IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 28, 2020, accepted June 30, 2020, date of publication July 13, 2020, date of current version July 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3008698

Fish Shoals Behavior Detection Based on Convolutional Neural Network and Spatiotemporal Information

FANGFANG HAN, JUNCHAO ZHU, BIN LIU, BAOFENG ZHANG, AND FUHUA XIE

Tianjin Key Laboratory for Control Theory and Applications in Complicated Systems, School of Electrical and Electronic Engineering, Tianjin University of Technology, Tianjin 300384, China

Corresponding authors: Fangfang Han (fangfanghan2004@163.com) and Junchao Zhu (zhujunchao_tjut@163.com)

This work was supported in part by the Major Project of Tianjin Internet Cross Border Integration and Innovation Science and Technology under Grant 18ZXRHSF00240, in part by the Project of Tianjin Science and Technology Plan under Grant 18ZLZNGX00040, and in part by the Project of Tianjin Science and Technology Plan under Grant 18ZLGCGX00080.

ABSTRACT Behavior is the first visible change in an animal species after exposure to its own or environmental stressors and is a sensitive indicator. Fish are social animals, and the abnormality of group behavior is more indicative about a particular event than individual behavior, providing more effective informeqation about environmental or group social changes. The group behavior is not only reflected in the spatial distribution, but also reflected in the temporal behavior of the group and individual movement changes under the influence of pressure factors. This paper proposes a group behavior discrimination method based on convolutional neural network and spatiotemporal information fusion, which intends to make use of the prominent performance of convolutional neural network in image recognition and state classification, and imitating the attentional mechanism of ventral channel and dorsal channel when the human brain processes visual signals. Some pressure environments are made in laboratory, the behavior states of fish shoals are recorded, and the sample database of shoals' behavior state is established by combining the spatial information of shoals' spatial distribution with the time information reflected in the movement behavior. A simple convolutional neural network is constructed to quickly identify the behavior state of fish shoals. The effects of bath size and training epoch on network training speed and recognition accuracy are discussed, and the visualization of the intermediate data of the convolutional neural network is studied. Shown from the results of experiments of this paper, different behavior states of fish shoals can be recognized and classified effectively by using the simple convolutional neural network and spatiotemporal fusion images. What's more, from the visualization of network intermediate data, it is found that the convolutional neural network has a higher discrimination power to the image edge feature than the image gray-value feature.

INDEX TERMS Intelligent agriculture, fish behavior, deep learning, convolutional neural network, spatiotemporal information fusion.

I. INTRODUCTION

Fish belong to the group like, underwater activities, breathing with gills hypothermic animals, and individual of fish is sensitive to changes and disturbances in water quality and surrounding environmental factors, so all these factors determine the particularity of fish breeding [1]. In the process of breeding, fish can not directly touch and arbitrary fishing, can only be judged by observation. On the other hand, behavior is the first visible change after exposure to one's own or environmental stressors, which has been shown to be a sensitive indicator with the most direct mapping of health, water quality, inventory density, physical disturbance and other factors. Group behavior, compared with individual behavior, is more likely to indicate significant information about an event, and to provide more effective information about environmental or group social changes. By observing and studying the group behavior state, its correlation with surrounding environment as well as the health of fish body can be found, thus providing reference value for fish breeding [2], [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Jihwan P. Choi.

In recent years, computer vision technology has made many breakthroughs in fish size measurement, shape analysis, quality estimation, body color analysis, disease diagnosis, identification and classification. Computer vision as a high-precision and contactless technology can monitor the size, quality and status of fish by recording equipment. It can avoid the influence on the living environment, colony habits and individual growth of fish shoals by some other physical and chemical methods, and it became an important monitoring method in aquaculture [4], [5].

Kato (1999) [6] developed a computer image processing system to quantify the behavior of goldfish, opening the way for the detection of fish behavior. Since then, Ishibashi Y (2002) [7] studied on the goldfish group as the research object, using two cameras placed in front of and vertically upper the breeding to record. A uniform coordinate axis was established in all the collected images to obtain the coordinates of the fish body, and to calculate the coordinates and standard deviations of the fish body in three directions of the spatial coordinate system, and study the behavior response of the fish under hypoxic stress (1mg/L). Pinkiewicz (2008) [8] took the Atlantic salmon population as the research object, used the background difference method to detect the fish body, and extracted the central coordinates, peripheral dimensions, eccentricity and other characteristics of the fish body, and quantified these characteristics into the swimming speed and direction, so as to study the relationship between the fish body speed and movement direction with tidal cycle. Sadoul (2014) [9] took rainbow trout as the research object, calculated the dispersion degree of the shoal by calculating the circumference of the shoal, and estimated the swimming behavior of the shoal. Yu Xin (2014) [10] studied the abnormal behavior of zebrafish populations by using the method based on the statistics of shoal movement characteristics. In this study, the velocity and angle of shoals were obtained by optical flow method, and then the abnormal behaviors of shoals were evaluated by two characteristic factors: standard mutu al information and local distance anomaly factor. Zhu Peiru (2015) [11], taking zebrafish as the research object, proposed a detection method based on the extraction of salient features of fish's eyes and head, and designed a fish head detector using statistical methods to track individual fish. Zhao Jian (2016) [12] took shoals of fish as research objects and proposed a method to quantify changes in feeding intensity of shoals by using an improved kinetic energy model. The model was obtained by combining the dispersion degree of shoals with spatial behavior characteristics. The Lucas-Lanade optical flow method was used to determine the velocity and angle of shoals, so as to determine the dispersion degree of shoals.

With the development of computational power and algorithm technology, deep learning provides an efficient method for information mining hidden in massive image (video) data. Artificial neural network (ANN) has a profound impact on the research of information processing in computer vision. Visual information processing most has the characteristics of multiple input and nonlinear. Artificial neural network can form a self-learning and self-adaptive architecture when the theory is not perfect, and form a nonlinear mapping or nonlinear dynamic system in the interaction with external information, so as to correctly reflect the relationship between input and output without knowing the precise mathematical model of this relationship in advance. Especially in recent years, the CNN (CNN, convolution neural network) model has made a breakthrough in solving abstract cognitive problems [13]–[15].

Rathi (2018) [16] took fish and crustaceans from natural beaches and estuaries as research objects, and combined R-CNN with three classification networks (ZFNeT, CNN-M and VGG16) to construct regional prediction method by combining regional prediction network and classifier. Mandal (2018) [17] took Fish4Knowledge data set as the research object, and used 3-layer CNN to classify 21 types tropical fish. Marini (2018) [18] took marine fish stocks as the research object, constructed a k-layer feature selection framework combining image segmentation and cross validation to realize two-classification, and used the collected data to evaluate the richness of fish. Konovalov (2019) [19] designed a CNN based fish detector (based on the Xception CNN) for fish swarm, and used the water image set to realize the underwater fish detection under the supervision of multiple water areas.

At present, most of the existing research methods of fish behavior state detection are image processing methods to classify individuals from the group (by region, edge, background difference, etc.), and identify the group state according to the movement trajectory, direction and speed of each individual. The discrimination of spatial distribution state is also based on the statistical feature method, such as the fish individual perimeter, area, centroid coordinates and calculated fish aggregation, dispersion and other indicators. What's more, now most of the existing research on fish problems based on deep learning model aims at fish species classification and recognition; the existing public image database related to fish is also a single fish image database aiming at fish species recognition. There are few deep learning models aiming at the recognition of fish shoals and behavior state, and there is no public image library with such problems.

This paper proposed using a CNN model to distinguish the behavior of fish shoals, which do not need to extract the statistical features from the image in the early stage, but directly send the fish image of various states to the network model for training, learning and discrimination. Considering the behavior of fish shoals is not only reflected in the spatial distribution, but also reflected in the changes of the population or individual movement with time, this paper combining the outstanding performance of CNN in image recognition and state classification, and imitating the characteristics of attention fusion of ventral channel and dorsal channel when human brain processes visual signals, proposes a fish shoals behavior discrimination method based on CNN and space-time information fusion. In the experiments,

the pressure source environments are created to record the behavior state of the fish group. Combining the spatial information of the fish spatial distribution and the time information of the movement behavior, the sample database is established. A simple CNN is constructed to realize the fast discrimination of fish shoals' behavior. The visualization of intermediate data of CNN and image feature extraction by convolution kernels when the network training is stable are studied, and the influence of space-time information fusion strategy, sample number and training rounds on network training speed and recognition accuracy are also discussed. Shown from the results of experiments, different behavior states of fish shoals can be recognized and classified effectively by using the simple convolutional neural network and spatiotemporal fusion images. This is the novelty and contribution of this paper.

II. ESTABLISHMENT OF EXPERIMENT IMAGE DATABASE

A. EXPERIMENTAL FISH SAMPLES AND PRESSURE ENVIRONMENT CREATION

Zebrafish is one of the standard experimental organisms recommended by ISO. The red zebrafish, which is 6-8 months old and (30 ± 2) mm long, is selected as the research object. The color of the fish is bright and is easy to be photographed by the collection equipment. The fish is with small size, easy to survive, good environmental adaptability, sensitive behavior, and almost all changes in the environment can affect its behavior and is easy to collect experimental data [20].

In the laboratory environment, dechlorinated tap water was used for running water feeding. During normal feeding, the temperature of water body was kept at (26±1) °, dissolved oxygen was (6.5±0.5) mg·L^{-1}mg · L^{-1}, and filtration device was used for water purification. The duration of illumination was 14h every day, and the duration of non-illumination was 10h. From 8:00 a.m. to 22:00 p.m., the white light source was used for illumination, and the purified water was stopped. From 22:00 p.m. to 8:00 p.m. the next day, there was no light environment, and open the filter device to treat sewage. The experiment chooses to collect data under the condition of light. Before the experiment, zebrafish had been living in the breeding environment for several weeks and fully adapted to the environment and were fed regularly. Feed once a day between 15:00-16:00 in the afternoon and feed the eggs of harvest shrimp. The feeding amount was based on the fish group eating in ten minutes.

