Journal of Imaging Science and Technology® 65(4): 040403-1–040403-10, 2021.
© Society for Imaging Science and Technology 2021

Automatic Fish Segmentation and Recognition in Taiwan Fish Market using Deep Learning Techniques

Ching-Han Chen and Lu-Hsuan Chen
Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan, R.O.C.

Ching-Yi Chen
Department of Information and Telecommunications Engineering, Ming Chuan University, Taoyuan, Taiwan, R.O.C.
E-mail: chingyi@mail.mcu.edu.tw

Abstract. *Taiwan fish markets sell a wide variety of fish, and laypeople may have difficulty recognizing the fish species. The identification of fish species is still mostly based on illustrated handbooks, which is time-consuming when users lack experience. Automatic segmentation and recognition of fish images are important for the field of oceanography. However, in fish markets, the instability of light sources and changes in illumination influence the brightness and colors of fish. Moreover, fish markets often arrange fish together and cover them with ice to keep them fresh, thus increasing the difficulty of automatic fish recognition. This study presents a fish recognition system that combines a state-of-art instance segmentation method along with ResNet-based classification. An input image is first passed through the fish segmentation model, which crops the image into several images containing specific objects with a plain black background. Then the cropped images are assigned to a class by the fish classification model, which returns the predicted label of each image. A database of real fish images was collected from a fish market to verify the system. The experimental results revealed that the system achieved 85% Top-1 accuracy and 95% Top-5 accuracy on the test data set.* © *2021 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2021.65.4.040403]

1. INTRODUCTION

Conventional fish field guides contain photographs that are shot manually and classified by experts. These images are then printed into books to serve as a reference for academics, fishermen, and others. However, identifying fish using a field guide is laborious, time-consuming, and inconvenient. Following the progress of information technology, computer vision technology has been widely used in image analysis [1, 2], object detection [3, 4], biometrics recognition [5, 6], and other different application fields, where all of them achieved good results. In addition to the above applications, some researchers have used computer vision to develop automatic fish recognition systems [7–16]. These studies have introduced image recognition technologies as a replacement for conventional fish field guides. Such fish recognition systems can help people who are unfamiliar with fish to easily identify various species, thereby substantially promoting the recognition and use of marine resources. However, although automatic fish recognition is critical in oceanography, it remains highly challenging. The fish displayed for sale in fish markets typically have different body sizes and are arranged randomly. Sometimes fish of the same species exhibit considerable differences in terms of body characteristics or morphology. In addition, fish markets in Taiwan operate at different times during the morning, noon, or evening, and the environments of individual fish markets also vary. Therefore, differences in illumination conditions cause variations in the brightness and color of fish in the images. Moreover, vendors in fish markets often stack various fish species in a single location and cover them with chipped ice to ensure freshness. These factors increase the difficulty in developing high-accuracy automatic fish identification systems.

Fish recognition studies have mostly adopted body size, shape, and color texture measurements as features and then used machine learning classifiers to obtain identification results. Lee et al. [7] used curvature analysis to locate critical landmark points. Contour segments of interest were then extracted based on these landmark points. According to the experimental results, shape-based fish identification methods yielded satisfactory classification results. Alsmadi et al. [8] used fish size and shape measurements as the features for backpropagation neural network (BPNN) classifier. Alsmadi et al. in 2010 [9] used color texture measurements to distinguish target objects from the background; then, the color texture characteristics were applied to a BPNN for identification. Sharmin et al. [10] used four types of characteristics, and conducted principal component analysis for dimensionality reduction. A support vector machine (SVM) was employed as the classifier. This system enabled users to use a handheld mobile device to identify captured freshwater fish. Qin et al. [11] used spare and low-rank matrix decomposition to remove complex backgrounds from fish images. Then, a deep neural network was used to extract the foreground characteristics, and a linear SVM was adopted for classification. In recent years, improvement in the performance of central processing units and the development of graphics processing units (GPU) have increased the feasibility of real-time deep learning with high complexity and time-consuming computation. Krizhevsky et al. proposed AlexNet [17], which achieved exceptional

Received Dec. 20, 2020; accepted for publication May 6, 2021; published online May 27, 2021. Associate Editor: Jia-Shing Sheu.
1062-3701/2021/65(4)/040403/10/$25.00

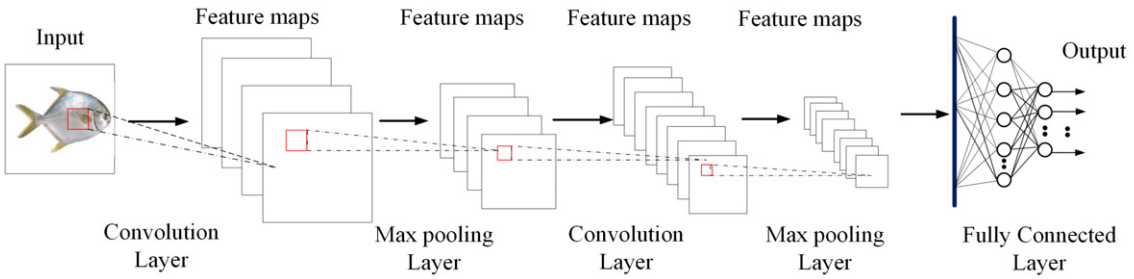


Figure 1. Typical CNN framework.

