Implementation of an Underwater Gesture Recognition and Communication System for Divers Based on Computer Vision

Min Tao[1], Wenpei Liu[2]
[1]School of Physical Sciences, Lingnan Normal University, Zhanjiang 524048, Guangdong, China
[2]School of Computer Science and Intelligence Education, Lingnan Normal University, Zhanjiang 524048, Guangdong, China
Corresponding author: Wenpei Liu (e-mail: liuwenpei198904@163.com)

ABSTRACT Underwater environment factors such as optical attenuation, color distortion, low contrast and dynamic background interference are highly challenging to diver gesture recognition and the current methods have a weakness in their inaccuracy in underwater complex environment. This paper suggests a deep learning-based diver gesture recognition and communication system. Initially, dark channel prior and white balance algorithms are implemented to preprocess underwater images to regain the color of the image and increase contrast. Second, an improved YOLOv5 (You Only Look Once v5) network is built that detects gestures, and a Convolutional Block Attention Module (CBAM) is implemented to increase the ability to extract features. Next, the features of gestures are extracted with ResNet50 (Residual Network 50) and a temporal model is trained with the help of LSTM (Long Short-Term Memory) network in order to identify dynamic gesture sequences. Lastly, a communication scheme is developed to translate the recognition outputs into standard diving commands. Experiments are conducted on a self-built dataset containing 8600 images of 12 commonly used diving gestures. The system achieves an average precision of 96.3%, a recall rate of 95.8%, and an F1 score of 96.0% in clean water environments. Its processing speed is stable at 35 frames per second, meeting realtime requirements and providing effective technical support for safe communication in underwater operations. The system links underwater optical recognition with acoustic modem communication, making it relevant to underwater information-transmission engineering.

INDEX TERMS underwater gesture recognition, computer vision, YOLOv5-LSTM, underwater communication, acoustic modem

I. INTRODUCTION

SEVERAL underwater operations are in dire need of effective communication systems. Marine resource exploration, underwater archaeology and diving rescue are some of these tasks that depend on the transmission of information among divers in an accurate manner. Multipath effects and environmental noise interference influence acoustic communication, and the range of activities is restricted in wired communication. Gestures are a natural and intuitive mode of communication and have special advantages. However, the underwater environment is significantly affected by optical degradation phenomena, including light absorption, scattering, and color shift. The presence of suspended particles and the dynamic nature of water flow result in image blurring and motion jitter, making conventional recognition methods based on simple color and shape features inadequate. The notion of terrestrial gesture recognition has also witnessed a breakthrough in the deep learning technology. Convolutional neural networks discern discriminative features automatically to eliminate the constraints of manual design [1-2], and recurrent neural networks model time-varying information to learn dynamic gesture trajectories, which give new concepts to address the underwater gesture recognition. Current literature is concerned with image enhancement and optimization of target detection. Dark channel prior algorithms can predict the medium transmittance to reconstruct the depth of the scene, whereas Retinex theory can isolate the component of illumination and reflection to fix color distortion. R-CNN and SSD detection networks are faster and find targets using region candidates and multi-scale feature pyramids, but are computationally complex, requiring them to be not only fast but also efficient in real-time. YOLO algorithms are based on a single-stage detection architecture and are designed to be faster, as well as the attention mechanism makes the network pay more attention to important areas in order to

reduce the interference of the background. With regard to time modeling, LSTM networks are able to store long-term dependencies with the help of gating units, which is why they are appropriate to work with continuous gesture sequences. In this paper, image enhancement, target detection, and temporal recognition are combined to create an end-to-end system and develops a protocol of communication to realize the automatic conversion of gestures into diving commands.

