making and ensuring a better understanding of medical information.
- Precision Medicine: Providing therapies tailored to each patient is, essentially, the vision of Precision Medicine, particularly in the fields of Genetics & Genetic Counselling, where each individual's genetic makeup is inherently unique and vital for personalized healthcare. Such systems can leverage personalized data and individualized analytics to support Precision Medicine, delivering targeted interventions and therapies tailored to individuals' unique characteristics.
- Scalability & Accessibility: As Their displays become more accessible and affordable, such systems could be deployed across various healthcare settings, benefiting a broader population and addressing healthcare disparities.

15 Conclusion

Virtual and holographic health information systems offer a promising avenue for transforming healthcare delivery. By providing interactive and personalized medical information, these systems can enhance patient engagement, improve communication, enable Precision Medicine, and facilitate scalability and accessibility. This

paper aims to provide an overview of the use of such technologies in healthcare, and explore their potential to revolutionize the healthcare industry. It also seeks to serve as an insightful resource to researchers and healthcare professionals with a vested interest in leveraging these systems to enhance healthcare practices.

References

1. Hutmacher, F.: Why is there so much more research on vision than on any other sensory modality? Front. Psychol. **10**, 2246 (2019). https://doi.org/10.3389/fpsyg.2019.02246
2. Amin, H.U., Ousta, F., Yusoff, M.Z., Malik, A.S.: Modulation of cortical activity in response to learning and long-term memory retrieval of 2D versus stereoscopic 3D educational contents: Evidence from an EEG study. Comput. Human Behav. **114**, Article 106526 (2021). https://doi.org/10.1016/j.chb.2020.106526
3. Lindal, P.J., Miri, H., Johannsdottir, K.R., Hartig, T., Vilhjalmsson, H.: Cities that sustain us: using virtual reality to test the restorative potential of future urban environments. In: 11th Biennial Conference on Environmental Psychology (BCEP), Groningen, Netherlands (2015)
4. Lindal, P.J., Miri, H., Kristjansson, U., Johannsdottir, K.R., Hartig, T., Vilhjalmsson, H.: Testing the restorative potential of future urban environments using virtual reality technology—the "cities that sustain us" project. In: 24th Conference for People–Environment Studies (IAPS) Lund, Sweden (2016)
5. Coufal, H.J., Psaltis, D., Sincerbox, G.T.: Holographic data storage. Springer (2000). https://link.springer.com/book/10.1007/978-3-540-47864-5
6. Yaraş, F., Kang, H., Onural, L.: J. Disp. Technol. **6**(10), 443–454 (2010)
7. Kervegant, C., Raymond, F., Graeff, D., Castet, C.: Touch hologram in mid-air. In: ACM SIGGRAPH 2017 Emerging Technologies, pp. 1–2 (2017)
8. Oliveira, S., Richardson, M.: The future of holographic technologies and their use by artists. In: Journal of Physics: Conference Series, p. 012007. IOP Publishing (2013)
9. Peddie, J.: Introduction. In: The History of the GPU—Steps to Invention. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-10968-3-1
10. Edward, S.I.: Sketchpad: a man-machine graphical communication system [Ph.D. thesis/Preprint]. Massachusetts Institute of Technology (1963)
11. Krull, F.N.: The origin of computer graphics within General Motors [PDF file] (1994). https://courses.cs.washington.edu/courses/cse490h1/19wi/resources/gm-origins.pdf
12. Ambs, P., Huignard, J.-P., Loiseaux, B.: Holography. Reference Module in Materials Science and Materials Engineering. Elsevier (2022). https://doi.org/10.1016/B978-0-323-90800-9.00002-0
13. Leith, E.N., Upatnieks, J.: Reconstructed wavefronts and communication theory. J. Opt. Soc. Am. **52**(10), 1123–1130 (1962). https://doi.org/10.1364/JOSA.52.001123
14. Tsang, P.W.M., Poon, T.-C.: Review on the state-of-the-art technologies for acquisition and display of digital holograms. IEEE Trans. Industr. Inf. **12**(3), 886–901 (2016). https://doi.org/10.1109/TII.2016.2550535
15. Jeong, T.H.: Basic principles and applications of holography [Preprint]. Lake Forest College, Lake Forest, Illinois (2010)
16. Bryndin, E.: Ensembles of intelligent agents with expanding communication abilities. Res. Intell. Manuf. Assem. **1**, 35–40 (2022). https://doi.org/10.25082/RIMA.2022.01.005
17. Singh, A., Liu, J., Cortes, C.A.T., Lin, C.-T.: Virtual global landmark: An augmented reality technique to improve spatial navigation learning. In: CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2021)
18. Liu, B., Ding, L., Meng, L.: Spatial knowledge acquisition with virtual semantic landmarks in mixed reality-based indoor navigation. Cartogr. Geogr. Inf. Sci. **48**(4), 305–319 (2021)

19. Holuša, V., Vaněk, M., Beneš, F., Švub, J., Staša, P.: Virtual reality as a tool for sustainable training and education of employees in industrial enterprises. Sustainability **15**, 12886 (2023). https://doi.org/10.3390/su151712886
20. Oren, M., Carlson, P., Gilbert, S., Vance, J. M. (2012). Puzzle assembly training: Real world versus virtual environment. In: 2012 IEEE Virtual Reality Workshops (VRW), pp. 27–30. Costa Mesa, CA, USA. https://doi.org/10.1109/VR.2012.618087
21. Wang, Q.-H., Huang, Z.-D., Li, J.-R., Liu, J.-W.: A force rendering model for virtual assembly of mechanical parts with clearance fits. Assem. Autom. **38**2017). https://doi.org/10.1108/AA-12-2016-175
22. Borsci, S., Lawson, G., Salanitri, D., Jha, B.: When simulated environments make the difference: the effectiveness of different types of training of car service procedures. Virtual Reality **20**, 1–14 (2016). https://doi.org/10.1007/s10055-016-0286-8
23. Bhatti, A., Nahavandi, A., Khoo, S., Anticev, D., Zhou, J.: Haptically enabled interactive virtual assembly training system development and evaluation (2012)
24. Hoedt, S., Claeys, A., Van Landeghem, H., Cottyn, J.: The evaluation of an elementary virtual training system for manual assembly. Int. J. Prod. Res. **55**, 1–13 (2017). https://doi.org/10.1080/00207543.2017.1374572
25. Sudeep, U.: Use of 3D hologram technology in engineering education. In: IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE) (2013). ISSN: 2278-1684, 62-67.
26. Kim, B.-H., Jung, M.-Y., Kim, J.: Development and application of 3D-hologram maker education materials for high school students in Korea. Adv. Sci. Lett. **24**(3), 2114–2117 (2018)
27. Lee, H.S.: 3D holographic technology and its educational potential. TechTrends **57**, 34–39 (2013)
28. Siang, C.V., Mohamed, F.: BoBoiBoy interactive holographic action card game application (2017)
29. Fadzli, F.E., Ismail, A.W., Rosman, M.F.A, Suaib, N.M., Rahim, M.S.M., Ismail, I. (2020). Augmented reality battleship board game with holographic display. In: IOP Conference Series: Materials Science and Engineering, p. 012013. IOP Publishing.
30. Lindsay, T.: Advantages of Holograms and What it Means for the Future of Marketing. Future of Marketing Institute (2020). https://futureofmarketinginstitute.com/advantages-of-holograms-and-what-it-means-for-the-future-of-marketing/. Accessed 2 March 2023
31. Han, H.-W., Jeong, A.-R.: Analysis on the spectacles of K-POP hologram concerts-focus on contents of SM entertainment. J. Korea Contents Assoc. **16**(7), 740–749 (2016)
32. Sachs, G.: Virtual & augmented reality: the next big computing platform? (2016). https://www.goldmansachs.com/insights/pages/virtual-and-augmented-reality-report.html. Accessed 2 March 2023
33. Mishra, S.: Hologram the future of medicine–from star wars to clinical imaging. Indian Heart J. **69**(4), 566 (2017)
34. Pathania, D., Im, H., Kilcoyne, A., Sohani, A.R., Fexon, L., Pivovarov, M., Abramson, J.S., Randall, T.C., Chabner, B.A., Weissleder, R., Lee, H., Castro, C.M.: Holographic assessment of lymphoma tissue (HALT) for global oncology field applications. Theranostics **6**(10), 1603 (2016)
35. Furlong, C., Dobrev, I., Rosowski, J., Cheng, J.: Assessing eardrum deformation by digital holography [Preprint]. SPIE newsroom (2013)
36. Abdelazeem, R.M., Youssef, D., El-Azab, J., Hassab-Elnaby, S., Agour, M.: Three-dimensional visualization of brain tumor progression based accurate segmentation via comparative holographic projection. PLoS One **15**(7), e0236835 (2020)
37. Heimbeck, M.S., Everitt, H.O.: Terahertz digital holographic imaging. Adv. Opt. Photonics **12**(1), 1–59 (2020). https://doi.org/10.1364/AOP.12.000001
38. Rong, L., Latychevskaia, T., Chen, C., Wang, D., Yu, Z., Zhou, X., Li, Z., Huang, H., Wang, Y., Zhou, Z.: Terahertz in-line digital holography of human hepatocellular carcinoma tissue. Sci. Rep. **5**(1), 8445 (2015)
39. Brudfors, M., García-Vázquez, V., Sesé-Lucio, B., Marinetto, E., Desco, M., Pascau, J.: ConoSurf: open-source 3D scanning system based on a conoscopic holography device for acquiring surgical surfaces. Int. J. Med. Robot. Comput. Assist. Surg. **13**(3), e1788 (2017)

40. Simpson, A.L., Sun, K., Pheiffer, T.S., Rucker, D.C., Sills, A.K., Thompson, R.C., Miga, M.I.: Evaluation of conoscopic holography for estimating tumor resection cavities in model-based image-guided neurosurgery. IEEE Trans. Biomed. Eng. **61**(6), 1833–1843 (2014)
41. Müller, F., Roner, S., Liebmann, F., Spirig, J.M., Fürnstahl, P., Farshad, M.: Augmented reality navigation for spinal pedicle screw instrumentation using intraoperative 3D imaging. Spine J. **20**(4), 621–628 (2020). https://doi.org/10.1016/j.spinee.2019.10.012
42. Plascencia-Villa, G., Ponce, A., Collingwood, J.F., Arellano-Jiménez, M.J., Zhu, X., Rogers, J.T., Betancourt, I., José-Yacamán, M., Perry, G.: High-resolution analytical imaging and electron holography of magnetite particles in amyloid cores of Alzheimer's disease. Sci. Rep. **6**(1), 24873 (2016)
43. Merrill, D., An, R., Turek, J., Nolte, D.D.: Digital holography of intracellular dynamics to probe tissue physiology. Appl. Opt. **54**(1), A89–A97 (2015)
44. Kumar, M., Matoba, O., Quan, X., Rajput, S.K., Morita, M., Awatsuji, Y.: Quantitative dynamic evolution of physiological parameters of RBC by highly stable digital holographic microscopy. Opt. Lasers Eng. **151**, 106887 (2022)
45. Salvetti, F., Bertagni, B.: Interactive holograms and tutorials in healthcare education: case studies from the e-REAL® experience. Int. J. Adv. Corp. Learn. **9**(2) (2016)
46. Patel, E., Mascarenhas, A., Ahmed, S., Stirt, D., Brady, I., Perera, R., Noël, J.: Evaluating the ability of students to learn and utilize a novel telepresence platform Proximie. J. Robot. Surg. **16**(4), 973–979 (2022). https://doi.org/10.1007/s11701-021-01330-4
47. Cheikh Youssef, S., Sabbubeh, B., Haram, K., Noël, J., Aydın, A., Challacombe, B., Reeves, F., Hachach-Haram, N., Dasgupta, P.: Augmented reality robot-assisted radical prostatectomy with PROXIMIE: preliminary clinical experience. Urol. Video J. **16**, 100187 (2022). https://doi.org/10.1016/j.urolvj.2022.100187
48. Braithwaite, C.: Future of healthcare education technology developed by Leeds Beckett and CAE Healthcare [Preprint]. Leeds Beckett University Website (2016)
49. Alexandrova, I.v. , Rall, M., Breidt, M., Tullius, G., Kloos, U., Bülthoff, H.H. & Mohler B.J(2012). Enhancing medical communication training using motion capture, perspective taking and virtual reality. Med. Meets Virtual Reality **19**(16–22). IOS Press.
50. Ward, P.: Trust and communication in a doctor-patient relationship: a literature review. J. Healthc. Commun. **3**(3), 36 (2018)
51. Uldry, E., Schäfer, M., Saadi, A., Rousson, V., Demartines, N.: Patients' preferences on information and involvement in decision making for gastrointestinal surgery. World J. Surg. **37**, 2162–2171 (2013)
52. Stewart, M.A.: Effective physician-patient communication and health outcomes: a review. CMAJ: Can. Med. Assoc. J. **152**(9), 1423 (1995)
53. Williams, S., Weinman, J., Dale, J.: Doctor-patient communication and patient satisfaction: a review. Fam. Pract. **15**(5), 480–492 (1998)
54. Hachem, F., Canar, J., Fullam, F.M.A., Gallan, A.SPh.D., Hohmann, S., Johnson, C.: The relationships between HCAHPS communication and discharge satisfaction items and hospital readmissions. Patient Exp. J. **1**(2), 71–77 (2014)
55. Zolnierek, K.B.H., DiMatteo, M.R.: Physician communication and patient adherence to treatment: a meta-analysis. Med. Care **47**(8), 826 (2009)
56. Ha, J.F., Anat, D.S., Longnecker, N.: Doctor-patient communication: a review. Ochsner J. **10**(1), 38–43 (2010)
57. Choonara, Y.E., du Toit, L.C., Kumar, P., Kondiah, P.P.D., Pillay, V.: 3D-printing and the effect on medical costs: a new era? Expert Rev. Pharmacoecon. Outcomes Res. **16**(1), 23–32 (2016)
58. Seely, K.D., Higgs, J.A., Nigh, A.: Utilizing the "teach-back" method to improve surgical informed consent and shared decision-making: a review. Patient Saf. Surg. **16**(1), 1–9 (2022)
59. Antel, R., Abbasgholizadeh-Rahimi, S., Guadagno, E., Harley, J.M., Poenaru, D.: The use of artificial intelligence and virtual reality in doctor-patient risk communication: a scoping review. Patient Educ. Couns. **105**(10), 3038–3050 (2022). https://doi.org/10.1016/j.pec.2022.06.006
60. Ramachandran, N.: First steps: PEACH reality—genomics and proteomics. PEACH (2017). https://medium.com/ucl-peach/first-steps-peach-reality-genomics-and-proteomics-caceb1af685a

61. Todd, S., Todd, P., McGowan, S., Hughes, J.R., Kakui, Y., Leymarie, F.F., Latham, W., Taylor, S.: CSynth: A dynamic modelling and visualisation tool for 3D chromatin structure (2019). https://doi.org/10.1101/499806
62. Shifman, A.R., Johnson, R.M., Wilhelm, B.T.: Cascade: an RNA-seq visualization tool for cancer genomics. BMC Genomics (2016). https://doi.org/10.1186/s12864-016-2389-8
63. Tang, B., Li, X., Li, G., Tian, D., Li, F., & Zhang, Z. (2021). Delta.AR: An augmented reality-based visualization platform for 3D genome. *Innovation(Camb)*. https://doi.org/10.1016/j.xinn.2021.100149
64. Wish-Baratz, S., Gubatina, A.P., Enterline, R., Griswold, M.A.: A new supplement to gross anatomy dissection: HoloAnatomy. Med. Educ. (2019). https://doi.org/10.1111/medu.13845

Design and Implementation of 3 MHz Co-site for Ultra-short Wave Link

Zongwei Gao

Abstract The 3 MHz co-site design has become a critical issue in ultra-short wave link communication, enhancing the sensing ability of a flying aircraft. However, current 3 MHz co-site systems face challenges from space isolation, broadband noise, and anti-blocking considerations. In this study, we attempt to provide an optimized 3 MHz co-site system that deals with the mentioned challenges. We primarily present a typical 3 MHz co-site system's structure. Then, antenna position, transmitter, and receiver optimizations are proposed for space isolation, broadband noise, and anti-blocking problems, respectively. Finally, we conducted a system test in a laboratory, and the experimental results demonstrated the effectiveness of the designed co-site system. In addition, differentiated design methods for different application requirements are summarized to satisfy the needs of various co-site systems.

Keywords Co-site · Isolation · Broadband noise · Anti-blocking ability

1 Introduction

With the development of avionics technology, there is an increasing trend of communication links onboard aircraft. On an airborne platform, the challenge is supporting multiple wireless devices in the same or different frequency bands to work simultaneously under limited space conditions. Because the ultra-short wave frequency band has the characteristics of stable and reliable communication, airborne platforms are basically equipped with two or more ultra-short wave links. Therefore, ultra-short wave link co-site design is a common requirement for current airborne platforms and has become an essential indicator for communication system applications.

Co-site of ultra-short wave links means that in the same position, two or more ultra-short Wave links at the same address (or platform) work simultaneously, causing Co-site interference [1]. Co-located interference refers to multiple ultra-short wave links working in the exact location. When one link transmits, typically, it interferes with the

Z. Gao (✉)
Southwest China Institute of Electronic Technology, Chengdu 610036, China
e-mail: gzw5mm@qq.com

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_9

regular communication of another link on the same platform. The voice reception and transmission of the interfered link will be noisy, intermittent, or interrupted, which affects the normal operation of the ultra-short Wave link. Therefore, co-site work intends to avoid co-site interference under certain conditions.

The performance characteristics of the transmitter and receiver primarily cause co-site interference. Several factors determine whether an Ultra-short Wave link can work co-located, including link broadband noise, anti-blocking ability, harmonics, spurious, intermodulation, intermodulation, transmit power, noise coefficient and sensitivity, and other indicators [2]. However, through the summary and analysis, it is studied that the Co-site interference of ultra-short wave links can be summarized into the following three fundamental factors:

1. Link transmitter performance: mainly refers to the noise level when the transmitter is transmitting, namely, the transmitter broadband noise;
2. Signal isolation: principally refers to the radio frequency signal isolation between links, specifically the spatial isolation between the antennas of the two links and the separation of the operating frequencies of the two links;
3. Link receiver performance mainly refers to the ability to resist out-of-band signal interference, that is, the receiver's anti-blocking ability.

Compared with other related techniques, we analyze the key factors affecting ultra-short wave links' co-site capability, according to the theory and engineering practice. We take the 3 MHz co-site operation of ultra-short wave communication links as an example; it is proposed and verified that the performance indicators related to the 3 MHz co-site operation of ultra-short wave links. The remainder of this paper is organized as follows: Sect. 2 presents our 3 MHz co-site design in detail. Section 3 shows the experimental result of our proposed method. We give the conclusions in the last section.

2 Ultra-short Wave Link with 3 MHz Co-site System

2.1 System Composition

The factors related to the co-site of ultra-short wave links principally consist of receivers/transmitters, co-site filters, and antennas (see Fig. 1). The receiver/transmitter completes baseband signal processing and mid-RF filtering and amplification processing functions. The co-site filter mainly completes the suppression of out-of-band signals to reduce the interference of transmitted harmonic signals and useless signals and improves the broadband noise indicators of the transmitting link and receiving link [3]. The antenna is responsible for receiving radiation signals from space and radiating the radio frequency signals to be transmitted into space. In the same platform, if the distance between the transmitting antenna and the receiving antenna is more significant, the impact of the co-site will be minor.

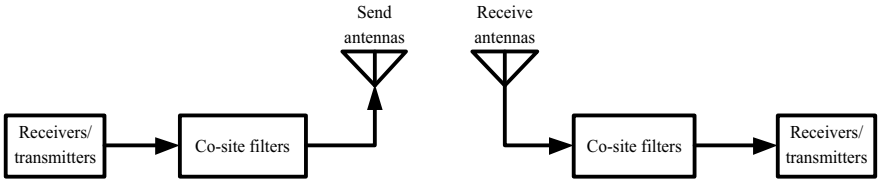


Fig. 1 System block diagram

However, this distance is limited by the size of the installation platform (aircraft, mobile vehicle, etc.). Therefore, combining the specific platform and application requirements, comprehensively considering the performance and position factors of the receiver and transmitter, filter, and antenna, and exploring the best balance point is the focus of system co-site design.

2.2 Space Isolation Design

Optimizing the antenna positions of the two Ultra-short Wave links through layout and increasing the spatial distance of the antennas can enhance signal isolation and effectively weaken the RF signal strength emitted by the transmitting link. Increasing the frequency interval of transceivers also increases the signal isolation. After passing through space radiation, interference signal strength is formed in the receiving link. The space loss of ultra-short wave propagation can be expressed as follows:

$$L_p = 32.4 + 20 \log f + 20 \log d \tag{1}$$

where f is the signal operating frequency in megahertz and d is the transmission distance in kilometers. The free space propagation loss $L_p = 35.3$ dB can be obtained through the reasonable intention of the antenna space layout and installation location, the distance between the antennas $d = 0.01$ km, and the operating frequency $f =$ 140 MHz.

The antenna's gain in a specific direction is often different for different frequency points in the working frequency band. The installation position of the antenna on the aircraft skin and the occlusion of the accessory body in all directions are different, and the antenna gain pattern is also different [4]. The antenna simulation gain diagram can roughly evaluate its performance. In this platform, the antenna transmits gain $G_t = -2$ dB, the antenna receives gain $G_r = -1$ dB. Combined with the RF cable length between the co-site filter and the antenna, we can estimate the transmit link cable loss $L_t = 0.7$ dB and receiver link cable loss $L_r = 1.0$ dB. Finally, according to the produce estimate, the spatial isolation can be expressed as follows:

$$L_p - G_t - G_r + L_t + L_r = 41.2 \text{ dB} \tag{2}$$

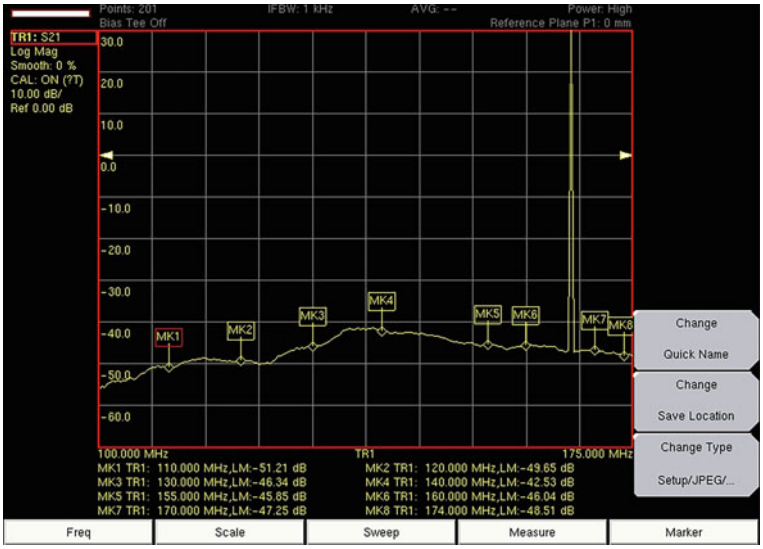


Fig. 2 Antenna isolation

Through actual testing on the platform, the lowest antenna isolation is about 42.5 dB (see Fig. 2).

2.3 Designed to Emit Broadband Noise

Broadband noise is the integrated value of noise power within a unit frequency that deviates from a certain frequency of the carrier signal. Broadband noise interference covers the entire communication spectrum because of its broadband characteristics. It will enter the receiving link together with the proper signal. When its power is comparable to the environmental noise of the receiving link, it will cause the receiver noise floor to increase and reduce the receiver output S/N. Noise ratio directly affects its receiving sensitivity [5]. The following formula can calculate the receiving circuit noise power spectral density:

$$N_p = -174 + NF, \tag{3}$$

where NF represents the noise figure. When $NF = 9$ dB, we get $N_p = -165$ dBm/Hz. Therefore, when the broadband noise of the transmitting link radiated through space and falling into the receiving link is less than -165 dBm/Hz, the broadband noise will not cause the receiver to malfunction. Therefore, how to develop the broadband noise index of the transmitter to meet the above requirements is one of the necessary conditions for system co-site operation.

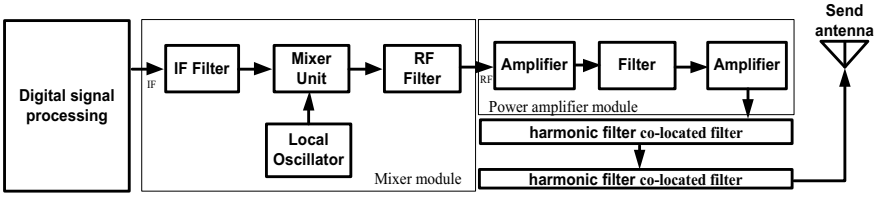


Fig. 3 Transmit link block diagram

In a co-located system link, broadband noise is mainly determined by the performance of the frequency conversion module and power amplifier module in the transceiver and the co-located filter's performance. The main block diagram is shown in Fig. 3.

In the design of the frequency conversion module, the module filters the intermediate frequency signal of digital signal processing and then upconverts the frequency to reach the working frequency. In this module, the output broadband noise mainly consists of two parts: one part is the phase noise of the local oscillator signal, and the other part is the thermal noise generated by the excitation link itself. Therefore, the optimal design of the two indexes can improve the transmit broadband noise. Among them, the broadband noise introduced by the thermal noise of the frequency conversion module can be calculated as -152 dBm/Hz through ADS software. The local oscillator phase noise is ≤ -140 dBc/Hz@3 MHz, and the phase noise deteriorates by 3 dB after mixing. The out-of-band suppression level of the frequency hopping filter is about 10 dB. Therefore, the local oscillator phase noise contributed to the broadband noise level of the entire module can be calculated as $-140 + 3 - 10 = -147$ dBm/Hz. The output power of the module is 2 dBm. Finally, it can be calculated that the contribution of phase noise to the module broadband noise is about -145 dBm/Hz. It can be seen from the above analysis that the noise introduced by the local oscillator phase noise (-145 dBm/Hz) and the noise introduced by the thermal noise of the excitation link itself (-152.5 dBm/Hz) work together, and broadband noise of output is -144.3 dBm/Hz.

In the design of the power amplifier module, when the input signal of the power amplifier module is 2 dBm, the output power is designed as 20 W, the overall gain is 41 dB, and the output broadband noise could be smaller than $-103.3 = (-144.3 + 41)$ dBm/Hz.

A first-level medium-power filter in the power amplification module suppresses $\geq$15 dBm beyond the carrier frequency $\pm$3 MHz. The broadband noise of the entire cascade is less than -119 dBm/Hz.

In co-site filter design, its out-of-band performance suppression capability is required to achieve better performance: $\geq$30 dB@ $\pm 3\% f_0$. After the transmitter/receiver output signal passes through the co-site filter, its broadband noise is further improved to -149 dBm/Hz. After the broadband noise passes through the signal

space isolation degree (41.2 dB), its noise level is lower than the circuit thermal noise power spectral density −174 dBm/Hz. Its impact on receiver sensitivity has reached a negligible level, so the design of transmitting broadband noise requires meeting co-site working conditions.

2.4 Anti-clogging Ability

The maximum power of an interference signal that deviates from a particular carrier signal frequency and enters the receiver together with the useful signal without being affected represents the anti-blocking capability. The interference signal may come from the valuable frequency signal of other transmitters or unwanted signals such as harmonics and spurious signals from other transmitters. The receiver's anti-blocking capability depends on the radio frequency front-end's ability to resist out-of-band suppression and linear operation capabilities. The stronger the anti-out-of-band suppression capability, the more the interference signals entering the receiver will be weakened or eliminated by filtering and the stronger the receiver's anti-blocking ability. Therefore, improving the filter performance on the receiving link can reduce the interference signal power and increase the anti-blocking ability of the link. The linear working capability of the device is determined by judging whether the interference signal causes the receiving circuit to be saturated. Working at the 1 dB compression point is used as a criterion. The higher the linear working capability, the stronger the anti-blocking capability. The structure of the receiving end RF link is shown in Fig. 4.

Since the filter on the frequency conversion link significantly attenuates signals outside the center frequency band, the blocking capability is mainly determined by the actual input power of the active device before frequency conversion and its own 1 dB compression point. By establishing a blocking level model and analyzing based on the input signal power, the gain value of each device, and the 1 dB compression point, it is obtained that the anti-blocking capability of the receiving link at a deviation of ±3 MHz from the operating frequency is −28.6 dBm. The effective power of this platform's transmitting connection is 43 dBm. After passing the spatial isolation of 42.5 dB, the interference signal power reaching the receiving link is 0.5 dBm. The co-site filter is designed to suppress out-of-band signals ≥30 dB@ ± 3%f_0, so the final blocking level entering the receiver is less than −29.5 dBm, which does not exceed the anti-blocking capability of the receiving end.