The following six groups of states were manufactured by experiment and recorded by camera.

Group 1 was the normal state. When the fish were fed normally and regularly, the behavior of the fish in the normal state was recorded by video.

Group 2 was the feeding state. During the normal and regular feeding, the behavior of fish feeding state was recorded by video.

Group 3 was the group stimulated state. Under the condition of normal regular feeding, the large disturbance environment of water surface was artificially created, and the excited state of fish shoals was also created. The behavior of fish shoals in the excited state was recorded by video.

Group 4 was the individual disturbance state. Under the condition of normal regular feeding, the interference of a certain individual in the fish group was produced by touching the slender needle, and the individual interference state was produced. The behavior of individual disturbance state was recorded by video.

Group 5 was the anoxic state. At the beginning of the experiment, the fish were fed regularly for five days. On the sixth day, the amount of feeding was increased, and the oxygenation and purification of water were not carried out, which was recorded until the twentieth day. The behavior of fish shoals in anoxic state was recorded by video.

Group 6 was the starvation state. At the beginning of the experiment, the feeding was stopped for ten days, and the normal feeding was resumed on the eleventh day until the twentieth day. The first ten days were to create starvation environment, and the last ten days were to recover. The behavior of fish shoals in starvation state was recorded by video.

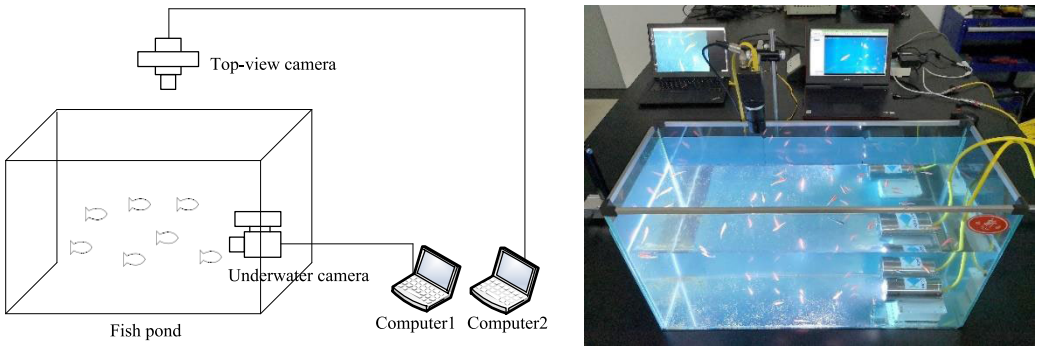
B. FISH SHOALS IMAGE ACQUISITION DEVICE

Because the fish live in the water environment of cube, in order to realize all-round collection of fish behavior, the design of fish image collection system structure scheme is shown in Figure 1-(a).

The top-view camera is installed with the optical axis vertical downward to the horizontal plane, and the behavior of the fish is observed from the top of water surface, in order to survey the horizontal distribution characteristics of the fish. The underwater camera is installed with the optical axis parallel to the horizontal plane, and the behavior of the fish is observed with the side view angle, in order to survey the vertical distribution characteristics of the fish. For the experiments of this paper, the acquisition device in the laboratory environment is shown in Figure 1-(b). Parameters of the cameras used in the experiment are shown in Table 1.

C. OPTICAL FLOW EXTRACTION OF SEQUENCE IMAGE

Fish is a kind of sensitive and movable social animal. In its living water environment, its normal state is swimming all time. The behavior state of fish shoals is not only reflected in the spatial distribution, but also in the movement speed and other time behaviors of the group and individual. In this paper, fish shoals' behaviors are studied by method of CNN. For the two-dimensional image space information part, the traditional image features are not extracted, but directly input the image into CNN network. The time information which can reflect the performance of motion behavior can be obtained through the calculation relationship of sequence images. In this paper, the optical flow energy map is obtained by calculating the optical flow information of the sequence images first, then the spatial distribution map and the optical flow energy map are fused into one image, and finally sent the image to a CNN for learning.



(a) Image collection system structure scheme (b) Image acquisition device in laboratory

FIGURE 1. Fish shoals image acquisition system scheme and experiment device. Adopting top-view camera and underwater to catch fish behavior in three-dimensions.

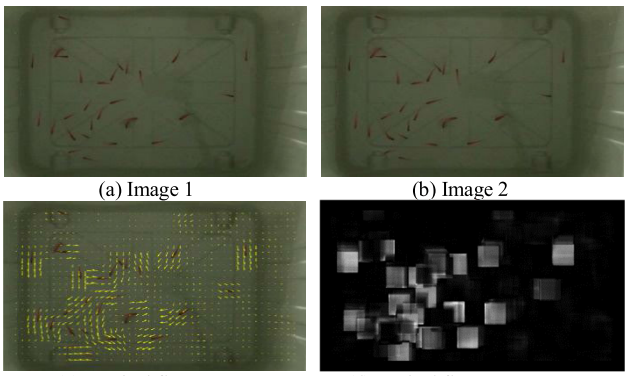
TABLE 1. Parameters of the cameras.

Item	Top-view camera	Underwater camera
Brand	SENTECH, Japan	ROBOTFISH, China
Type	STC-SC152POE	LBF-C50HD
Resolution	SXGA	1080P
Effective pixels	1360×1040	1920×1080
Frame rate	19fps	25fps
Sensor size	1/2"CCD	1/3" CMOS

Optical flow analysis is a common method in the process of image dynamic information acquisition. The concept of optical flow mainly refers to the change of the instantaneous velocity of the moving object in the space-time field at each pixel on the imaging plane observed by human. It can represent the motion information of the target and the motion trend of the target. There will be different methods to calculate optical flow when different constraints are introduced. Among them, the classic optical flow algorithm based on gradient is also called differential method. Its idea is to get the motion vector of each pixel in the original image according to the gradient function of the gray image. The common gradient based optical flow algorithms include Horn-Schunck algorithm to extract dense optical flow field [21] and Lucas-Kanada algorithm to extract sparse optical flow method [22].

Lucas-Kanada algorithm assumes that the motion vector remains constant in a certain spatial neighborhood, instead of the smooth term of velocity, uses the weighted least square method to estimate the optical flow, which is widely used in the extraction of sparse optical flow field. The optical flow vector of the region can be obtained by solving the equation with the least square method. However, the L-K optical flow method needs to include more than two edges in the above $n \times n$ size window to ensure that the matrix is reversible and that the optical flow constraint equation has solutions. Therefore, the calculation process of L-K optical flow method needs to combine corner and edge information. The optical flow vector obtained by this way is sparse optical flow vector.

Considering the color and size of the fish are consistent and with a uniform water background, the L-K algorithm of sparse optical flow extraction is adopted in this paper. Pictures in Figure 2 are the effects of fish image optical flow extraction using L-K algorithm in this paper. In the optical flow energy map, the larger the gray value of the pixel, the greater the motion energy, that is, the faster the motion speed.



(a) Image 1 (b) Image 2

(c) Optical flow vector (d) Optical flow energy map

FIGURE 2. L-K optical flow vector and energy map of fish shoals' image. Figure 2-(a) and 2-(b) are two adjacent frames of sequence images collected by the aerial camera; Figure 2-(c) is an optical flow vector obtained by two images, and Figure 2-(d) is the optical flow energy map obtained by calculating the motion amplitude information through the optical flow vector.

D. SPATIAL AND TEMPORAL INFORMATION FUSION OF IMAGE DATA

The behavior of fish shoals is not only reflected in the spatial distribution, but also reflected in the changes of time behavior

of the group or individual movement under the influence of pressure factors. Some states of fish shoals cannot be distinguished only by spatial information or temporal information, which need to be combined to make a judge.

For example, in the normal state, the fish shoals are scattered in the space, and every fish body moves around idly with normal speed. When the environment changes dramatically, such as water surface disturbance, the shoals are excited, and every fish body moves around in panic, with the swimming speed increases dramatically. Under normal conditions, if one fish suddenly touches another fish, it will cause individual interference, and the fish that is touched will suddenly swim faster and faster, resulting in individual abnormal speed. In the feeding state, the fish are concentrated in the feeding area, and the spatial distribution and optical flow map are concentrated.

Figure 3 shows several fish shoals' states collected from the top view.

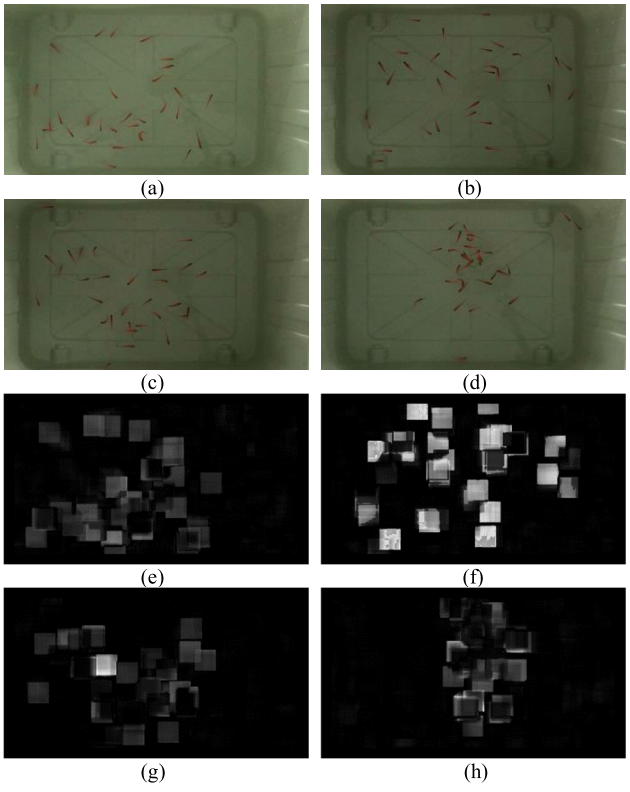


FIGURE 3. Spatial distributions and corresponding optical flow energy maps of four fish shoals' states from the top view. Figure 3-(a), 3-(b), 3-(c) and 3-(d) show the spatial distribution of fish shoals in the normal state, the group stimulated state, the individual disturbance state and the feeding state, respectively; Figure 3-(e), 3-(f), 3-(g) and 3-(h) are the optical flow energy maps corresponding to the above states respectively.