performance in the ImageNet large scale visual recognition competition (ILSVRC). This not only demonstrates that the convolutional neural network (CNN) can yield superior performance compared with conventional machine learning but also inspired many researchers to employ CNN for image recognition. Moreover, many researchers have applied CNN characteristics to detect objects and used bounding boxes to locate and annotate specific objects in the image. Refs. [12–14] have used CNNs to identify fish. Li et al. [15] used Fast region-based CNN method (Fast R-CNN) and a CNN to detect and recognize fish in underwater images with favorable results. Deep et al. [16] proposed a hybrid CNN framework, in which a CNN is used for feature extraction, and SVM and k-nearest neighbor (k-NN) classifiers are applied for classification.

After the publication of AlexNet, CNN models such as VGGNet, GoogLeNet [18], and Residual CNN (ResNet) [19] have improved in terms of accuracy and performance. The ResNet framework includes a shortcut connection mechanism to enable the eigenvalues of shallow layers to be connected with deeper layers through shortcuts. Batch normalization [20] has been widely applied to solve the degradation problem of neural networks in deep layers, with results exhibiting superior performance in image classification. In addition, CNNs can be used to locate the position of an object in an image. However, to extract additional object information from images, the morphology and appearance of objects must be further processed. That is, object segmentation methods must be incorporated to separate objects from the background. By integrating object detection and semantic segmentation techniques, the Mask region-based CNN (Mask R-CNN) [21] method can effectively detect targets in images, generate a segmentation mask for each instance, and complete object detection and shape segmentation tasks simultaneously.

Because the collection of fish image data is difficult, in this study, we developed a ResNet-based automatic fish recognition system to yield robust and high-accuracy recognition based on a limited training data set with minimal data labeling. This study included Mask R-CNN object segmentation and ResNet-based image classification modules, and early stopping and data augmentation mechanisms were applied to improve system operational performance for the test data set. First, the Mask R-CNN extracts fish objects belonging to the foreground. Then, the obtained objects are input to the ResNet-based image classifier, which assigns a category to each object.

2. RELATED APPROACHES

2.1 *Deep CNN Model*

The CNN is a mainstream deep neural network that possesses the features of local connectivity and parameter sharing. It has demonstrated extremely effective feature extraction for images. The convolutional layers, which are the core of the CNN, normally comprise tens to hundreds of $n \times n$ filters. Each filter performs a function of feature extraction and enhances specific image patterns. In a CNN, convolution layer filters are applied to generate feature maps. The convolution layers are generally followed by pooling layers to downsample the feature maps. In the CNN applied for image recognition, input data passes through one to three convolutional layers and a pooling layer followed by two or more fully connected layers, which yield the prediction result. Figure 1 presents a typical CNN framework. The input data are forward propagated layer by layer. The errors between the target and predicted output values are then calculated, and the backpropagation algorithm is used to update the network weights.

2.2 *Residual CNN Model*

The CNNs can be roughly classified into various typical frameworks, such as LeNet-5, AlexNet, VGGNet, GoogLeNet, and ResNet, according to the network pattern. For the VGGNet and GoogLeNet frameworks, sufficient depth is critical for excellent model performance. However, with increasing network depth, accuracy may begin to plateau or even decrease; this is referred to as the network degradation problem. ResNet uses identity mapping to increase the network depth and address this degradation problem [20]. For a residual block, if the input is x and $H(x)$ is the residual block output, then the residual function can be defined as $F(x) = H(x) - x$, as shown in Figure 2. When the residual is 0 (i.e., $F(x) = 0$), then the residual block yields the identity mapping (i.e., $H(x) = x$). Under this circumstance, the number of identity mapping blocks is increased, and the performance is not reduced. By contrast, under other circumstances, the deep neural network may yield superior performance because of the residual blocks.

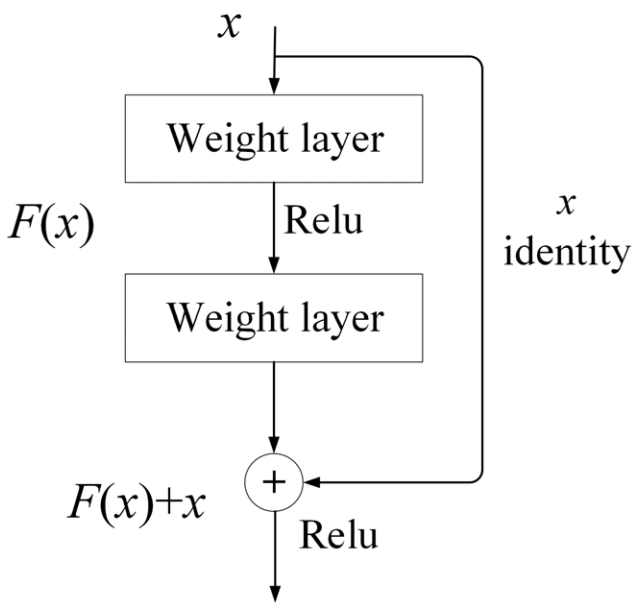


Figure 2. Basic residual block.