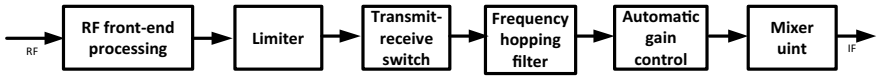


Fig. 4 Receiver link block diagram

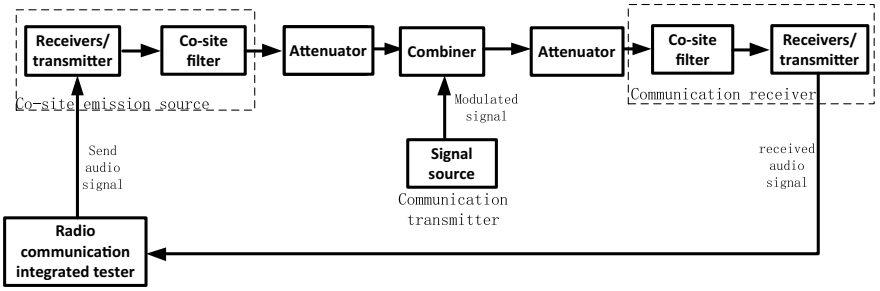


Fig. 5 Laboratory test block diagram

3 Ultra-short Wave Co-site Testing and Verification

According to the design analysis and requirements, the two ultra-short wave links were tested and verified in the EMC laboratory to avoid radio frequency interference. With the dashed box is the equipment we designed and others are standard instruments. The test block diagram is shown in Fig. 5.

The communication transmitter frequency f_1 is 140 MHz, and power is sensitivity. The transmit power of the co-located transmitter is 43 dBm, and the frequency f_2 of the co-located transmitter is between 136 and 144 MHz at certain frequency intervals. When the audio signal-to-noise ratio of the communication receiver is lower than 10 dB, it indicates that the co-located transmitting source is affecting receiver work. The co-site frequency interval of the system is max ($|f_2 - f_1|$) when co-site emission will not affect receiver work, and the test results show that co-site working requirements of 3 MHz are met.

4 Conclusion

The ultra-short wave co-site system will be more complex when different platforms face multiple ultra-short wave transmitters and receivers. Nevertheless, based on the above design and analysis methods, each transmitter's broadband noise, spatial isolation, and receiver blocking capabilities are analyzed separately, which is still adequate for developing and evaluating co-site work in complex systems.

By analyzing the co-site capability factors of ultra-short wave radio stations and considering the signal isolation characteristics of the antenna layout, this paper completes the design and verification of co-site performance indicators for ultra-short wave links. Experimental results show that this co-site system can achieve the platform's 3 MHz Co-site operating requirements. However, limited by the spatial of the co-site platform, transceivers performance needs to be improved when the space

isolation is smaller. That must increase the volume, weight and power consumption of the system. Therefore, based on the performance level of transceivers, limitations and requirements of platform, co-site system design is still needed to be comprehensively evaluated by various systems.

References

1. Zhao, L.: Co-site design of multiple radio stations with small frequency intervals based on electrically tuned filters. Command Inf. Syst. Technol. **4**(2), 59–62 (2013)
2. Luzzatto, A., Haridim, M.: Wireless Transceiver Design, 1st edn. Tsinghua University Press, Beijing (2019)
3. Hou, F., Zhou, Q.: Research on the anti-broadband noise interference performance of frequency hopping communication systems. Inf. Commun. **175**(5), 195–196 (2017)
4. Qin, W.: Layout simulation analysis of tank-mounted multi-antenna system. Firepower Command Control **4**(4), 98–101 (2015)
5. Zhou, W., Chen, Z.: Analysis of helicopter ultra-short wave communication link. Helicopter Technol. **172**(3), 34–37 (2012)

Soft Consensus Under Weighted Average Aggregation Operator and Its Effect on Consensus

Yilei Li, Dongjie Guo, Yifeng Ma, and Huanhuan Zhang

Abstract Facilitating consensus in group decision-making involves some decision-makers or groups reaching a certain agreement after thorough discussion.Extensive research has been conducted on achieving consensus regarding costs in group decision-making, and they often utilize aggregation operators to combine individual opinions into a group opinion. However, the influence of aggregation operators on cost consensus remains uncertain. This study establishes cost consensus models using three different weighted average aggregation operators and a model without any aggregation operator. The analysis explores how consensus results are affected by various weighted average aggregation operators. This paper indicates that employing aggregation operator in cost consensus model raises the total consensus cost, and reveals its effect on the consensus-building process.

Keywords Soft consensus · Minimum cost · Aggregation operator

1 Introduction

Consensus decision-making in management activities needs to achieve an agreement through decision maker's discussion.In consensus decision-making, the process aims to strengthen agreement among decision-makers, necessitating a moderator for dynamic exchange of opinions. In recent years, consensus decision-making has gained wide attention [1–6], but it can be challenging and resource-consuming due

Y. Li
Office of Academic Affairs, Southwestern University of Finance and Economics, Chengdu 611130, China

D. Guo · Y. Ma
School of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, China

H. Zhang (✉)
School of Management Science and Engineering, Southwestern University of Finance and Economics, Chengdu 611130, China
e-mail: hhzhang6@163.com

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_10

to the need for effective communication between decision-makers and the moderator. Given the complexity and cost of the consensus process, consensus cost is an important issue to study.

The introduction of the minimum cost consensus concept originated in Ben-Arieh et al. [7, 8], as a novel approach to model cost in consensus process. It outlines a linear cost associated with altering decision-makers' opinions in the context of both single and multi-criteria decision consensus. After that, a variety of models of consensus cost were proposed [9–14], and The minimum cost consensus has found practical application in various real-world scenarios. Aggregation operators play a role in GDM problems by deriving a unified group opinion from individual perspectives. Zhang et al. [13] initially integrated aggregation operators into the cost consensus model, introducing a minimum cost consensus model grounded in these operators. Liu et al. [12] utilized fuzzy multiplication operator to aggregate individuals' fuzzy opinions, and Suggested an optimal consensus approach minimizing costs within the framework of fuzzy GDM. In Li et al. [11], the authors adopted OWA operator to combine the viewpoints of decision-makers in the devised consensus process, considering minimal and uncertain costs in GDM. Using OWA operator and linguistic quantifiers, the authors [15] proposed a straightforward consensus framework utilizing the minimum cost consensus model to address consensus challenges involving multiple alternatives. Liu et al. [3] proposed the minimum cost strategic weight manipulation model in group decision-making, investigating how interval attribute weight information influences the cost of manipulating strategic weights.

Although aggregation operators are useful to aggregate decision-makers' opinions, the impact of aggregation operator on consensus results is unclear. How does aggregation operators affect the consensus results? What is the influence of varying weighted average operator on consensus cost when using aggregation operators? The contributions and novelty of the paper lies in its examination of the impact of aggregation operators on cost consensus in group decision-making.This paper introduces a novel approach by constructing cost consensus models employing three distinct weighted average aggregation operators, along with a model without any aggregation operator. The analysis of different weighted average aggregation operators reveals that their utilization leads to an increase in the overall consensus cost. The study sheds light on the intricate relationship between aggregation operators and the consensus-building process, contributing valuable insights to reveal the influence mechanism of aggregation operator on consensus and the role it plays in consensus.

The aim of this paper is to analyze the impact of aggregation operators on the consensus results. We compare and analyze the effects of using weighted average operator (WAO) and not using any WAO on the results of consensus. The influence of varying WAO on consensus results and the role of aggregation operator in soft cost consensus are revealed.

The rest of this paper is structured as follows: Sect. 2 provides an introduction to the basics of related models, while Sect. 3 establishes soft cost consensus models, including those with aggregation operators and a model without aggregation

operators. Section 4 shows two numerical examples based on the consensus models. The influence of WA aggregation operators on consensus results is studied. Section 5 concludes the paper.

2 Preliminaries

Suppose there are m decision-makers, denoted as $E = \{e_1, \ldots, e_m\}$, involved in a GDM problem. Let $o_i \in R$ represent the opinion of decision-maker $e_i (i \in M = \{1, 2, \ldots, m\})$, and the consensus opinion be o'. The unit cost that a moderator is willing to pay to reach a consensus for decision-maker e_i is denoted as c_i, and $c_i f_i(o')$ represents the total cost paid to decision-maker e_i. Ben-Arieh and Easton [7, 8] introduced the concept of minimum cost consensus, proposing a consensus model with a minimum cost, subject to an ε threshold constraint on opinion deviation:

$$\begin{aligned} &\min \varphi = \sum_{i=1}^{m} c_i \left| o' - o_i \right| \\ &s.t. \left| o' - o_i \right| \leq \varepsilon \end{aligned} \tag{1}$$

Model (1) characterizes the opinions of all decision-makers using precise numerical values. Where c_i represents the cost per unit paid by a moderator to a decision-maker e_i, and o_i is the initial opinion of decision-maker e_i and o' denotes the group consensus opinion. In the objective function, $c_i \left| o' - o_i \right|$ is the cost paid to e_i, and $\sum_{i=1}^{m} c_i \left| o' - o_i \right|$ is the overall payment made by a moderator, and the goal is to minimize this cost.

In some decision-making problems, the opinions are expressed as interval numbers, based on multi-objective programming, the minimum cost consensus (MCC) model is constructed [9]:

$$\begin{aligned} &\min \varphi = \sum_{i=1}^{m} c_i \left| o' - o_i \right| \\ &s.t. \begin{cases} o_i \in [o_{li}, o_{ri}], & i \in M \\ o' \in O \end{cases} \end{aligned} \tag{2}$$

where c_i denotes a unit cost that the moderator pays to individual d_i, M is the set of the decision-makers, and $M = \{1, 2, \ldots, m\}$. o_i is the initial opinion of decision-maker d_i, which is uncertain and presented as an interval opinion $o_i \in [o_{li}, o_{ri}]$. o_{li} and o_{ri} represents the minimum and maximum limits of d_i. O is the feasible set of consensus opinions for a group. $c_i \left| o' - o_i \right|$ is the cost paid to d_i, and $\sum_{i=1}^{m} c_i \left| o' - o_i \right|$ is the overall payment made by a moderator, with the objective of minimizing it.

Zhang et al. [13] firstly incorporated the aggregation operators into the MCC problem, and suggested multiple minimum cost consensus models featuring a linear cost function under widely used aggregation operators. Specifically, the minimum cost consensus model under the arithmetic average operator is formulated as follows:

$$\min \varphi = \sum_{i=1}^{m} c_i |o' - o_i|$$
$$s.t. \begin{cases} |o' - o_i| \leq \varepsilon, \quad i \in M & (3-1) \\ o' = \sum_{i=1}^{m} w_i o_i' & (3-2) \\ o_i', o' \geq 0 \end{cases} \tag{3}$$

The objective function in Model (3) is $\varphi = \sum_{i=1}^{m} c_i \left| o' - o_i \right|$, which represents the group's minimum total consensus cost. The limiting condition (3–1) denotes an individual's original opinion and the group's opinion is within a certain tolerance range ε. Constraint (3–2) indicates that the model employs the arithmetic weighted average operator to combine decision-makers' opinions and achieve group consensus, striving to minimize the total cost under these constraints.

3 Model Construction

Let $w_i, i \in M = \{1, 2, \ldots, m\}$ denote the weights of the decision-makers' opinions. Let $o_i', i \in M$ be the adjusted opinion of the ith decision-maker and o' be the group consensus opinion. We first construct soft consensus cost model with aggregation operator, and then build a consensus model that does not contain any aggregation operators.

3.1 Soft Consensus Cost Model with Aggregation Operator

Aggregation operators play a crucial role in forming collective opinions for GDM problems. When there is a specific relationship requirement between individual opinions and the group consensus, the use of aggregation operators becomes essential in cost consensus models. To study the impact of aggregation operators on resource costs in the MCC problem, this subsection introduces WAOs and a special OWA operator to construct consensus models for different decision-making scenarios.

(1) **Weighted average operator**

Weighted average operator is a kind of widely used aggregation operator, and we denote it as $WA(o_1, o_2, \ldots, o_m)$ The following are three common types of weighted average operators:

(a) Suppose there are m decision-makers in a GDM, and their corresponding opinions are $o_i, i \in M = 1, 2, \ldots, m$. The weight vector is $\vec{w} = (w_1, w_2, \ldots, w_m)^T$, $\sum_{i=1}^{m} w_i = 1$. If the arithmetic weighted average operator (AWAO) is used to aggregate the opinions of all decision-makers, it can be expressed as: $WA(o_1, o_2, \ldots, o_m) = \sum_{i=1}^{m} w_i o_i$. If $w_i = w_j (i, j = 1, 2, \ldots, m)$ and

$\sum_{i=1}^{m} w_i = 1$, then aggregate the opinions of all decision-makers under equal weight can be expressed as $WA(o_1, o_2, \ldots o_m) = \sum_{i=1}^{m} w_i o_i = \frac{1}{m} \sum_{i=1}^{m} o_i$.

(b) Suppose there are m decision-maker in a GDM, and their corresponding opinions are $o_i, i \in M = \{1, 2, \ldots, m\}$. The weight vector is $\vec{w} = (w_1, w_2, \ldots, w_m)^T$, $\sum_{i=1}^{m} w_i = 1$. If the geometric weighted average operator (GWAO) is used to aggregate the opinions of all decision-makers, it can be expressed as: $WA(o_1, o_2, \ldots, o_m) = \prod_{i=1}^{m} w_i o_i$. If $w_i = w_j (i, j = 1, 2, \ldots, m)$ and $\sum_{i=1}^{m} w_i = 1$, then to aggregate the opinions of all decision-makers under equal weight can be expressed as $WA\,(o_1, o_2, \ldots o_m) = \prod_{i=1}^{m} w_i o_i = \left(\prod_{i=1}^{m} o_i\right)^{\frac{1}{m}}$.

(c) Suppose there are m decision-maker in a GDM, and their corresponding opinions are $o_i (i \in M = \{1, 2, \ldots, m\})$. The weight vector is $\vec{w} = (w_1, w_2, \ldots, w_m)^T$, $\sum_{i=1}^{m} w_i = 1$. If the harmonic weighted average operator (HWAO) is used to aggregate the opinions of all decision-makers, it can be expressed as: $WA(o_1, o_2, \ldots o_m) = \frac{1}{\sum_{i=1}^{m} \frac{w_i}{o_i}}$. If $w_i = w_j (i, j = 1, 2, \ldots, m)$ and $\sum_{i=1}^{m} w_i = 1$, then to aggregate the opinions of all decision-makers under equal weight can be expressed as $WA\,(o_1, o_2, \ldots o_m) = \frac{1}{\sum_{i=1}^{m} \frac{w_i}{o_i}} = \frac{m}{\sum_{i=1}^{m} \frac{1}{o_i}}$.

(2) **OWA operator**

The OWA operator was defined by Yager [16]. When OWA operator is introduced to the consensus models, we can rank decision-makers' opinions according to their importance first, then build a variety of consensus models that indicate the relationships between individual decision-maker's opinion and the group consensus opinion. We denote it as $OWA(o_1, o_2, \ldots, o_m) = F(w_i, o_{(i)})$, where $o_{(i)}$ is the largest (smallest) element in the set $\{o_1, o_2, \ldots, o_m\}$. A particular kind of OWA is with the weight vector $\vec{w} = \left(0, \frac{1}{m-2}, \ldots, \frac{1}{m-2}, 0\right)^T$, In GDM problems, individuals have their own purposes, and frequently manipulate the preference information to achieve their goals. This OWA operator is widely used to prevent such manipulation in reality. For example, in some sports (such as gymnastics), the arithmetic average score is calculated after removing the highest and lowest scores provided by judges. This scoring process uses this OWA operator to reduce referees' bias.

Utilizing various WAOs to formulate the optimal group consensus opinion, we can construct various soft consensus model minimizing total cost. For convenient analysis, inspired by Zhang et al. [14], here we establish a comprehensive consensus model founded on the aggregation operator. The soft cost consensus model, relying on the aggregation operator at the γ level, is defined as follows:

$$\begin{aligned} &\min \varphi = \sum_{i=1}^{m} c_i \left| o_i' - [o_{li} + \alpha_i (o_{ri} - o_{li})] \right| \\ &s.t. \begin{cases} 1 - \frac{|o_i' - o'|}{o'} \geq \gamma & (4\text{–}1) \\ o' = \sum_{i=1}^{m} w_i \otimes o_i & (4\text{–}2) \\ o_i', o' \geq 0 \end{cases} \end{aligned} \tag{4}$$

In model (4), if $u_i \geq 0$, $v_i \geq 0$ and $u_i * v_i = 0$, such that $|o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})]| = u_i + v_i$, $o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})] = u_i - v_i$. If we let $u_i = \frac{1}{2}|o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})]| + \frac{1}{2}\{o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})]\}$, $v_i = \frac{1}{2}|o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})]| - \frac{1}{2}\{o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})]\}$, the above conditions hold. The linear programming representation of the nonlinear model (4) is outlined as follows:

$$\min \ \varphi = \sum_{i=1}^{m}(c_i u_i + c_i v_i)$$

$$s.t. \begin{cases} o_i' - u_i + v_i - \alpha_i * (o_{ri} - o_{li}) = o_{li}, i \in M & (5-1) \\ o_i' - (2-\gamma)o' \leq 0, i \in M & (5-2) \\ o_i' - \gamma o' \geq 0, i \in M & (5-3) \\ \alpha_i \leq 1, i \in M & (5-4) \\ \sum_{i=1}^{m} w_i \otimes o_i - o' = 0 & (5-5) \\ o' \geq 0, o_i' \geq 0, u_i \geq 0, v_i \geq 0, \alpha_i \geq 0, i \in M \end{cases} \quad (5)$$

The objective function in Model (5) is $\varphi = \sum_{i=1}^{m}(c_i u_i + c_i v_i)$, which represents a group's the minimum overall consensus cost. The limiting condition (5–1) denotes the limits of the deviation between an individual's original opinion and adjusted opinion. Constraints (5–2) and (5–3) indicate that the group consensus is attained at a specific consensus level and delineate the bounds for the deviation between an individual's adjusted opinion and the group's optimal opinion, respectively. Constraint (5–5) specifies that the model employs the arithmetic weighted average operator to combine decision-makers' opinions and derive the group's consensus opinion, where $\otimes$ denotes the general aggregation operator. Subject to these constraints, the objective is to minimize the total cost of achieving a group consensus.

3.2 Soft Consensus Cost Model with No Aggregation Operator

Within certain consensus decision-making scenarios, group consensus opinions can be accomplished using consensus degree functions, without aggregation operators. Therefore, a MCC model under a specific degree of consensus may contain no aggregation operator. In this case, the degrees of decision-makers' opinions adjustments are more relaxed without the constraint of aggregation operators. Therefore, the total compensations using models without aggregation operators are less than the models that contain aggregation operators.

From the construction of the above minimum cost consensus model, it can be found that the importance of a decision-maker is reflected in the number of unit compensation price that the decision-maker can obtain, which indicates the difficulty of modifying a decision-maker's opinion and to some extent reflects the weight of the decision-makers in a consensus. However, an aggregation operator does not reflect

the importance of decision-makers in the consensus, it represents the relationship between decision-makers' opinions and group opinion in GDM. It indicates the proportion of an decision-maker's adjusted opinion to the group's ideal opinion, and reflects the influence of an decision-maker's adjusted opinions on the group's ideal consensus opinions. To facilitate the analysis of impact of the aggregation operator on consensus results, we construct a soft consensus cost model without aggregation operator.

The consensus model of minimum cost with no aggregation operator under the level is defined as:

$$\begin{aligned} &\min \varphi = \sum_{i=1}^{m} c_i |o_i' - [o_{li} + \alpha_i (o_{ri} - o_{li})]| \\ &s.t. \begin{cases} 1 - \frac{|o_i' - o'|}{o'} \geq \gamma & (6\text{–}1) \\ o_i', o' \geq 0 \end{cases} \end{aligned} \tag{6}$$

In model (6), if $u_i \geq 0, v_i \geq 0$ and $u_i * v_i = 0$, such that $u_i = \frac{1}{2}|o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})]| + \frac{1}{2}o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})]$, $v_i = \frac{1}{2}|o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})]| - \frac{1}{2}o_i' - [o_{li} + \alpha_i * (o_{ri} - o_{li})]$, the aforementioned conditions hold. The linear programming representation of the nonlinear model (7) is as follows:

$$\begin{aligned} &\min \varphi = \sum_{i=1}^{m} (c_i u_i + c_i v_i) \\ &s.t. \begin{cases} o_i' - u_i + v_i - \alpha_i * (o_{ri} - o_{li}) = o_{li}, i \in M & (7\text{–}1) \\ o_i' - (2 - \gamma) o' \leq 0, i \in M & (7\text{–}2) \\ o_i' - \gamma o' \geq 0, i \in M & (7\text{–}3) \\ \alpha_i \leq 1, i \in M & (7\text{–}4) \\ o' \geq 0, o_i' \geq 0, u_i \geq 0, v_i \geq 0, \alpha_i \geq 0, i \in M \end{cases} \end{aligned} \tag{7}$$

The objective function in Model (7) is $\varphi = \sum_{i=1}^{m} (c_i u_i + c_i v_i)$, which represents the minimum total consensus cost of a group. Condition (7–1) sets the boundaries for the deviation between an individual's original opinion and adjusted opinion. Constraints (7–2) and (7–3) indicate that the group consensus is reached at a specific consensus level and define the limits for the deviation between an individual's adjusted opinion and the group's optimal opinion. Within these constraints, the goal is to minimize the total cost of achieving a group consensus.

For the proposed models, an important Theorem is given as follows.

Theorem 1 *Let $o_i^*, i \in M$ are the decision-makers' optimal opinions, and o^* is the group's optimal opinion. The weight vector is $\vec{w} = (w_1, w_2, \ldots, w_m)^T$, $\sum_{i=1}^{m} w_i = 1$. If the soft consensus cost model MCC_1 contains constraints $\sum_{i=1}^{m} w_i o_i' - o' = 0$, and its minimum consensus cost is denoted as Φ_1^*. The soft consensus cost model MCC_2 does not contain constraints $\sum_{i=1}^{m} w_i o_i' - o' = 0$, and its minimum consensus cost is denoted as Φ_2^*. Under the same degree of consensus, there must be $\Phi_1^* \geq \Phi_2^*$, and when $o^{**} = \sum_{i=1}^{m} w_i o_i^{**}$, $\Phi_1^* = \Phi_2^*$ holds.* □

Proof Suppose the optimal solution of consensus model MCC_1 is $X_1^* = o_i^*, o^*, u_i^*, v_i^*, i \in M$, and for model MCC_2 is $X_2^{**} = o_i^{**}, o^{**}, u_i^{**}, v_i^{**}, i \in M$.

(1) First, we need to prove that there exists $\Phi_1^* \geq \Phi_2^*$ holds.

Since X_1^* is the optimal solution of the model MCC_1, then we have $o_i^* - u_i^* + v_i^* = o_i, i \in M$ (1), $o_i^* - (2-\gamma)o^* \leq 0$ (2), and $o_i^* - \gamma o^* \geq 0$ (3) holds. From Eqs. (1) to (3), it can be seen that the optimal solution X_1^* is also the basic feasible solution of the model MCC_2.

Because X_2^{**} is the optimal solution of the model MCC_2, and X_1^* is only the basic feasible solution of the model MCC_2, according to the property of linear programming, ${\Phi_2}_{(x_1^*)} \geq {\Phi^*}_{2(x_2^{**})}$ holds.

Because ${\Phi_2}_{(x_1^*)} = \sum_{i=1}^m (c_i u_i^* + c_i v_i^*) = {\Phi^*}_{1(x_1^*)}$, ${\Phi^*}_{1(x_1^*)} \geq {\Phi^*}_{2(x_2^{**})}$ holds, that is, $\Phi_1^* \geq \Phi_2^*$ (8) holds.

(2) The following we proof $o^{**} = \sum_{i=1}^m w_i o_i^{**} \Rightarrow \Phi_1^* = \Phi_2^*$.

If the equation $o^{**} = \sum_{i=1}^m w_i o_i^{**}$ (4) holds, then $\sum_{i=1}^m w_i o_i^{**} - o^{**} = 0$ holds.

Since X_2^{**} is the optimal solution of the model MCC_2, then we have the constraint conditions $o_i^{**} - u_i^{**} + v_i^{**} = o_i, i \in M$ (5), $o_i^{**} - (2-\gamma)o^{**} \leq 0$ (6), $o_i^{**} - \gamma o^{**} \geq 0$ (7) holds.

From Eqs. (4) to (7), it can be seen that the optimal solution X_2^{**} is the basic feasible solution of the model MCC_1.

Because X_1^* is the optimal solution of the model MCC_1, and X_2^{**} is only the basic feasible solution of the model MCC_1, according to the property of linear programming, ${\Phi^*}_{1(x_1^*)} \leq {\Phi_1}_{(x_2^{**})}$ holds.

Because ${\Phi_1}_{(x_2^{**})} = \sum_{i=1}^m \left(c_i u_i^{**} + c_i v_i^{**}\right) = {\Phi^*}_{2(x_2^{**})}$, ${\Phi^*}_{1(x_1^*)} \leq {\Phi^*}_{2(x_2^{**})}$ holds.

That is, $\Phi_1^* \leq \Phi_2^*$ holds.

From (8), the formula $\Phi_1^* \geq \Phi_2^*$, so we have $\Phi_1^* = \Phi_2^*$ holds.

That is, $o^{**} = \sum_{i=1}^m w_i o_i^{**} \Rightarrow \Phi_1^* = \Phi_2^*$. □

From a theoretical viewpoint, Theorem 1 proves that the introduction of a aggregation operator in the cost consensus model will increase the consensus cost. In addition, Theorem 1 provides the condition that the cost does not increase when aggregation operator is introduced, that is, the group's optimal opinion and the decision-makers' optimal opinions in the model do not contain aggregation operator constraints, which satisfy the relationship $o^{**} = \sum_{i=1}^m w_i o_i^{**}$. This relationship shows that the group's optimal opinion is obtained through the aggregation of decision-makers' optimal opinions. The essence is that the model satisfies the constrained conditions of the aggregation operator, and the model at this time is actually equivalent to the model containing the aggregation operator.