Fish live in the three-dimensional water space, so the behavior state is not only reflected in the horizontal distribution, but also in the vertical distribution of water space. If the water is not changed for a long time and the fish are in the state of anoxia, all the fish swim near the water surface, showing a vertical upper distribution. Otherwise, if the fish are not fed for a long time and in starvation state, there will be a lot of excreta under the water, and the fish will often swim under the water looking for food in the excreta below, showing a vertical lower distribution, as shown in Figure 4.

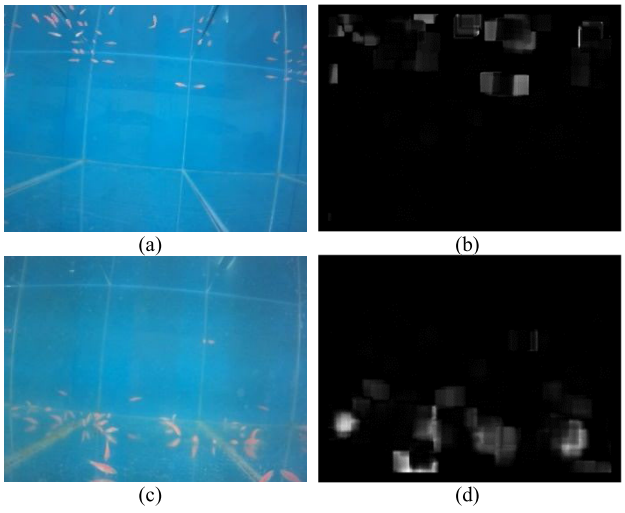


FIGURE 4. Spatial distributions and corresponding optical flow energy maps of two fish shoals' states from the side view. Figure 4-(a) and 4-(b) are the spatial distribution and the optical flow energy map of fish shoals in anoxic state; Figure 4-(c) and 4-(d) are the spatial distribution and the optical flow energy map of fish shoals in starvation state.

In the normal state, group stimulated state and individual disturbed state, the spatial distribution images shown in Figure 3 all show the scattered distribution of the fish shoals and no difference can be seen; however, the optical flow energy maps of the three states do have obvious difference. In the normal state, the optical flow energy map shows the state of dispersion but the overall motion energy is weak; in the group stimulated state, the optical flow energy map shows the state of dispersion but the overall motion energy is strong; in the individual disturbed state, the optical flow energy map shows the state of individual motion energy outstanding. But in the state of feeding, anoxic and starvation, no matter in the horizontal or in the vertical distribution, there are obvious differences. According to the experimental results shown in Figure 3 and Figure 4, under different environmental conditions, the behavior of fish shoals shows great changes. Some of these changes are reflected in the spatial distribution, some in the optical flow energy map with time. Therefore, only from a single information of space or time, it is impossible to distinguish the behavior state of fish shoals, and it needs to integrate two aspects of information to distinguish.

Design of spatiotemporal fusion strategy is based on the following several considerations.

(1) In this paper, the fusion strategy proposed is only a simple superposition of spatial image and optical flow energy map (time information), because the emphasis of this paper is to design a CNN to realization discrimination of fish shoals' states and carry out visualization analysis of intermediate data of CNN. A simple fusion strategy is helpful for our first study of this problem. In the follow-up researches, we can further

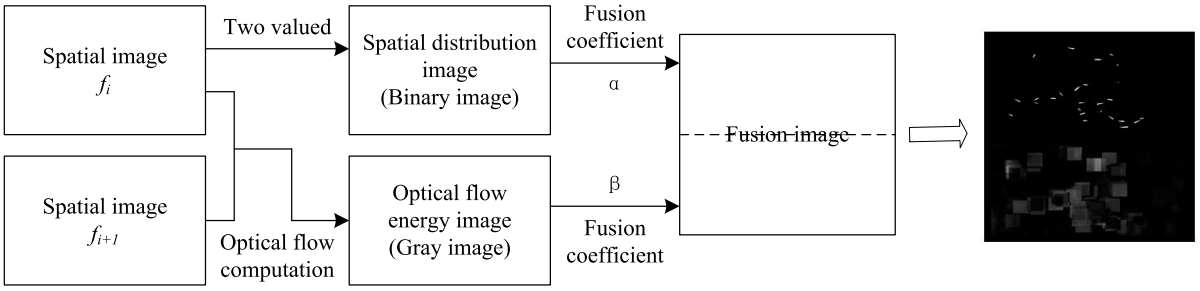


FIGURE 5. Spatiotemporal fusion strategy. The binary image representing the spatial distribution information of fish shoals and the optical flow intensity gray image representing the motion energy information are simply up-down overlapped into one image.

discuss some other fusion strategies such as information weighting and pixel level hybrid fusion, etc.

(2) In this paper, the color of the fish is same, and it is clearly distinguished relative to the environment; the information involved is only spatial distribution (location coordinates) information and optical flow energy intensity information. Color information is not important in this paper. Therefore, in order to reduce the redundant information of image samples and the complexity of neural network, this paper uses gray-scale image for spatial image and optical flow image.

(3) In this paper, the spatial image only uses the spatial distribution (location coordinate) information of fish shoals, which is independent of the gray intensity of pixels, while the optical flow energy image mainly uses the motion energy information, which is reflected by the gray intensity of pixels. Therefore, binary image is used in spatial image and gray image is used in optical flow energy image.

Based on the above considerations, the spatiotemporal fusion strategy adopted in this paper is shown in Figure 5.

E. IMAGE SIZE STANDARDIZATION

Because of the differences of illumination angle, hardware parameters, sampling time and sampling method, the image parameters such as image size (resolution), color information (pixel depth) are also different. In order to train the learning network effectively, improve the accuracy of classification and reduce the sample antagonism brought by the difference of original sampling, it is necessary to standardize the image sample data firstly [23].

Because this paper uses binary image and gray image to analyze, so it does not involve the problem of color standardization, only involves the problem of image size standardization. There are three common methods of image size standardization: nearest neighbor method, bilinear interpolation method and cubic interpolation method. Considering the factors of accuracy and operation speed, this paper adopts bilinear interpolation method [24] to standardize the image size.

For an image with size of $m \times n$, if the zoom factor is t, the size of the target image is $(m \times t) \times (n \times t)$. That is, when $t > 1$, the image size is enlarged; when $t < 1$, the image size is reduced. The gray value of pixel (x, y) in target image can be calculated by Equation (1):

$$P(x, y) = P'(x', y') = P'\left(\frac{x}{t}, \frac{y}{t}\right) \tag{1}$$

Here $P'(x', y')$ is the gray value of pixel (x', y') in original image.

Since x' and y' are floating-points, there are no real coordinate values on the original image, so it is necessary to use the gray value of the four adjacent points, and use Equation (2) to get it. Therefore, a complete image can be scaled by bilinear interpolation, shown as in Figure 6.

$$\begin{aligned} &f(i+u, j+v) \\ &= (1-u) \times (1-v) \times f(i,j) + (1-u) \times v \times f(i,j+1) \\ &\quad + u \times (1-v) \times f(i+1,j) + u \times v \times f(i+1,j+1) \end{aligned} \tag{2}$$

F. ESTABLISHMENT OF IMAGE DATABASE OF FISH SHOALS' BEHAVIORS

Through continuous experimental observation, this paper classified the behaviors of fish shoals into six behavior states, which are normal state, group stimulated state, individual disturbed state, feeding state, anoxic state and starvation state. The space and motion behaviors of the six states under the observation of top-view and side-view cameras are summarized as shown in Table 2.

According to the representations in Table 2, those repeated and non-separable states are eliminated, and the states with pink color shading in Table 2 are retained, and the image data base is established. Twelve groups of images are used to distinguish six type states.

The six type states are normal state, group stimulated state, individual disturbed state, feeding state, anoxic state and starvation state. In normal state, fish shoals are scattered representing on spatial image, fish individual moves around idly with normal speed, as well as the optical flow energy map shows dispersion but the overall motion energy is weak. In group stimulated state, every fish individual moves around in panic, fish shoals are scattered on spatial image but swimming speed increases dramatically, and the overall motion energy of optical flow energy map is strong. In individual

TABLE 2. Classification of behavior state of fish shoals and the corresponding characteristics.