II. RELATED WORK

The enhancement technologies of underwater images have come up with different remedies to the optical attenuation issue. Physical model approaches define light propagation models using underwater imaging principles, and repair degraded images by estimating the medium transmittance and the background light intensity. Dark channel prior algorithms take into account the statistical consistency of low-intensity pixels in at least one color channel in local regions of fog-free images to compute scene depth information to offset light attenuation. White balance algorithms remove shifts in color according to the gray-world assumption or perfect reflection assumption, to remove the tint of blue-green induced by the fast attenuation of the red channel in water. Retinex theory breaks down images into illumination and reflection parts and boosts the features that are important by reducing illumination variations. Data-based approaches use the generative adversarial networks to train the mapping relationship between underwater and clear images, where the discriminator directs the generator to optimize the image quality over time but this approach uses large-scale paired-data to train and is very costly [3-4]. Gesture recognition algorithms have developed since the old techniques to the deep learning. The early techniques divided the regions of hands according to the skin color detection, contours, texture and geometric characteristics to generate the classifiers, whereas those techniques are vulnerable to changes in illumination and complexity of the background. CNNs automatically learn hierarchic representations of features using multi-level nonlinear transformations, eliminating the need to manually design features. Convolutional neural networks produce candidate regions to be classified and regressed, and single-detection networks solve a regression problem by jointly performing localization and classification [5-6]. Attention mechanisms point out important information, and reduce redundant features by distributing weights adaptively, channel attention measures the significance of various feature maps, and spatial attention finds salient areas. Recurrent neural networks, in terms of temporal modeling, handle sequential data to learn temporal interactions and gating mechanisms address the gradient vanishing problem and long-term memory, which make them the ideal choice in identifying continuous dynamic gestures. Three-dimensional convolutional networks extract both spatial and temporal features, but have a large number of parameters and high computational cost.

III. METHODS

A. SYSTEM OVERALL ARCHITECTURE DESIGN

To implement the real-time recognition and communication conversion of diver gestures, a four-layer progressive architecture is followed: The image acquisition layer reads video streams with an underwater camera. The sensor is set up with 1920×1080 resolution, 30fps frame rate and 8mm lens focus to address the requirements of underwater wide-angle. The preprocessing layer solves the issue of underwater optical attenuation, implementing dark channel prior algorithm to estimate the medium transmittance, calculating the minimum value of the local dark channel map to estimate the atmospheric light intensity, guided filtering to optimise the distribution of transmittance to avoid halo effect, and white balance algorithm to eliminate the attenuation of the red channel by the grey world hypothesis. The enhanced image is fed to a better YOLOv5 network through the detection layer. Multi-scale features are extracted by the backbone network CSPDarknet53 and a CBAM attention mechanism will be added following the C3 module. The channel attention branch produces feature weights by using global average and max pools and the spatial attention branch concentrates on the gesture area to reduce the background noise [7-8]. ResNet50 is employed in the recognition layer to obtain gesture feature vectors. Specifically, the network takes the localized hand regions (bounding boxes) generated by the improved YOLOv5 detector as input, ensuring that feature extraction and temporal modeling are focused solely on the gesture area to minimize computational overhead. The output provides 512-dimensional feature codes to a two-layer LSTM network. The hidden layer has 256 neurons, and the time window is 16 frames to get dynamic gesture trajectories. The communication layer establishes a mapping table between recognized gestures and corresponding diving commands, including 12 standard underwater signals: Breathe, Danger, Down, Ears not clearing, Get with your buddy, Go up, How much air do you have, Low on air, Okay, Out of air, Something's wrong, and Stop. The recognition results are converted into communication commands and transmitted to the surface monitoring station through a wireless acoustic modem.

B. UNDERWATER IMAGE PREPROCESSING AND ENHANCEMENT

Underwater image preprocessing uses a multi-level enhancement method to eliminate the influence of optical degradation. Considering the wavelength-dependent attenuation in water, the dark channel prior (DCP) algorithm is adapted to the underwater optical imaging model (UDCP). Unlike atmospheric haze, underwater degradation is dominated by red-light absorption; therefore, the DCP is utilized here primarily to estimate the medium transmission of the blue and green channels to reconstruct scene depth. Dark channel prior algorithm partitions input image into local blocks of size 15×15 pixels, and computes the minimum value of the R, G and B three channels in each block to build

dark channel map. Global atmospheric light intensity is statistically analyzed from the top 0.1% brightness values of the dark channel map, and attenuation of light in water is estimated using medium transmittance model. Guide filter window radius is set as 60 pixels, and regularization parameter is set as 0.001 to smooth the edge of coarse transmittance map. White balance correction uses gray-world assumption, and mean value of red, green and blue channels are calculated respectively. Gain coefficient of red channel is multiplied by 1.8 to compensate long wave attenuation, and gain of blue channel is multiplied by 0.9 to suppress scattering [9-10]. Table 1 shows enhancement effect of different enhancement algorithms on underwater image quality. Our method achieves best performance in PSNR and SSIM, and information entropy reaches 7.42 bits, which means enough details are recovered.