4 Numerical Examples

Case 1 Suppose there are four decision-makers denoted as (e_1, e_2, e_3, e_4) in a GDM scenario. Their respective opinions are as follows: $o_1 = [14, 37], o_2 = [22, 30], o_3 = [64, 153]$, $o_4 = [8, 61]$ [9]. The assigned weights for their opinions are $w_1 = 0.4$, $w_2 = 0.3$, $w_3 = 0.2$, $w_4 = 0.1$. Additionally, the unit costs paid to these decision-makers by the moderator are $c_1 = 1$, $c_2 = 2$, $c_3 = 3$, and $c_4 = 1$ [9]. Following consultations and discussions, the moderator establishes a minimum consensus level for the group, set at $\gamma = 0.8$. Let the consensus opinion of the group be denoted as o', and the final opinion of a decision-maker after several adjustments be o_i', where $i = 1{-}4$. We can construct various soft consensus models based on different aggregation operators. This example uses the widely used WAOs as aggregation operator to interpret the consensus models that have been suggested. If we use the arithmetic weighted average operator to assemble the collective consensus opinion of the group, the optimization model under consensus level $\gamma = 0.8$ for the whole group is constructed as follow:

$$\begin{aligned} min\ \varphi = {} & 1 * |o_1' - (14 + 23a_1)| + 2 * |o_2' - (22 + 8a_2)| \\ & + 3 * |o_3' - (64 + 89a_3)| + 1 * |o_4' - (8 + 53a_4)| \\ s.t. \begin{cases} 1 - \frac{|o_1' - o'|}{o'} \geq 0.8 \quad 1 - \frac{|o_2' - o'|}{o'} \geq 0.8 \\ 1 - \frac{|o_3' - o'|}{o'} \geq 0.8 \quad 1 - \frac{|o_4' - o'|}{o'} \geq 0.8 \\ o' = 0.4o_1' + 0.3o_2' + 0.2o_3' + 0.1o_4' \\ o_i', o' \geq 0, 0 \leq a_i \leq 1, i = 1, 2, 3, 4 \end{cases} \end{aligned} \tag{8}$$

Let $|o_i' - [o_{li} + a_i * (o_{ri} - o_{li})]| = u_i + v_i$ and $o_i' - [o_{li} + a_i * (o_{ri} - o_{li})] = u_i - v_i$, Model (8) is equivalent to the following linear programming model:

$$\begin{aligned} min\ \varphi = {} & u_1 + v_1 + 2u_2 + 2v_2 + 3u_3 + 3v_3 + u_4 + v_4 \\ s.t. \begin{cases} o_1' - u_1 + v_1 - 23a_1 = 14 \\ o_2' - u_2 + v_2 - 8a_2 = 22 \\ o_3' - u_3 + v_3 - 89a_3 = 64 \\ o_4' - u_4 + v_4 - 53a_4 = 8 \\ 1 - \frac{|o_1' - o'|}{o'} \geq 0.8 \quad 1 - \frac{|o_2' - o'|}{o'} \geq 0.8 \\ 1 - \frac{|o_3' - o'|}{o'} \geq 0.8 \quad 1 - \frac{|o_4' - o'|}{o'} \geq 0.8 \\ a_1 \leq 1, a_2 \leq 1, a_3 \leq 1, a_4 \leq 1 \\ o' = 0.4o_1' + 0.3o_2' + 0.2o_3' + 0.1o_4' \\ o', o_i', u_i, v_i, a_i \geq 0, i = 1, 2, 3, 4 \end{cases} \end{aligned} \tag{9}$$

The model is solved using MATLAB, a widely-used computational tool for model solution. MATLAB provided a flexible and efficient platform to model the cost consensus scenarios and analyze the results. The unique solution of Model (9) is $\bar{X}^* = (17.08, 0, 12.67, 0, 0, 0, 0, 0, 54.08, 42.67, 64, 61, 53.33, 1, 1, 0, 1, 42.42)^T$, and the optimal value of the objective function for Model (9) is $min\ \varphi = 42.42$. The optimal consensus opinions for the four decision-makers are $o_1^* = 54.08$,

$o_2^* = 42.67$, $o_3^* = 64$, and $o_4^* = 61$. The optimal consensus opinion for the group is $o^* = 53.33$, meeting the desired consensus level of $\gamma = 0.8$.

Case 2 Let the data be taken from Case 1, the optimization model under consensus level for the whole group with no aggregation operator is constructed as follow:

$$\begin{aligned} min\ \varphi &= 1 * |o_1' - (14 + 23a_1)| + 2 * |o_2' - (22 + 8a_2)| \\ &\quad + 3 * |o_3' - (64 + 89a_3)| + 1 * |o_4' - (8 + 53a_4)| \\ s.t. &\begin{cases} 1 - \frac{|o_1' - o'|}{o'} \geq 0.8 \quad 1 - \frac{|o_2' - o'|}{o'} \geq 0.8 \\ 1 - \frac{|o_3' - o'|}{o'} \geq 0.8 \quad 1 - \frac{|o_4' - o'|}{o'} \geq 0.8 \\ o_i', o' \geq 0, 0 \leq a_i \leq 1, i = 1, 2, 3, 4 \end{cases} \end{aligned} \tag{10}$$

Let $|o_i' - [o_{li} + a_i * (o_{ri} - o_{li})]| = u_i + v_i$ and $o_i' - [o_{li} + a_i * (o_{ri} - o_{li})] = u_i - v_i$, Model (10) is equivalent to the following linear programming model:

$$\begin{aligned} min\ \varphi &= u_1 + v_1 + 2u_2 + 2v_2 + 3u_3 + 3v_3 + u_4 + v_4 \\ s.t. &\begin{cases} o_1' - u_1 + v_1 - 23a_1 = 14 \\ o_2' - u_2 + v_2 - 8a_2 = 22 \\ o_3' - u_3 + v_3 - 89a_3 = 64 \\ o_4' - u_4 + v_4 - 53a_4 = 8 \\ 1 - \frac{|o_1' - o'|}{o'} \geq 0.8 \quad 1 - \frac{|o_2' - o'|}{o'} \geq 0.8 \\ 1 - \frac{|o_3' - o'|}{o'} \geq 0.8 \quad 1 - \frac{|o_4' - o'|}{o'} \geq 0.8 \\ a_1 \leq 1, a_2 \leq 1, a_3 \leq 1, a_4 \leq 1 \\ o', o_i', u_i, v_i, a_i \geq 0, i = 1, 2, 3, 4 \end{cases} \end{aligned} \tag{11}$$

The unique solution of Model (11) is $\bar{X}^* = (5.67, 0, 12.67, 0, 0, 0, 0, 0, 42.67, 42.67, 64, 42.67, 53.33, 1, 1, 0, 0.65, 31)^T$, and the optimal value of the objective function of Model (11) is min $\varphi = 31$. The optimal consensus opinions of the four decision-makers $o_1^* = 42.67$, $o_2^* = 42.67$, $o_3^* = 64$, $o_4^* = 42.67$, and the optimal consensus opinion for the group is $o^* = 53.33$, which satisfies the expected level $\gamma = 0.8$.

We calculated the results of using different WA aggregation operators under different consensus levels. The consensus results when aggregation operator is not used are also calculated for comparison. Table 1, 2, 3 and 4 summarize the results.

Tables 1, 2, 3 and 4 present the results of consensus models under different aggregation operators. The analysis from the tables reveals the following insights: (1) Consensus level is a crucial parameter influencing the total cost of consensus models. Across varying consensus levels (ranging from 0.35 to 1), the models in the four tables exhibit a similar trend. As consensus level increases, the total consensus cost also rises, indicating a positive and nonlinear correlation between consensus level and total consensus cost. (2) The choice of aggregation operator is a significant factor impacting the total cost of different consensus models. Under the same consensus level, diverse consensus models exhibit variations in total consensus cost. The models with harmonic weighted average operator (HWAO) show the highest total cost, while models without any aggregation operator demonstrate the lowest total

Table 1 Results of the consensus model with AWAO

Consensus	$\gamma =$ 0.5	$\gamma =$ 0.55	$\gamma =$ 0.6	$\gamma =$ 0.65	$\gamma =$ 0.7	$\gamma =$ 0.75	$\gamma =$ 0.8	$\gamma =$ 0.85	$\gamma =$ 0.9	$\gamma =$ 0.95	$\gamma =$ 1
o_1^*	37	40.59	44.54	48.16	49.98	51.94	54.08	56.4	58.93	58.1	61
o_2^*	30	30	30	30.81	34.46	38.4	42.67	47.3	52.36	55.19	61
o_3^*	64	64	64	64	64	64	64	64	64	61	61
o_4^*	60.67	61	61	61	61	61	61	61	61	61	61
o^*	42.67	44.14	45.71	47.41	49.23	51.2	53.53	55.65	58.18	58.1	61
Total cost	0	3.59	7.54	12.79	21.9	31.75	42.42	54.01	66.66	80.48	95

Table 2 Results of the consensus model with GWAO

Consensus	$\gamma =$ 0.4	$\gamma =$ 0.45	$\gamma =$ 0.5	$\gamma =$ 0.55	$\gamma =$ 0.6	$\gamma =$ 0.65	$\gamma =$ 0.7	$\gamma =$ 0.75	$\gamma =$ 0.8	$\gamma =$ 0.85	$\gamma =$ 0.9	$\gamma =$ 0.95	$\gamma =$ 1
o_1^*	36.7	38.23	41.49	45.16	49.1	52.69	53.53	54.76	55.55	57.19	59.2	58.11	61
o_2^*	29.71	30	30	30	30	30.81	34.46	38.4	42.67	47.3	52.36	55.19	61
o_3^*	64.29	64	64	64	64	64	64	64	64	64	64	61	61
o_4^*	57.72	61	61	61	61	61	61	61	61	61	61	61	61
o^*	40.32	41.29	42.67	44.14	45.71	47.41	49.23	51.2	53.33	55.65	58.18	58.1	61
Total cost	0	1.23	4.49	8.16	12.1	17.32	25.45	34.56	43.88	54.8	66.93	80.49	95

Table 3 Results of the consensus model with HWAO

Consensus	$\gamma =$ 0.35	$\gamma =$ 0.4	$\gamma =$ 0.45	$\gamma =$ 0.5	$\gamma =$ 0.55	$\gamma =$ 0.6	$\gamma =$ 0.65	$\gamma =$ 0.7	$\gamma =$ 0.75	$\gamma =$ 0.8	$\gamma =$ 0.85	$\gamma =$ 0.9	$\gamma =$ 0.95	$\gamma =$ 1
o_1^*	36.87	39.08	42.31	46.12	49.73	52.01	54.5	57.23	57.52	57.52	58.29	59.76	59.08	61
o_2^*	29.89	30	30	30	30.46	31.85	33.37	35.05	38.4	42.67	47.3	52.36	55.78	61
o_3^*	64.12	64	64	64	64	64	64	64	64	64	64	64	61.65	61
o_4^*	59.19	61	61	61	61	61	61	61	61	61	61	61	61	61
o^*	38.92	40	41.29	42.67	44.14	45.71	47.41	49.23	51.2	53.33	55.65	58.18	58.71	61
Total cost	0	2.08	5.31	9.12	13.65	18.71	24.24	30.33	37.32	45.85	55.9	67.48	80.69	95

cost. Additionally, the impact of HWAO is more pronounced than that of AWAO and GWAO, suggesting that the use of aggregation operators increases the overall cost of consensus. (3) The consensus models with different aggregation operators show similar trends for different consensus values, but with some differences. This illustrates that the consensus models using different aggregation operators have different advantages and disadvantages, and the appropriate aggregation operator needs to be selected according to the specific situation and objectives in practical applications.

As the consensus level rises, the consensus cost for the four models gradually increases. When the consensus level reaches 1, the consensus cost for all four mod-

Table 4 Results of the consensus model without WA

Consensus	$\gamma =$ 0.6	$\gamma =$ 0.65	$\gamma =$ 0.7	$\gamma =$ 0.75	$\gamma =$ 0.8	$\gamma =$ 0.85	$\gamma =$ 0.9	$\gamma =$ 0.95	$\gamma = 1$
o_1^*	27.43	30.81	34.46	38.4	42.67	47.3	52.36	57.9	61
o_2^*	27.43	30.81	34.46	38.4	42.67	47.3	52.36	57.9	61
o_3^*	64	64	64	64	64	64	64	64	61
o_4^*	27.43	30.81	34.46	38.4	42.67	47.3	52.36	57.9	61
o^*	45.71	47.41	49.23	51.2	53.33	55.65	58.18	60.95	61
Total cost	0	1.63	8.92	18.2	31	44.91	60.09	76.71	95

els equals 95. This implies that decision-makers in the group have achieved full consensus, with each decision-maker's opinion matching the consensus opinion.

Comparing models with aggregation operators with the model that is without any aggregation operators, it can be found that, under equal consensus levels, the minimum consensus cost of models with aggregation operators is always greater than that of the model without any aggregation operators. This is because the aggregation operator in the consensus model increases the consensus cost. When the consensus level equals to 1, the aggregation operator has no impact on the consensus results. The consensus model utilizing the aggregation operator will degenerate into a model without the aggregation operator, and the minimum consensus cost of the final model will remain the same.

5 Conclusion

Consensus decision-making in management activities needs to achieve an agreement through decision makers, and involves the reaching of a certain consensus after thorough discussion. Aggregation operators have been widely used in cost consensus model to facilitate decision-makers in achieving consensus. However, the impact of aggregation operators on cost consensus remains unclear. This paper constructs cost consensus models using three different weighted average aggregation operators and a model without any aggregation operator. The impact of different WAOs on consensus results and the role of aggregation operator in cost consensus are revealed. We find that the use of aggregation operator in soft consensus increases the consensus cost.

The compensation strategy employed by the moderator for decision-makers affects the results of consensus achievement in consensus decision-making, which is a worthwhile issue to study. One potential avenue for our future research involves investigating the impact of moderator's compensation strategy for the decision-makers on consensus cost.

Funding This research received funding support from the National Natural Science Foundation of China (#72201213).

References

1. Chao, X., Kou, G., Li, T., Peng, Y.: Jie Ke versus AlphaGo: a ranking approach using decision making method for large-scale data with incomplete information. Eur. J. Oper. Res. **265**(1), 239–247 (2018)
2. Herrera-Viedma, E., Cabrerizo, F.J., Kacprzyk, J., Pedrycz, W.: A review of soft consensus models in a fuzzy environment. Inform Fusion **17**, 4–13 (2014)
3. Liu, Y., Dong, Y., Liang, H., Chiclana, F., Herrera-Viedma, E.: Multiple attribute strategic weight manipulation with minimum cost in a group decision making context with interval attribute weights information. IEEE Trans. Syst. Man Cybern.: Syst. 1–12 (2018)
4. Shen, Y., Ma, X., Xu, Z., Herrera-Viedma, E., Maresova, P., Zhan, J.: Opinion evolution and dynamic trust-driven consensus model in large-scale group decision-making under incomplete information. Inform. Sci. **657**, 119925 (2024)
5. Wu, Z., Song, Y., Ji, Y., Qu, S., Gong, Z.: Data-driven distributionally robust support vector machine method for multiple criteria sorting problem with uncertainty. Appl. Soft Comput. **149**, 110957 (2023)
6. Zhang, H., Kou, G., Peng, Y.: Minimum cost consensus models measuring moderator's preference on consensus levels. IEEE Trans. Syst. Man Cybern.: Syst. **53**(5), 2938–2948 (2023)
7. Ben-Arieh, D., Easton, T.: Multi-criteria group consensus under linear cost opinion elasticity. Decis. Support Syst. **43**(3), 713–721 (2007)
8. Ben-Arieh, D., Easton, T., Evans, B.: Minimum cost consensus with quadratic cost functions. IEEE Trans. Syst. Man Cybern.: Syst. **39**(1), 210–217 (2009)
9. Gong, Z.W., Xu, X.X., Zhang, H.H., Ozturk, U.A., Herrera-Viedma, E., Xu, C.: The consensus models with interval preference opinions and their economic interpretation. Omega-Int. J. Manag. Sci. **55**, 81–90 (2015)
10. Guo, W., Zhang, W.-G., Gong, Z., Kou, G., Xu, X.: Multi-round minimum cost consensus model with objectivity-fairness driven feedback mechanism. Inform. Fusion **104**, 102185 (2024)
11. Li, Y., Zhang, H.J., Dong, Y.C.: The interactive consensus reaching process with the minimum and uncertain cost in group decision making. Appl. Soft Comput. **60**, 202–212 (2017)
12. Liu, J., Chan, F.T., Li, Y., Zhang, Y., Deng, Y.: A new optimal consensus method with minimum cost in fuzzy group decision. Knowl.-Based Syst. **35**, 357–360 (2012)
13. Zhang, G., Dong, Y., Xu, Y., Li, H.: Minimum-cost consensus models under aggregation operators. IEEE Trans. Syst. Man Cybern.: Syst. **41**(6), 1253–1261 (2011)
14. Zhang, H., Kou, G., Peng, Y.: Soft consensus cost models for group decision making and economic interpretations. Eur. J. Oper. Res. **277**(3), 964–980 (2019)
15. Xu, W., Li, J., Huang, S.: A direct consensus framework based on extended MCCM for multiperson decision making problem with different preference representation structures. J. Intell. Fuzzy Syst. **33**(2), 1173–1186 (2017)
16. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision-making. IEEE Trans. Syst. Man Cybern. **18**(1), 183–190 (1988)

Multi-layer Cross-Scale Coupling Feature Pyramid Network for Food Logo Detection

Baisong Zhang, Sujuan Hou, Songhui Zhao, Qiang Hou, Xiaojie Li, and Wuxia Yan

Abstract Food logos typically serve as a visual representation of a food brand or product by using unique and recognizable images or graphics that are related to the brand or product, such as food items, utensils, or cooking equipment. The study of food logos has important real-world applications, such as in food safety, self-service shops, food recommendation systems, and food brand management. Although detection technology is developing rapidly, these techniques generally suffer from the issue of small logos. Small food logos usually have fewer pixels, making it difficult to extract discriminative features. To address this problem, a Multi-layer Cross-Scale Coupling Feature Pyramid Network (MCCFP-Net) is proposed, which can enhance semantic information by coupling multi-layer features. Specifically, the cross-scale connections break through the limitations imposed by neighboring nodes, enabling the learning of more fundamental features and higher-level semantics. Moreover, to address the problem of similar appearance of different logo categories in classification tasks, the Feature Transformation Offset and Weight (FTOW) is designed to adaptively change the feature regions and assign feature weights. Additionally, Side-Aware Boundary Localization (SABL) is adopted to locate the small food logo objects more precisely. Extensive experimental results demonstrate the effectiveness of our proposed MCCFP-Net.

Keywords Food logo detection · Feature pyramid · Small food logo detection · Feature coupling

B. Zhang · S. Hou (✉) · S. Zhao · Q. Hou · X. Li (✉)
School of Information Science and Engineering, Shandong Normal University, Jinan 250358, Shandong, China
e-mail: sujuanhou@sdnu.edu.cn

X. Li
e-mail: xli162011@163.com

W. Yan
School of Mathematics and Information Science, Nanjing Normal University of Special Education, Nanjing 210038, Jiangsu, China

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_11

1 Introduction

Recently, food computing and food recommendation have drawn much attention in the field of multimedia, as well as food health and diet [1–5]. The detection of food is of great significance for ensuring food safety, protecting consumer rights, and preventing issues such as low-quality, fake, and deceptive products. Inspecting food logos and labels can ensure that they are truthful, accurate, clear, and comprehensive. Food logo detection can help consumers identify authentic products and fundamentally eliminate counterfeit and inferior products. Moreover, it can promote competition among businesses and foster the healthy development of the food industry. In food safety supervision, food trademark testing can also function as a regulatory tool to strengthen the supervision and management of food production and sales. Food logo detection is a special form of logo detection that can be applied to food recommendations, food computing and various food-related applications and services. The development of deep learning has expanded the application range of computer vision [6–11], and the technology of logo detection and food logo detection holds increasing significant commercial value [12–16].

Food logo detection, however, faces several challenges in reality, mainly due to three aspects, as illustrated in Fig. 1. Firstly, fewer pixels in an image result in weakly related semantics in low-level feature maps used for detecting small food logo objects. Extracting effective semantic information directly from fewer pixels is difficult. Without adequate semantic information, detecting small food logo objects and achieving high performance becomes challenging. Secondly, the similar appearance of different food logo categories can lead to incorrect classification, which can degrade the overall classification performance. Finally, complex scenes with noisy regions can interfere with the localization of food logo objects and lead to more challenging tasks for the accurate detection of small food logo objects.

In order to address the challenges posed by small objects, several methods have been proposed using feature pyramid structures. FPN [17] introduces a top-down pathway to build a feature pyramid by sequentially combining high-level and low-level feature maps via top-down and lateral connections. The low-level feature maps are used to detect small objects, while high-level feature maps are used to detect large objects. PANet [18] augments an extra bottom-up pathway to improve the performance of detecting large objects by propagating information from low-level to high-level features. However, its effects on feature extraction for small objects are limited. BFP [19] integrates balanced multiscale features and refines the original features by adding a non-local block, focusing on the balance between feature maps at all levels. While BFP can effectively detect small objects using low-level feature maps, the weakly related semantics in these feature pyramid structures still hamper the overall detection performance.

To address the limitations of previous methods, we propose a Multi-layer Cross-Scale Coupling Feature Pyramid Network (MCCFP-Net). Our model builds a powerful representation feature pyramid with fine resolution by coupling multi-layer cross-scale features, which enhances discriminative abilities in extracting features

Fig. 1 The main challenges of food logo detection

of small food logo objects. Additionally, our model introduces Feature Transformation Offset and Weight (FTOW) to adaptively assign feature weight, increasing the accuracy of small food logo classification. We also adopt Side-Aware Boundary Localization (SABL) [20] to locate food logo objects more precisely, which is effective in handling the localization of small food logo objects in complex scenes and improving localization performance. We evaluate our method on three logo datasets and the results confirm its effectiveness.

This work makes several contributions as follows:

- We propose a new feature fusion strategy, named Multi-layer Cross-Scale Coupling Feature Pyramid, to effectively extract discriminative features from small food logos with fewer pixels.
- We design Feature Transformation Offset and Weight to address the problem of misclassification caused by similar appearing logo categories. This approach adaptively adjusts feature transformation offset and assigns feature weight to improve classification accuracy.

– We adopt Side-Aware Boundary Localization to precisely locate small logo objects in complex scenes.
– We evaluate our proposed MCCFP-Net on the FoodLogoDet-1500 dataset and compare it with a wide range of state-of-the-art detection methods. The extensive evaluations demonstrate the effectiveness of our proposed method. Furthermore, experiments on QMUL-OpenLogo and FlickrLogos-32 datasets further validate the versatility of MCCFP-Net.

2 Related Work

Feature Pyramid Architecture. Size variation is a well-known challenge in the detection task, especially for small objects. Feature pyramid architecture is an effective way to alleviate the problem arising from small objects, and many detectors [21–26] often utilize feature pyramid architecture to improve detection performance. FPN [17] is one of the most classical feature fusion structures, it consists of a top-down pathway and lateral connections to merge multi-level representations into high-resolution feature maps. Zhao et al. [27] proposed the M2Det to fuse multi-scale features via using a U-shape module. FPG [28] built a regular grid of parallel pyramid pathways to fuse multi-directional features. These feature pyramid variants improve the performance of detection. However, they are not efficient enough for our tasks, because it is hard to extract effective semantics information from fewer pixels, and features in low-level feature maps are weakly related to semantics.

Logo Detection. Early logo detection methods are established on hand-crafted visual features (e.g., SIFT [29], HOG [30] and Harr-like [31]) and traditional classification models (e.g., SVM [32]). With the recent advancements in deep learning-based object detection, there has been significant progress in logo detection. A large number of current logo detection methods employ object detection models for this purpose directly. Bao et al. [33] utilized the classic Faster R-CNN for logo detection, while Velazquez et al. [34] adopted FPN to enhance the detection accuracy.

There are also some methods that make some adjustments to improve the logo detection performance. Wilms et al. [13] proposed a detection system to detect logo objects in real-world scenarios. Xu et al. [35] proposed a baseline to solve the problem of long-tail category on logo detection. Eggert et al. [36] put forward a generating anchor proposals scheme for detecting small logo objects. Ke et al. [37] proposed an augmentation strategy for logo datasets to improve logo detection performance.

Food Logo Detection. Food logo detection is a significant subfield within logo detection that has been garnering increasing attention in recent times. Wang et al. [15] introduced the LogoDet-3K dataset, which has 3000 categories with 932 categories belonging to food logos. Nevertheless, the detection solution proposed by Wang et al. is not specifically designed for food logo detection. Hou et al. [12] expanded the food logos to 1500 categories, and presented the FoodLogoDet-1500 dataset

which is the largest food logo dataset currently available. They also designed a decoupling network to distinguish between multiple food logo categories. However, their network lacked attention to small food logo objects.

In contrast to these methods, our approach relies on the multi-layer cross-scale coupling feature pyramid to improve the semantic information of fine resolution feature maps and obtain better feature representation of small food logo objects. Meanwhile, feature transformation offset and weight is designed to solve the problem of misclassification caused by the similar appearance, which can increase the accuracy of small food logo classification. We also adopt side-aware boundary localization to accurately locate bounding boxes in complex scenes.

3 Data and Methods

3.1 Datasets

In this work, the performance comparison was conducted on three datasets, and their detailed statistics of them are shown in Table 1. The experiments were mainly conducted on the FoodLogoDet-1500 [12], which is currently the largest food logo detection dataset. It comprises 1500 categories of food logos and includes 99,768 images. More specifically, FoodLogoDet-1500 contains 145,400 food logo objects, out of which 16,463 are small logo objects, accounting for 11.32% of the total. To evaluate the generality of the MCCFP-Net, experiments were also performed on two other publicly available logo datasets, namely QMUL-OpenLogo [38], and FlickrLogos-32 [39].

3.2 Evaluation and Implementation Details

In this study, we have utilized the commonly employed mean Average Precision (mAP) metric to evaluate the performance. Additionally, we have adopted the following evaluation metrics to showcase the effectiveness of our method across logos of various sizes: AP_S as the Average Precision (AP) for small logo objects (area $<$

Table 1 Statistics of three logo datasets

Datasets	Supervision	#Classes	#Images	#Objects	#Trainval	#Test
FoodLogoDet-1500 [12]	Object-level	1500	99,768	145,400	80,280	19,488
QMUL-OpenLogo [38]	Object-level	352	27,083	51,207	18,752	8,331
FlickrLogos-32 [39]	Object-level	32	2,240	3,405	1,478	762

32^2), AP_M as the AP for medium logo objects ($32^2 <$ area $< 96^2$), and AP_L as the AP for large logo objects (area $> 96^2$). The threshold of Intersection over Union (IoU) between the predicted bounding box and ground-truth bounding box is set to 0.5.

The proposed framework was implemented based on the ResNet-50 backbone, which was pre-trained on the ImageNet [40]. For fair comparisons, all baseline detectors were re-implemented based on the publicly available mmdetection toolbox [41], and the GPU used in the experiment is NVIDIA 2080Ti. These detectors were trained with an initial learning rate of 2.5×10^{-3} and the input images were resized to 1000×600.

3.3 Methods

Overall Architecture of the Approach. As depicted in Fig. 2, the framework of MCCFP-Net comprises three primary components: MCCFP, FTOW, and SABL. Specifically, the image is first processed by the backbone network of ResNet-50 to extract preliminary features (B_i), which are then input into MCCFP to obtain discriminative features (P_i). After RPN and RoI pooling, and the classification is performed using FTOW, the localization is achieved through SABL. Furthermore, we provide the overall algorithm of MCCFP-Net in Algorithm 1.

Multi-layer Cross-Scale Coupling Feature Pyramid. Different from previous feature pyramid methods, the local fusion operation of MCCFP simultaneously integrates both high-level and backbone representations. Let B_i denote the feature map of the ith stage in the backbone network. C_i is the feature map of ith stage in the intermediate process, P_i is the output of the feature pyramid at level i, and f_i is the ith multi-scale feature fusion operation. In short, the operations in the multi-layer cross-scale coupling of MCCFP is described as:

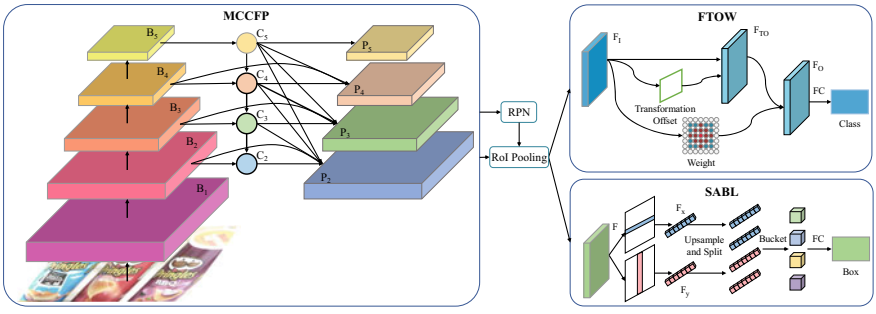


Fig. 2 The framework of Multi-layer Cross-Scale Coupling Feature Pyramid Network (MCCFP-Net)

Algorithm 1 Multi-layer Cross-Scale Coupling Feature Pyramid Network (MCCFP-Net)

Input: Training images I_{train} with class label C_{train} and localization L_{train}
Output: Predicted localization L_{test} and class labels C_{test} on testing images I_{test}
1: Extract feature maps FM by MCCFP on I_{train}
2: Generate anchor boxes L_{train_anchor} by RPN on FM
3: Train a localization F_{reg} by SABL on FM with L_{train_anchor}
4: Train a classification F_{cls} by FTOW on FM with C_{train}
5: Use F_{reg} to predict L_{test} on I_{test}
6: Use F_{cls} to predict C_{test} on I_{test}
7: **Return**: L_{test}, C_{test}

$$P_i = \begin{cases} f_i(C_i) & i = 5 \\ f_i(C_i, C_{i+1}, \cdots C_{max}, B_i) & i = 4, 3, 2 \end{cases} \tag{1}$$

where high-level features $C_{i+1}, \cdots C_{max}$ are propagated through the classical nearest interpolation function for upsampling.