		Spatial representation	Time (motion) representation
Normal state	Top-view	Dispersive distribution	Moderate speed
	Side-view	Dispersive distribution	Moderate speed
Group stimulated state	Top-view	Dispersive distribution	Overall fast speed
	Side-view	Dispersive distribution	Overall fast speed
Individual disturbed state	Top-view	Dispersive distribution	Individual fast speed
	Side-view	Dispersive distribution	Individual fast speed
Feeding state	Top-view	Concentrated distribution	Moderate speed
	Side-view	Concentrated distribution	Moderate speed
Anoxic state	Top-view	Dispersive distribution	Slow speed
	Side-view	Vertical-upper distribution	Slow speed
starvation state	Top-view	Dispersive distribution	Slow speed
	Side-view	Vertical-lower distribution	Slow speed

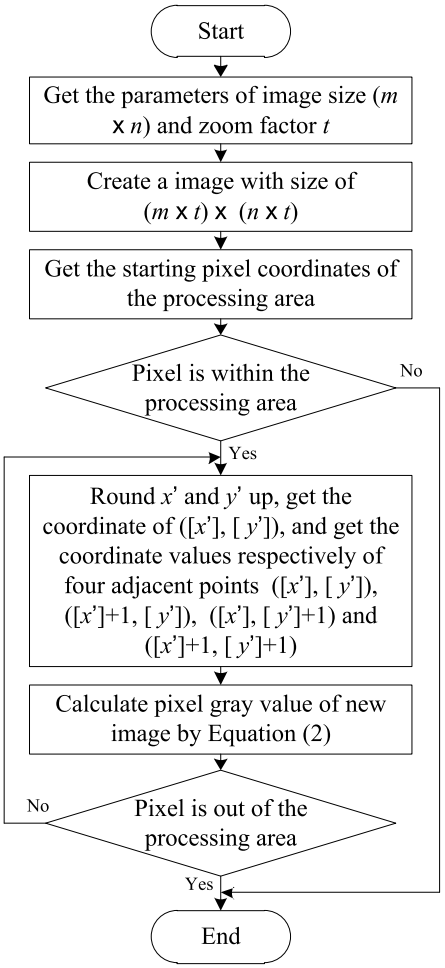


FIGURE 6. Flow diagram of bilinear interpolation image processing.

disturbed state, fish shoals are also scattered on spatial image, but because of individual interference, and the fish that is touched will suddenly swim faster and resulting in individual abnormal speed, the optical flow energy map shows individual motion energy outstanding. In feeding state, the fish are concentrated in the feeding area and represented as a centralized distribution with normal speed, and the spatial distribution and optical flow map are both on concentrated representing. In anoxic state, all the fish swim near the water surface for oxygen and showing a vertical upper distribution both on spatial image and optical flow map. In starvation state, all the fish often swim under the water looking for food in the excreta below and showing a vertical lower distribution both on spatial image and optical flow map.

These six states can be distinguished either in the spatial distribution of the fish shoals, or in the overall or individual swimming speed. Combined with the spatial distribution characteristics (space information) and the motion information (time information) embodied in the swimming speed energy, these six states can be and distinguished classified.

There are six types of fish shoals' behavior states in the self-made database. Each kind of behavior state contains 100 spatial images and 100 optical flow energy images. According to the fusion strategy, each spatial image and its corresponding optical flow energy image are combined into one sample image, so there are 600 sample images in the self-made database. When making training data sets and test data sets, the images in the original database are scrambled, and 80% of the data are randomly selected as the training data sets and 20% of the data as the test data sets.

An example of image samples after spatiotemporal information fusion and image size standardization is shown in Figure 7.

III. DESIGN OF CNN

A. STRUCTURE OF CNN

In this paper, the object of study is simple, and several kinds of fish behavior have a clear and separable state in the space

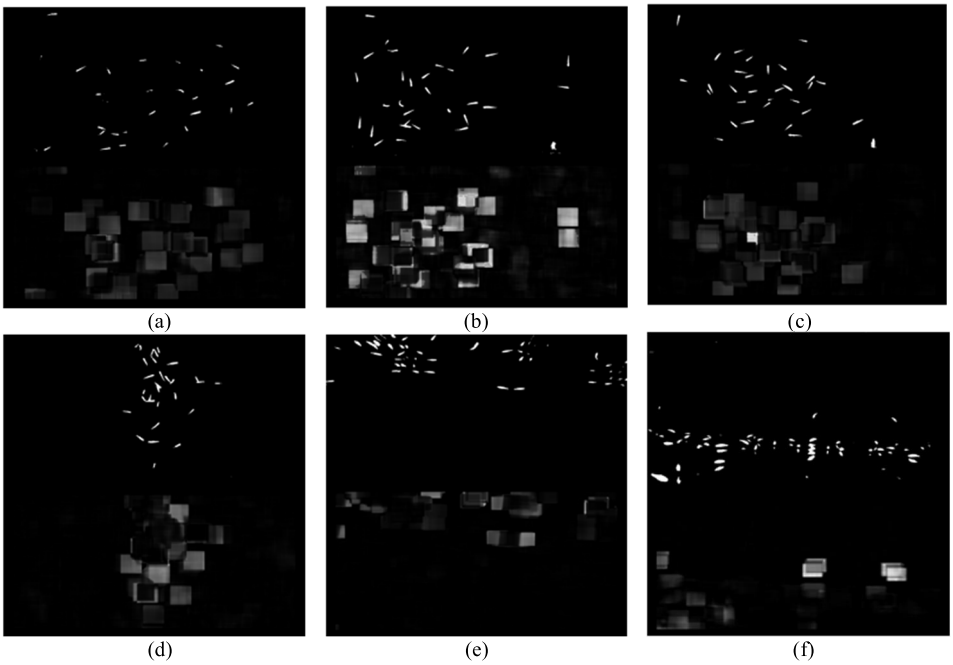


FIGURE 7. Image samples for image database. Figure 7-(a), 7-(b), 7-(c), 7-(d), 7-(e) and 7-(f) represent the image samples of fish shoals' behavior in normal state, group stimulated state, individual disturbed state, feeding state, anoxic state and starvation state respectively.

TABLE 3. Parameter configuration of CNN in this paper.

Input layer	W_1	Convolution layer	Pooling layer	W_2	Fully connected layer	W_o	Output layer
Input image pixels	[9, 9, 20]	20 convolution kernels with size 9 X 9	2X2 average pooling	[100, 307520]	Hidden layer nodes	[6, 100]	Types of fish shoals' behavior state
256 X 256		9 X 9 X 20	((256-(9-1))/2) X ((256-(9-1))/2) X 20		100		6

or time information image. The purpose of this paper is to quickly identify the behavior state of fish shoals, and make a discussion on intermediate data visualization of neural network. Therefore, this paper self-constructs a simple CNN to realize the fast discrimination of fish behavior state, studies the visualization of intermediate data of CNN, and discusses the influence of bath size and training epoch on the network training speed and recognition accuracy.

The simple CNN is constructed as shown in Figure 8 [25], and the specific parameter configuration is shown in Table 3. The input of CNN is the image data after spatiotemporal information fusion and image size standardization. The input image is processed firstly by convolution layer, which consists of 20 convolution kernels with 9×9 size. Data outputted from convolution layer will be processed by pooling layer, which adopts 2×2 average pooling. Data outputted from pooling layer will be input into a fully connected neural network, which includes 100 hidden layer nodes. At last, the output layer of the CNN is set as six output nodes, corresponding to the six behavior states of the fish shoals to be distinguished in this paper.

The neural network structure proposed in this paper has three parameter matrices to be determined by operation, which are W_1, W_2 and W_o as shown in Figure 8. W_1 is the parameter matrix of 20 convolution kernels of the convolution layer, with the dimension of [9 9 20]; W_2 is the parameter matrix between the pooling layer and the fully connected hidden layer, with the dimension of [100 307520] (Because the original input image size is 256×256, without image edge expansion, after a convolution operation by 9×9 kernel, the image size is 248×248; after 2×2 average pooling, the image size is 124×124; operated by 20 convolution kernels, there are $124 \times 124 \times 20 = 307520$ output data in the pooling layer; hidden layer of the fully connected neural network sets the number of nodes to 100); W_o is the parameter matrix between the hidden layer of fully connected neural network and the output layer, and the dimension is [6 100].

B. ACTIVATION FUNCTION OF NEURAL NETWORK NODE

1) ACTIVATION FUNCTION OF HIDDEN LAYER NEURONS

The activation function of hidden layer neuron node adopts Rectifier Linear Unit function (ReLU, Rectified Linear Unit).

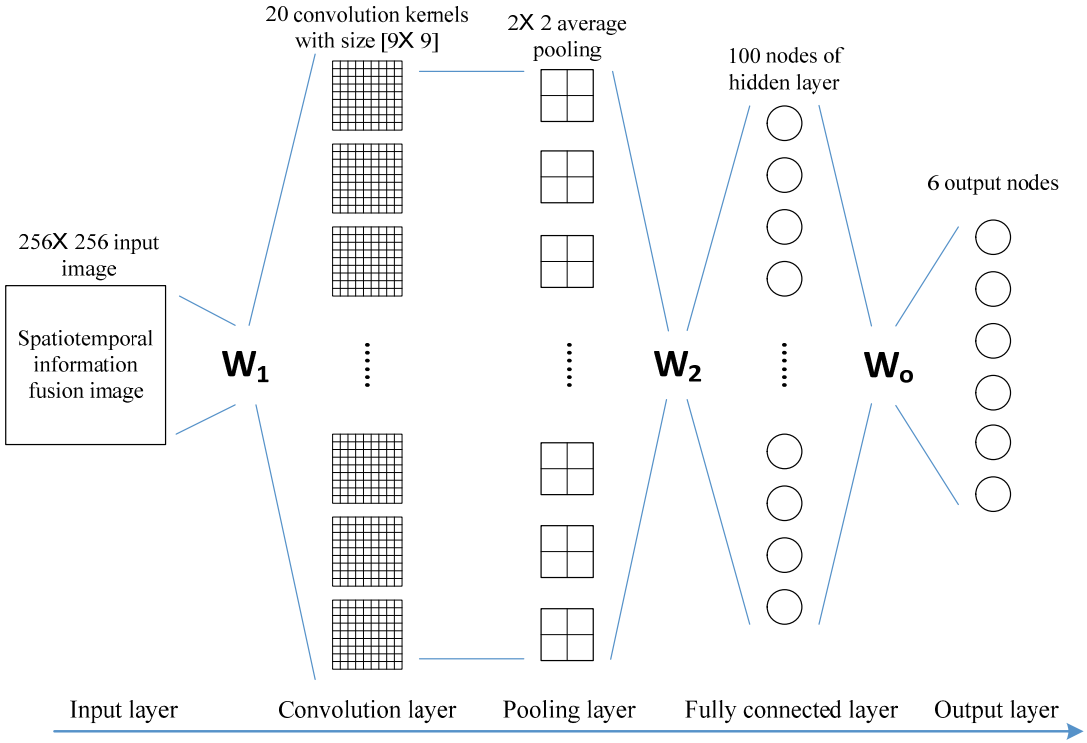


FIGURE 8. Structure of CNN used in this paper. The CNN is composed of one input-layer with 256 × 256 image, one convolution layer with 20 convolution kernels of 9 × 9 size, one pooling layer with 2 × 2 average pooling algorithm, and one fully connected neural network with 100 hidden layer nodes and six output nodes.