2.3 *Mask R-CNN*

Faster region-based CNN (Faster R-CNN) is a two-stage object detection framework. Rather than using a selective search algorithm to extract region proposals, Faster R-CNN uses a region proposal network (RPN) to generate region proposals and then detect objects in each of the proposed regions. This approach considerably increases the processing speed. Mask R-CNN [21] is a state-of-the-art model for instance segmentation. The model inherits and improves on the Faster R-CNN framework. In Mask R-CNN, the region of interest (ROI) pooling layer is replaced with the more accurate ROI alignment layer to solve the image shift phenomenons observed in ROI pooling when executing max pooling. The ROI alignment layer connects an additional branch to output pixel-wise masks of each object in the image, which accurately correspond to the object shapes, as shown in Figure 3.

3. DESIGNING THE FISH RECOGNITION SYSTEM

The fish recognition system introduced in this study uses an instance segmentation method to segment the fish objects in the input images. Then, the images are cropped into fish images with the background removed. After obtaining these cropped images, the fish classification model assigns the corresponding category.

The fish recognition system developed in this study comprised two modules: the fish segmentation module and the fish classification module. IDEF0, from the Integrated Computer-Aided Manufacturing definition (ICAM) [22], was used to establish a hierarchical and modular design for the fish recognition system. Structural analyses were conducted to illustrate the relationships between the functional modules in the system. Figure 4 illustrates the IDEF0 structure of the proposed fish recognition system; every block in the figure represents a functional module in the system architecture. A0, the main controller in the uppermost layer of the system, collected fish images and output their category. The A0 module comprised two functional submodules, namely A1 and A2.

3.1 *Fish Segmentation Module*

Figure 5 presents the Grafcet discrete-event model [23] of the fish segmentation module. RPN network was used to generate region proposals for the input images, and the ROI alignment layers gererate pixel-wise masks corresponding to fish shapes. These fish shape images were cropped to facilitate processing in the subsequent recognition steps.

To realize the aforementioned fish segmentation module, a data set with sufficiently many fish images were required to train a Mask R-CNN model to accurately segment the fish shapes. In the present study, the VGG image annotator (VIA) tool was employed to complete image annotation and labeling for the training data.

3.1.1 *Using the VIA Tool to Annotate Fish Images*

The instance segmentation algorithm calculates the pixel-wise segmentation mask for all objects in the image. It not only locates independent objects, but also predicts their

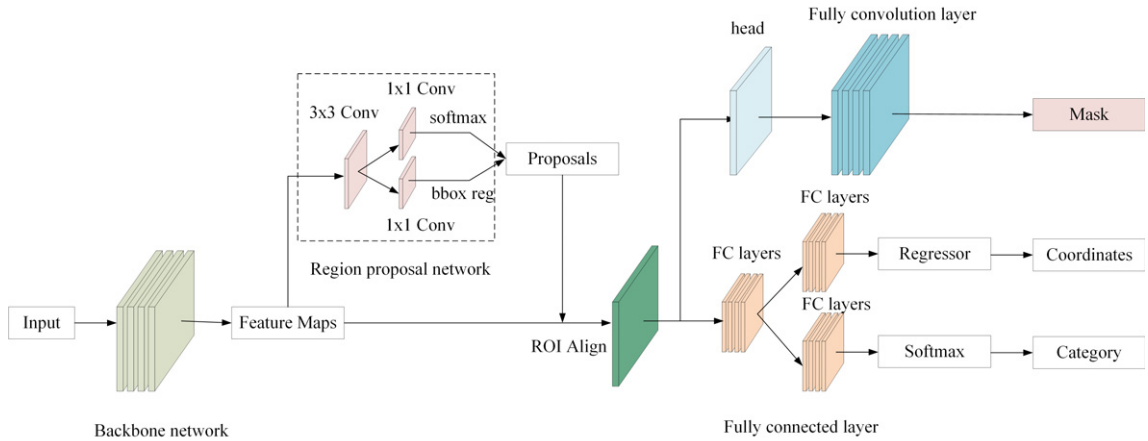


Figure 3. The framework of Mask R-CNN [21].