As shown in Fig. 3, we employ parallel-cross, vertical-cross, and upsampling to build the multi-layer cross-scale coupling pathway. The concrete operations are as follows:

– Parallel-Cross pathway: In order to obtain original features from the backbone stage, we introduce a parallel-cross pathway. It basically consists of a 1×1 convolutional layer. Therefore, the feature hierarchy from the backbone consists of multi-scale feature maps with the same channels.
– Vertical-Cross pathway: In order to directly obtain high-level features, we introduce the vertical-cross pathway. We use multiple upsampling to get the corre-

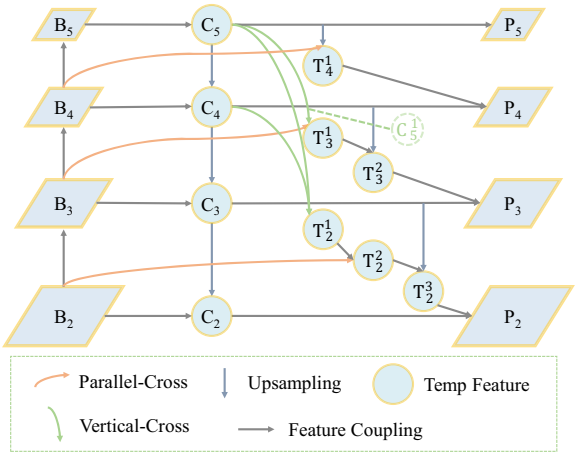


Fig. 3 The detailed structure of MCCFP

sponding size of the higher-level feature map and then fuse it with the low-level feature map.
- Upsampling: We upsample the $(i + 1)$th feature map by a scaling factor of 2 with the nearest neighbor upsampling for simplicity to obtain the same resolution.
- Feature Coupling: We merge two different scale feature maps by adding them together, and use a 3×3 convolutional layer to make this operation learnable.

Take P_3 as an example, we can get it through the following operations. As shown in Fig. 3, the vertical-cross method is used to obtain C_5^1 from the cross-scale high-level feature map (C_5 here). We use the feature coupling operation on the C_5^1 and the B_3 to get T_3^1. Then we take the upsampling operation on the high-level feature map C_4, and merge it with the feature map T_3^1 by the feature coupling operation to get T_3^2. Finally, the feature map T_3^2 is coupled into C_3, and through a 3×3 convolutional layer, we get the feature map P_3. The fusion operation is described as:

$$P_3 = f_3\{f_3[f_3(C_5, B_3), C_4], C_3\} \tag{2}$$

f_3 is the third multi-scale feature fusion operation. Firstly, it resizes multi-scale feature maps to the resolutions of P_3, and then these feature maps are coupled in pairs. Finally, it uses a 3×3 convolutional layer to reduce the aliasing effect. The feature maps of the other levels are acquired in a similar manner.

To address the classification difficulties caused by similar small food logos, and the challenges of localization due to the complex background of small food logos, we integrated modulated deformable convolution [42] into FTOW and employed SABL in MCCFP-Net to address the problems of classification and localization, respectively.

Feature Transformation Offset and Weight. The obstacle caused by similar appearance may decrease the performance of logo classification. To address this issue, FTOW is introduced to adaptively assign feature weight and increase the accuracy of small food logo classification. We have implemented FTOW using modulated deformable convolution [42].

As shown in Fig. 2, we apply deformable convolution to the RoI feature map to predict the transformation offset and weight. F_{TO} can be obtained from the RoI feature map modified by transformation offset, and F_{TO} is multiplied by weight to get the final feature map F_O. Therefore, the feature maps in the classification branch have the ability to adjust features and increase the food logo object's feature weight, it thus can better focus on pertinent feature content and deliver greater classification confidence.

Specifically, RoI pooling divides the RoI feature map into I spatial bins. Within each bin, sampling grids of even spatial intervals are applied. The sampled values on the grids are averaged to compute the bin output. The output bin feature $f_{OUT}(i)$ is computed as:

$$f_{OUT}(i) = \sum_{j=1}^{j \le n_i} f_{IN}(L_{ij} + TO_i) \cdot W_i / n_i \tag{3}$$

where L_{ij} is the sampling location for the jth grid in the ith bin, and n_i denotes the number of the sampled grid. TO and W are the Transformation Offset and Weight for the ith bin, respectively. Both TO_i and W_i are obtained via a separate convolution layer applied over the same input feature maps f_{IN}.

Side-Aware Boundary Localization. Complex scenes make it harder for localization of small food logo objects. We follow the Side-Aware Boundary Localization (SABL) [20] for the conventional bounding box regression to locate the small food logo objects more precisely. As shown in Fig. 2, SABL first extracts horizontal and vertical features F_x and F_y by aggregating the RoI features F along X-axis and Y-axis, respectively. Then we upsample F_x and F_y and split them into four side-aware features F_{left}, F_{right}, F_{top} and F_{down}. Each object boundary is divided into the bucket horizontally and vertically. It firstly estimates in which bucket the boundary resides. Specifically, given a proposal box, i.e., (B_{left}, B_{right}, B_{top} and B_{down}), we relax the candidate region of boundaries by a scale factor of σ ($\sigma > 1$) to cover the entire object. The candidate regions are divided into $2k$ buckets on both X-axis and Y-axis, with k buckets corresponding to each boundary. The width of each bucket on X-axis and Y-axis are l_x and l_y defined as follows:

$$\begin{aligned} l_x &= (\sigma B_{right} - \sigma B_{left})/2k \\ l_y &= (\sigma B_{down} - \sigma B_{top})/2k \end{aligned} \tag{4}$$

We use the centerline of the selected bucket as a rough estimate, then perform a fine regression with the predicted offset. The bucketing scheme adopts localization reliability to guide the rescoring, which helps maintain the best box with an accurate localization. SABL could reduce the regression variance and ease the difficulties of localization, and thus can locate small food logo objects more precisely in complex scenes.

Loss Function. The proposed MCCFP-Net is optimized in an end-to-end manner. In the training of the framework, the overall optimization loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \mathcal{L}_{reg} \tag{5}$$

where $\mathcal{L}_{rpn}$, $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ are RPN loss, classification loss and localization loss, respectively.

The $\mathcal{L}_{rpn}$ remains the same as Cascade R-CNN. Here we adopt Cross Entry Loss for $\mathcal{L}_{cls}$,

$$\mathcal{L}_{cls} = \mathcal{L}_{CrossEntry} = -\sum_{k=1}^{N}(p_k \cdot log q_k) \tag{6}$$

where N denotes the number of categories, k denotes one of the categories, p_k is the ground-truth category, and q_k is the predicted category. Moreover, we adopt a variant of Smooth $L1$ Loss to optimize the $\mathcal{L}_{reg}$.

$$\mathcal{L}_{reg} = \mathcal{L}_{SmoothL1} = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{7}$$

where x is the elementwise difference between the prediction box and the ground-truth box.

4 Results

4.1 Comparison of State-of-the-Art Methods

We compare MCCFP-Net with the state-of-the-art detection approaches on the FoodLogoDet-1500 dataset and the results are presented in Table 2. It is noteworthy that the proposed MCCFP-Net outperforms other baselines under all metrics, including mAP, AP_S, AP_M, and AP_L.

Table 2 Detection results on the FoodLogoDet-1500 dataset

Methods	Backbone	$mAP(\%)$	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
One-stage					
ATSS [43]	ResNet-50-FPN	81.5	75.0	78.4	85.3
FSAF [44]	ResNet-50-FPN	79.7	73.9	76.3	83.5
GFL [21]	ResNet-50-FPN	80.0	72.9	76.8	83.7
Two-stage					
Faster R-CNN [26]	ResNet-50-FPN	84.0	75.3	81.0	88.0
PANet [18]	ResNet-50-PAFPN	84.2	75.8	82.1	88.4
Complete IoU [45]	ResNet-50-FPN	83.3	74.2	80.5	87.3
Libra R-CNN [19]	ResNet-50-BPN	83.7	78.0	81.2	87.2
Cascade R-CNN [23]	ResNet-50-FPN	84.3	74.5	80.2	88.5
MFDNet [12]	ResNet-50-BFP	86.6	–	–	–
Dynamic R-CNN [46]	ResNet-50-FPN	85.9	77.9	82.9	90.2
Sparse R-CNN [47]	ResNet-50-FPN	81.7	76.9	80.2	85.1
MCCFP-Net	**ResNet-50-MCCFP**	**86.8**	**79.9**	**82.9**	**90.6**

Fig. 4 Visualization of detection results of MCCFP-Net

Moreover, it can be observed that, compared to existing two-stage methods, MCCFP-Net significantly outperforms these state-of-the-art frameworks and reaches an mAP value of 86.8%. In particular, MCCFP-Net improves AP_S by 4.6, 4.1, 5.7, 1.9, 5.4, 2.0 and 3.0% compared to Faster R-CNN, PANet, Complete IoU, Libra R-CNN, Cascade R-CNN, Dynamic R-CNN and Sparse R-CNN, respectively. Similar observations have been made in comparison with ATSS, FSAF, and GFL, where MCCFP-Net achieves the best results and obtains 4.9, 6.0 and 7.0% improvement on AP_S, respectively.

Furthermore, to verify the effectiveness of MCCFP-Net in detecting small food logo objects, some visual detection results are given in Fig. 4. It can be seen that our model has good detection results for images with a single small food logo object, such as "Weidendorf" and "Budweiser". Moreover, when there are multiple small food logo objects in a test image, such as "Campbelli" and "Del Monte", our model can also detect all small food logo objects accurately. It is worth mentioning that in the third image in the first row, which contains a logo object with a tilt angle on the image, our model achieves an accuracy value of 59%. This demonstrates that MCCFP-Net is robust and can detect hard small food logo objects.

4.2 Experiments on Other Logo Datasets

Experiments on the QMUL-OpenLogo. From Table 3, it can be observed that MCCFP-Net performs best on all evaluation metrics (53.9% mAP, 31.8% AP_S, 54.5% AP_M and 67.5% AP_L). Compared with Cascade R-CNN, MCCFP-Net obtains 3.5 and 2.1% improvement on mAP and AP_S, respectively. Our framework also improves the mAP by 1.1% and 4.2% compared with Libra R-CNN and PANet, respectively. In addition, compared with the best performing one-stage method GFL, our method obtains 6.0% improvement on mAP.

Experiments on the FlickrLogos-32. From Table 4, MCCFP-Net is significantly superior to other baselines on all evaluation metrics with 88.2% mAP, 27.7% AP_S, 88.6% AP_M and 94.3% AP_L. Specifically, MCCFP-Net obtains 0.8, 6.6, 4.8 and 1.3% improvement compared with Cascade R-CNN in terms of mAP, AP_S, AP_M, and AP_L respectively. MCCFP-Net outperforms Libra R-CNN by 0.2% and 0.5% in terms of mAP and AP_S respectively. It also achieves 2.2% improvement in mAP and 7.8% improvement in AP_S compared to PANet. When compared to the best performing one-stage method ATSS, our method outperforms it by 2.3 and 6.7% improvement on mAP and AP_S, respectively. These results further demonstrate the advantages of MCFFP-Net in detecting small logo objects.

4.3 Ablation Study

We conduct empirical analysis for each of the three components in MCCFP-Net, namely MCCFP, FTOW, and SABL. Ablation studies are performed on Cascade R-CNN with ResNet-50 and the overall ablation results are reported on the FoodLogoDet-1500 dataset in Table 5.

MCCFP. We designed MCCFP to address the challenge of extracting semantic features from images with fewer pixels. We evaluate the effect of the MCCFP by comparing it with FPN. MCCFP is mainly used to enrich the feature representation of low-level feature maps. The MCCFP brings 0.7% mAP improvement than the Cascade R-CNN on the FoodLogoDet-1500 dataset in Table 5, especially increases the AP_S by 2.7%. It is worth mentioning that AP_M and AP_L are also improved, as the Parallel-Cross pathway makes a positive impact. In addition, based on FTOW, MCCFP brings 1.2% mAP and 3.5% AP_S improvement. And MCCFP brings 1.1% mAP and 3.0% AP_S improvement than the Cascade R-CNN with SABL. Based on FTOW and SABL, MCCFP brings 1.0% mAP and 3.7% AP_S improvement. Taken together, these results demonstrate the effectiveness of MCCFP.

Moreover, we have generated visual representations of the detection results for two images that contain small logo objects in Fig. 5. In the first column, MCCFP-Net is able to correctly identify two food logo objects below while Cascade R-CNN cannot. In the second column, Cascade R-CNN misses a food logo in the upper left

Table 3 Detection results on the QMUL-OpenLogo

Methods	Backbone	$mAP(\%)$	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
One-stage					
ATSS [43]	ResNet-50-FPN	47.3	29.6	49.5	58.8
FSAF [44]	ResNet-50-FPN	40.2	22.0	42.2	49.8
GFL [21]	ResNet-50-FPN	47.9	28.3	49.4	59.8
Two-stage					
Faster R-CNN [26]	ResNet-50-FPN	51.1	29.5	51.5	65.9
PANet [18]	ResNet-50-PAFPN	49.7	30.5	50.9	63.6
Complete IoU [45]	ResNet-50-FPN	49.2	30.4	51.9	60.5
Libra R-CNN [19]	ResNet-50-BFP	52.8	31.0	54.4	66.1
Cascade R-CNN [23]	ResNet-50-FPN	50.4	29.7	51.1	63.6
MFDNet [12]	ResNet-50-BFP	51.3	–	–	–
Dynamic R-CNN [46]	ResNet-50-FPN	**52.2**	**31.4**	**53.4**	**64.9**
Sparse R-CNN [47]	ResNet-50-FPN	50.5	31.3	51.7	63.9
MCCFP-Net	ResNet-50-MCCFP	**53.9**	**31.8**	**54.5**	**67.5**

corner, but MCCFP-Net could detect all food logo objects well. It is shown that MCCFP can build strong feature representations in high-resolution feature maps.

FTOW. This module is designed to obtain the feature transformation offset and assign feature weight to improve the reliability of classification. As shown in Table 5, FTOW obtains 0.5, 0.5, 0.9, and 0.2% improvement compared with the baseline in terms of mAP, AP_S, AP_M, AP_L respectively. Besides, FTOW improves the mAP from 85.0% to 85.8% and increases AP_S from 75.9 to 76.2% than Cascade R-CNN with SABL. And based on MCCFP, FTOW brings 1.0% mAP and 1.2% AP_S improvement. Besides, FTOW brings 0.7% mAP and 1.0% AP_S improvement than the Cascade R-CNN both with MCCFP and SABL. These results can indicate the advantage of FTOW.

Similarly, we have generated visualizations of the detection results of three images with small logo objects to verify the effectiveness of FTOW. As shown in Fig. 6, we can see category “Chipsmore” (first column) and category “Chipsahoy” (second

Table 4 Detection results on the FlickrLogos-32

Methods	Backbone	$mAP(\%)$	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
One-stage					
ATSS [43]	ResNet-50-FPN	85.9	21.0	79.3	82.4
FSAF [44]	ResNet-50-FPN	82.9	24.0	75.5	89.2
GFL [21]	ResNet-50-FPN	83.3	20.9	80.5	88.6
Two-stage					
Faster R-CNN [26]	ResNet-50-FPN	86.3	22.8	84.9	91.5
PANet [18]	ResNet-50-PAFPN	86.0	19.9	83.0	93.4
Complete IoU [45]	ResNet-50-FPN	87.2	23.4	79.1	93.0
Libra R-CNN [19]	ResNet-50-BFP	88.0	27.2	87.8	92.6
Cascade R-CNN [23]	ResNet-50-FPN	87.4	21.1	83.8	93.0
MFDNet [12]	ResNet-50-BFP	86.2	–	–	–
Dynamic R-CNN [46]	ResNet-50-FPN	87.6	22.3	82.9	93.5
Sparse R-CNN [47]	ResNet-50-FPN	81.6	16.5	72.5	88.6
MCCFP-Net	ResNet-50-MCCFP	**88.2**	**27.7**	**88.6**	**94.3**

Table 5 Evaluating individual component on the FoodLogoDet-1500 dataset

Method	MCCFP	FTOW	SABL	$mAP(\%)$	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
Baseline				84.3	74.5	80.2	88.5
Ours	✓			85.0	77.2	81.1	89.2
		✓		84.8	74.9	81.1	88.7
			✓	85.0	75.9	81.4	89.3
	✓	✓		86.0	78.4	82.8	89.5
		✓	✓	85.8	76.2	82.2	89.8
	✓		✓	86.1	78.9	82.3	90.0
	✓	✓	✓	**86.8**	**79.9**	**82.9**	**90.6**

Fig. 5 Ablation Studies on MCCFP. Qualitative examples (small food logo objects) comparison on MCCFP between Cascade R-CNN and MCCFP-Net. The top in each pair denotes Cascade R-CNN results, while the down denotes MCCFP-Net results. The blue boxes represent ground-truth boxes, the green boxes represent correct detection boxes

column) are extremely similar in appearance, and the classification confidence of MCCFP-Net is significantly higher than that of Cascade R-CNN. In the second column, Cascade R-CNN incorrectly classifies the text as the category "Bikanervala". However, MCCFP-Net can correctly take this text as background. In the third column of Fig. 6, MCCFP-Net also achieves higher classification accuracies than Cascade R-CNN. These results demonstrate the effectiveness of FTOW in classifying small food logo objects.

SABL. We adopt SABL to handle the localization of small food logo objects in complex scenes. As shown in Table 5, SABL improves mAP, AP_S, AP_M, and AP_L by 0.7, 1.4, 1.2, and 0.8%, respectively. SABL improves the mAP from 85.0 to 86.1% and increases AP_S from 77.2 to 78.9% compared with Cascade R-CNN implementing MCCFP, and improves the mAP from 84.8 to 85.8% and increases AP_S from 74.9 to 76.2% compared with Cascade R-CNN implementing FTOW. Based on MCCFP

Fig. 6 Ablation studies on FTOW. Qualitative examples (small food logo objects) comparison on FTOW between cascade R-CNN and MCCFP-Net. The yellow boxes represent mistaken detection boxes

and FTOW, SABL improves the mAP from 86.0 to 86.8% and increases AP_S from 78.4 to 79.9%. As shown in Fig. 7, it is obvious that the detection boxes (green) of MCCFP-Net are closer to ground-truth boxes (blue) than Cascade R-CNN, which proves the effectiveness of SABL in locating small food logo objects.

4.4 Feature Pyramid Architecture

What's more, we compared the performance of MCCFP with other feature pyramid structures on the FoodLogoDet-1500 dataset in Table 6. MCCFP-Net improves mAP by 0.7%, 2.2%, 1.9% and 1.1% compared with FPN, PAFPN, HRFPN and BFP, respectively. MCCFP also improves the AP_S by 2.7, 5.2, 2.9 and 1.7% compared with FPN, PAFPN, HRFPN and BFP, respectively. The experimental results verify the superiority of MCCFP in detecting small food logos.

Fig. 7 Ablation studies on SABL. Qualitative examples (small food logo objects) comparison on SABL between cascade R-CNN and MCCFP-Net. The top in each pair denotes cascade R-CNN results, while the down denotes MCCFP-Net results

Table 6 Detection results on different feature pyramid

Methods	$mAP(\%)$	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
FPN [17]	84.3	74.5	80.2	88.5
PAFPN [18]	82.8	72.0	78.4	87.0
HRFPN [48]	83.1	74.3	79.1	87.7
BFP [19]	83.9	75.5	80.3	87.8
MCCFP	**85.0**	**77.2**	**81.1**	**89.2**

5 Conclusions

In this work, we propose the Multi-layer Cross-Scale Coupling Feature Pyramid Network (MCCFP-Net) to solve the challenges of small food logo objects detection. MCCFP-Net includes three components, namely Multi-layer Cross-Scale Coupling Feature Pyramid (MCCFP), Feature Transformation Offset and Weight (FTOW), and Side-Aware Boundary Localization (SABL). MCCFP builds a multi-layer cross-scale

coupling pathway to enhance the extraction of semantic features of low-level feature maps. FTOW solves the problem of misclassification caused by the similar appearance of different logo categories. SABL is adopted to precisely locate small logo objects in complex scenes. The experiments on three datasets indicate the advantages of MCCFP-Net over the state-of-the-art baseline methods, ranging from one-stage modes to two-stage modes. As for future work, we will continue to work on addressing other challenges, such as large aspect ratios and rotating food logos, to further improve the overall performance of food trademarks. Meanwhile, we are committed to making extensive efforts to develop more effective food logo detection models in real-world scenarios to ensure food safety and quality, promote the development of the healthy food market, and safeguard consumer rights and interests.

References

1. Min, W., Jiang, S., Liu, L., Rui, Y., Jain, R.: A survey on food computing. ACM Comput. Surv. (CSUR) **52**(5), 1–36 (2019)
2. Phanich, M., Pholkul, P., Phimoltares, S.: Food recommendation system using clustering analysis for diabetic patients. In: 2010 International Conference on Information Science and Applications, pp. 1–8. IEEE (2010)
3. Wang, W., Duan, L.-Y., Jiang, H., Jing, P., Song, X., Nie, L.: Market2dish: health-aware food recommendation. ACM Trans. Multimedia Comput. Commun. Appl. (TOMM), **17**(1), 1–19 (2021)
4. Min, W., Jiang, S., Jain, R.: Food recommendation: framework, existing solutions, and challenges. IEEE Trans. Multimedia **22**(10), 2659–2671 (2019)
5. Gao, X., Feng, F., Huang, H., Mao, X.-L., Lan, T., Chi, Z.: Food recommendation with graph convolutional network. Inform. Sci. **584**, 170–183 (2022)
6. Liu, H., Tang, X., Shen, S.: Depth-map completion for large indoor scene reconstruction. Pattern Recogn. **99**, 107112 (2020)
7. Qiao, Y., Cui, J., Huang, F., Liu, H., Bao, C., Li, X.: Efficient style-corpus constrained learning for photorealistic style transfer. IEEE Trans. Image Process. **30**, 3154–3166 (2021)
8. Liu, H., Zhang, Q., Fan, B., Wang, Z., Han, J.: Features combined binary descriptor based on voted ring-sampling pattern. IEEE Trans. Circ. Syst. Video Technol. **30**(10), 3675–3687 (2019)
9. Liu, H., Jin, F., Zeng, H., Pu, H., Fan, B.: Image enhancement guided object detection in visually degraded scenes. IEEE Trans. Neural Netw. Learn. Syst. (2023)
10. Gao, X., Zhu, L., Xie, Z., Liu, H., Shen, S.: Incremental rotation averaging. Int. J. Comput. Vis. **129**, 1202–1216 (2021)
11. Yan, L., Fan, B., Liu, H., Huo, C., Xiang, S., Pan, C.: Triplet adversarial domain adaptation for pixel-level classification of VHR remote sensing images. IEEE Trans. Geosci. Remote Sens. **58**(5), 3558–3573 (2019)
12. Hou, Q., Min, W., Wang, J., Hou, S., Zheng, Y., Jiang, S.: Foodlogodet-1500: a dataset for large-scale food logo detection via multi-scale feature decoupling network. In: Proceedings of the ACM International Conference on Multimedia, pp. 4670–4679 (2021)
13. Wilms, C., Heid, R., Sadeghi, M.A., Ribbrock, A., Frintrop, S.: Which airline is this? Airline logo detection in real-world weather conditions. In: International Conference on Pattern Recognition, pp. 4996–5003. IEEE (2021)
14. Kuznetsov, A., Savchenko, A.V.: A new sport teams logo dataset for detection tasks. In: International Conference on Computer Vision and Graphics, pp. 87–97. Springer (2020)
15. Wang, J., Min, W., Hou, S., Ma, S., Zheng, Y., Jiang, S.: LogoDet-3K: a large-scale image dataset for logo detection. ACM Trans. Multimedia Comput. Commun. Appl. **18**(1), 1–19 (2022)

16. Jin, X., Su, W., Zhang, R., He, Y., Xue, H.: The open brands dataset: unified brand detection and recognition at scale. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4387–4391. IEEE (2020)
17. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
18. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
19. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra R-CNN: towards balanced learning for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 821–830 (2019)
20. Wang, J., Zhang, W., Cao, Y., Chen, K., Pang, J., Gong, T., Shi, J., Loy, C.C., Lin, D.: Side-aware boundary localization for more precise object detection. In: Proceeding of the European Conference on Computer Vision, pp. 403–419. Springer (2020)
21. Li, X., Wang, W., Lijun, W., Chen, S., Xiaolin, H., Li, J., Tang, J., Yang, J.: Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. Adv. Neural Inform. Process. Syst. **33**, 21002–21012 (2020)
22. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9627–9636 (2019)
23. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)
24. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
25. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural Inform. Process. Syst. **28**, 91–99 (2015)
27. Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., Ling, H.: M2Det: a single-shot object detector based on multi-level feature pyramid network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9259–9266 (2019)
28. Chen, K., Cao, Y., Loy, C.C., Lin, D., Feichtenhofer, C.: Feature pyramid grids (2020). arXiv preprint arXiv:2004.03580
29. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE (2005)
30. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
31. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
32. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
33. Bao, Y., Li, H., Fan, X., Liu, R., Jia, Q.: Region-based CNN for logo detection. In: Proceedings of the International Conference on Internet Multimedia Computing and Service, pp. 319–322 (2016)
34. Velazquez, D.A., Gonfaus, J.M., Rodriguez, P., Xavier Roca, F., Ozawa, S., Gonzàlez, J.: Logo detection with no priors. IEEE Access **9**, 106998–107011 (2021)
35. Xu, W., Liu, Y., Lin, D.: A simple and effective baseline for robust logo detection. In: Proceedings of the ACM International Conference on Multimedia, pp. 4784–4788 (2021)
36. Eggert, C., Zecha, D., Brehm, S., Lienhart, R.: Improving small object proposals for company logo detection. In: Proceedings of the ACM International Conference on Multimedia, pp. 167–174 (2017)

37. Ke, X., Du, P.: Vehicle logo recognition with small sample problem in complex scene based on data augmentation. Math. Prob. Eng. 1–10 (2020)
38. Su, H., Zhu, X., Gong, S.: Open logo detection challenge (2018). arXiv preprint arXiv:1807.01964
39. Romberg, S., Pueyo, L.G., Lienhart, R., Van Zwol, R.: Scalable logo recognition in real-world images. In: Proceedings of the ACM International Conference on Multimedia, pp. 1–8 (2011)
40. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
41. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open MMLab detection toolbox and benchmark (2019). arXiv preprint arXiv:1906.07155
42. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable Convnets v2: more deformable, better results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern, pp. 9308–9316 (2019)
43. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9759–9768 (2020)
44. Zhu, C., He, Y., Savvides, M.: Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 840–849 (2019)
45. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-IoU loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12993–13000 (2020)
46. Zhang, H., Chang, H., Ma, B., Wang, N., Chen, X.: Dynamic R-CNN: towards high quality object detection via dynamic training. In: Proceeding of the European Conference on Computer Vision, pp. 260–275. Springer (2020)
47. Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse R-CNN: end-to-end object detection with learnable proposals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 14454–14463 (2021)
48. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions (2019). arXiv preprint arXiv:1904.04514

Identification of Imaging Genetics Association for Mild Cognitive Impairment Based on Adaptive Constrained Canonical Correlation Analysis

Ruolan Du and Wei Luo

Abstract Mild cognitive impairment (MCI) is a progressive neurodegenerative disease, with primary clinical manifestations including memory deterioration, declining cognitive abilities, and behavioral issues. Considering the progression over time, the disease advances with regional variations in different brain areas, possibly influenced by genetic molecular functions. Therefore, to determine the correlation between imaging genetics data, this study explores the association between multiple time point brain regions and genetic data based on the adaptive sparse multi-view canonical correlation analysis algorithm, aiming to fully utilize data from multiple time points. The study fully utilizes voxel data of brain regions at two time points of MCI and corresponding gene expression data of samples, thereby exploring high-order correlated features of brain imaging genetic data. This includes conducting bioinformatics analysis on identified high-risk brain regions and top-ranking genes to validate the supplemental effects of time point data. This research offers a new perspective on MCI and contributes to the discovery of potential biomarkers and early intervention strategies.

Keywords Imaging genetics · Association analysis · Mild cognitive impairment

1 Introduction

Mild cognitive impairment (MCI) is a progressive neurodegenerative disorder that can have significant financial and emotional consequences for individuals and their families [1, 2]. The main clinical manifestation of early cognitive impairment is declining in cognitive function, which worsen over time [3–5]. Disease progression occurs at different time points and brain regions [6–8]. Besides, the regional differences were influenced by many genetic molecules. Understanding complex genetic

R. Du · W. Luo (✉)
South China Agricultural University, Guangzhou 510642, China
e-mail: dsnora@163.com

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_12

factors that affect the association patterns of imaging genetics has been on the rise in recent years [9–11]. Identifying biomarkers of the disease development process is crucial for early intervention, treatment, and halting the progression of early cognitive impairment [12].