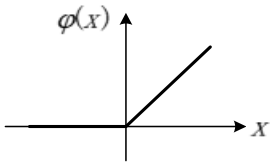


FIGURE 9. Function curve of ReLU.

The function curve is shown in Figure 9, and the function formula is shown in Equation (3).

Using the ReLU function, the activity of neurons in the artificial neural network (i.e. the output is positive) can be adjusted. In contrast, if using S-type function (Sigmoid function) as the activation function of neuron node, when the input is 0, it reaches 1 / 2, that is to say, it is already semi-saturated and stable, which is not in line with the expectation of actual biology to simulate neural network.

In the deep neural network structure, the use of ReLU function can more effectively carry out the error back propagation, avoiding gradient explosion and gradient disappearance [26]. Because there is no other complex activation function, such as exponential function, it can simplify the calculation process, and the dispersion of activity makes the overall calculation cost of neural network reduce.

$$ReLU(x) = \max(0, x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{3}$$

2) ACTIVATION FUNCTION OF OUTPUT LAYER NEURONS

The problem of fish shoals' behavior detection discussed in this paper is a multi-classification problem, so the output layer neurons use Normalized Exponential function (Softmax function). This function can "compress" a K-dimensional vector Z with any real number into another K-dimensional real vector σ (Z), so that the range of each element is between (0,1), and the sum of all elements is 1. Because the function considers the relative size of all output values, it is a suitable choice for all multi-classification neural networks.

The Softmax function calculates the output value of the *i*th output node as follows:

$$y_i = \phi(v_i) = \frac{e^{v_i}}{e^{v_1} + e^{v_2} + e^{v_3} + \ldots + e^{v_M}} = \frac{e^{v_i}}{\sum_{k=1}^{M} e^{v_M}} \tag{4}$$

where v_i is the weighted sum of the *i*th output node and M is the number of output nodes. Therefore, the Softmax function satisfies the following conditions:

$$\phi(v_1) + \phi(v_2) + \phi(v_3) + \ldots + \phi(v_M) = 1 \tag{5}$$

After the Softmax function, each output node is mapped into a category vector by using "one hot coding" or "1-of-N" coding. Only "1" is generated on the corresponding node and "0" is generated on the other nodes.

For example, the output vectors corresponding to the six categories of classification problem are shown in Figure 10.

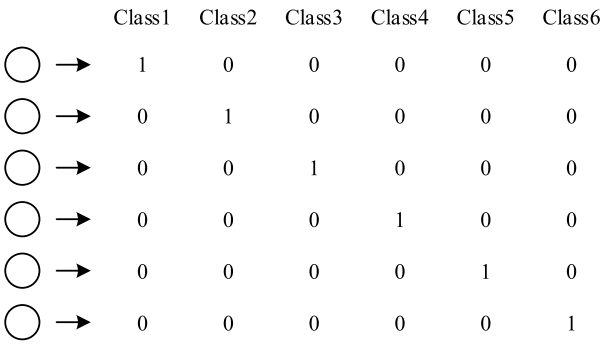


FIGURE 10. Output vector of "one hot coding" for six categories of classification problems.

C. MINIMUM MEAN SQUARE LEARNING RULE

The CNN designed in this paper has a simple structure and fewer layers. Therefore, the minimum mean square learning rule can minimize the quadratic variance between the actual output and the expected output.

The learning signals is:

$$r = d_j - W_j^T X \tag{6}$$

The weight vector adjustment amount is:

$$\Delta W_j = \eta(d_j - W_j^T X)X \tag{7}$$

Components of ΔW_j are:

$$\Delta w_{ij} = \eta\left(d_j - W_j^T X\right)x_j \quad i = 0, 1, .., n \tag{8}$$

In Equation (6)∼(8), X is the input vector, d_j is the teacher signal, W_j is the weight vector, ΔW_j is the adjustment of the weight vector, and η is the learning rate.

The least mean square learning rule is independent of the transformation function adopted by the neuron, so it does not need to calculate the derivative of the transformation function, which is not only fast in learning, but also has high accuracy. The weight can be initialized to any value.

D. WEIGHT ADJUSTMENT OF MOMENTUM METHOD

In the process of neural network training, the advantage of using better weight adjustment method is that it can get higher stability and faster speed, which are especially suitable for deep learning. Momentum is added to the incremental rule to adjust the weight. Momentum, to some extent, pushes the weight to a certain direction, rather than making the weight change immediately.

$$\begin{cases} \Delta w = \alpha\delta x \\ m = \Delta w + \beta m^- \\ w = w + m \\ m^- = m \end{cases} \tag{9}$$

The formula of adjusting weight by momentum is as Equation (9). Here, m^- is the previously calculated momentum, and β is a positive constant less than 1. The momentum adjustment process is as follows:

$$\begin{aligned} m(0) &= 0 \\ m(1) &= \Delta w(1) + \beta m(0) = \Delta w(1) \\ m(2) &= \Delta w(2) + \beta m(1) = \Delta w(2) + \beta\Delta w(1) \\ m(3) &= \Delta w(3) + \beta m(2) = \Delta w(3) + \beta\left[\Delta w(2) + \beta\Delta w(1)\right] \\ &= \Delta w(3) + \beta\Delta w(2) + \beta^2\Delta w(1) \\ &\vdots \end{aligned} \tag{10}$$

As the process progresses, the previous weight update values, such as $\Delta w(1)$, $\Delta w(2)$, and $\Delta w(3)$, are added to each momentum. Because β is a value less than 1, the earlier weight update value will have a smaller impact on momentum, which gradually decreases over time, but the earlier weight update is still stored in momentum. Therefore, the weight is not only affected by one weight update value, so the stability of learning is improved. In addition, as the weight continues to update, the momentum becomes larger and larger. As a result, the weight update value also becomes larger and larger. In this way, the learning speed is also improved.

IV. EXPERIMENT

A. OPERATION PLATFORM

The operation platform of this paper is Dell laptop precision7730. See Table 4 for specific configurations.

TABLE 4. Operation platform configuration information.

Part	Configuration
CPU	Intel(R) Core(TM) i7-8750H @2.20GHz 2.21GHz
Memory	32G DDR4
Hard disk	256G Solid-state disk +2T
System	Windows10
Software platform	MATLAB 2017b

B. INTRODUCTION FOR EXPERIMENT PROCESS AND BASIC PARAMETERS

There are 600 images in the self-made image database, including 6 types of fish behavior states, each of which has 100 images. 80% of every behavior state image samples are used for network training and 20% for network testing. When training the network, it is usually necessary to initialize the parameters of the network according to a certain distribution. The appropriate initial value can make the loss function in the training converge quickly and get a better network training effect. Improper initial value may cause loss function to fall into local minimum state. Network training should update the weight parameters according to the actual training situation.

1) ADOPT SMALL BATCH TRAINING, AND SELECTION FOR BATCH_SIZE

Small batch algorithm is a hybrid form between SGD (stochastic gradient descent) algorithm and batch algorithm.

It can not only avoid the sensitivity of SGD algorithm to training data, but also save operation time compared with batch algorithm to some extent. In this paper, the number of training samples for each type is 80, and the total number of training samples is $6 \times 80 = 480$. This paper discusses the influence of the selection of small batch capacity on training speed and training accuracy. The experiment was carried out by selecting 8, 16, 20 and 40 as Batch_size respectively.

2) SELECTION FOR EPOCH

One data training round is to complete an Epoch, which means that all training data is transmitted through neural network and back propagation once. However, in the process of neural network training, it can't achieve the best effect only to transfer the complete data set once. It is necessary to adjust the network parameters to the best by transferring the entire data set in the neural network for many times. In this paper, Epoch was selected 20, 30, 50, 80, 100, 150, 200 for experiments, and the training results were compared and discussed.

3) SELECTION FOR MOMENTUM

In Section III-D of this paper, the application principle of Momentum method is explained. Momentum is based on the principle of energy conversion between potential energy and kinetic energy in physics. The larger the Momentum, the greater the energy converted into potential energy, and the more likely it is to get rid of the local minimum. In order to avoid the network oscillation caused by large Momentum, the value of Momentum is set to 0.95, which is also a commonly used value.

4) SELECTION OF LEARNING RATE

Learning rate is a super parameter, which indicates the degree of adjusting network weight in the process of calculating gradient loss. The larger the learning rate is, the faster the learning speed of the network is, the more likely it is to produce gradient explosion and oscillation; the smaller the learning rate is, the slower the convergence speed of the network is, the more likely it is to have over fitting. The CNN proposed in this paper has fewer layers and adopts static learning rate, which is set as 0.01.

C. IMPACT OF BATCH_SIZE ON NETWORK PERFORMANCE

This paper discusses the classification of six types of fish behavior. The number of training image samples for one type is 80, and the total number of training samples is $6 \times 80 = 480$. Select 8, 16, 20 and 40 for Batch_size for the experiment. Under the same training conditions, the number of Epoch is all in 50, and the training time and accuracy are shown in Table 5.

It can be seen from the data in Table 5 that for the training samples in this paper, when Batch_size is selected as 16, the training time and training accuracy reach the highest cost-performance. For deep neural network, generally, when the integer power of 2 is selected for Batch_size, the network training will achieve the highest cost-performance [27].

TABLE 5. Impact of Batch_size on network training.