TABLE 1. Performance Comparison of Underwater Image Enhancement Algorithms

Enhancement Algorithm	Peak signal-to-noise ratio (dB)	Structural similarity	Information entropy (bit)	Processing time (ms)
Original image	18.32	0.621	5.87	-
Histogram Equilibrium	21.45	0.704	6.53	12
Dark passage prior	24.68	0.812	7.09	35
White balance correction	22.91	0.758	6.88	8
This article's method	26.83	0.879	7.42	12

C. GESTURE DETECTION BASED ON IMPROVED YOLOV5

The backbone network CSPDarknet53 contains five convolutional layers and three C3 layers. The kernel size of convolution gradually increases from 3×3 to 9×9. The receptive field is extended. Feature Pyramid Network (FPN)uses top-down way to fuse features of different scales. The output layer has three detectors:80×80 and 40×40 and 20×20 to accommodate the near-range,middle-range and far-range gestures. CBAM attention module is applied in the backend of C3 structure. Channel attention branch performs average pooling and max pooling on 512-dimensional feature map. The two outputs are fed into shared multilayer perceptron to output vector of weight, and the output of Sigmoid is clipped to the range of 0-1. Spatial attention branch performs summation on feature along channel dimension. 7×7 convolution produces the spatial weight map. The hand region is enhanced, and the water flow disturbance is suppressed. To address the challenge of backscatter and moving debris in murky port waters, the system employs a two-stage filtering mechanism: the YOLOv5 detector uses a high confidence threshold to filter out non-hand-shaped floating objects, while the LSTM temporal model distinguishes purposeful gesture sequences from the stochastic motion of suspended particles or small marine life. Loss function contains three parts:bounding box regression loss,confidence loss,classification loss. The bounding box uses CIoU(Intersection over Union),which takes into account the center point distance and aspect ratio. The confidence threshold is set to 0.6 to suppress the poor box. The non-maximum suppression-IoU threshold is set to 0.45 to suppress the overlapping detection. Training: Use SGD optimizer,initial learning rate is 0.01,momentum parameter is 0.937,weight decay is 0.0005, batch size is 16,it will converge after 300 iterations.

D. GESTURE RECOGNITION AND COMMUNICATION PROTOCOL BASED ON RESNET50-LSTM

ResNet50 network extracts gesture features in the detection bounding box. The network includes one convolutional layer, four residual modules and one fully connected layer. Firstly, resizing the input image to the size of 224×224 pixels, then convolving the image with convolutional layer of size 7×7 to obtain 64-dimensional feature map as output. The four residual modules output feature vectors with dimensions of 256, 512, 1024 and 2048 sequentially. Global average pooling layer shrinks the two dimensions. Finally, one fully connected layer maps to output 512-dimensional feature vector. A two-layer LSTM network takes 16 consecutive frames of feature sequence as input to build its temporal model. The two hidden layers contain 256 hidden units. The forget gate decides what information to be ignored, the input gate decides what information to be added to the memory, and the output gate generates the current state. At last, a dropout layer to avoid over-fitting drops 50% of the neurons randomly, and a Softmax classifier outputs the probability distribution on 12 classes of gestures.

Communication protocol builds a two-way mapping relationship between hand gestures and diving commands. One hand raised represents one ascent command (coded as 0x01). Two hands pressed down represents one descent command (coded as 0x02). Clenched fist represents one stop command (coded as 0x03). Thumb pointing downwards represents danger command (coded as 0x04). Other dynamic gestures are recognized by trajectory features. Horizontal wave for three consecutive frames appears as directional guidance (coded as 0x08). Vertical wave appears as distress signal (coded as 0x09).

Acoustic modem adopts Frequency Shift Keying (FSK) modulation method, and the carrier frequency is 25kHz and data rate is 1200bps. Hamming code is used in channel coding for data error correction. Bit error rate is controlled below 0.001, and the communication reliability is guaranteed.