Sparse canonical correlation analysis (SCCA)-based methods were recently applied to mine the association of imaging genetics data [13, 14]. It aims to extract the significant association patterns based on the data weights. Building upon SCCA, Fang et al. introduced the Joint Sparse Canonical Correlation Analysis (JSCCA) technique, which incorporates sparsity modeling and graph theory [15]. JSCCA improves the accuracy of joint analysis algorithms for imaging data. Later, the study proposed connectivity based JSCCA method. It is demonstrated its robustness in selecting biomarkers related to Parkinson's disease using multiple data [16]. The quantitative traits (QTs) have been tested by considering the nonlinear effects of SNPs and produces more realistic impacts compared to traditional linear models [17]. However, most existing methods only pay attention to the association of genetic sites at one time point, disregarding vast amount carried by the continuous evolution of QTs over multiple time points. Moreover, several researchers have considered the temporal information of imaging data. One common approach is to enhance the SCCA algorithm by incorporating sparse terms. This helps in identifying correlated markers and constructing diagnostic models, allowing the algorithm to identify a small number of strongly association of multiple data. Hao et al. proposed a valuable algorithm called time-constrained group SCCA. This algorithm aims at identifying the relationship between SNPs and longitudinal sMRI phenotypes [18]. Similarly, Du et al. introduced the multi-task SCCA to analyze the affect changes of SNPs in brain QT [10]. Later, several studies have started considering multiple regression or classification to uncover associations between biomarkers of Alzheimer's disease and different modal data [19].

Considering the limitations of linear models in uncovering the complex nonlinear relationships between genetic factors and brain structure/function, as well as the current lack of methods for analyzing image genetic data across multiple time points, it is crucial to explore how to leverage data from multiple time points to explore the association of imaging genetics. Thus, our study will employ an algorithm called AdaSMCCA to analyze bivariate correlations between longitudinal structural MRI data and gene expression data at two time points [20]. By utilizing voxel data from specific brain regions and gene expression data from corresponding samples at both MCI time points, we aim to identify higher-order correlation features in brain imaging genetic data. Furthermore, we will conduct bioinformatics analysis on the identified high-risk brain regions and top-ranked genes to validate the complementary effect of multi-time-point data.

2 Materials and Methods

2.1 Data Collection and Preprocessing

The longitudinal brain imaging data were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) website (https://adni.loni.usc.edu), which included 147 participants, comprising 144 individuals with mild cognitive impairment (MCI) and 56 normal controls (NC). Neuroimaging was retrieved from sMRI data from baseline (T1), 6 months (T2), and excluded brain imaging samples with unclear structures and distorted images, a total of 210 samples. 144 MCI patients and 56 NC patients were screened out. Then, the SPM12 program in MATLAB was used to segment, register, and extract features from 2 brain images of each sample. After 122 ROIs were extracted using the Anatomical Automatic labeling (AAL) template, 32 cerebellar structures were removed, and 90 ROIs were retained as the final features of the brain images.

2.2 Adaptive Sparse Multi-view Canonical Correlation Analysis Model

Neuroimaging was retrieved from sMRI data from baseline. In this study, to explore high-order correlations between MRI and gene features, the study introduces hypergraph Laplacian matrices for each of the three original matrices, the objective function of the AdaSMCCA algorithm is formulated as follows in Eq. (1):

$$\begin{aligned} &\min_{\mathrm{w}_k} \sum_{1\le i<j\le 3} \frac{1}{2\sigma_{ij}^2} \left\| \mathrm{X}_i \mathrm{w}_i - \mathrm{X}_j \mathrm{w}_j \right\|_2^2 + \log \sigma_{ij} + \sum_{1\le i\le 3} \lambda_i \|\mathrm{w}_1\|_i, \\ &\text{s.t. } \| \mathrm{X}_k \mathrm{w}_k \|_2^2 = 1, \forall k = 1, 2 \end{aligned} \tag{1}$$

The study employs an alternating iterative algorithm to solve for two weight vectors. In order to minimize Eq. (1), the derivative of this objective with respect to w_1 is set to zero. This leads to the formulation in Eq. (2):

$$\left(\left(\frac{1}{\sigma_{12}^2} + \frac{1}{\sigma_{13}^2} \right) \mathrm{X}_1^\top \mathrm{X}_1 + \lambda_1 \mathrm{D}_1 \right) \mathrm{w}_1 = \frac{1}{\sigma_{12}^2} \mathrm{X}_1^\top \mathrm{X}_2 \mathrm{w}_2 + \frac{1}{\sigma_{13}^2} \mathrm{X}_1^\top \mathrm{X}_3 \mathrm{w}_3 \tag{2}$$

Here, $\mathrm{D}_i \ (i = 1, 2)$ is a diagonal matrix, and its i th diagonal element is $\frac{1}{|W_{1i}|}$ $(i = 1, \ldots, p)$. Consequently, the iterative rule for w_1 is given by Eq. (3):

$$\mathrm{w}_1 = \left(\left(\frac{1}{\sigma_{12}^2} + \frac{1}{\sigma_{13}^2} \right) \mathrm{X}_1^\top \mathrm{X}_1 + \lambda_1 \mathrm{D}_1 \right)^{-1} \left(\frac{1}{\sigma_{12}^2} \mathrm{X}_1^\top \mathrm{X}_2 \mathrm{w}_2 + \frac{1}{\sigma_{13}^2} \mathrm{X}_1^\top \mathrm{X}_3 \mathrm{w}_3 \right) \tag{3}$$

Similarly, the iterative rules for w_2 and w_3 can be obtained, as shown in Eq. (4):

$$\begin{aligned} w_2 &= \left(\left(\frac{1}{\sigma_{12}^2}+\frac{1}{\sigma_{23}^2}\right)X_2^\top X_2+\lambda_2 D_2\right)^{-1}\left(\frac{1}{\sigma_{12}^2}X_2^\top X_1 w_1+\frac{1}{\sigma_{23}^2}X_2^\top X_3 w_3\right) \\ w_3 &= \left(\left(\frac{1}{\sigma_{13}^2}+\frac{1}{\sigma_{23}^2}\right)X_3^\top X_3+\lambda_3 D_3\right)^{-1}\left(\frac{1}{\sigma_{13}^2}X_3^\top X_1 w_1+\frac{1}{\sigma_{23}^2}X_3^\top X_2 w_2\right) \end{aligned} \tag{4}$$

In this context, Adaptive Multi-constraint Sparse Canonical Correlation Analysis is an algorithm proposed by Ref. [20]. It alleviates the problem of gradient domination by calculating the iteratively changing weights κ_{ij}. Here, λ_k is a balancing parameter that controls the sparsity of the model. The definition of the AdaSMCCA algorithm is as follows:

$$\min_{\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_K} \sum -\kappa_{ij}\mathbf{w}_i^\top \mathbf{x}_i^\top \mathbf{x}_j \mathbf{w}_{j_{i<j}} + \sum \lambda_k \|\mathbf{w}_k\|_{1_k}$$

$$\text{s.t. } \|\mathbf{w}_k\|_2^2 = 1, \forall k = 1, \ldots, K, \tag{5}$$

where $\mathbf{w}_k$ is employed to induce sparsity, λ_k is the non-negative adjustment parameters, λ_i is used to control the strength of regularization, where β adjusts the balance between individual and joint feature selection.

3 Results

3.1 The Performance of AdaSMCCA Method

To verify that the AdaSMCCA algorithm is feasible on MCI data, the effect of typical correlation coefficients $\boldsymbol{CCC_i}$ (i = 1, 2, 3) was calculated in this study for sMRI data and gene expression data at two time points of MCI as shown in Fig. 1. For $\boldsymbol{CCC_i}$ (i = 1, 2, 3), $\boldsymbol{CCC_1} = \boldsymbol{corr}(\mathbf{X_1 w_1}, \mathbf{X_2 w_2})$, $\boldsymbol{CCC_2} = \boldsymbol{corr}(\mathbf{X_1 w_1}, \mathbf{X_3 w_3})$, $\boldsymbol{CCC_3} = \boldsymbol{corr}(\mathbf{X_2 w_2}, \mathbf{X_3 w_3})$. As depicted in Fig. 1, the correlation coefficients between the two datasets exceed 0.9, demonstrating the efficacy of the AdaSMCCA algorithm in multimodal data association mining of MCI data.

3.2 Top Brain Regions and Genes Related to MCI

Following a typical correlation analysis, this study identified three data features with the highest weights: sMRI features at T1 time point, sMRI features at T2 time point, and feature genes. Additionally, an intersection analysis of the top brain regions at T1 and T2 time points was conducted, as shown in Fig. 2a. Furthermore, Fig. 2b

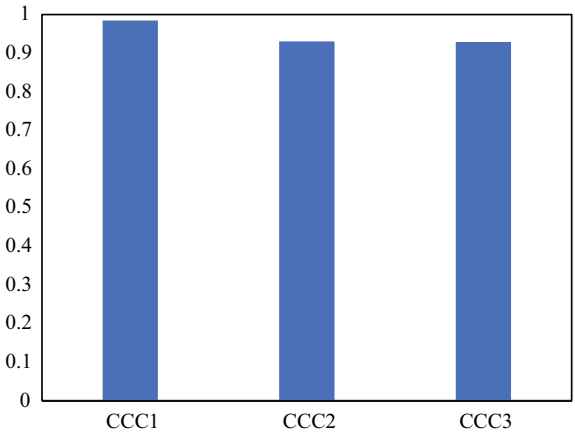


Fig. 1 Typical correlation coefficients in three types of data

illustrates the top 20 genes, with the horizontal axis representing the corresponding weight values.

From the Fig. 2, the top ranked features in both time points have 80% duplications, of which 16 brain regions include rFus, lRolOpe, lLin, rRolOpe, rLin, rRec, rOlf, rSupMedFro, lMidCin, lRec, rPCu, lFus, lSupTem, lIns, rCal and rMidCin. For top genes, the top three weighted genes were DIO2, MAP9, and FAM110B.

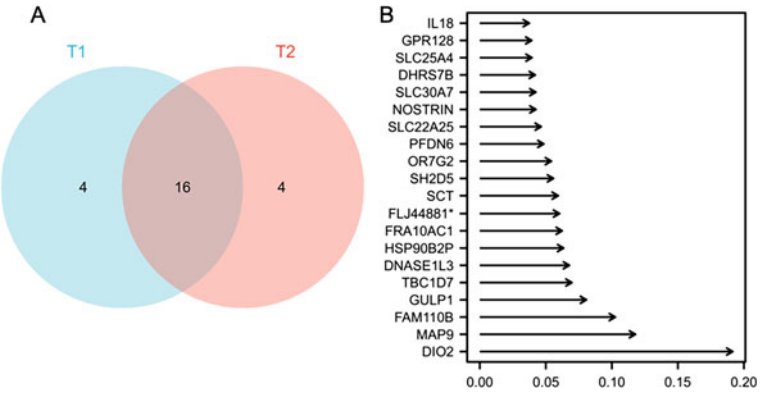


Fig. 2 **a** Intersection analysis of top 20 brain regions at T1 time point and T2 time point **b** Top 20 feature genes

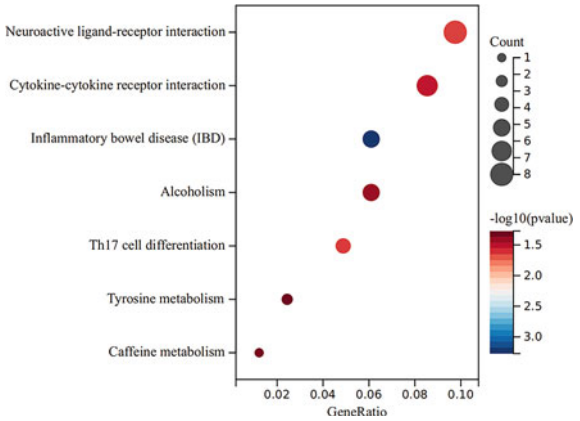


Fig. 3 KEGG enrichment analysis

3.3 Pathway Enrichment Analysis of Top Genes Related to Brain Regions

Exploring the genetic features associated with significant brain regions from the perspective of signaling pathways can tap the association information from a more comprehensive view. Therefore, in this study, the top20% gene sets were enriched and analyzed based on the Clusterprofile R package, as shown in Fig. 3.

As can be seen in Fig. 3, the top 200 weighted genes are most significantly enriched in neural activity-related pathways, followed by pathways significantly enriched with neurological disorders such as AD, such as Th17 cell differentiation, Tyrosine metabolism, and so on.

3.4 Association Analysis of Featured Brain Regions and Genes

To validate the characterized genes associated with the brain regions, the correlation between the characterized brain regions and top genes at T1 and T2 time points was calculated in this study, and the Pearson's correlation between the brain regions and the genes is demonstrated in Fig. 4.

The set of genes significantly associated with brain regions changed with the change in time points as seen in Fig. 4. At the BASELINE time point (Fig. 4a), genes significantly associated with significant brain regions were mainly significantly involved in human disease pathways, such as genes DCC, CCND1 and FN1. at the T2 time point (Fig. 4b) after a time delay of 6 months, metabolism-associated genes appeared, such as the gene UGDH, which were significantly involved in multiple pathways, including Ascorbate and aldarate metabolism and Pentose and glucuronate interconversions.

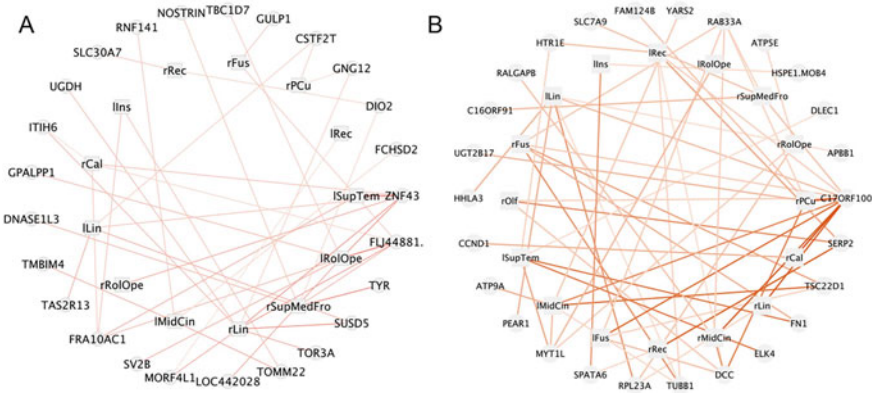


Fig. 4 Correlation network of 14 brain regions with top genes. Gray circles indicate genes and gray squares indicate brain regions. Line colors indicate correlation size. **a** T1 point; **b** T2 point

4 Conclusion

In this study, we used MCI imaging genetic data, including sMRI imaging data at two time points and gene expression data from the same batch of samples, after which we mined the imaging genetic association features, including multiple brain regions and characteristic genes, using the AdaSMCCA algorithm, and at the same time explored the imaging genetic association patterns at different time points through correlation calculations, to identify the significant brain regions and genetic features that evolved over time. We also explored the genetic association patterns of images at different time points by correlation calculation and identified significant brain region-associated genetic features over time. These studies provide new ideas for future biomarker mining and identification and provide new perspectives for personalized treatment of MCI.

References

1. Stringer, G., Couth, S., Heuvelman, H., et al.: Assessment of non-directed computer-use behaviours in the home can indicate early cognitive impairment: a proof of principle longitudinal study. Aging Ment. Health **27**(1), 193–202 (2023)
2. Sanctis, P.D., Wagner, J., Molholm, S.: Neural signature of mobility-related everyday function in older adults at-risk of cognitive impairment. Neurobiol. Aging **122**, 1–11 (2023)
3. Ju, Y.J., Tam, K.Y.: Pathological mechanisms and therapeutic strategies for Alzheimer's disease. Neural Regen. Res. **17**(3), 543–567 (2022)
4. Porsteinsson, A.P., et al.: Diagnosis of early Alzheimer's disease: clinical practice in 2021. J. Prev. Alzheimer's Dis. **8**(3), 371–386 (2021)
5. Chang, J., et al.: Neural stem cells promote neuroplasticity: a promising therapeutic strategy for the treatment of Alzheimer's disease. Neural Regen. Res. **19**(3), 619–628 (2024)

6. Grubman, A., et al.: A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat. Neurosci. **22**(12), 2087–2098 (2019)
7. Castillo-Ordoñez, W.O., Cajas-Salazar, N., Velasco-Reyes, M.A.: Genetic and epigenetic targets of natural dietary compounds as anti-Alzheimer's agents. Neural Regen. Res. **19**(4), 846–854 (2024)
8. Dubois, B., et al.: Clinical diagnosis of Alzheimer's disease: recommendations of the International Working Group. Lancet Neurol. **20**(6), 484–496 (2021)
9. Cheng, B., et al.: Robust multi-label transfer feature learning for early diagnosis of Alzheimer's disease. Brain Imaging Behav. **13**(1), 138–153 (2019)
10. Du, L., et al.: Identifying progressive imaging genetic patterns via multi-task sparse canonical correlation analysis: a longitudinal study of the ADNI cohort. Bioinformatics **35**(14), I474–I483 (2019)
11. Huang, M.Y., et al.: Imaging genetics study based on a temporal group sparse regression and additive model for biomarker detection of Alzheimer's disease. IEEE Trans. Med. Imaging **40**(5), 1461–1473 (2021)
12. Huang, M., et al.: Alzheimer's disease neuroimaging initiative. Incorporating spatial-anatomical similarity into the VGWAS framework for AD biomarker detection. Bioinformatics **35**(24), 5271–5280 (2019)
13. Witten, D.M., Tibshirani, R.J.: Extensions of sparse canonical correlation analysis with applications to genomic data. Stat. Appl. Genet. Mol. Biol. **8**(1), 28 (2009)
14. Gupta, R., Das, A.K.: Some variants of strong normality in closure spaces generated via relations. J. Math. 6917297 (2021)
15. Fang, J., et al.: Joint sparse canonical correlation analysis for detecting differential imaging genetics modules. Bioinformatics **32**(22), 3480–3488 (2016)
16. Kim, M., et al.: Joint-connectivity-based sparse canonical correlation analysis of imaging genetics for detecting biomarkers of Parkinson's disease. IEEE Trans. Med. Imaging **39**(1), 23–34 (2020)
17. Wang, X., Chen, H., Yan, J., et al.: Quantitative trait loci identification for brain endophenotypes via new additive model with random networks. Bioinformatics **34**(17), i866–i874 (2018)
18. Hao, X.K., et al.: Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis. Bioinformatics **33**(14), I341–I349 (2017)
19. Brand, L., et al.: Joint multi-modal longitudinal regression and classification for Alzheimer's disease prediction. IEEE Trans. Med. Imaging **39**(6), 1845–1855 (2020)
20. Du, L., Zhang, J., Liu, F., et al.: Identifying associations among genomic, proteomic and imaging biomarkers via adaptive sparse multi-view canonical correlation analysis. Med. Image Anal. **70**, 102003 (2021)

Multitasking Evolutionary Algorithm with SVM-Based Knowledge Discriminator

Xin Zong, Lei Zhao, Zhongwen Cheng, Jia Chen, and Lizhong Yao

Abstract Evolutionary multitasking (EMT) algorithms simultaneously address multiple optimization problems while enhancing problem-solving through knowledge migration. However, inappropriate knowledge transfer can lead to negative consequences, affecting search dynamics and algorithm convergence. In this regard, we propose an SVM-based knowledge discriminator (SVMKD) integrated into the EMT algorithm to reduce the likelihood of negative knowledge transfer. SVMKD utilizes task population experiences to train the SVM classifier, discerning solution quality during knowledge transfer and reducing the probability of negative transfer. The effectiveness of SVMKD is evaluated through a comprehensive empirical study on nine single-objective multitasking benchmark problems.

Keywords Evolutionary multitasking · Knowledge transfer · SVM

1 Introduction

Commonly, humans draw upon existing or analogous solution experiences when addressing problems. Building on this insight, Gupta et al. [1] introduced the Evolutionary Multitasking (EMT) paradigm, along with the algorithm MFEA, to simultaneously address multiple optimization tasks. EMT facilitates the sharing of evolutionary experiences across tasks, accelerating their search processes and endowing them with superior problem-solving and optimization capabilities.

In recent years, EMT has emerged as a prominent research direction in evolutionary computation. Numerous studies have introduced diverse EMT algorithms, achieving success across various intricate optimization problems [2–4]. Noteworthy examples include Gupta et al.'s introduction of a multi-objective variant, MO-MFEA,

X. Zong (✉) · Z. Cheng · J. Chen
School of Electrical Engineering, Chongqing University of Science and Technology, Chongqing, China
e-mail: I.2021204005@cqust.edu.cn

L. Zhao · L. Yao
College of Physics and Electronic Engineering, Chongqing Normal University, Chongqing, China

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_13

based on NSGA2 [5]. Feng et al. [6] presented an EMT algorithm within an explicit framework using self-coding . Wu et al. [7] diversified the transferred knowledge by employing various transfer modes. Li et al. [8] addressed sensor coverage using an adaptive solver framework.

Undoubtedly, the aforementioned literature has demonstrated success in the realm of multitasking optimization. Nevertheless, it is imperative to acknowledge that the algorithmic efficacy of EMT may be compromised by negative migration, particularly in scenarios where tasks exhibit low similarity. Specifically, during selective mating, if individuals from disparate tasks engage in crossover operations due to meeting the specified random mating rate (rmp) criteria, the resulting offspring may face suboptimal performance in the context of the target task. Consequently, a compelling research imperative arises to devise a real-time and effective classifier strategy that mitigates negative migration and its deleterious impact on algorithmic performance.

Given the above motivations, this study introduces an SVM-based knowledge discriminator mechanism, denoted as SVMKD. Firstly, task population evolutionary experiences are employed to train a task-specific knowledge quality discriminator. Subsequently, the SVM classifier is deployed to discriminate solutions during knowledge migration, retaining high-quality solutions for migration and discarding low-quality ones. Finally, the classifier undergoes retraining based on the efficacy of knowledge migration, thereby enhancing its classification success rate. Empirical results across nine benchmark tests substantiate the effectiveness of SVMKD in significantly improving the convergence performance of EMT algorithm.

2 Support Vector Machine (SVM)

SVM attains classification by identifying the optimal hyperplane under conditions of linear differentiability. Precisely, considering a given set of samples $S = \{(x_i, y_i)_{i=1}^N | x_i \in R^d, y_i \in \{+1, -1\}\}$, where N represents the number of sample points, p denotes the dimension of x_i, and y_i signifies the category to which sample x_i belongs. If the hyperplane equation $w \cdot x_i + b = 0$ can effectively discriminate between two classes of samples, maximizing the interval, the task of determining the parameters of this equation can be formulated as follows:

$$\begin{cases} \min \frac{1}{2} w^2 + c \sum_{i=1}^{N} \xi_i, i = 1, 2, \cdots, N \\ \quad \text{s.t.} \begin{cases} y_i (w \cdot x_i + b) > 1 - \xi_i \\ C \geq 0, \xi_i \geq 0 \end{cases} \end{cases} \tag{1}$$

where w is the weight vector, b is the bias, ξ is the relaxation factor, and C is the penalty factor. Using the Lagrangian function construction to solve and according to the Kuhn-Tucker theorem, Eq. (1) can be transformed into:

$$\begin{cases} \max\limits_{\alpha} \sum\limits_{i=1}^{N} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{j} \sum\limits_{j=1}^{N} y_i y_j \alpha_i \alpha_j \quad < x_i, x_j > \\ \text{s.t} \quad \sum\limits_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \end{cases} \tag{2}$$

In this study, task populations are initially treated as input samples. Subsequently, individuals are sorted based on scalar fitness to establish sample categories. Finally, a trained SVM classifier discriminates the transferred individuals, ensuring that only high-quality solutions undergo knowledge migration. This precautionary measure aims to prevent the onset of negative migration. The detailed program design is elucidated in Sect. 3.

3 Proposed Methods

In this section, we delve into the integration of the knowledge migration strategy with SVM classifiers, presenting a novel knowledge migration approach grounded in SVM classification. This strategic amalgamation aims to effectively address the challenge of negative migration within multitasking evolutionary algorithms.

3.1 SVM-Based Knowledge Discriminator (SVMKD)

Take multitasking single-objective optimization algorithm as an example, assuming that actual curve of the current objective function is shown in Fig. 1.

Depicted in Fig. 1, the blue curve represents task 1, and the red curve represents task 2. Assuming a specific iteration process, consider task 1, where the green five points (P1–P5) denote the parent population, and the red five points (O1–O5) denote the offspring population. The new population is then formed by selecting the five points with the smallest target values from the combined set of ten points.

The SVM classifier is employed along the x-axis, where the y-values of individuals are treated as 0. Assuming a total of ten individuals, with five newly selected individuals, it is presumed that these five individuals are superior, and the values in their proximity are more likely to be optimal. As a result, the label "+" is assigned to these five individuals. Conversely, the five individuals eliminated by the elite strategy are regarded as inferior, and the label "-" is applied to them. The resulting classification is presented in Table 1.

Subsequently, the classification model for recognizing knowledge migration at the next iteration is established by training the aforementioned data using a nonlinear kernel in the SVM classifier, depicted in Fig. 2.

In practical application, the selection of the kernel function type and hyperparameters holds considerable influence on the classification efficacy of SVM classifiers.

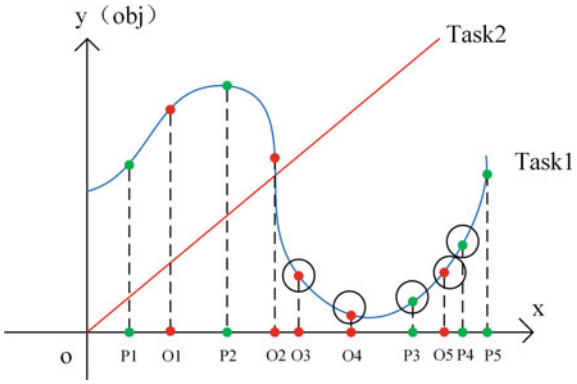
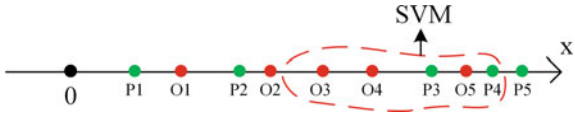


Fig. 1 Actual curve of the objective functions

Table 1 SVM classifier training

Individual name	P1	P2	P3	P4	P5	O1	O2	O3	O4	O5
Whether retention	No	No	Yes	Yes	No	No	No	Yes	Yes	Yes
Labels	–	–	+	+	–	–	–	+	+	+

Fig. 2 SVM classifier model training diagram



Hence, their correct choice is crucial, aligning with the actual task complexity and dataset characteristics. Simultaneously, when confronted with unfamiliar challenges, despite the potential trade-off in overall algorithmic efficiency with the utilization of hyperparameters for model construction, it nonetheless yields superior outcomes for classification purposes.

It is noteworthy that maintaining the original population size requires the random generation and addition of new individuals when eliminating those exhibiting negative migration.

3.2 Framework of Algorithms

The algorithm introduced in this paper extends the MFEA basic framework, incorporating an SVMKD module for knowledge migration classification. This augmentation enables discernment of knowledge during migration, subsequently enhancing algorithm's convergence performance.

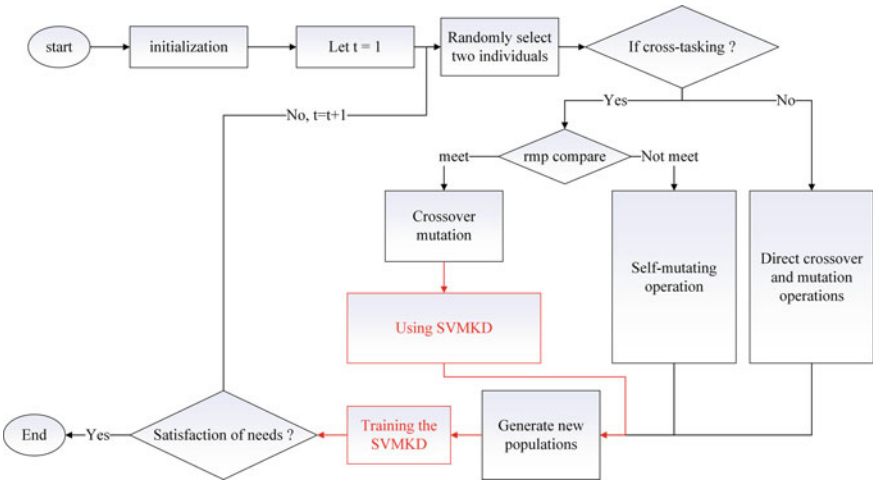


Fig. 3 Overall framework diagram of proposed algorithm

Initially, an N-sized population is randomly generated within a uniform search space Y, with each individual being assigned a specific skill factor to ensure equitable representation across tasks. Subsequently, the population undergoes regular iteration, and during cross-task communication, continuation depends on satisfying the rmp; otherwise, each parent generation independently mutates to generate new individuals. When conducting cross-task crossover to produce a new individual, SVMKD is employed for classification. If recognized as a positive migration, the individual is retained; in case of negative migration, individual is eliminated, and a new individual is randomly generated to maintain a constant population size. The algorithmic flow is illustrated in Fig. 3.