Batch_size	Accuracy	Training time (s)
8	0.708333	4217.051
16	0.825	4089.668
20	0.775	4074.389
40	0.741667	4260.141

D. IMPACT OF EPOCH ON NETWORK PERFORMANCE

According to the results of Table 5, two Batch_size with relatively high training accuracy and short training time are selected, i.e. Batch_size is selected as 16 and 20. Then make a comparison on accuracy and training time corresponding to training samples and test samples respectively.

It can be seen from the data in Table 6 that no matter Batch_size is selected as 16 or 20, there are the trends as follows: the error of the training samples has been significantly reduced with the increase of the number of Epoch; the accuracy of the test samples first increases with the increase of the number of Epoch, but it is difficult to improve the accuracy more after reaching a certain number of Epoch. It shows that in the process of network training, with the increase of Epoch, the network parameters have been continuously optimized and adjusted, which makes the discrimination error of training samples reduce constantly; but for the ''never seen'' test samples, when the network training reaches a certain degree, the accuracy of the test samples is difficult to be improved. As shown in Table 6, when Batch_size is selected as 20, if the accuracy of training samples reaches the highest value, with the increase of Epoch, the accuracy begins to decline, indicating that the network training has been over fitted.

What's more, according to the experimental data, when the Batch_size is selected as 16, the network can quickly achieve a stable state with high recognition accuracy of test samples, and the recognition accuracy has obvious advantages over other Batch_size options. In general, when Batch_size is chosen as the integer power of 2, the deep neural network has a higher efficiency.

E. DISCUSSION ON THE VALIDITY AND DATA VISUALIZATION OF CNN FOR FISH BEHAVIOR DETECTION

According to the above discussion, Batch_size of 16 and Epoch of 100 were selected for fish behavior state detection experiment. The relationship between Epoch and training samples detection error is shown in Figure 11. It can be seen from Figure 11, with the increase of Epoch, the detection error rate of training samples continues to decline. When Epoch is 100, the error rate reaches the level of - 5 power of 10. The Mean Square Error function is selected as the error calculation function.

TABLE 6. Impact of Epoch on network training.

	Batch_size=16			Batch_size=20		
Epoch	Training samples error	Test samples accuracy	Time (s)	Training samples error	Test samples accuracy	Time (s)
20	1.65×10^{-1}	0.750	1.67×10^{3}	3.54×10^{-2}	0.742	1.66×10^{3}
30	5.6×10^{-3}	0.808	2.50×10^{3}	3.97×10^{-3}	0.767	2.48×10^{3}
50	5.27×10^{-4}	0.825	4.08×10^{3}	5.50×10^{-4}	0.775	4.11×10^{3}
100	6.16×10^{-5}	0.825	8.63×10^{3}	6.43×10^{-5}	0.783	8.51×10^{3}
150	1.60×10^{-5}	0.825	1.24×10^{4}	2.10×10^{-5}	0.775	1.24×10^{4}
200	7.34×10^{-6}	0.825	1.64×10^{4}	9.82×10^{-6}	0.775	1.64×10^{4}

TABLE 7. Test samples classification accuracy.

	1	2	3	4	5	6	Total accuracy
1	$\underline{\mathbf{0.342 \pm 0.141}}$	$\underline{\mathbf{0.251 \pm 0.080}}$	$\underline{\mathbf{0.329 \pm 0.097}}$	0.039 ± 0.014	0.013 ± 0.004	0.026 ± 0.009	0.825
2	0.080 ± 0.054	$\underline{\mathbf{0.578 \pm 0.147}}$	$\underline{\mathbf{0.338 \pm 0.153}}$	0.003 ± 0.002	$3 \times 10^{-4} \pm 3 \times 10^{-4}$	$1 \times 10^{-4} \pm 1 \times 10^{-4}$	
3	$3 \times 10^{-5} \pm 1 \times 10^{-5}$	0.003 ± 0.001	$\underline{\mathbf{0.995 \pm 0.003}}$	0.002 ± 0.002	$2 \times 10^{-7} \pm 9 \times 10^{-8}$	$1 \times 10^{-6} \pm 4 \times 10^{-7}$	
4	0.012 ± 0.011	0.034 ± 0.030	0.051 ± 0.022	$\underline{\mathbf{0.902 \pm 0.039}}$	$4 \times 10^{-4} \pm 2 \times 10^{-4}$	$6 \times 10^{-5} \pm 3 \times 10^{-5}$	
5	0.012 ± 0.010	0.037 ± 0.028	0.009 ± 0.008	0.013 ± 0.008	$\underline{\mathbf{0.924 \pm 0.055}}$	0.003 ± 0.003	
6	0.114 ± 0.075	0.041 ± 0.032	0.028 ± 0.009	0.008 ± 0.003	0.004 ± 0.002	$\underline{\mathbf{0.806 \pm 0.093}}$	

The vertical index number represents the actual class number of the test sample; the horizontal index number represents the class number outputted from the CNN for test sample.

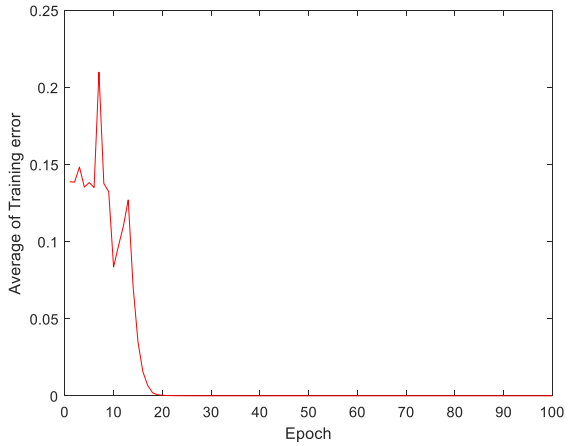


FIGURE 11. Training samples detection error with training epoch With the increase of Epoch, the detection error rate of training samples continues to decline.

At the end of 100 Epoch of training, the test sample data of this paper is detected and distinguished. The ''student'' distribution (''t'' distribution) method [28] for small sample measurement is used to calculate the probability and confidence limit of each kind of samples. The formula for calculating the confidence limit is shown in Equation (11), where, t_a is the confidence coefficient, $\sigma_{\bar{x}}$ is the standard deviation of the arithmetic mean of n measurements of parameter x. In the experiment of this paper, the significance level is $\alpha = 0.05$, so the confidence probability is $P = 1 - \alpha = 0.95$, and the freedom degree is $v = n-1 = 19$ (the number of test samples for each state are 20, so $n = 20$). Finally, the confidence coefficient is obtained by looking up the table [28], that is $t_a = 2.09$. The classification accuracy of each type of sample is shown in Table 7.

$$\delta_{lim}\bar{x} = \pm t_\alpha \sigma_{\bar{x}} \tag{11}$$

It can be seen from Table 7 that the simple CNN constructed in this paper can basically realize the classification of fish shoals' behavior states. For states of 3, 4 and 5, it can be accurately classified with a probability of more than 90% of the mean value and with a lower confidence limit; for state of 6, it can also be accurately classified with a probability of more than 80% of the mean value. For the state of 2, it can be classified with probability of more than 50%; but for the state of 1, the classification accuracy is relatively low. Adopting ''1-of-N'' coding method, the state with the largest output discrimination probability is the state to be classified. Then under the condition of 100 Epoch of training samples in this paper, the overall classification accuracy of the test samples is 82.5%. The experiment results shown as in Table 7 indicate that states of 3, 4, 5 and 6 can always be identified well, but states of 1 and 2 are easily confused in other states.

Examples of six states' sample-images are shown in Figure 7. By analyzing the image characteristics of six kinds

of samples, it can be seen that in the 4, 5 and 6 categories of samples which represent the state of feeding, anoxia and starvation respectively, fish shoals have the characteristics of spatial aggregation, or distribution above the water surface, or distribution below the water body, that is, both spatial images and optical flow energy images have obvious characteristics of spatial distribution. On the other hand, the 1, 2 and 3 categories of samples which represent the normal state, group stimulated state and the individual disturbed state respectively, fish shoals all have the characteristics of spatial dispersive distribution, and the difference is mainly reflected in the optical flow energy map, which shows generally weak for state 1, generally strong for state 2, and individually prominent for state 3.

According to the characteristics of fish shoals' behavior states in Figure 7 and the results of classification shown in Table 7, it can be seen that the CNN constructed in this paper has better classification results for the images with strong spatial distribution characteristics and especially prominent for outstanding individual targets; for the images with no obvious spatial distribution and gray intensity (e.g. for Class 1 images), the classification results are the worst and easy to be discriminated into other categories. On the other hand, the experimental results also show that the CNN has a higher discrimination power to image edge features (representing spatial distribution features) than image gray features (representing energy intensity features).

In order to verify the above experimental results, this paper further studies the intermediate process of image feature extraction based on CNN. The visualization of 20 convolution kernels of 9 × 9 after the training corresponding are shown in Figure 12.

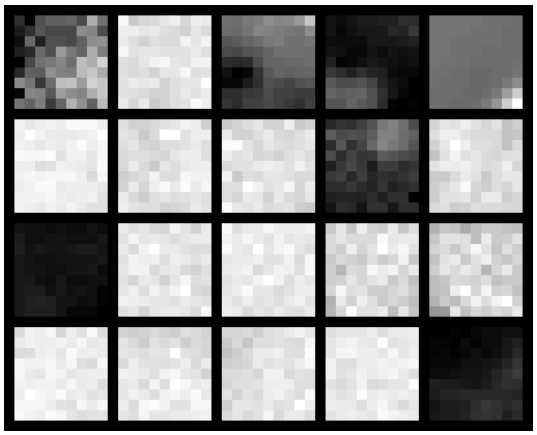


FIGURE 12. Visualization of convolution kernel coefficients There are 20 convolution kernels totally, each of which is 9 × 9 in size. The coefficient intensity of each convolution kernel is displayed as a gray-scale image.