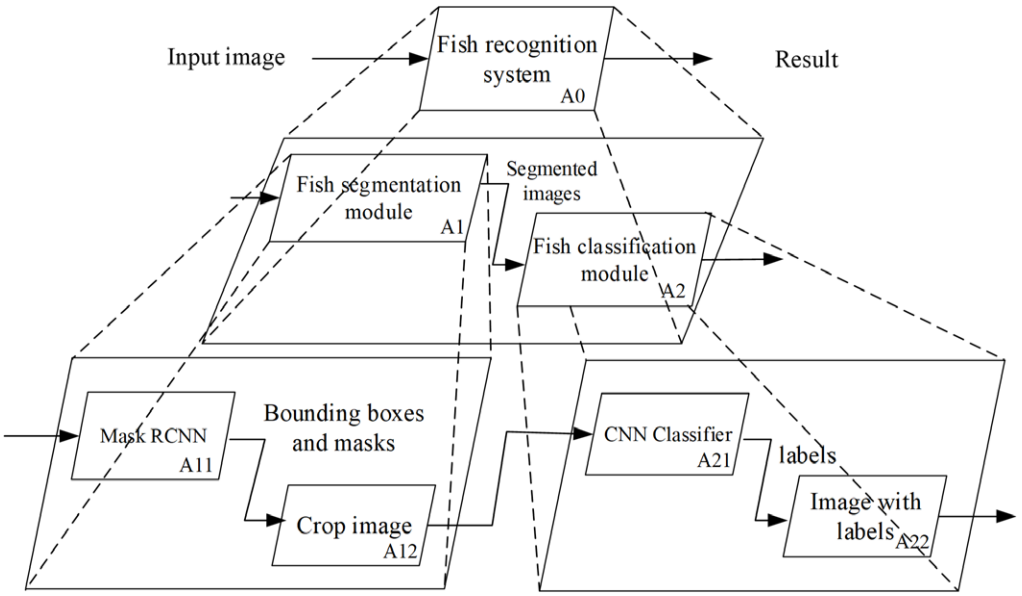


Figure 4. IDEF0 structure of the proposed system.

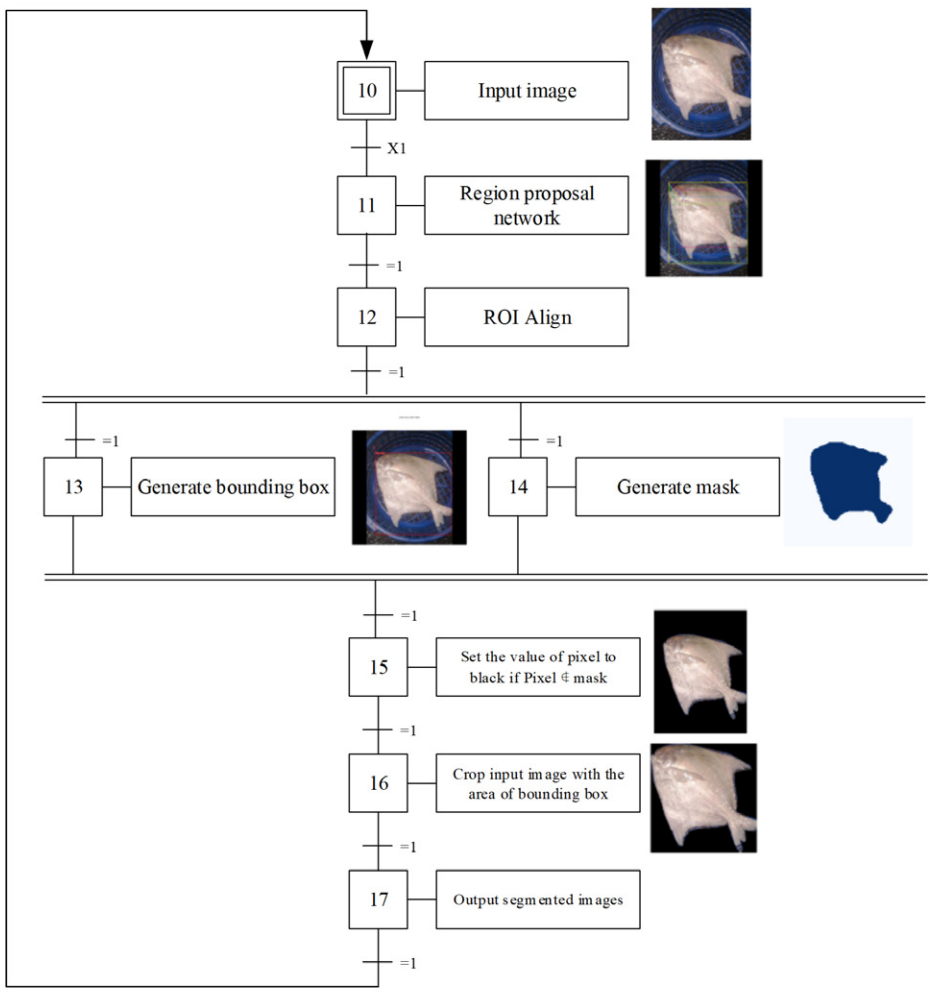


Figure 5. The Grafcet discrete-event model of the fish segmentation module.