Following the completion of each elite selection strategy, individuals in the new population are designated with the label "+", while eliminated individuals receive the label "−". This data is subsequently utilized as training data for SVMKD. The trained SVMKD contributes to enhanced negative migration recognition for the subsequent generation. It is imperative to highlight that each task in the multitasking optimization algorithm necessitates training a distinct SVMKD. In the context of concurrently addressing two tasks, an SVMKD must be trained for task 1 to classify knowledge-migrated individuals from task 2, and vice versa for task 2.

4 Experimental Results

4.1 Problems and Parameter Settings

To assess SVMKD's efficacy, we selected nine well-established multitasking single-objective optimization problems as benchmark tests [9].

The configurations for the experiment are outlined below:

- Hardware environment: 13th Gen Intel(R) Core (TM) i9-13900HX CPU @5.40 GHz; RAM 16.0 GB; SSD 2048 GB.
- Software environment: Win10 Professional Edition; MATLAB R2021a.

The algorithm is configured with the following specific parameters:

- Population size: 100.
- Maximum number of runs: MaxGen = 500.
- Number of independent runs: 20.
- Random mating rate: rmp = 0.3.
- Additional parameters include pil = 1 (probability of individual learning), Pc = 1 (simulated binary crossover probability), mu = 10 (simulated binary crossover parameter), and sigma = 0.02 (standard deviation of Gaussian variance model).

Significantly, for method validation, we integrate SVMKD into MFEA and MFEA-II, denoting them as MFEA-SVM and MFEA-II-SVM, respectively.

4.2 Results

Tables 2 and 3 present the average outcomes derived from 20 independent runs of MFEA-SVM and MFEA-II-SVM across nine tasks. The optimal results are emphasized in bold. Notably, both MFEA-SVM and MFEA-II-SVM, augmented with SVM classifiers, exhibit noteworthy enhancements in convergence outcomes compared to both the MFEA and MFEA-II algorithms. This substantiates the positive impact of the SVM component on augmenting the performance of multitasking evolutionary algorithms.

To visually illustrate the effectiveness of the SVM component, representative convergence comparison plots are provided in Fig. 4. In contrast to MFEA and MFEA-II, both MFEA-SVM and MFEA-II-SVM, integrated with SVMKD, consistently exhibit enhanced convergence performance across the entire iterative process. This robustly validates the efficacy of our proposed method in mitigating negative knowledge transfer incidents during the knowledge migration process.

Table 2 Data comparison of MFEA and MFEA-SVM

Algorithm	M1-T1	M1-T2	M2-T1	M2-T2	M3- T1	M3- T2
MFEA	9.02E-01	**2.84E+02**	5.88E+00	3.32E+02	2.04E+01	4.99E+03
MFEA-SVM	**8.07E-01**	2.93E+02	**5.50E+00**	**3.26E+02**	**2.03E+01**	**4.43E+03**
	M4-T1	M4-T2	M5-T1	M5-T2	M6-T1	M6-T2
MFEA	6.89E+02	1.46E+02	5.21E+00	1.37E+04	2.02E+01	1.76E+01
MFEA-SVM	**6.45E+02**	**8.91E+01**	**5.00E+00**	**8.26E+03**	**1.94E+01**	**1.90E+01**
	M7-T1	M7-T2	M8-T1	M8-T2	M9-T1	M9-T2
MFEA	1.40E+04	4.00E+02	1.02E+00	2.85E+01	8.32E+02	5.23E+03
MFEA-SVM	**1.02E+04**	**3.81E+02**	**9.59E-01**	**2.75E+01**	**7.63E+02**	**4.27E+03**

Table 3 Data comparison of MFEA-II and MFEA-II-SVM

Algorithm	M1-T1	M1-T2	M2-T1	M2-T2	M3- T1	M3- T2
MFEA-II	2.79E-01	2.99E+02	**2.06E+00**	2.83E+02	**2.12E+01**	3.12E+03
MFEA-II-SVM	**2.17E-01**	**8.83E+01**	2.08E+00	**1.06E+02**	2.12E+01	**1.03E+03**
	M4-T1	M4-T2	M5-T1	M5-T2	M6-T1	M6-T2
MFEA-II	3.56E+02	**4.03E-01**	1.68E+00	3.49E+02	**1.15E+00**	3.91E+00
MFEA-II-SVM	**1.78E+02**	9.98E-01	**1.58E+00**	**2.46E+02**	2.31E+00	**3.63E+00**
	M7-T1	M7-T2	M8-T1	M8-T2	M9-T1	M9-T2
MFEA-II	**6.60E+02**	3.43E+02	**9.88E-02**	**8.66E+00**	3.06E+02	2.76E+03
MFEA-II-SVM	7.91E+02	**1.30E+02**	1.49E-01	1.37E+01	**1.64E+02**	**1.53E+03**

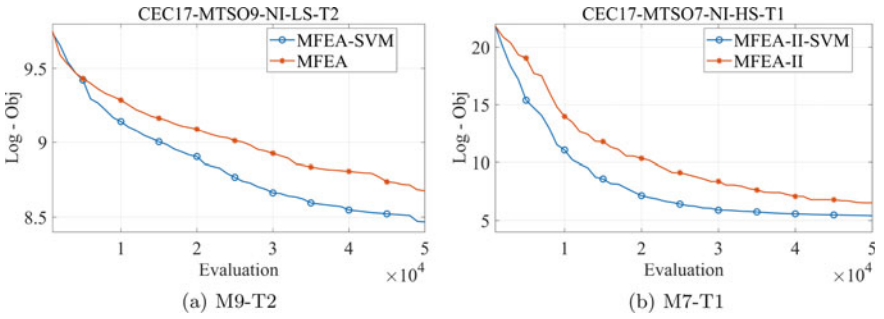


Fig. 4 Iteration comparison graph

5 Conclusion

This paper introduces a knowledge transfer mechanism utilizing an SVM-based knowledge discriminator. The proposed mechanism enhances knowledge transfer between diverse tasks by discerning solution quality during the migration process, retaining high-quality solutions, and eliminating inferior ones. The outcomes demonstrate that the proposed scheme effectively mitigates negative migration, enhancing the performance of MFEA and MFEA-II. In our upcoming research endeavors, we aim to employ deep learning methodologies in the classification process of knowledge transfer, with the goal of more effectively mitigating adverse transfer effects.

Acknowledgements This work was supported in part by Natural Science Foundation of Chongqing, China (No. 2023NSCQ-MSX2158), in part by Foundation Program of Chongqing Normal University (No. 22XLB014), in part by Science and Technology Research Program of Chongqing Municipal Education Commission (No. KJQN202200531).

References

1. Gupta, A., Ong, Y.S., Feng, L.: Multifactorial evolution: toward evolutionary multitasking. IEEE Trans. Evol. Comput. **20**(3), 343–357 (2015)
2. Jiang, Y., Zhan, Z.H., Tan, K.C., Zhang, J.: Block-level knowledge transfer for evolutionary multitasking optimization. IEEE Trans. Cybern. **54**(1), 558–571 (2023)
3. Wang, X., Kang, Q., Zhou, M., Yao, S., Abusorrah, A.: Domain adaptation multitask optimization. IEEE Trans. Cybern. **53**(7), 4567–4578 (2023)
4. Li, S., Gong, W., Wang, L., Gu, Q.: Evolutionary multitasking via reinforcement learning. IEEE Trans. Emerg. Top. Comput. Intell. **8**(1), 762–775 (2023)
5. Gupta, A., Ong, Y.S., Feng, L., Tan, K.C.: Multiobjective multifactorial optimization in evolutionary multitasking. IEEE Trans. Cybern. **47**(7), 1652–1665 (2016)
6. Feng, L., Zhou, L., Zhong, J., Gupta, A., Ong, Y.S., Tan, K.C., Qin, A.K.: Evolutionary multitasking via explicit autoencoding. IEEE Trans. Cybern. **49**(9), 3457–3470 (2018)
7. Wu, X., Wang, W., Yang, H., Han, H., Qiao, J.: Diversified knowledge transfer strategy for multitasking particle swarm optimization. IEEE Trans. Cybern. Early Access 1–14 (2023)
8. Li, Y., Gong, W., Li, S.: Multitasking optimization via an adaptive solver multitasking evolutionary framework. Inform. Sci. **630**, 688–712 (2023)
9. Da, B., Ong, Y.S., Feng, L., Qin, A.K., Gupta, A., Zhu, Z., Ting, C.K., Tang, K., Yao, X.: Evolutionary multitasking for single-objective continuous optimization: benchmark problems, performance metric, and baseline results (2017). arXiv preprint arXiv:1706.03470

The Design and Realization of the "Smart Ceramics" Virtual Ceramics Museum

Lei Zhang, Pengshuai Li, Yufan Hu, Shijie Rong, and Shuxin Chen

Abstract With the continuous advancement of virtual reality technology, an increasing number of traditional museums are embracing the utilization of virtual reality to exhibit cultural relics and artworks. Leveraging the IdeaVR tool, a realistic virtual interactive ceramic museum has been successfully developed, offering diverse interactive methods such as holographic projection, popular science quizzes, multimedia technology, and simulation displays. Visitors can manipulate the holographic projection to translate, rotate, and scale ceramics while experiencing the ceramic production process through simulated workshops and interactive screens. An intelligent interactive screen mounted on the wall provides historical insights into ceramic development along with video selections that enable visitors to comprehend the history and evolution of Chinese ceramic culture. The overall design is characterized by its uniqueness and aesthetic appeal, creating an exceptional ambiance suitable for showcasing ceramic culture while evoking profound artistic sentiments and cultural experiences. This paper presents the design scheme as well as details regarding its implementation process and effects achieved in constructing this virtual ceramic museum; furthermore, it offers prospects for future developments in this field.

Keywords Virtual museum · IdeaVR · Holographic projection · Ceramic culture

1 Introduction

Ceramic culture is an important part of China's 5,000 years of civilization, and it is not only an art form, but also a way of life and spiritual pursuit. However, due to historical reasons and technical limitations, many people do not have the opportunity to personally experience the charm of ceramic culture. Empowered by virtual reality technology, immersive design through technology to expand the application of spatial forms, immersive technology aspects of AR, VR, holographic projection, interactive physical examination technology and other digital technologies in the reform of

L. Zhang · P. Li · Y. Hu · S. Rong · S. Chen (✉)
School of Intelligent Computing Engineering, Tianjin Ren'ai College, Tianjin 301636, China
e-mail: shuxinfriend@tju.edu.cn

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_14

the original environmental form, based on the IdeaVR engine to design a virtual ceramic museum, the use of interactive design forms to enhance the authenticity of the physical object, so that visitors through the viewing of the huge visual effect and cultural sense of impact, allowing visitors to understand and feel the charm of Chinese ceramic culture [1]. The museum is based on IdeaVR engine.

2 The Significance of the Construction of the Virtual Ceramic Museum

Through the construction of ceramic digital virtual exhibition hall, can better display ceramic artwork and its technology, so as to deepen the understanding of visitors to the ceramic culture, ceramic culture heritage and dissemination. Integration of virtual reality technology, animation technology, audio and video technology, etc., can create a combination of static and dynamic, real and vivid visiting platform and experience environment [2]. At the same time, through the construction of virtual screening hall, it can create a combination of dynamic and static, real and vivid visiting platform and experience environment. At the same time, through the construction of virtual screening room, with the help of advanced interactive technology, multi-clue, multi-level content, multi-dimensional spatial integration, to achieve the exhibition of non-linear, virtualized, extended through, showing a diverse range of museum requirements [3] The exhibition is designed in such a way that it can better meet the needs of the visitors. Such a design can better meet the visitors' needs for immersive spatial experience [4–6].

In addition, the virtual ceramic museum can also increase the visitors in the museum to achieve a sense of satisfaction, which is an important part of the future of ceramic culture museum digitization. Virtual ceramic museum using virtual reality technology, so that visitors can be more immersed in three-dimensional virtual museum scenes, so that visitors have a more realistic experience. The charm of the virtual display is that visitors roaming in the virtual ceramic museum, can take in enough cultural information, access to ceramic culture level of knowledge and understanding.

3 "Smart Ceramics" Virtual Museum Overall Design

IdeaVR is a virtual reality engine software that provides a comprehensive cross platform and cross hardware collaborative experience for multiple people, and supports multiple VR hardware devices. In this project, IdeaVR2023 version is used to complete the design and realization of "Smart Ceramics" Virtual Museum.

3.1 Virtual Museum Layout Design

The construction of the virtual ceramic museum requires the use of virtual reality technology, digital twin technology, etc., through the integration and reconstruction of technology and content. In terms of virtual environment modelling, modelling is carried out through the IdeaVR engine, which is adjusted and optimized according to the actual venue. The IdeaVR engine is able to provide functions such as rapid scene construction and interactive logic editing, which is suitable for teaching and training, simulation training, marketing display and other applications in irreversible or inaccessible and other scenarios. In terms of interaction design, a variety of interaction methods are adopted, such as holographic projection, science quiz, etc., to improve user participation and experience. In terms of user interface design, the UI editor that comes with the IdeaVR engine is used for design, and some animation effects and prompt messages are added, as shown in Figs. 1 and 2.

The design of the "Smart Ceramics" museum mainly includes three parts: Smart Ceramics exhibition hall, Smart Ceramics experience hall and Smart Ceramics holographic screening hall, and the specific design process is shown in Fig. 3.

① Smart Ceramics Exhibition Hall: Combining traditional and modern display methods, it shows the beauty of ceramic art through the arrangement of the exhibition hall, lighting design, etc., and makes use of modern technology to let visitors get an unusual visiting experience.

Fig. 1 Interaction editor interface layout

Fig. 2 Exterior view of ceramic museum

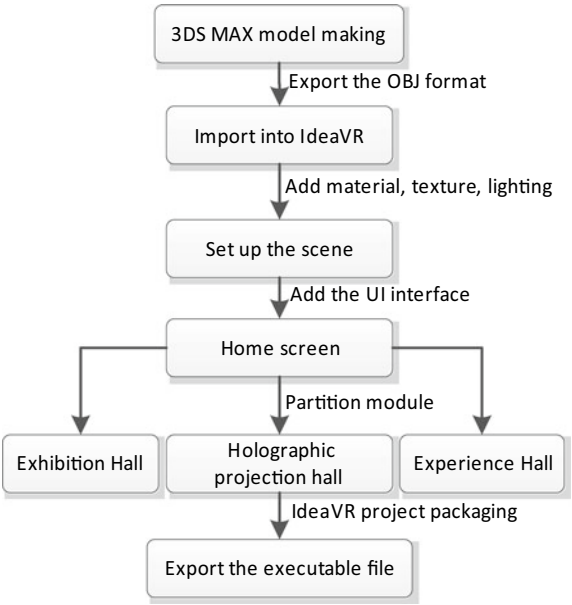


Fig. 3 Flowchart for the design of Smart Ceramics museum

② Smart Ceramic Experience Hall: Through space planning and interactive design, visitors can gain an in-depth understanding and experience of ceramic culture, including the setting of a simulation workshop that allows visitors to personally operate the ceramic production process, increasing the sense of participation and interactivity.

③ Holographic Screening Hall: Adopting wave pattern design, including ceiling and walls, showing a unified and harmonious design style, the walls are covered with multi-colored murals, integrating different art styles and ceramic elements, presenting a unique ceramic screening space.

In order to achieve the ceramic virtual museum without time and space limitations of cultural relics display, the use of IdeaVR engine designed a three-dimensional virtual ceramic museum, that is, "intelligent ceramic" museum. Specific layout is divided as shown in Figs. 4 and 5.

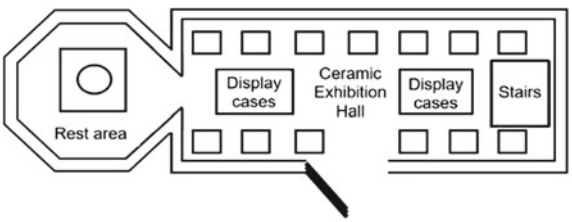


Fig. 4 Overhead view of the exhibition hall

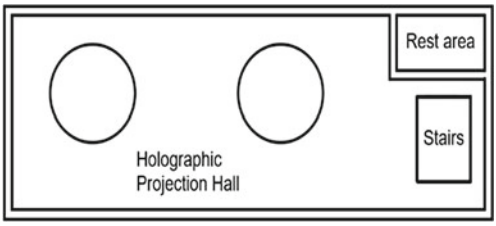


Fig. 5 Overhead view of the screening room

Fig. 6 Left side layout of exhibition hall

3.2 Smart Ceramics Exhibition Hall

Ceramic museum exhibition hall design not only inherits the tradition, but also innovation, taking its traditional ceramic museum exhibition hall layout, cabinet design, lighting layout and other aspects of the design of the essence of the design, pay attention to the unity of aesthetics and functionality. The booth adopts modern display methods, through the booth with the design of lighting, making ceramics show more outstanding. The aesthetics of the exhibition space and viewing angle can maximize the display of ceramic beauty of art, and meet the comfort and flow of the visitor's perception. In the display form, the use of modern technology, so that visitors can be more intuitive, image of ceramic culture, and give them an unusual experience. Specifically, as shown in Figs. 6, 7, 8 and 9.

3.3 Smart Ceramics Experience Hall

Ceramic museum design focuses on the interactive relationship between space and exhibits, through reasonable space planning and layout, so that visitors in the process of visiting the ceramic culture can deeply understand and feel. Set up a simulation workshop in the ceramic exhibition hall, so that visitors personally operate and

Fig. 7 Right side layout of exhibition hall

Fig. 8 Navigation bar of Smart Ceramics showroom

Fig. 9 Navigation map of Smart Ceramics

Fig. 10 Ceramics making

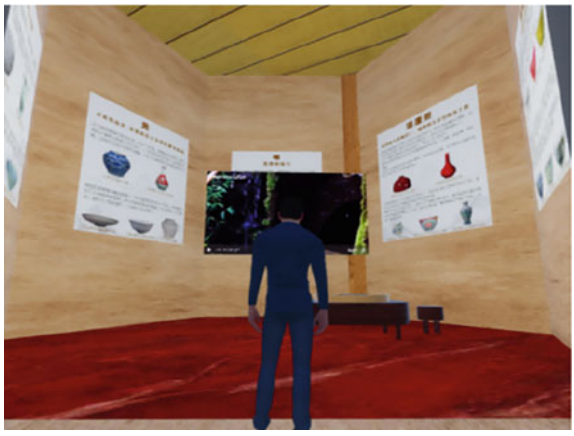


Fig. 11 Video projection

experience the process of ceramic production, increasing the sense of participation and interactivity. Specifically, as shown in Figs. 10 and 11.

3.4 Smart Ceramics Holographic Screening Room

The design of the screening room on the first floor of the Ceramics Museum uses a wave pattern, both on the ceiling and the walls, to give the space a unique and attractive appearance. The wave pattern not only appears on the ceiling but also extends to the walls, creating a unified and harmonious design style. The walls are covered with a variety of colorful murals, incorporating different art styles and ceramic elements to showcase a uniquely designed ceramic screening room. This is shown in Figs. 12, and 13.

Fig. 12 Design of the holographic screening room

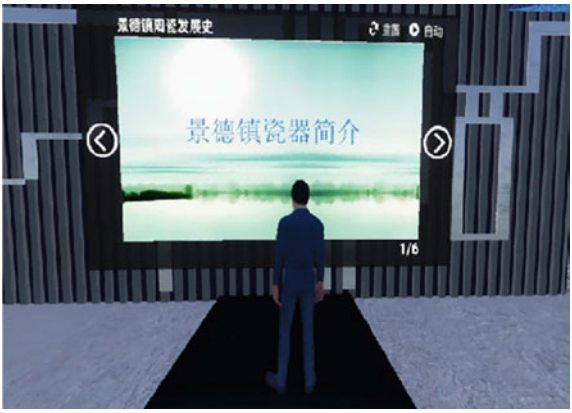


Fig. 13 Holographic screening room porcelain profile

The holographic screening room features a long, curved wall that creates a visually striking and futuristic appearance. The artwork on the wall further adds to the beauty of the building and is designed to attract attention. A gallery exists in the center, in which 3D ceramic models are displayed via holographic projection, capable of automatic three-dimensional rotation. Visitors can view the details of the 3D ceramic models and text presentations on both sides of the gallery. The walls of the wave are set up with intelligent interactive screens, which allow users to choose to watch the history of ceramic development or videos.

The screening hall is designed with a combination of natural light and artificial light, allowing visitors to feel the beauty of modern ceramics in the interplay of light and shadow. Through multimedia technology, simulation display and interactive display, visitors can understand the history, types, production process and artistic characteristics of ceramics. The whole atmosphere of the screening room echoes the ceramic culture, showing the magnificent beauty and flavor of the national culture. The blue and white design elements echo the white walls, adding to the overall aesthetic. This design may be intended to evoke feelings of serenity and high technology, as blue is often associated with tranquility and white with purity. By means

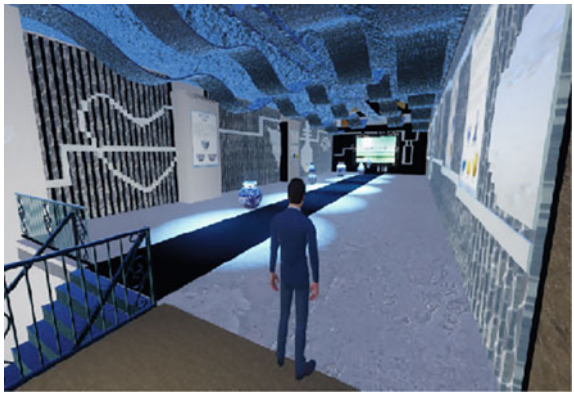


Fig. 14 Holographic screening room porcelain projection I

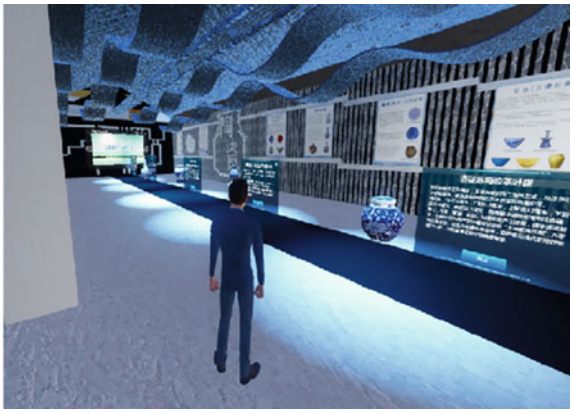


Fig. 15 Holographic screening room porcelain projection II

of lighting, music and visual effects, a unique atmosphere suitable for the display of ceramic culture is created to immerse visitors and create a strong artistic feeling and cultural experience. Specifically, as shown in Figs. 14 and 15.

4 IdeaVR Innovative Design Realization

4.1 Intelligent Interaction Experience

Ceramic museum design focuses on the interactive relationship between space and exhibits, able to beautiful ceramic models for close observation, able to pan, rotate, zoom effect has a better experience. Through reasonable space planning and layout, so that visitors in the process of visiting the ceramic culture can deeply understand and feel.

Spatially triggered 3D exhibit information display and hiding, roaming module to visitors as a first-person perspective into the pre-set spatial triggering area to display 3D buttons (exhibit name), click on the pop-up 3D panel (exhibit information), leave the spatial triggering area, the visibility of panels and buttons become False state, that is, the 3D panel disappears. Specifically, as shown in Figs. 16 and 17.

Fig. 16 Porcelain interaction 1

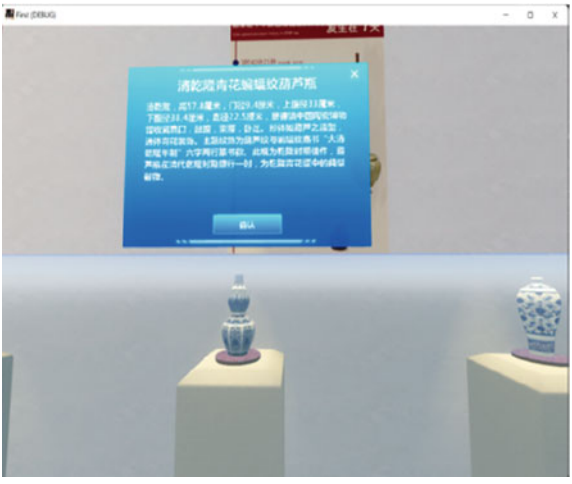


Fig. 17 Porcelain interaction II

4.2 Personalized Guide Service

Provide personalized tour guide services based on the interests and needs of tourists. Different thematic exhibition areas are set up to allow visitors to choose the content of the exhibitions that interest them according to their own interests; guided tours with different levels of difficulty and interpretations can also be provided according to factors such as the age and educational background of visitors.

Enter the door to the guide desk, click on the 3D button, pop-up 3D panel, click on the "tour map", pop-up virtual museum map of the first two floors, 3D panel disappears, click on the "Confirm", that is, the 3D panel disappears. Specifically, as shown in Figs. 18 and 19.

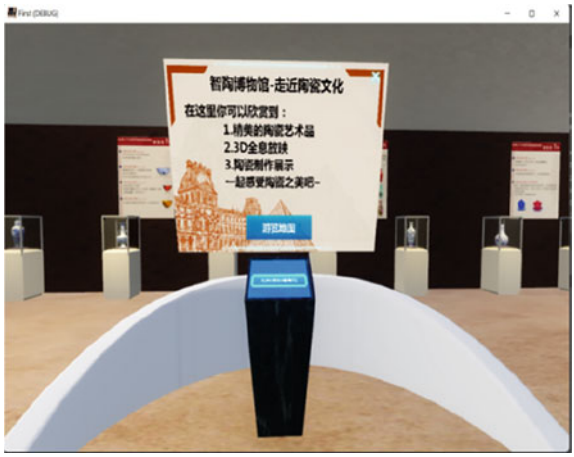


Fig. 18 Navigation interaction I

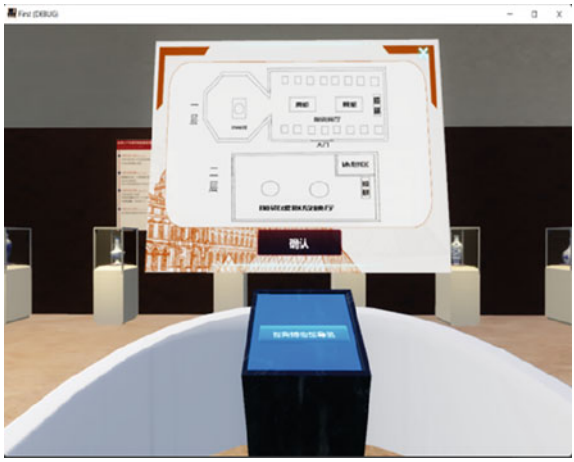


Fig. 19 Navigation interaction II

4.3 Innovative Interaction and Artistic Appearance Design

① Holographic projection technology: projection hall through holographic projection in the gallery to show 3D ceramic models, ceramics can automatically three-dimensional rotation, visitors through the viewing holographic projection, close observation, and zoom effects have a better experience. Can intuitively understand the details and artistic characteristics of ceramics, so that visitors have a deeper understanding of ceramic culture.
② Multimedia technology and interactive screen: the virtual museum uses the projection hall of multimedia technology and intelligent interactive screen to display the history, types, production process and artistic characteristics of ceramics in the form of images, audio and video. Visitors interact with the display content by touching the screen and selecting menus to get more information about ceramics and enhance the viewing experience.
③ Artistic aesthetic design: The screening room adopts blue and white design elements, echoing the white walls, creating an atmosphere of tranquility and purity. At the same time, through the means of lighting, music and visual effects, it creates a unique atmosphere suitable for the display of ceramic culture, which immerses the visitors and creates a strong artistic feeling and cultural experience.
④ Porcelain element design: the walls are covered with a variety of colorful murals, combining different artistic styles and ceramic elements to display a uniquely designed ceramic screening room. This design not only brings visual enjoyment to visitors, but also enables them to have a deeper understanding of the diversity and artistic charm of ceramic culture. It also incorporates modern art elements. The diversified design style makes the screening room richer and more interesting, attracting more attention and participation of visitors.