Select one typical image from each class for display analysis. The original sample images of six classes are shown in Figure 13; the visualization of the convolution results of the sample images through the 20 convolution kernels are shown in Figure 14; the visualization of average pooling function

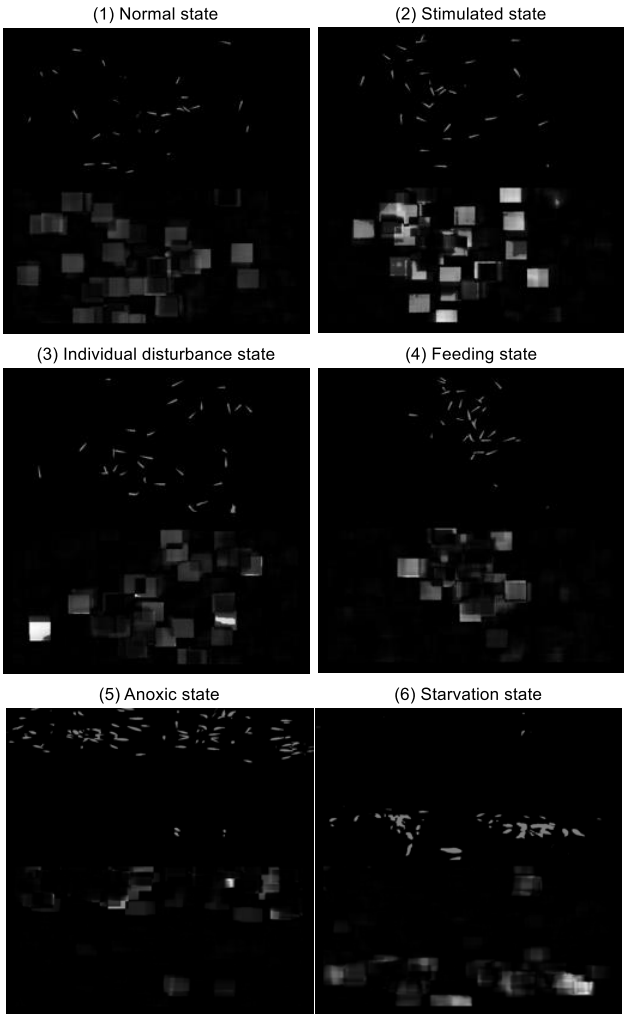


FIGURE 13. Original sample images of six classes. Figure 13-(a), 13-(b), 13-(c), 13-(d), 13-(e) and 13-(f) represent the original sample images of fish shoals' behavior in normal state, group stimulated state, individual disturbed state, feeding state, anoxic state and starvation state respectively, corresponding to the following data visualization discussions in Figure 14 and Figure 15.

results are shown in Figure 15, that is, the data visualization before being sent into the full connected neural network.

From the data visualization results in Figure 14, it can be seen that for the main area of the target (fish body and optical flow energy area), convolution results are all in negative gray-value (the gray-value of the background area is 0). After processed by the ReLU neuron activation function, the negative gray-value area is converted to 0.

From the data visualization results in Figure 15, it can be seen that after the ReLU neuron activation function, the edge information of the upper fish body is outlined as a high gray-value by the convolution kernel and retained; while the optical flow energy information, except the image corner information extracted by the fifth convolution kernel (as shown in Figure 12, the fifth convolution kernel coefficients present the characteristics of increasing gradually from the upper left corner to the lower right corner), almost no

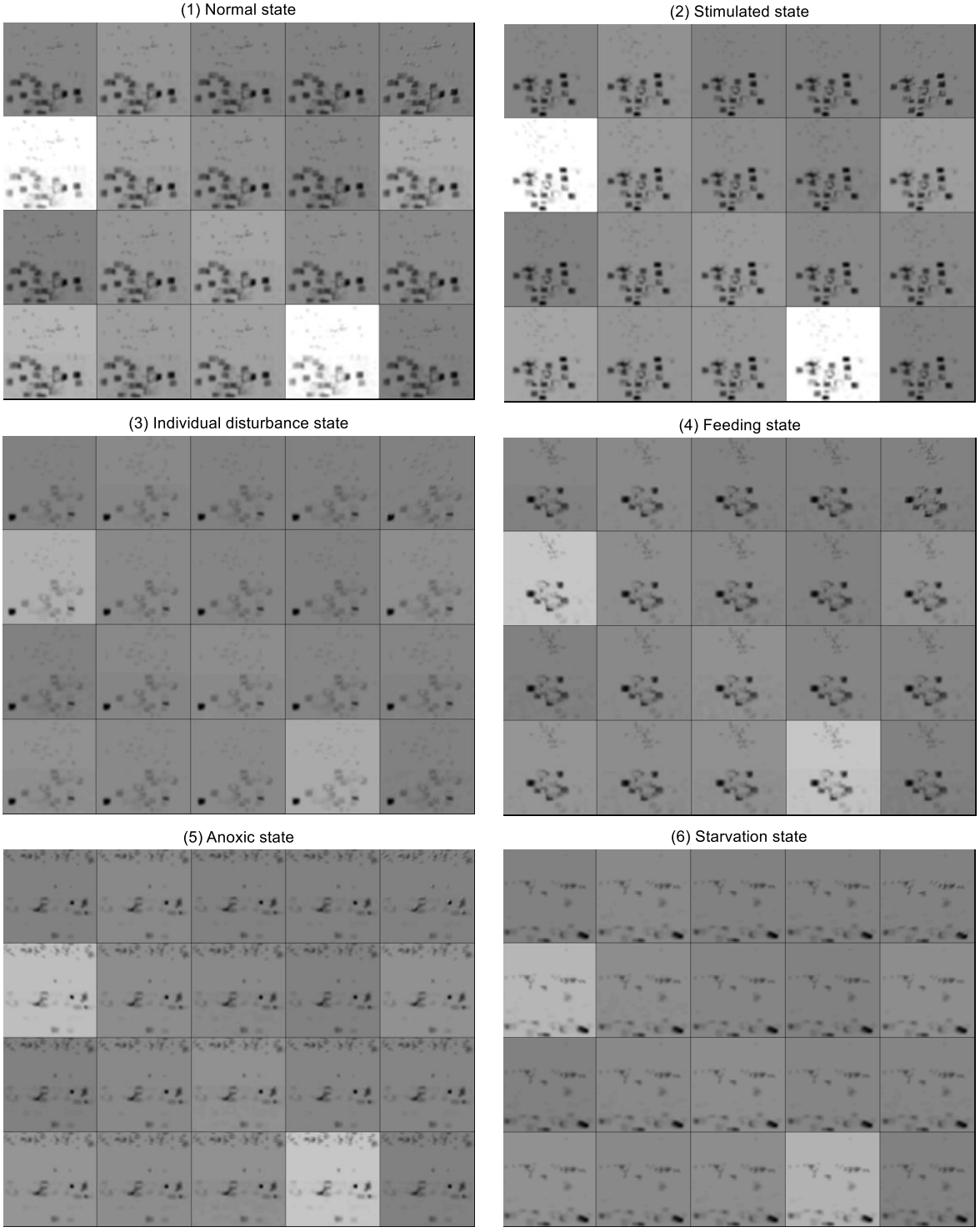


FIGURE 14. Visualization of data after convolution operation Figure 14-(a), 14-(b), 14-(c), 14-(d), 14-(e) and 14-(f) represent the data visualization after convolution operation for Figure 13-(a), 13-(b), 13-(c), 13-(d), 13-(e) and 13-(f) respectively.

other optical flow energy information is retained by the other convolution kernels. The data shown in Figure 15 is the data sent to the full connection neural network. It can be seen that the CNN is mainly based on the edge information of fish space distribution, and the optical flow energy information plays a very small role. Therefore, the result of the network is poor for the first and second classes of samples.

In the above experiments, ReLU function is used as the activation function of neurons, which can inhibit the negative value. In order to explore whether the increase