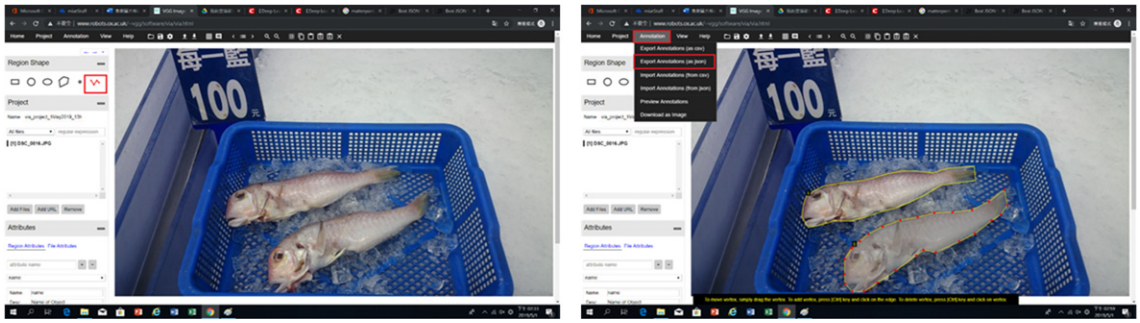


Figure 6. Using the VIA tool to annotate fish images.



Figure 7. Photograph of a box of fish taken at a fish market.

boundaries. Therefore, before training the segmentation model, the independent objects in the training images must be annotated and masks must be defined for these objects. The ROI corresponding to each mask is the collection of pixels that display the object. The bounding box is the smallest rectangular box containing the mask. Figure 6 shows the processing results for annotating fish images using the VIA tool.

3.1.2 *Annotation Protocols for the Fish Segmentation Data Set*
An entire fish image includes characteristics such as the eye, pectoral fin, dorsal fin, and caudal fin. However, in the fish markets, fish are often stacked over one another or covered by ice or other objects, which hinders the presentation of the entire fish body. We annotated regions containing at least two of the aforementioned characteristics. In Figure 7, the fish body in the red box included the expected characteristics; however, a portion of the fish body was covered by the caudal fin of another fish. For the fish in the blue box, only the eye and pectoral fin are visible. The results of the annotated regions are shown in Figures 8(a) and 8(b).

3.2 *Fish Classification Module*
Cropped fish body images are the inputs for the fish classification module; an image first passes through the convolutional layers for feature extraction, and these features are fed into the fully connected layers to yield the final classification. Figure 9 presents the Grafcet model of the fish classification module.

In this study, ResNet-50 was adopted for the classifier (Fig. 9). This deep neural network comprises the input, convolutional, and output layers. The model includes five convolutional layers, namely Conv1, Conv2, Conv3, Conv4, and Conv5. The specifications are shown in Table I.

Table I. Modified ResNet-50 architecture.

Layer name	Output size	50-layer
Input	224 × 224	—
Conv1	112 × 112	7 × 7, 64, stride 2
Conv2	56 × 56	3 × 3 max pool, stride 2
—	—	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
Conv3	28 × 28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$
Conv4	14 × 14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$
Conv5	7 × 7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$
Softmax (output)	1 × 1	41

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 *Fish Image Database*
4.1.1 *Data Set for the Fish Segmentation Module*
This study used 811 images of fish for sale in a fish market to train the fish segmentation module. Among these images, 685 were used as the training data set and the remaining 126 were used as the test data set. The four images shown in Figure 10 are part of the sample.



Figure 8. Segmentation for overlapping fish bodies: (a) upper fish body and (b) lower fish body.