Enter the holographic projection hall with the visitor as the first-person perspective, click the button whether to view the holographic, slide player display, holographic projection mode open, porcelain model projected to the middle gallery, exhibits rotating playback, 3D panels showing ceramic information, mouse select holographic exhibits, mouse controller setup click left button, zoom in, click right button to zoom out. Keyboard controller: K, Z, X, C, V from left to right to control the rotating ceramic exhibits. Specifically shown in Figs. 20 and 21.

4.4 Simulation Demonstration

Combined with the characteristics of the virtual museum, visitors can learn about ceramic cultural knowledge in the game. For example, visitors can experience the process of ceramic production. Screening room set up a simulation workshop, visitors can personally operate and experience the process of ceramic production. Through touch, simulation and other ways, visitors can feel the texture of ceramics and the production process, increasing the sense of participation and interactivity.

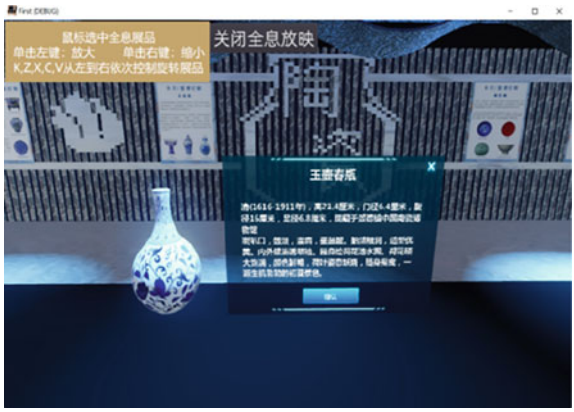


Fig. 20 Holographic screening room porcelain display

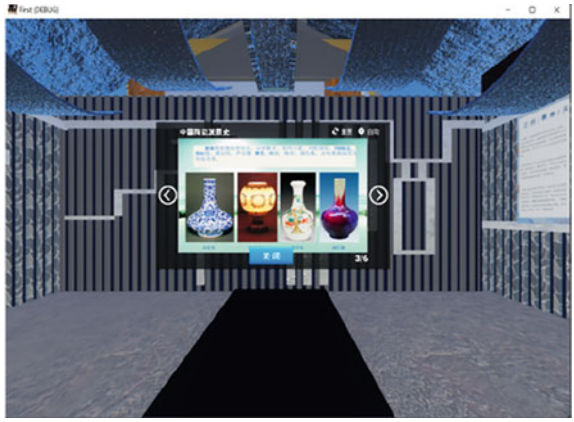


Fig. 21 Holographic screening room porcelain history

Visitors can enter the space set in advance to trigger the region to display the UI interface, while the ceramic production process animation is automatically played, 3D video player automatically play the ceramic production process, click "close", 3D video player disappears. Specifically, as shown in Figs. 22 and 23.

Fig. 22 Smart Ceramics experience hall interaction 1

Fig. 23 Smart Ceramics experience hall interaction II

5 Conclusion

This project introduces an interaction design scheme for a virtual ceramic museum based on IdeaVR and its implementation process. In the virtual scene, the visual interaction design of the virtual ceramic museum content, visitors can immerse themselves in the history and development of ceramic culture, to improve user experience and participation. This ceramic museum design can highlight the connotation of ceramic culture, combine innovation and tradition, adopt high-tech means, realize intelligent interaction, unify aesthetics and function, and create a unique atmosphere of ceramic culture display. In the future, different forms of immersive interactive experience will be pursued in different fields, such as history museums and science and technology museums. With the continuous development of virtual reality technology, the virtual museum will have a broader development prospect.

Acknowledgements This paper is supported by the joint project of the National Natural Science Foundation of China under Grant No. u2031142, and the research plan project of undergraduate teaching reform and quality construction in Tianjin colleges and universities (Project No. b231403805). Thank you for the co construction and cultivation of the digital culture Joint Laboratory of Tianjin Ren'ai College.

References

1. Lei, G., Bing, H.: Exploring the development of digital virtual museum in the post epidemic era. Daguan (6), 102–103 (2021)
2. Wang, J., Liang, X., Wang, Z.: Design and development of virtual museum based on VR technology. Mod. Inf. Technol. **7**(15), 29–34 (2023)
3. Zhang, T.-S., Man, Y., Yang, Z.-L.: Research on user experience design in virtual museum. China Ethnic Expo (13), 250–252 (2023)
4. Pivec, M., Kronberger, A.: Virtual museum: playful visitor experience in the real and virtual world. In: 2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), pp. 1–4, Barcelona, Spain (2016)
5. Azzahra, W., Munir, W.: Compustory: a virtual museum game for modeling the history of computer evolution. In: Badioze Zaman, H., et al. (eds.) Advances in Visual Informatics. IVIC 2023
6. Kim, K., Kwon, O., Yu, J.: Evaluation of an HMD-based multisensory virtual museum experience for enhancing sense of presence. IEEE Access **11**, 100295–100308 (2023)

Advancing Colon Cancer Detection: A YOLOv5-Based Approach with Emphasis on Precision, Interpretability, and Real-World Deployment Considerations

Tushar H. Jaware, Jitendra P. Patil, and Ravindra D. Badgujar

Abstract Cancer, characterized by uncontrolled cell division, necessitates early detection to alleviate the threat it poses. Research posits that uncovering treatable cancer in its initial stages can potentially save up to 99% of an individual's life. This study introduces a YOLOv5-based model designed for colon cancer detection, achieving an impressive 99.94% training accuracy and 99.75% validation accuracy. Notably, the model addresses the critical concern of overfitting through the incorporation of metrics such as precision, recall, and F1 score. Emphasizing the significance of visual inspection and early stopping during training, the YOLOv5s model underscores the need for meticulous evaluation. Researchers are encouraged to scrutinize model predictions on the validation set to ensure its precision in detecting colon cancer features. The study advocates for the implementation of early stopping based on validation metrics to mitigate overfitting risks. In conclusion, the proposed YOLOv5s model presents substantial advancements in the realm of colon cancer detection. The call for rigorous evaluation on testing sets, examination of confusion matrices, and ongoing refinement through potential fine-tuning strategies contributes to the continual progress of deep learning in medical image analysis, particularly in the context of early colon cancer diagnosis. This research serves as a pivotal step towards leveraging sophisticated techniques for improved medical image analysis, with a specific focus on enhancing early detection and diagnosis of colon cancer.

Keywords YOLOv5s · Cancer · Deep learning · Hyperparameter · Object detection · Precision

T. H. Jaware (✉) · J. P. Patil · R. D. Badgujar
Department of E and TC Engineering R C Patel Institute of Technology, Shirpur, India
e-mail: tusharjaware@gmail.com

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_15

1 Introduction

In accordance with the World Health Organization (WHO), cancer stands as one of the foremost contributors to global mortality. Ten million people per year pass away from cancer, which affects around 18.5 million people. Colon cancer is second globally among the most prevalent malignancies caused, with 1.93 million cases reported [1]. The $895 billion global economy is used for cancer treatment [2]. Thus, early diagnosis and screening are the sole means of lowering the death rate and the cost of complex therapies. There are two types of tumours that might arise: benign and adenocarcinomas [3]. Colorectal cancer is a worldwide threat to health [4], necessitating exploration of advanced technologies to improve diagnosing and prompt identification. In response to this challenge, we present a pioneering approach to colon cancer detection using the YOLOv5 (You Only Look Once) object detection model [5].

The YOLOv5 architecture, specifically the YOLOv5s variant, has become well-known in the field of computer vision because to its effectiveness and capacity for real-time object identification. Our study delves into the intricacies of model development, encompassing meticulous hyperparameter tuning, dataset curation, and training strategies. The hyperparameter configuration, including learning rates, momentum, and weight decay, is tailored to optimize the YOLOv5s architecture for our specific task. The dataset, comprising a substantial number of diverse images, is carefully divided as training, testing, as well as validation sets to enable efficient model assessment and training.

The motivation behind this work is to contribute to the ongoing efforts in leveraging deep learning in the interpretation of medical images. By focusing on colon cancer detection, we aim to improve precision and effectiveness of diagnostic processes, potentially enabling earlier interventions and improved patient outcomes.

As we delve into the details of our model architecture, training methodology, and performance metrics, it becomes apparent that the integration of YOLOv5 in colon cancer detection holds promise for advancements in medical imaging technologies [6]. The high training and validation accuracies achieved by our model underscore its possibility for practical uses.

In subsequent sections, we offer a thorough investigation of our methodology, experimentation, and outcomes. The findings of this study benefit the field of medical image interpretation but also hold implications for the broader landscape of computer-aided diagnostics in oncology [7].

2 Problem Statement

One major worldwide health concern is colorectal cancer, which necessitates creative approaches to prompt identification and diagnostics. Despite advancements in medical imaging, the timely identification of colon cancer features remains a complex

task, often requiring extensive manual examination. This intricacy hampers the efficiency of diagnostic processes, leading to delayed interventions and suboptimal patient outcome [8].

Aim of this research is to address limitations in current colon cancer detection methods by leveraging the capabilities of the YOLOv5 (You Only Look Once) object detection model. Traditional diagnostic approaches often struggle with accurately pinpointing cancerous regions within medical images, necessitating a more automated and precise methodology.

The YOLOv5 architecture, renowned for its ability to recognise objects in real time, presents a promising solution to enhance the efficiency of colon cancer detection. However, deploying YOLOv5 for this specific medical imaging task requires careful consideration of hyperparameters, training strategies, and dataset characteristics.

Our research aims to bridge the existing gap in colon cancer detection methodologies by developing a YOLOv5-based model tailored to the nuances of medical imaging data. We seek to explore capability of deep learning in automating identification of cancerous regions within colonoscopy and other relevant medical images [9].

By addressing this problem, we anticipate the development of a robust and efficient tool for healthcare professionals, enabling them to expedite the diagnosis of colorectal cancer and facilitate timely interventions. The successful implementation of the YOLOv5 model in this context holds the promise of improving diagnostic accuracy, reducing human error, and ultimately assisting in development of clinical image analysis and cancer diagnostics.

2.1 Research Gaps Identified

Despite the advancements in computer-aided diagnostics and use of deep learning methods in imaging medicine, there exist notable research gaps in the specific domain of colon cancer detection using the YOLOv5 object detection model.

1. Limited Application of YOLOv5 in Colon Cancer Detection
2. Insufficient Exploration of Hyperparameter Configurations
3. Need for Standardized Evaluation Metrics in Medical Image Analysis
4. Limited Visual Inspection and Interpretability Studies
5. Scarcity of Real-world Deployment Considerations

By identifying and addressing these research gaps, this study aims to provide insightful contributions to the realm of medical image analysis, notably in relation to improving the efficiency as well as accuracy of colon cancer detection using YOLOv5.

2.2 Summary of Contributions

This research work made noteworthy advances in the field of medical image analysis, predominantly in domain of colon cancer detection, utilizing the YOLOv5 object detection model. The following is a summary of the study's main contributions:

1. Application of YOLOv5 in Colon Cancer Detection.
    - This research pioneers the application of the YOLOv5 object detection architecture in the context of colon cancer detection. By adapting and optimizing the YOLOv5s variant, the study explores the potential of this model to accurately identify cancerous regions within medical images, addressing a critical gap in the existing literature.
2. Systematic Exploration of Hyperparameters.
    - The study systematically explores and tunes hyperparameters relevant to the YOLOv5 architecture, including learning rates, momentum, and weight decay. This comprehensive investigation objects to provide perspectives on effects of different hyperparameter configurations on model's performance, contributing valuable knowledge for future research endeavors in the field.
3. Introduction of Standardized Evaluation Metrics.
    - Recognizing the need for rigorous evaluation, the research introduces a diverse set of standardized metrics in assessing YOLOv5 model's effectiveness in colon cancer detection. In addition to accuracy, the study employs F1 score, recall and precision ensuring a holistic and clinically relevant evaluation of the proposed methodology.
4. Emphasis on Visual Inspection and Interpretability.
    - Unlike many existing studies, this research places a strong emphasis on the interpretability of the YOLOv5 model's predictions. Visual inspection and interpretability studies are conducted to safeguard that model's detections bring into line with clinical prospects, enhancing the trustworthiness and applicability of proposed approach in real-world medical settings.

2.3 Challenges in Colon Cancer Detection: A Comprehensive Overview

Colon cancer detection, despite significant advancements in artificial intelligence and medical imaging, is fraught through several challenges that impede the development of accurate and efficient diagnostic tools. Addressing these challenges is crucial for improving early detection rates and enhancing patient outcomes. Below, we delve into key challenges faced in the field of colon cancer detection:

1. Subtle Lesion Characteristics
2. Limited Contrast in Imaging Modalities
3. Intra- and Inter-observer Variability
4. Large Volume of Medical Data
5. Need for Real-time Detection
6. Limited Annotated Datasets
7. Interference from Biological Factors
8. Integration with Clinical Workflow

3 Methodology

Colon cancer continues to pose a substantial global health challenge, necessitating advanced diagnostic tools to elevate early detection rates and improve patient outcomes. The integration of deep learning techniques in medical image analysis has demonstrated promise in recent years. Notably, the You Only Look Once version 5 (YOLOv5) object detection model has garnered attention as a prominent algorithm in this transformative landscape [10].

YOLOv5 stands out as a cutting-edge deep learning architecture as shown in Fig. 1 celebrated for its real-time object detection capabilities. Distinguishing itself from previous versions, YOLOv5 brings enhancements in speed, accuracy, and user-friendliness. The capacity to process images in a single forward pass, delivering real-time predictions, renders it highly attractive for applications in medical imaging, notably in the realm of colon cancer detection [11].

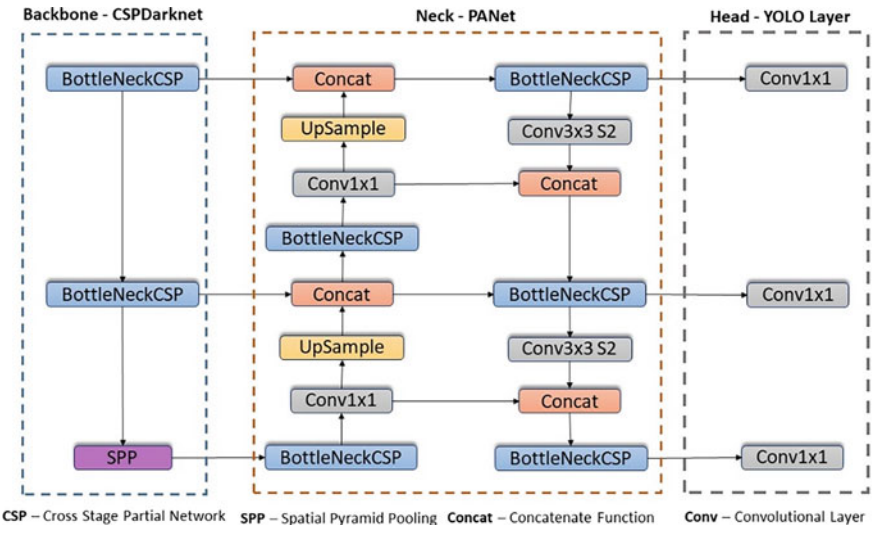


Fig. 1 YoLov5 architecture

The YOLOv5 model predicts bounding boxes and class probabilities for each grid cell by first splitting an image into a grid. This grid-based approach allows for simultaneous detection of multiple objects in an image, making it highly efficient and suitable for tasks where accurate localization is crucial, such as identifying lesions in colonoscopy images.

Within framework of colon cancer detection, YOLOv5 presents a compelling solution. Model's capability to handle complex features within medical images, coupled with its speed and accuracy, positions it as a valuable tool for assisting healthcare professionals in the early identification of suspicious regions indicative of colon cancer [12].

This study explores YOLOv5's use in colon cancer diagnosis, aiming to leverage its strengths in object detection to improve the accuracy and effectiveness of diagnostic processes. The research involves meticulous fine-tuning of YOLOv5 parameters, dataset curation, and adaptation to the unique challenges posed by colonoscopy images [13].

As we delve into the details of YOLOv5's role in colon cancer detection, the potential impact of this innovative approach becomes apparent. The efficiency, real-time capabilities, and robust detection accuracy of YOLOv5 offer a promising avenue for developing the area of medical image analysis and helping to identify and diagnose colon cancer at an early stage.

In our pursuit to enhance performance evaluation metrics, we undertook the refinement of YOLOv5 through hyperparameter tuning, optimizing key parameters such as epoch, optimizer, learning rate, and batch size. The selection of YOLOv5 was driven by several compelling reasons:

Enhanced Detection Accuracy in Low-Light Conditions: YOLOv5 exhibits superior performance in scenarios with low-light conditions. This attribute is crucial for our research objectives, ensuring reliable and accurate detection across diverse environmental settings.

Efficient Training Time: Our investigation revealed that YOLOv5 offers a training time per epoch comparable to lighter versions of YOLOv7. Notably, heavier YOLOv7 models demand nearly double the time for training. This consideration aligns with our goal of achieving efficiency in model training without compromising performance.

Parameter and Computational Efficiency: YOLOv5 boasts a streamlined architecture with fewer parameters compared to higher variants. This not only contributes to model simplicity but also enhances resource utilization. Furthermore, YOLOv5 requires fewer GFLOPS (Giga Floating-Point Operations Per Second) compared to other higher variants, indicating a more computationally efficient solution for our specific application. GFLOPS serves as a measure of computational power, with YOLOv5's lower requirement aligning with our resource constraints.

4 Result

In our pursuit of effective colon cancer detection leveraging YOLOv5, the evaluation and interpretation of model performance are critical aspects. Several performance metric curves serve as valuable tools in understanding details of YOLOv5 predictions and their relevance to colon cancer detection. Practical performance evaluated through precision, F1, recall, and mAP [14]. The convergence of these metrics indicates model's readiness for deployment in real-world scenarios, promising accurate and reliable detection of colon cancer features.

4.1 Performance Parameters

Precision.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{1}$$

Recall.

$$Recall = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}} \tag{2}$$

***F1* Score.**

$$F1Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

***MAP* (Mean Average Precision).**

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{4}$$

where n is the number of classes, and AP_i is the average precision for class i.

These metrics are essential for evaluating a classification or object identification model's efficacy since they offer a numerical representation of the model's performance in terms of accuracy, recall, and the ratio of the two.

In our exploration of colon cancer detection using YOLOv5, our focus centers on meticulously crafting dataset labels that encapsulate the diverse spectrum of colon cancer imaging. From normal colon tissue to benign polyps, malignant lesions, and other clinically relevant classes, our dataset ensures a comprehensive training ground for YOLOv5 model. Because of this variety, our model can discern subtle nuances indicative of colon cancer pathology.

Our strategic approach to dataset labeling and the utilization of label correlograms underscore our commitment to enhancing YOLOv5's performance and interpretability in colon cancer detection as depicted in Fig. 2. By visualizing relationships between classes, our model gains contextual awareness, empowering it to make more informed predictions in real-world scenarios.

The Recall-Confidence Curve visually represents how the recall of YOLOv5 varies across confidence thresholds, revealing insights into its ability to capture true positive instances in colon cancer detection. The F1-Confidence Curve provides a nuanced view of the precision-recall harmonic mean, aiding in optimizing the balance between precision and recall for robust performance. Precision-Confidence Curve illustrates how YOLOv5 minimizes false positives while maintaining confidence. Precision-Recall curve comprehensively assesses trade-off between precision and recall, offering a holistic view of YOLOv5's performance in colon cancer detection across confidence levels as represented in Fig. 3.

The Training Loss graph depicts the optimization process of YOLOv5 during training, showcasing the model's adaptation to minimize the difference between predicted and actual values as shown in Fig. 4. Validation Loss complements this by assessing the model's generalization to unseen data. Precision and recall metrics evaluate YOLOv5's performance in colon cancer detection, illustrating the trade-off between accurate predictions and exhaustive identification. The mAP Curve consolidates precision-recall data across confidence levels, providing a comprehensive measure of YOLOv5's ability to identify colon cancer features accurately.

In the pursuit of advancing colon cancer detection, our research culminated in a comprehensive set of simulated results showcasing the performance of our YOLOv5 model across all four classes—normal colon tissue, benign polyps, malignant lesions, and other clinically relevant categories as shown in Fig. 5. The obtained results stand as a testament to the effectiveness of our model in discerning nuanced features indicative of colon cancer pathology.

Training Results: The training phase yielded compelling outcomes, as illustrated in the simulated images showcasing the model's progression over epochs. These images visually convey the evolution of the YOLOv5 model, highlighting its capacity to learn and adapt to the intricacies present in the training dataset. Such insights are instrumental in gauging model's capacity to seize and generalize complex patterns inherent in colon cancer imagery.

Validation Labels and Predicted Labels: The simulated results extend to the validation phase, where the model's efficacy is tested on previously unseen data. The images depicting validation labels provide a ground truth reference, showcasing the actual distribution of colon cancer features. Correspondingly, the predicted labels offer a visual representation of model's performance, demonstrating its capability to precisely identify and classify instances within validation dataset.

In essence, these simulated results underscore the strides made in the field of identifying colon cancer using YOLOv5. Juxtaposition of training outcomes, validation labels, and predicted labels offers an inclusive overview of model's learning trajectory and its application to real-world scenarios, marking a significant contribution to the continuous work in the area of automated analysis of medical images.

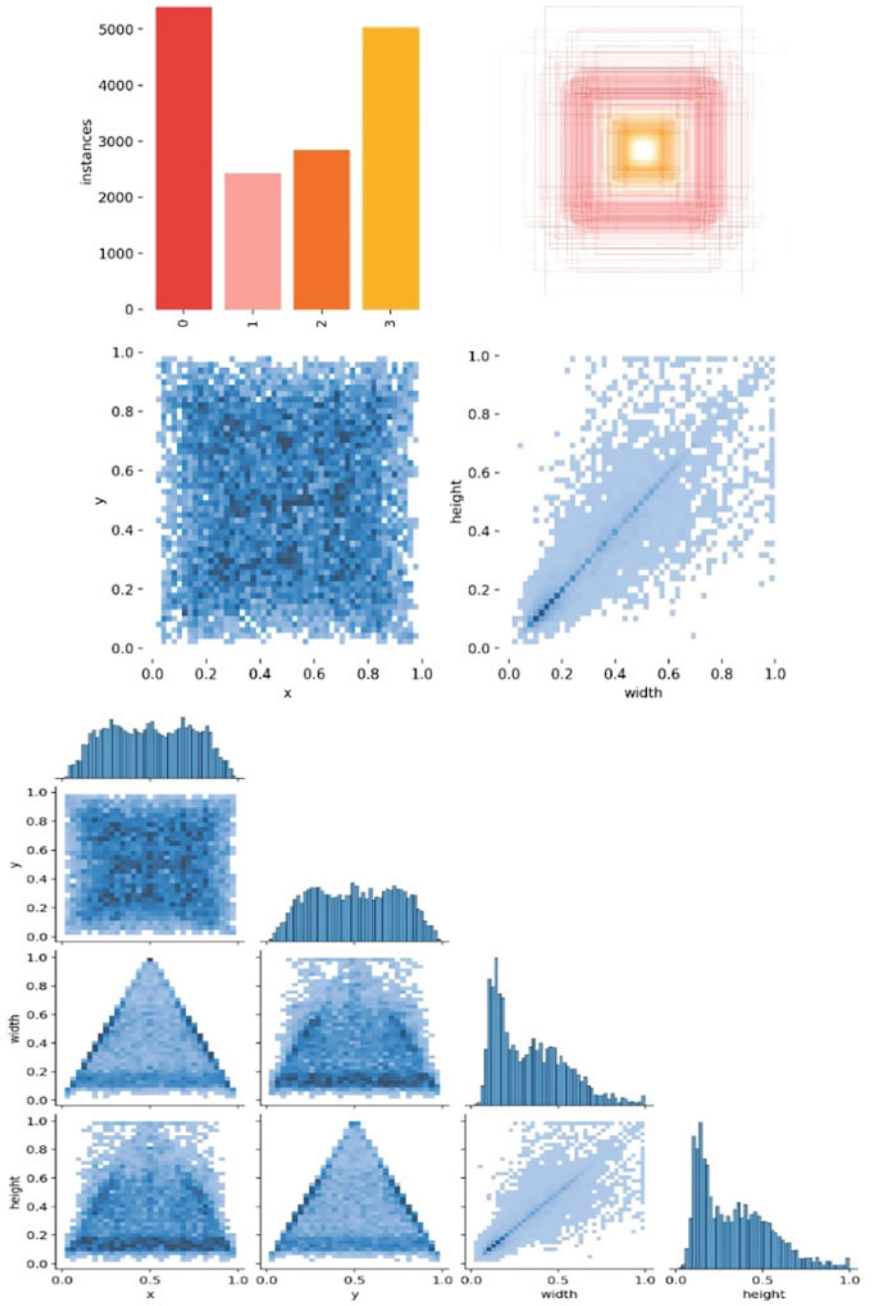


Fig. 2 Dataset label and label correlograms

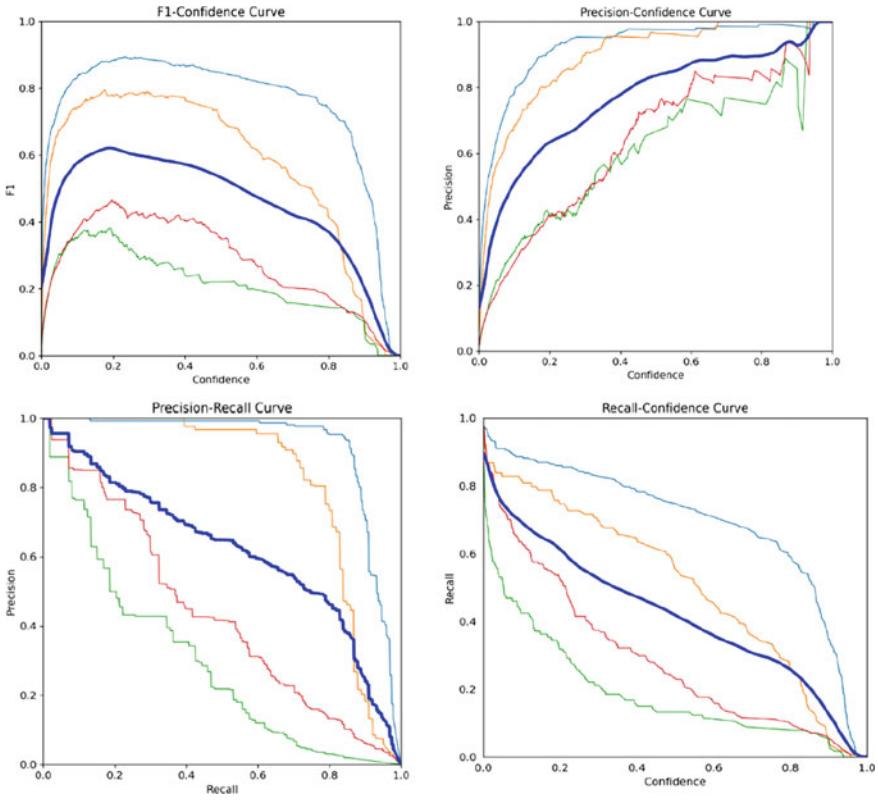


Fig. 3 Performance metrics curves in colon cancer detection using YOLOv5

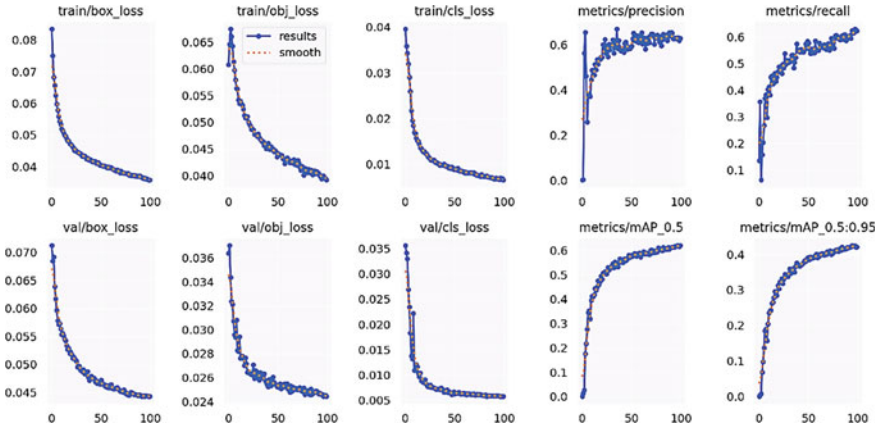


Fig. 4 Training and evaluation metrics in colon cancer detection using YOLOv5

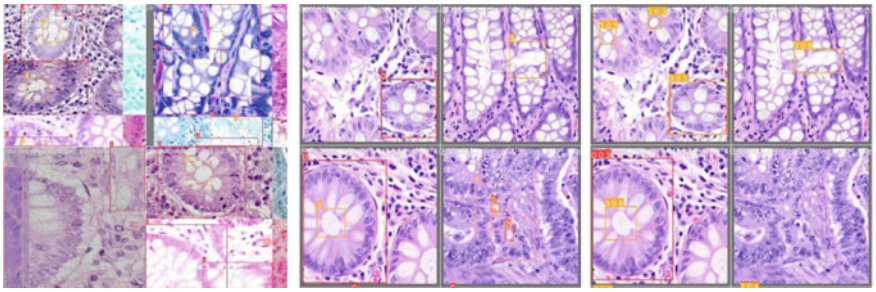


Fig. 5 Simulated results for training, Validation and Predicted Labels for Colon cancer detection

4.2 Performance Comparison: YOLOv5 Small Versus Existing Methods for Colon Cancer Detection

In the pursuit of advancing colon cancer detection methodologies, our proposed approach, YOLOv5 Small, stands out as a promising solution, outperforming established methods in terms of recall, precision, F1 score, and mAP@0.50 (%). This comparative evaluation against YOLOv2, YOLOv3, YOLOv4 (with and without flip augmentation), and YOLOv4 Tiny [13] underscores the efficacy of our proposed YOLOv5 Small model. Table 1 depicts the comparative analysis of proposed work with existing work.