IV. RESULTS AND DISCUSSION

A. DATASET CONSTRUCTION AND EXPERIMENT SETUP

The dataset consists of 8600 photographs of 12 different categories of diving hand signs, including Breathe, Danger, Down, Ears not clearing, Get with your buddy, Go up, How much air do you have, Low on air, Okay, Out of air, Something's wrong, and Stop, which represent common underwater communication commands used by divers. The images were taken in three aquatic contexts namely 3200 images in a clear-water swimming pool, 2800 in nearshore

waters, and 2600 in a turbid harbor, and the water depth of these three locations was 5 to 15 meters and the light conditions included the natural light, artificial light and the low-light scenes. The annotation of the data was carried out with the help of the LabelImg tool to identify the gesture bounding boxes. The labeling process involved three professional divers and two computer vision researchers. To ensure high quality, a double-blind cross-verification method was adopted, achieving an inter-annotator agreement (Cohen's Kappa) of 0.92. The final accuracy was verified to be 98.5% after a manual secondary review. The dataset is partitioned into training, validation and test sets at a ratio of 7:2:1. The experimental platform is set up with an Intel i9-10900K processor with an NVIDIA RTX 3090 with 24GB VRAM and 64GB RAM, Ubuntu 20.04, PyTorch 1.12 deep learning framework, and CUDA 11.6 to perform accelerated computations. To compare, the performance of Faster R-CNN and SSD were chosen. Faster R-CNN has a ResNet101 backbone network with 8 anchor box sizes whereas SSD has a VGG16 base network with 6 layers of features. The two algorithms have been trained on the same set of data to 300 rounds.

B. IDENTIFICATION PERFORMANCE ANALYSIS UNDER DIFFERENT AQUATIC ENVIRONMENTS

The accuracy of different gesture recognition methods was evaluated with the help of the confusion matrix. Precision and recall were determined by enumerating the true positive, false positive, true negative and false negative samples. The nearshore sea area was a little turbid, and visibility was 6 to 8 meters, in a clear water swimming pool setting, where the visibility was over 10 meters with uniform and stable lighting, and where light scattering occurred due to suspended particulate matter. The water in port waters was turbid with a 3 to 5 meter visibility and organic matter and sediment strongly affected the image quality. Table 2 shows the recognition performance of various gestures under different water environments.

TABLE 2. Gesture recognition performance in different water environments

Performance indicators	Clear water swimming pool	nearshore waters	murky port
Visibility (m)	> 10	6-8	3-5
Accuracy (%)	96.3	92.4	89.7
Accuracy (%)	96.3	92.7	90.1
Recall rate (%)	95.8	91.9	87.8
F1 score (%)	96.0	92.3	88.9
Processing speed (fps)	35	33	31

In clean water environments, the system achieved an average precision of 96.3%, a recall of 95.8%, and an F1 score of 96.0%, with a stable processing speed of 35 frames per second. Although the precision decreased in turbid water, the F1 score remained at 88.9%, demonstrating that the preprocessing algorithm effectively compensated for water degradation, and the CBAM attention mechanism enhanced feature robustness. The system still possesses practical value in complex environments.

C. SYSTEM ROBUSTNESS AND REAL-TIME PERFORMANCE EVALUATION

The robustness testing was used to simulate a number of conditions of interference to assess the stability of the systems. There were four conditions in the experiment, namely, lighting changes, disruption of water flow, interference of occlusion, and distance of the target. The intensity of the underwater supplemental lighting (100 lux) was modified to 20 lux to produce lighting changes. A thruster was used to create a flow velocity of 0.5 to 1.5 m/s which was used to disturb the flow of water. Aquatic plants, suspended objects, or other limbs of the divers were obstructing 10-40 percent of the gesture area and caused occlusion interference. The range of targets was 2 to 8 meters. The time taken by every module, i.e., image preprocessing, gesture detection, feature extraction, and sequence recognition, was logged in real-time and statistically analyzed changes in end-to-end latency and frame rate. Table 3 shows the system's performance under different interference conditions.

TABLE 3. System robustness and real-time performance under different interference conditions

Interference type	Interference intensity	Accuracy (%)	Average latency (ms)	Frame rate (fps)	F1 score (%)
Standard Environment	-	96.3	28.6	35	96.0
Light changes	20 lux	91.8	30.2	33	91.5
water flow disturbance	1.5m/s	88.4	31.5	32	88.1
Obstruction and interference	40% occlusion	82.6	29.8	34	81.9
Target distance	8 meters	87.9	32.1	31	87.3

When illumination decreased from 100 lux to 20 lux, accuracy dropped from 96.3% to 91.8%. Image enhancement algorithms compensated for the degradation caused by low illumination, maintaining a high recognition rate. Water flow disturbance caused blurring of gesture edges, and accuracy dropped to 88.4% at a flow rate of 1.5 m/s. LSTM temporal modeling smoothed inter-frame jitter and improved robustness. Even with 40% occlusion, accuracy still reached 82.6%. CBAM spatial attention focused on the unoccluded features, maintaining discriminative ability. The system had an average end-to-end latency of 38.6 milliseconds and a stable frame rate of around 26 fps, meeting the requirements for real-time underwater communication, demonstrating that algorithm optimization effectively balanced accuracy and computational efficiency.