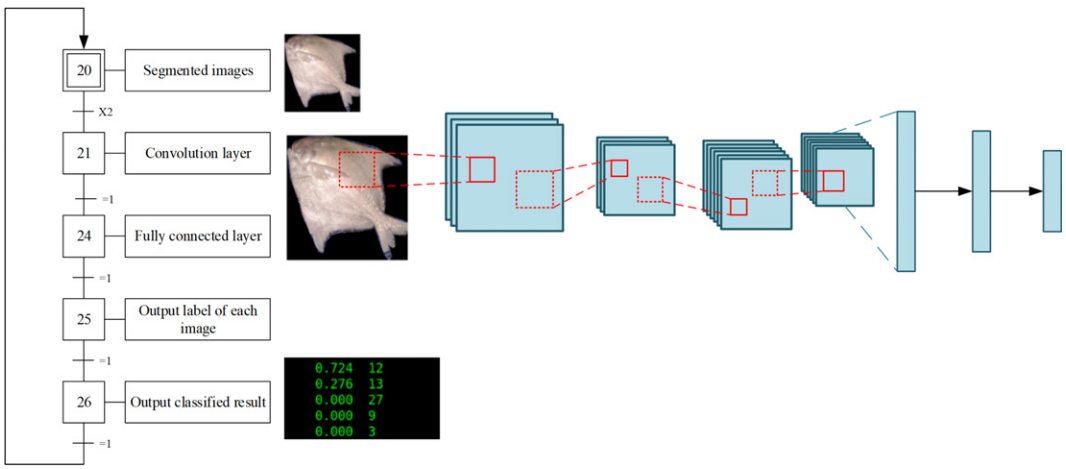
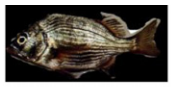
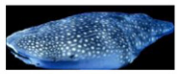
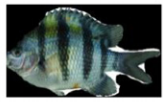
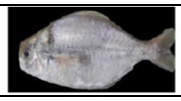
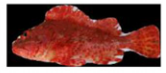
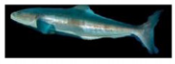
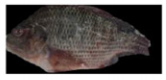
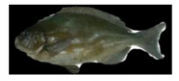
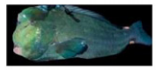
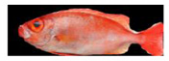
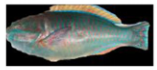
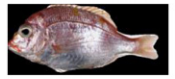


Figure 9. Grafcet discrete-event model of the fish classification module.



Figure 10. Examples of images used for training the fish segmentation module.

Table II. Example images for each species of fish in the data set.

1		11		21		31	
2		12		22		32	
3		13		23		33	
4		14		24		34	
5		15		25		35	
6		16		26		36	
7		17		27		37	
8		18		28		38	
9		19		29		39	
10		20		30		40	
						41	

4.1.2 *Data Set for the Fish Classification Module*

The data set used for training the fish classification module contained 1025 images. For each class of images in the data set, we selected 20 of these images as the training data set and

Table III. Training hyperparameters for the Mask R-CNN model.

Backbone architecture	ResNet-101
Epochs	500
Number of GPU	1
Image per GPU (batch size)	2
Learning rate	0.0001
Pre-trained weights	MS COCO pre-trained weights

Table IV. Training configuration for the ResNet-50 classifier.

Loss function	Categorical cross-entropy
Optimizer	Adam
Learning rate	0.00001
Batch size	8
Input image size	224 × 224
Pre-trained weights	ImageNet pre-trained weights
Dropout percentage	50%
Number of freezing layers	2
Stop criteria of early stopping	Stop training when validation loss has stopped improving after *100 epochs*

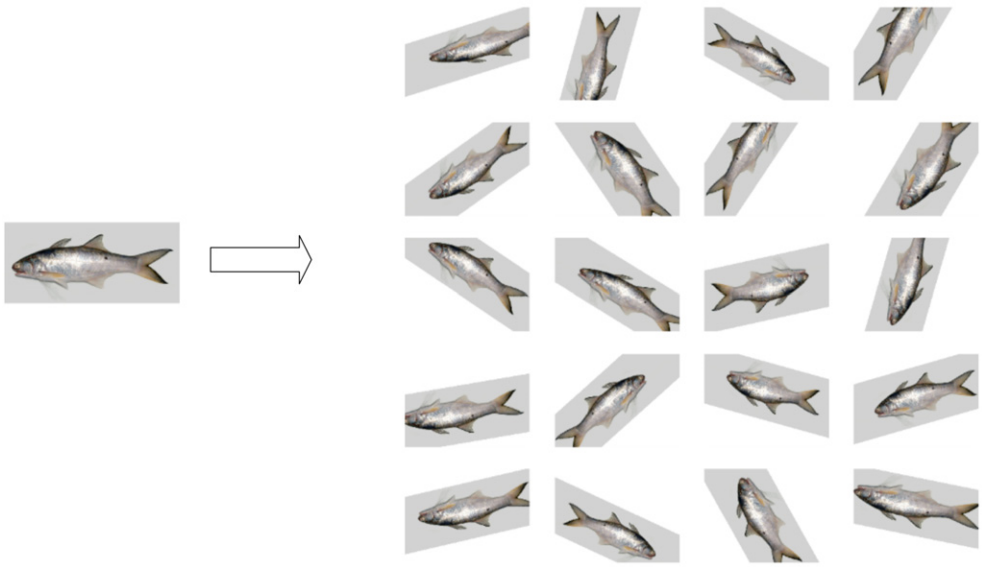


Figure 11. Data augmentation for fish images.