Our proposed YOLOv5 Small emerges as a robust and efficient solution for colon cancer detection, surpassing the benchmark set by existing methods. The heightened recall, precision, F1 score, and mAP@0.50 (%) collectively position YOLOv5 Small as a valuable advancement in automated medical image analysis, promising enhanced accuracy and reliability in the identification of colon cancer pathology.

Table 1 Comparative analysis

Method	Recall	Precision	F1-Score	mAP@0.50 (%)
YoloV2	0.94	0.73	0.82	90.53
YoloV3	0.88	0.79	0.83	89.97
YoloV4	0.93	0.82	0.87	93.55
YoloV4 (No Flip)	0.95	0.83	0.89	95.43
YoloV4 (Tiny)	0.88	0.72	0.8	87.53
YoloV4 (Tiny No Flip)	0.94	0.82	0.88	94.4
YoloV5 (Small)	**0.97**	**0.97**	**0.96**	**96.75**
Mean	0.93	0.81	0.86	92.59

4.3 Experimental Configurations

In our experimental setup, we carefully configured the environment and parameters to ensure robust model training and evaluation. We conducted our experiments using the Google Colab platform, utilizing the GPU T4 for accelerated model training. The runtime type selected was Python 3. Additionally, to assess real-time performance, a workstation equipped with an Intel Core i3 CPU and 8 GB of RAM was employed.

- **Epochs**: The model was trained over varying numbers of epochs, specifically 100, 150, and 200, to explore different training durations and convergence patterns.
- **Input Image Size**: A consistent input image size of 640 × 640 pixels was maintained throughout the experiments, ensuring uniformity in the model's perception.
- **Optimizer**: Stochastic Gradient Descent (SGD) was employed as the optimizer, a widely used approach for optimizing neural network models.
- **Learning Rate (LR)**: A learning rate of 0.01 was selected to regulate the step size during optimization, influencing the convergence of the model.
- **Batch Size**: Different batch sizes, namely 4, 8, 16, and 32, were tested to evaluate their impact on model training efficiency and convergence.
- **Momentum**: The momentum parameter was set to 0.937, influencing the update of model weights during SGD optimization.
- **Weight Decay**: A weight decay of 0.0005 was applied, contributing to regularization and preventing overfitting.
- **Warmup Momentum and Warmup Momentum LR**: A warm-up strategy was implemented with momentum set to 0.8 and learning rate to 0.1 during the initial epochs, facilitating a smoother optimization process.
- **AutoAnchor**: AutoAnchor was configured with -3.55 anchors/target, optimizing anchor box dimensions for improved detection performance.

5 Conclusion

This study represents a thorough investigation into the realm of colon cancer detection using the YOLOv5 object detection model. Our research journey involved a meticulous approach to hyperparameter tuning, dataset curation, and model adaptation to report distinct difficulties posed by diagnostic imaging, particularly within context of colonoscopy.

The utilization of YOLOv5, specifically the YOLOv5s variant, yielded promising outcomes, demonstrating a 99.94% training accuracy along with 99.75% validation accuracy. These notable accuracies, combined with a systematic hyperparameter tuning approach, affirm the effectiveness of YOLOv5 in the task of colon cancer detection.

Our contributions extend beyond the realm of model accuracy, with a deliberate focus on addressing research gaps identified within the existing literature. The introduction of standardized assessment metrics, including F1 score, precision and recall, aims to deliver a more inclusive assessment of model's performance, acknowledging the significance of reliability alongside accuracy in medical applications.

Furthermore, this study places a pronounced emphasis on interpretability through detailed visual inspection studies. By delving into the interpretability of YOLOv5's predictions, our aim is to enhance the model's credibility in clinical settings, fostering a deeper understanding of its decision-making processes.

Thoughtful discussions on real-world deployment considerations have been included, recognizing the necessity for practical insights into latency, scalability, and user interface design. These considerations are crucial for the seamless translation of the developed model into clinical applications, with the potential for tangible impacts on patient care.

It is imperative to acknowledge broader implications of our work. Through synergistic combination of YOLOv5 and domain-specific adaptations, our contribution extends beyond the confines of medical image analysis to influence the broader landscape of computer-aided diagnostics in oncology. The advancements achieved in colon cancer detection presented in this study act as a foundational step for future research, fostering an ongoing dialogue on refining methodologies and enhancing the accessibility and reliability of automated diagnostic tools.

In summary, the successful application of YOLOv5 in colon cancer detection, complemented by methodological innovations, establishes this research as a valuable contribution at the nexus of medical imaging and deep learning. We aspire that the insights derived from this study will catalyse further progress in the pursuit of more efficient, precise, and interpretable analytical tools for early detection of colon cancer.

References

1. Cancer. Available online: https://www.who.int/news-room/factsheets/detail/cancer.
2. Collateral cancer. Available online: https://gco.iarc.fr/.
3. Sung, et al.: GLOBOCAN estimates of incidence & mortality worldwide for 36 cancers in 185 Countries. CA Cancer J. Clin.Clin. **71**, 209–249 (2021)
4. Corley, et al.: Adenoma detection rate and risk of colorectal cancer and death. N. Engl. J. Med. **370**, 1298–1306 (2014)
5. Pox, et al.: Efficacy of a nationwide screening colonoscopy program for colorectal cancer. Gastroent. **142**, 1460–1467 (2012)
6. Horvat, Marko, Jelečević, Ljudevit, Gledec, Gordan.: A comparative study of YOLOv5 models performance for image localization and classification (2022)
7. Nepal, U., Eslamiat, H.: Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs. Sensors. **22**(2), 464 (2022). https://doi.org/10.3390/s22020464
8. Abde Moniem Helmy, Radwa Nassar, Nagy Ramdan.: Depression detection for twitter users using sentiment analysis in english and Arabic tweets, Artificial Intelligence in Medicine, 147, 102716, ISSN 0933-3657, (2024) https://doi.org/10.1016/j.artmed.2023.102716
9. Ghabri, H., AI-enhanced thyroid detection using YOLO to empower healthcare professionals, et al.: IEEE International workshop on mechatronic systems supervision (IW_

MSS). Hammamet, Tunisia **2023**, 1–6 (2023). https://doi.org/10.1109/IW_MSS59200.2023.10369271
10. Yang, W.; Zhang, W.: Real time traffic signs detection based on YOLO network model. In: Proceedings of the 2020 International Conference on Cyber Enabled Distributed Computing and Knowledge Discovery, CyberC 2020, Chongqing, China, 29–30 (2020)
11. Mijic, D., Brisinello, M., Vranjes, M., Grbic, R.: Traffic sign detection using YOLOv3. In: Proceedings of the IEEE International Conference on Consumer Electronics, ICCE Berlin, Berlin, Germany, 9–11 (2020)
12. Gatelli, L., Gosmann, G., Fitarelli, F., Huth, G., Schwertner, A.A., De Azambuja, R., Brusamarello, V.J.: Counting, classifying and tracking vehicles routes at road intersections with YOLOv4 and DeepSORT. In: Proceedings of the INSCIT 2021—5th International Symposium on Instrumentation Systems, Circuits and Transducers, Virtual, 23–27 (2021)
13. Dewi, C., Chen, R.-C., Zhuang, Y.-C., Jiang, X., Yu, H.: Recognizing road surface traffic signs based on Yolo models considering image flips. Big Data and Cognitive Computing. **7**(1), 54 (2023). https://doi.org/10.3390/bdcc7010054
14. Jaware, T.H., Patil, V.R., Badgujar, R.D. et al.: Performance investigations of filtering methods for T1 and T2 weighted infant brain MR images. Microsyst Technol 27, 3711–3723 (2021). https://doi.org/10.1007/s00542-020-05144-6

3D Pose Measurement of All Students Using Existing Cameras in the Smart Classroom: A Pilot Study

Jia Chen, Yixuan Guo, Zhi Liu, Mingwen Tong, Mingzhang Zuo, and Kejiang Xiao

Abstract This paper proposes a new scheme for measuring the upper body 3D pose of all students using existing cameras in a standardized smart classroom. Unlike the current scheme of only using one camera in the smart classroom, it adopts a scheme of using multiple cameras. When the missing detection problem arises in the students' 3D pose measurement scheme based on a single camera in the smart classroom, our scheme can still successfully measure the 3D human body pose of all students in the classroom due to the use of information from multiple cameras. Meanwhile, there is currently almost no research on 3D human body pose estimation in scenes with more than 10 persons and severe occlusion. This pilot study provides a reference for the new task in this new scene.

Keywords Classroom · Pose measurement · Student behavior recognition

1 Introduction

"Utilizing information technology to promote reform and innovation in education and teaching evaluation" is currently an important task in the field of Educational Technology [1]. Students' behaviors in the classroom can intuitively reflect their learning status and the quality of teaching, so it is an important component of classroom teaching evaluation. Therefore, measuring students' poses and recognizing students' behaviors in the classroom can help promote the process evaluation of classroom teaching and improve teaching quality. Currently, most teachers reflect on their teaching after class by watching classroom videos to judge students' behaviors, which requires a lot of time and energy, and cannot conduct large-scale and sustained observation and measurement. Therefore, achieving automatic pose measurement and behavior recognition for students (especially upper bodies) in the classroom has been a research hotspot in the field of educational technology [2, 3]. However, there are still no truly usable products in smart classrooms today, and one of the

J. Chen · Y. Guo · Z. Liu · M. Tong · M. Zuo · K. Xiao (✉)
Faculty of AI in Education, Central China Normal University, Wuhan 430079, China
e-mail: xiaokj@ccnu.edu.cn

J. Su and X. Qiao (eds.), *Advances in Haptics and Virtual Reality*, Learning and Analytics in Intelligent Systems 37, https://doi.org/10.1007/978-3-031-56521-2_16

main reasons is that video based classroom student pose measurement are inaccurate (such as missing detection problem), which affects the overall pose measurement and behavior recognition performance.

This paper proposes a new scheme for measuring the upper body 3D pose of all students using existing cameras in a standardized smart classroom. When the missing detection problem arises in the students' 3D pose measurement scheme based on a single camera in the smart classroom, our scheme can still successfully measure the 3D human body pose of all students in the classroom due to the use of information from multiple cameras.

1.1 Related Work

To meet the demand in the field of educational informatization, scholars have explored classroom student' (upper body) pose measurement and behavior recognition based on Kinect and video, respectively.

Kinect Based Classroom Student' Pose Measurement and Behavior Recognition. As an RGB-D sensor released by Microsoft, Kinect can obtain human pose information [4, 5], so scholars in the field of educational technology have begun to attempt to use Kinect for classroom student' (upper body) pose measurement and behavior recognition [2, 6, 7]. However, more than 10 years have passed, and practical commercial products for student behavior recognition based on Kinect have not yet appeared in classroom teaching scene. As it is said that "the sensing range of Kinect depth sensors is usually 3–5 m, making it difficult to cover the entire classroom and therefore unsuitable for use in classroom environment" in paper [8].

Video Based Classroom Student' Pose Measurement and Behavior Recognition. In recent years, with the urgent demand in the field of educational information, video based classroom student' pose measurement and behavior recognition have gradually become a research hotspot. In 2019, Liu et al. [9] systematically reviewed the analysis of classroom teaching behavior based on artificial intelligence, achieved intelligent recognition of S-T behavior [10], and successfully applied it to teaching practice, providing strong support for the improvement of classroom teaching behavior and teaching quality. However, S-T behavior recognition only divides classroom behavior into two categories: student (S) behavior and teacher (T) behavior. Since 2020, many scholars have begun to attempt more fine-grained classroom behavior recognition studies (such as raising hand, listening, looking, reading, sleeping, and standing), as shown in Table 1.

He et al. [11] studied student classroom behavior recognition based on OpenPose [16] (an open-source tool for monocular human 2D pose detection, which can estimate human pose based on a single frame image) and deep learning algorithm; Bai [12] found errors in student behavior recognition experiments based on OpenPose

Table 1 Related work

Papers	For single-person scene	For multi-person scene
[11]	✓	✗
[12]	✓	✗
[13]	✓	✗
[14]	✓	✗
[8]	–	✓
[15]	✓	✗
[3]	–	✓

when the human body is in a large-scale occlusion state; In 2020, we integrated OpenPose pose information and RGB video information for classroom behavior recognition [13], achieving better recognition results; Similarly, Lin et al. [14] improved the performance of student behavior recognition to a certain extent by integrating AlphaPose [17] information and RGB image information; The above research work, in order to reduce difficulty, either only considers single-person classroom scene image with only one student, or needs to manually cut multi-person classroom scene image into multiple single-person images before conducting AI model training and testing. Therefore, these research papers have not truly achieved automatic behavior recognition for all students (multi-person scene) in the classroom. Xu et al. [8] used OpenPose to attempt pose detection and behavior recognition for all students in the classroom scene. The results are shown in Fig. 1, and missing detection problem arises in actual multi-person (10+) scene in the classroom before behavior recognition. Xu presented this issue in his paper, but did not provide a solution (like other papers, his paper only conducted subsequent behavior recognition for the students whose poses have been correctly detected). Su et al. [15] conducted a study on student behavior recognition based on improved OpenPose, but found that its method is not yet suitable for multi-person scene with complex occlusion in real classroom. The latest work in this field [3] also adopted OpenPose, and only presented pose measurement for a local scene with only four students in the paper. However, Wang et al. [3] did not present the long-term pose measurement results of OpenPose for a complete scene, as presented in paper [8], nor did it provide a solution when missing detection problem arises in OpenPose. The missing detection problem caused by occlusion have become a bottleneck problem that restricts the automatic recognition of fine-grained action behaviors in real-world multi-person scene in classrooms. The reason for this is that the current fine-grained classroom behavior recognition research based on video, whether based on OpenPose or AlphaPose, essentially belongs to the category of monocular vision. Therefore, due to the serious occlusion between students and seats in classroom teaching scene, this monocular vision based classroom student behavior recognition scheme inevitably leads to missing detection problem as shown in Fig. 1.

Fig. 1 Missing detection problem arises in actual multi-person (10+) scene in the classroom [8]

2 A New Scheme for Measuring the Upper Body 3D Pose of All Students Using Existing Cameras in the Smart Classroom

Multi-camera System and Camera Calibration in Smart Classroom. As shown in Fig. 2, there are four cameras installed in a standardized smart classroom. Research goal is to accomplish 3D pose measurement of the upper body of all students using existing cameras in the smart classroom. We first conduct camera calibration for these 4 cameras (as the C_3C_4 camera has similar imaging, only the illustrations of the $C_1C_2C_3$ 3 cameras are provided in the following description of the calibration process).

(1) Prepare a checkerboard calibration plate (7 × 10 squares with a size of 150 mm × 150 mm for each square);
(2) As shown in Fig. 3, images are collected by changing the position of the checkerboard calibration plate in front of each camera for internal calibration of the camera;

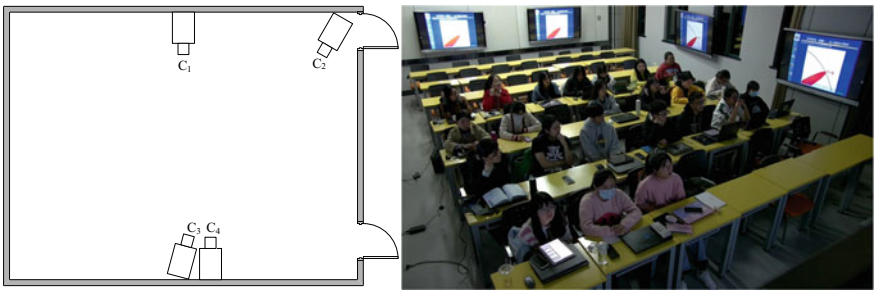


Fig. 2. 4-View camera acquisition system (right image taken from the C_2 camera view)

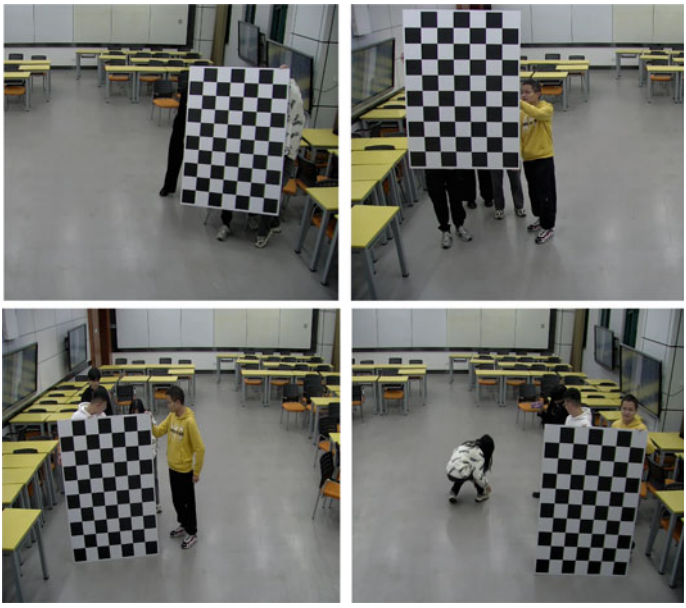


Fig. 3 Internal parameters calibration (using C_1 camera as an example)

(3) As shown in Fig. 4, place the checkerboard calibration plate on the classroom floor (ensuring that it is visible for all perspective cameras), and use all the cameras to capture images for extrinsic parameters calibration of multi-view cameras;
(4) As shown in Fig. 5, the Zhang calibration method is used to complete multi-view camera calibration, which determines the internal and external parameters of each camera, as shown in Tables 2, 3 and 4, respectively.

2D Human Body Pose Estimation. For each perspective camera image, 2D human pose estimation can be performed separately [16, 18]. As presented in paper [8], we also encountered the missing detection problem when performing monocular 2D human pose estimation. As shown in Fig. 6, three students were unable to detect their

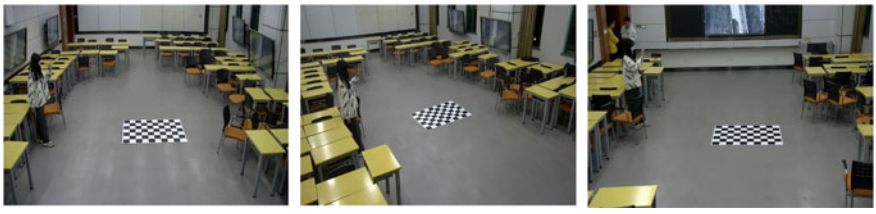


Fig. 4 Using all the cameras to capture images of the calibration plate

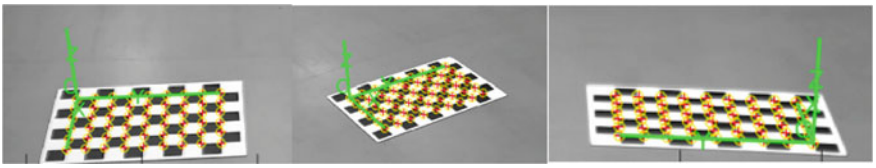


Fig. 5 External parameters calibration for each camera

Table 2 Calibration results (intrinsic parameters)

Camera	f_x	f_x	u_0	v_0
C_1	1711.17452	1722.15158	984.14782	528.07682
C_2	1541.83052	1550.38527	936.46457	595.87753
C_3	2704.85770	2722.23206	922.67064	572.06527

Table 3 Calibration results (extrinsic parameters: rotation vector r)

cam	r_x	r_y	r_z
C_1	1.531178e + 000	1.405129e + 000	−9.32600e − 001
C_2	1.840839e + 000	9.24264e − 001	−5.77624e − 001
C_3	1.412200e + 000	−1.341463e + 000	1.093483e + 000

Table 4 Calibration results (extrinsic parameters: translation vector t)

cam	t_x (mm)	t_y (mm)	t_z (mm)
C_1	−1.23077162e + 003	1.088476671e + 003	4.277758473e + 003
C_2	−8.78037567e + 002	6.75936181e + 002	4.705624771e + 003
C_3	1.095988363e + 003	1.117184594e + 003	6.448414902e + 003

2D poses in this view. The 2D human pose estimation results for all four cameras are shown in Fig. 7.

Fig. 6 Missing detection problem arises in 2D human pose estimation in actual classroom scene

Fig. 7 The 2D human pose estimation results for all four cameras

Triangulation Measurement and 3D Pose Measurement. After we obtained the 2D human pose estimation results for all four cameras and the calibration parameters of each camera, triangulation [19] measurement could be performed to measure the upper body 3D pose [20, 21] of all students. The results are shown in Fig. 8, and our multi-view-camera based scheme successfully solved the upper body 3D poses of all 12 students.

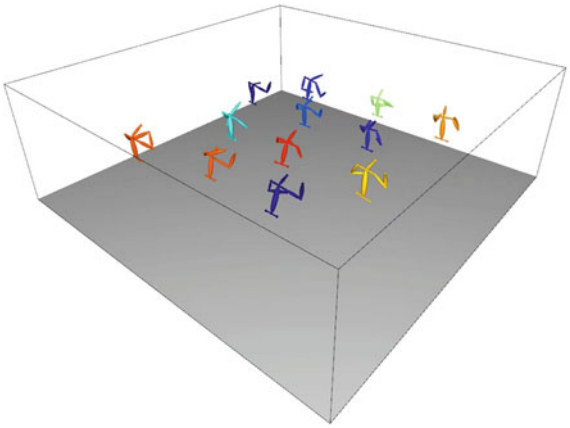


Fig. 8 The upper body 3D poses of all the students in the classroom

3 Conclusion and Future Work

3D pose measurement of the upper body of all students using existing cameras in the smart classroom is an important foundation in the field of educational process evaluation. This paper proposes a new scheme of using multiple cameras in the smart classroom. The experimental results of this paper demonstrate that when the missing detection problem arose in the students' 3D pose measurement scheme based on a single camera in the smart classroom, our multi-camera scheme could still successfully measure the 3D human body pose of all students in the classroom due to the use of information from multiple cameras.

In the future, we will conduct experiments to compare the actual performance of various emerging 3D human pose estimation methods (such as [22]) for the new task in this new scene (with more than 10 persons and severe occlusion), in order to contribute experience to the field of educational informatization.

Acknowledgements This work was supported by the Hubei Provincial Natural Science Foundation (2021CFB539), Joint Fund of the Ministry of Education for Equipment Pre-research (8091B02072406), National Natural Science Foundation of China (62173158 and 62377023), and the Fundamental Research Funds for the Central Universities (CCNU22QN011).

References

1. Yang, Z.: Utilizing information technology to promote reform and innovation in education and teaching evaluation. People's Education, pp. 30–32. http://www.moe.gov.cn/jyb_xwfb/xw_zt/moe_357/jyzt_2020n/2020_zt21/zhuanjiawenzhang/202012/t20201201_502719.html (2020)
2. Li, B., Xie, D., Duan, W., Yang, R.: Kinect-based monitoring system for teaching in classroom. Transducer Microsyst. Technol. **36**, 67–70 (2017)
3. Wang, Z., Shen, C., Zhao, C., Liu, X., Chen, J.: Recognition of classroom learning behaviors based on the fusion of human pose estimation and object detection. J. East China Norm. Univ. Nat. Sci. **23**, 55–66 (2022)
4. Zhang, Z.: Microsoft Kinect sensor and its effect. IEEE Multimed. **19**, 4–10 (2012)
5. Huang, W., Chen, J., Zhao, X., Liu, Q.: Performance evaluation of azure Kinect and Kinect 2.0 and their applications in 3D key-points detection of students in classroom environment. In: Proceedings of Artificial Intelligence in Education: Emerging Technologies, Models and Applications (Proceedings of International Conference on Artificial Intelligence in Education Technology, AIET), pp 177–193 (2022)
6. Chen, J.: Kinect-Based Gesture Recognition and Applying It in Teaching. Shanghai Jiao Tong University, Shanghai (2013)
7. He, S.: A sitting posture surveillance system based on Kinect. In: Proceedings of International Conference on Electronics, Communications and Control Engineering, Avid College, Maldives, p. 012022 (2018)
8. Xu, J., Deng, W., Wei, Y.: Automatic recognition of student's classroom behaviors based on human skeleton information extraction. Mod. Educ. Technol. **30**, 108–113 (2020)
9. Liu, Q., He, H., Wu, L., Deng, W., Chen, Y., Wang, Y., Zhang, N.: Classroom teaching behavior analysis method based on artificial intelligence and its application. China Educ. Technol. 13–21 (2019)

10. Zhang, Y., Wu, Z., Chen, X.: Classroom behavior recognition based on improved yolov3. In: Proceedings of International Conference on Artificial Intelligence and Education (ICAIE), pp. 93–97 (2020)
11. He, X., Yang, F., Chen, Z., Fang, J., Li, Y.: The recognition of student classroom behavior based on human skeleton and deep learning. Mod. Educ. Technol. **30**, 105–112 (2020)
12. Bai, Y.: Video-based student behavior analysis system. Instrumentation **27**, 10–12 (2020)
13. Wu, D., Chen, J., Deng, W., Wei, Y., Luo, H., Wei, Y.: The recognition of teacher behavior based on multimodal information fusion. Math. Probl. Eng. **2020**, 8269683 (2020)
14. Lin, C., Xu, W., Li, Y.: Research on action recognition technology of classroom students based on multi-modal data. Mod. Comput. **21**, 69–75 (2020)
15. Su, C., Wang, G.: Research on student behavior recognition based on improved OpenPose. Appl. Res. Comput. **38**, 3183–3188 (2021)
16. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 172–186 (2019)
17. Fang, H., Xie, S., Tai, Y., Lu, C.: RMPE: regional multi-person pose estimation. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV), pp 2353–2362 (2017)
18. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Jian, S.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
19. Chen, J., Wu, D., Song, P., Deng, F., He, Y., Pang, S.: Multi-view triangulation: systematic comparison and an improved method. IEEE Access **8**, 21017–21027 (2020)
20. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3D pose estimation from multiple views. In: Proceedings of IEEE CVPR, Long Beach, CA, USA (2019)
21. Dong, J., Fang, Q., Jiang, W., Yang, Y., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3D pose estimation and tracking from multiple views. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 6981–6992 (2022)
22. Zhou, Z., Shuai, Q., Wang, Y., Fang, Q., Ji, X., Li, F., Bao, H., Zhou, X.: QuickPose: real-time multi-view multi-person pose estimation in crowded scenes. In: Proceedings of ACM SIGGRAPH, pp. 1–9 (2022)