V. CONCLUSION

This paper uses a dark channel prior algorithm and white balance algorithm to recover the quality of images to counter the challenges of gesture recognition due to underwater optical degradation and dynamic background interference. A better YOLOv5 network with built-in CBAM attention mechanism provides a correct gesture localization. An acoustic modem is used to complete the transmission of commands to a ResNet50-LSTM architecture that creates

a temporal feature model to identify sequence of dynamic gestures and it is used to create a complete communication channel. The system shows excellent robustness and real-time functionality in diverse aquatic conditions and interference, confirming the suitability of deep learning models to underwater vision tasks. Nevertheless, the existing technique still requires enhancement in its flexibility to very turbid water and conditions with a high changing of lighting, and the variety of categories of gestures which should be developed to cover more professional diving movements remains to be studied. Future directions will bring multimodal sensors to combine depth and inertial data, study lightweight network models to run on a portable underwater system, create multi-person collaborative gesture recognition algorithms to aid team operation cases, and encourage the viable use of intelligent underwater communication systems.

AUTHOR CONTRIBUTIONS

Conceptualization – Min Tao and Wenpei Liu; methodology – Min Tao and Wenpei Liu; investigation – Min Tao and Wenpei Liu; writing-original draft preparation – Min Tao and Wenpei Liu. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by, 2025 Higher Education Teaching Reform Project of the Guangdong Provincial Department of Education: “Research on the Pathways of Diving Talent Cultivation and Curriculum System Innovation from the Perspective of Industry-Education Integration”

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] Mangalvedhekar S, Nahar S, Maskare S, et al., “Interpretable Underwater Diver Gesture Recognition,” IEEE, 2023, doi: 10.1109/ELEXCOM58812.2023.10370123.

[2] Antillon D W O, Walker C R, Rosset S, et al., “Glove-Based Hand Gesture Recognition for Diver Communication,” IEEE transactions on neural networks and learning systems, 2022, doi: 10.1109/TNNLS.2022.3161682.

[3] Walker C R, Na U, Antillon D W O, et al., “Diver-Robot Communication Glove Using Sensor-Based Gesture Recognition,” IEEE Journal of Oceanic Engineering, vol. 48, no. 3, pp. 778-788, 2023, doi: 10.1109/joe.2023.3265634.

[4] I. Kvasic, Nad, N. Miskovic, et al., “Diver-robot communication dataset for underwater hand gesture recognition,” Computer networks, no. May, pp. 245, 2024, doi: 10.1016/j.comnet.2024.110392.

[5] Joshi R, Sattar J, “One-Shot Gesture Recognition for Underwater Diver-To-Robot Communication,” 2025, doi: 10.1109/IROS60139.2025.11247305.

[6] Jiang Y, Zhao M, Wang C, et al., “Diver’s hand gesture recognition and segmentation for human-robot interaction on AUV,” Signal Image and Video Processing, 2021(5), doi: 10.1007/s11760-021-01930-5.

[7] Caiti A, A. Munafò, Petroccia R, “Underwater Communication,” Encyclopedia of Robotics, 2020, doi: 10.1007/978-3-642-41610-1_14-1.

[8] Maideen A J A, Achuthan B, Arunkumar B, et al., “Underwater Communication Using Li-Fi,” 2021, doi: 10.1109/ICSPC51351.2021.9451725.

[9] Chauhan D S, Kaur G, Kumar D, “Green Underwater Communication Systems,” Green Communication Technologies for, Future Networks, 2022:115-131, doi: 10.1201/9781003264477-7.

[10] Yall S T, Kavya W, “Cloud-IoT in Underwater Communication,” Integration of Cloud Computing and IoT, 2024:247-273, doi: 10.1201/9781032656694-14.