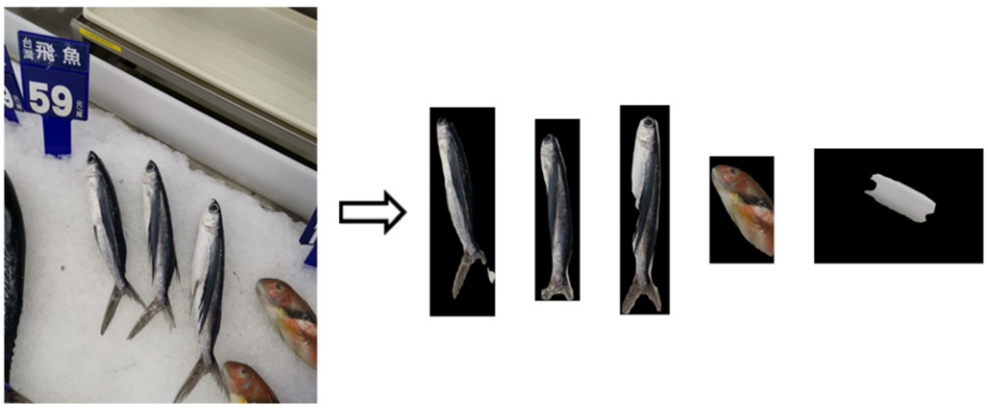


Figure 12. Example of image segmentation. The system output four complete and one incomplete region.

Table V. Classification accuracy.

Accuracy type	Testing accuracy
Top-1	85%
Top-5	95%

another five images as the test data set. This data set featured a total of 41 different species of fish.

For the fish recognition system developed in this study, the background of the input image is first removed to extract the fish body shape. Then, the fish body images are input into the fish classification module for classification. Therefore, the fish images used to train the fish classification module were all cropped using the fish segmentation module, and the fish bodies were separated from the backgrounds, as shown in Table II. Table II presents one image for each of the 41 categories in the data set.

Collecting sufficient training data set is essential for deep network training. To generate more labeled fish images, we adopted data augmentation methods including rotating, cropping, and flipping, as shown in Figure 11.

4.2 *Model Hyperparameters*

4.2.1 *Hyperparameters for the Mask R-CNN model*

The training hyperparameters have a substantial influence on the performance of Mask R-CNN. Table III shows the settings used for training Mask R-CNN in this study. ResNet-101 was selected as the backbone architecture for Mask R-CNN, and pre-trained weights of the MS COCO data set [24] were adopted as the initial parameters.

4.2.2 *Hyperparameters for the Fish Classification Architecture*

The fish image classification module was trained based on ResNet-50. The relevant parameter settings are shown in Table IV. Deep learning may result in overfitting due to an uneven distribution of samples between categories during training, thereby reducing the generalizability of the

Table VI. Examples of correctly classified fish.

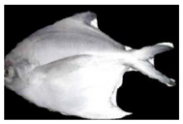
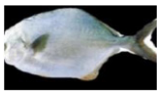
Input image	Output result	Scientific name
	Class 11	Eleutheronema tetradactylum
	Class 13	Pampus argenteus
	Class 25	Oreochromis niloticus

Table VII. Examples of misclassified fish.

Input image	Predicted result	Ground truth result
	Class 36	Class 11
	Class 6	Class 13
	Class 2	Class 25

model. Therefore, during classifier training, early stopping was applied to mitigate overfitting. When the validation loss ceased decreasing with the number of iterations, the model training was stopped.

4.3 *Performance Analysis of the Fish Recognition System*

The fish image-segmentation module achieved a mean average precision (mAP) of 78 for an intersection of union (IoU) of 0.5 with the test data set under the VOC 2012 measurement standard. Figure 12 shows an example of image segmentation, in which the entire and incomplete regions are output.

Table V lists the classification results for the test data set; the developed system yielded accuracy rates of 85% and 95% for Top-1 and Top-5, respectively. Figure 13 represents the confusion matrix for Top-1 accuracy with each class of fish.

Table VI shows examples of correct classification results. The cropped images retained the crucial features of the fish bodies. Table VII shows examples of incorrect classifications. The failure to retain the features of the fish body during cropping may have caused misclassification; moreover, fish species with similar shapes or colors might have been misclassified.

5. CONCLUSION

Fish are essential components of daily diets in Taiwan. However, because many species of fish are available in

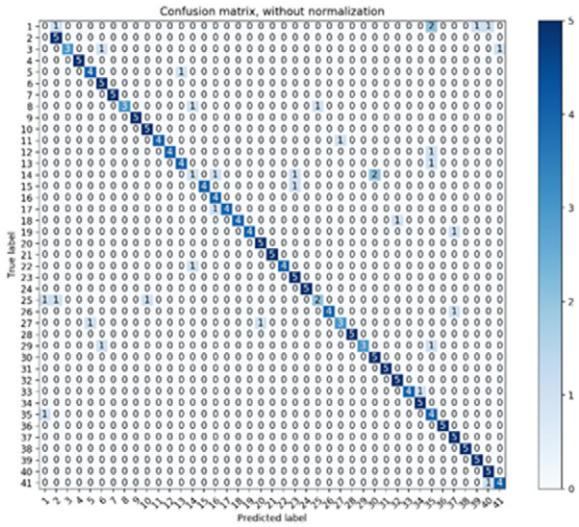


Figure 13. Confusion matrix for Top-1 accuracy.

fish markets, the identification of fish is difficult for nonprofessionals. In addition, searching fish field guides to identify fish is inconvenient. To provide users with a real-time and convenient method to identify fish species commonly seen in fish markets, this study established a fish image recognition system based on the ResNet model. The proposed system cropped the input fish photographs to obtain individual fish bodies and separated the foreground and background. This processing facilitated the subsequent classification of fish species. For the test data set, the proposed system achieved a Top-1 accuracy of 85% and a Top-5 accuracy of 95%. The system can satisfy the demand of Taiwanese people for identifying fish species in fish markets.

REFERENCES

[1] C. H. Hsia, S. H. Lin, and Y. Y. Chen, "Adaptive wavelet shrinkage based on intelligent FIS learned thresholding", J. Imaging Sci. Technol. **63**, 030407-1–030407-10 (2019).

[2] C. H. Hsia, T. Y. Lin, and J. S. Chiang, "An adaptive binarization method for cost-efficient document image system in wavelet domain," J. Imaging Sci. Technol. **64**, 030401-1–030401-14 (2020).

[3] C. H. Hsia, T. C. Wu, and J. S. Chiang, "A new method of moving object detection using adaptive filter," J. Real-Time Image Process. **13**, 311–325 (2017).

[4] C. H. Hsia, Y. J. Liou, and J. S. Chiang, "Directional prediction CamShift algorithm based on adaptive search pattern for moving object tracking," J. Real-Time Image Process. **12**, 183–195 (2016).

[5] C. H. Hsia, "Improved finger-vein pattern method using wavelet-based for real-time personal identification system," J. Imaging Sci. Technol. **62**, 030402-1–030402-8 (2018).

[6] C. H. Hsia, "New verification strategy for finger-vein recognition system," IEEE Sensors J. **18**, 790–797 (2018).

[7] D. J. Lee, S. Redd, R. Schoenberger, X. Xu, and P. Zhan, "An automated fish species classification and migration monitoring system," *Proc. 29th Annu. Conf. on IEEE Ind. Electron. Soc. (IECON)* (IEEE, Piscataway, NJ, 2003), pp. 1080–1085.

[8] M. K. Alsmadi, K. B. Omar, S. A. Noah, and I. Almarashdeh, "Fish recognition based on robust features extraction from size and shape measurements using neural network," J. Comput. Sci. **6**, 1088–1094 (2010).

[9] M. K. Alsmadi, K. B. Omar, S. A. Noah, and I. Almarashdeh, "Fish recognition based on robust features extraction from color texture measurements using back-propagation classifier," J. Theor. Appl. Information Technol. **18**, 11–18 (2010).

[10] I. Sharmin, N. F. Islam, I. Jahan, T. A. Joye, M. R. Rahman, and M. T. Habib, "Machine vision based local fish recognition," SN Appl. Sci. **1**, 1529 (2019).

[11] H. Qin, X. Li, J. Liang, Y. Peng, and C. Zhang, "Deepfish: Accurate underwater live fish recognition with a deep architecture," Neurocomputing **187**, 49–58 (2016).

[12] G. Ding, Y. Song, J. Guo, C. Feng, G. Li, B. He, and T. Yan, "Fish recognition using convolutional neural network," *OCEANS 2017* (IEEE, Piscataway, NJ, 2017), pp. 1–4.

[13] L. Jin and H. Liang, "Deep learning for underwater image recognition in small sample size situations", *OCEANS 2017* (IEEE, Piscataway, NJ, 2017) pp. 1–4.

[14] M. N. Rachmatullah and I. Supriana, "Low resolution image fish classification using convolutional neural network", *5th Int'l. Conf. on Advanced Informatics: Concept Theory and Applications (ICAICTA)* (IEEE, Piscataway, NJ, 2018) pp. 78–83.

[15] X. Li, M. Shang, H. Qin, and L. Chen, "Fast accurate fish detection and recognition of underwater images with Fast R-CNN", *OCEANS 2015* (IEEE, Piscataway, NJ, 2015) pp. 1-5.

[16] B. V. Deep and R. Dash, "Underwater fish species recognition using deep learning techniques", *6th Int'l. Conf. on Signal Processing and Integrated Networks (SPIN)* (IEEE, Piscataway, NJ, 2019) pp. 665–669.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM **60**, 84–90 (2017).

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", arXiv:1409.4842 (2014).

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv:1512.03385 (2015).

[20] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift", arXiv:1502.03167 (2015).

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," arXiv:1703.06870v3 (2018).

[22] R. J. Mayer, IDEF0 Function Modeling, Air Force Systems Command (1992).

[23] R. David, "Grafcet: A powerful tool for specification of logic controllers," IEEE Trans. Control Syst. Technol. **3**, 253–268 (1995).

[24] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," arXiv:1405.0312 (2014).