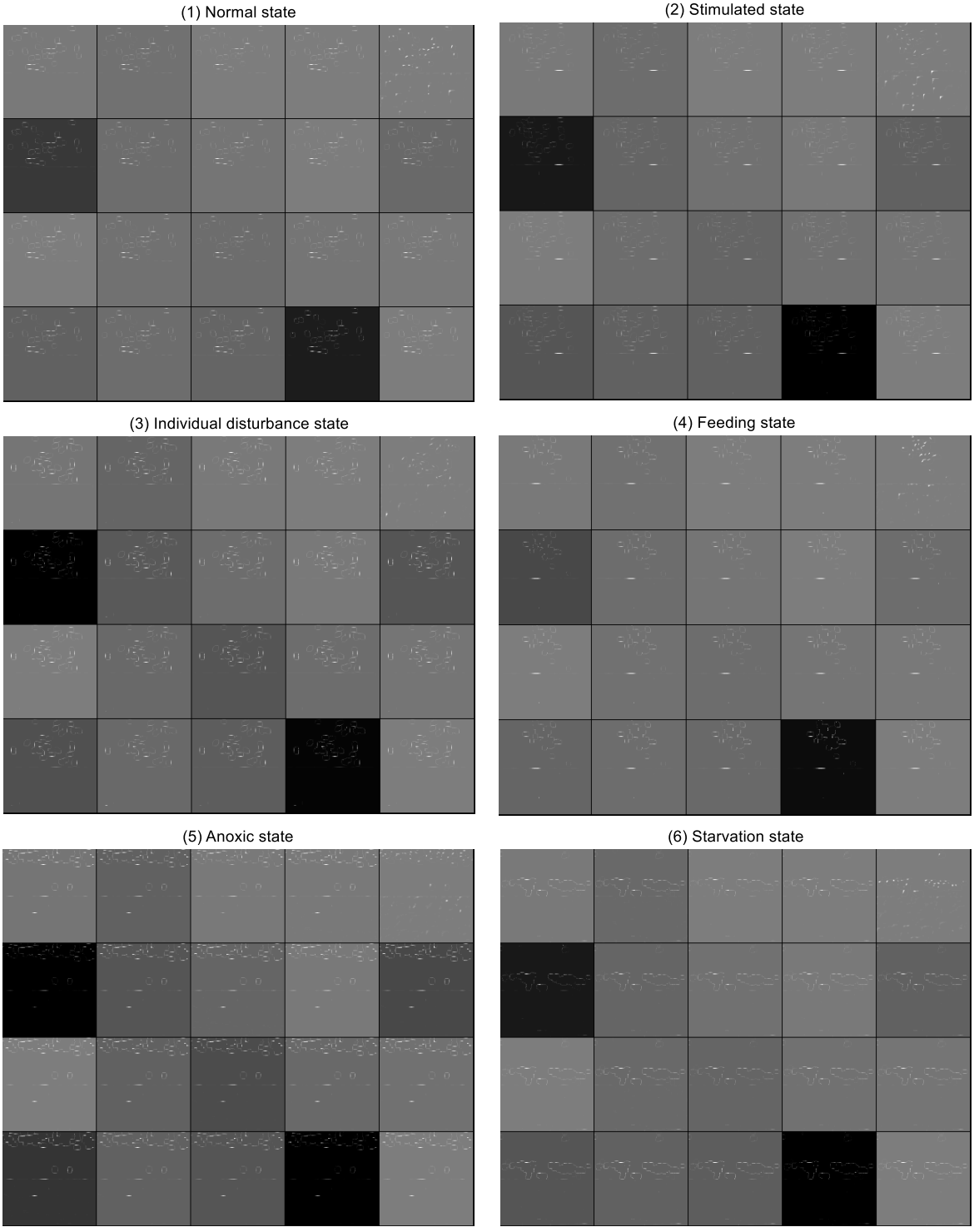


FIGURE 15. Visualization of data before sending into fully connected neural network. Figure 15-(a), 15-(b), 15-(c), 15-(d), 15-(e) and 15-(f) represent the data visualization before sending into fully connected neural network for Figure 13-(a), 13-(b), 13-(c), 13-(d), 13-(e) and 13-(f) respectively.

of negative target area value can increase the result of classification, the activation function of neural network is tested with S-type function, that is Log-Sigmoid function (normal S-type function) and Tan-Sigmoid function (Hyperbolic Tangent S-type function) respectively. The formula of Log-Sigmoid function and Tan-Sigmoid function are shown

in Equation (12) and Equation (13), respectively.

$$\varphi(x) = \frac{1}{1+e^{-x}} \tag{12}$$

$$\varphi(x) = \frac{2}{1+e^{-2x}} - 1 \tag{13}$$

Taking the first sample image shown in Figure 13-(a) as an example, the data of neural network neurons after activation functions of Log-Sigmoid and Tan-Sigmoid are visualized. Figure 16 and Figure 18, respectively, show the data visualization after using Log-Sigmoid and Tan-Sigmoid function and before sending into the fully connected neural network, as well as Figure 17 and Figure 19, respectively, show their training samples detection error curve with training Epoch.

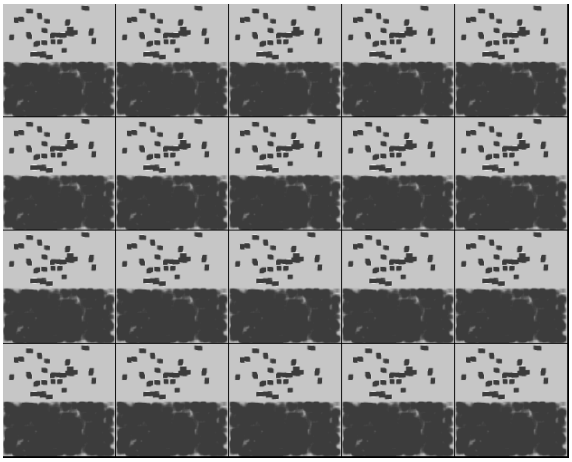


FIGURE 16. Visualization of data processed by Log-Sigmoid neuron activation function Using Log-Sigmoid activation function, most of the regional energy information is input into the neural network with the negative numerical value.

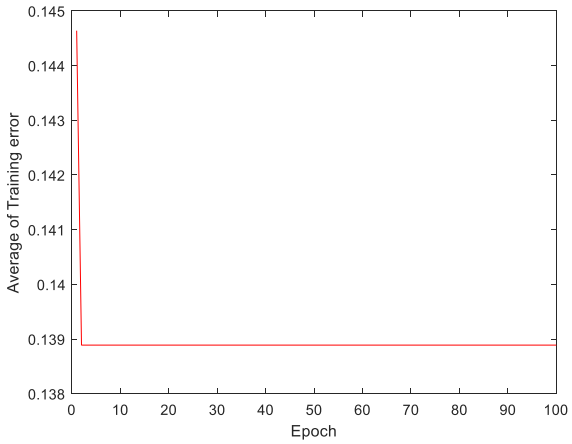


FIGURE 17. Training error curve processed by Log-Sigmoid neuron activation function Using Log-Sigmoid activation function, training of the neural network will stop after initial several Epoch.

From data visualization results in Figure 16 and Figure 18, it can be seen that after using the neuron activation function of Log-Sigmoid and Tan-Sigmoid, the data sent into the fully

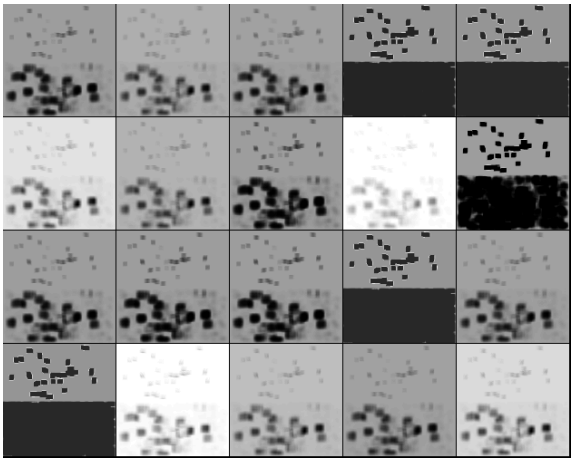


FIGURE 18. Visualization of data processed by Tan-Sigmoid neuron activation function Using Tan-Sigmoid activation function, most of the regional energy information is input into the neural network with the negative numerical value.

connected neural network contains the optical flow energy information in the lower half of the sample image, and most of the regional energy information is input into the neural network with the negative numerical value.

However, from the training sample error curve shown in Figure 17, and data from Table 8, it can be seen that if adopted Log-Sigmoid function, after the initial several parameter adjustments, the training of the neural network will stop, and the network parameters and the training error will no longer have any changes. From the training sample error curve shown in Figure 19, it can be seen that if adopted Tan-Sigmoid function, the neural network will be always in the state of vibration, never unable to settle down to a stable and effective state.

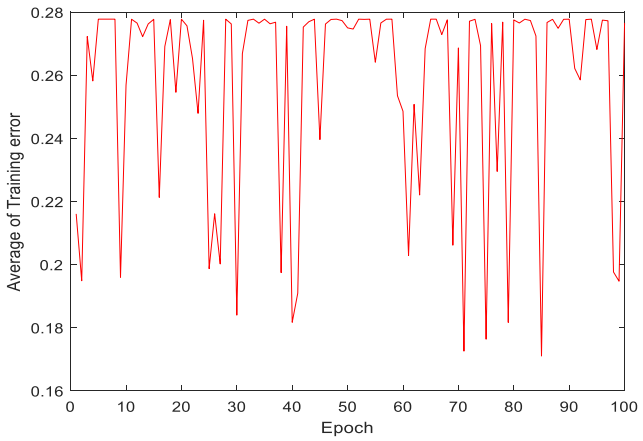


FIGURE 19. Training error curve processed by Tan-Sigmoid neuron activation function Using Tan-Sigmoid activation function, the neural network will never settle down to a steady state.

When Epoch is selected as some different typical value, training sample error, test sample accuracy and training time adopted Log-Sigmoid and Tan-Sigmoid neuron activation

TABLE 8. Classification accuracies for test samples adopted Log-Sigmoid and Tan-Sigmoid function.

	Log-Sigmoid activation function			Tan-Sigmoid activation function		
Epoch	Training samples error	Test samples accuracy	Time (s)	Training samples error	Test samples accuracy	Time (s)
20	66.67	0.1667	$1.70X10^3$	133.33	0.1667	$1.66X10^3$
30			$2.54X10^3$	88.31		$2.49X10^3$
50			$4.26X10^3$	132.00		$4.13X10^3$
100			$8.49X10^3$	132.77		$1.20X10^4$
150			$1.75X10^4$	133.30		$1.75X10^4$
200			$1.68X10^4$	133.33		$1.67X10^4$

function respectively are shown in Table 8. From the data shown in Table 8, it can be seen that with Log-Sigmoid function and Tan-Sigmoid function, the network is in a completely invalid state whether from the training samples error or the test samples accuracy, and this invalid state will not change with the increase of the number of Epoch. While with the ReLU activation function, after 100 Epoch, the classification accuracy of the network to the test sample is 0.825, which is an effective network.

V. DISCUSSION AND CONCLUSION

The experimental object of this paper is zebrafish raised in laboratory with an average body length of about 3-4cm. But the size parameters of individual fish do not need to bet concerned about, because this paper puts emphasis on the group behavior of fish shoals. This model has no limitation on the size of individual fish, however, it is required to adjust the construction structure of the visual imaging system (such as the parameters of focal length, object distance, image distance, field of view, and so on) according to the size of the monitored fish shoals, in order to make the system be able to image the whole group of fish, to observe the features of spatial distribution and group motion energy.

In the experiments of this paper, the number of fish used is about 30. According to the experiments conducted in the research process of this paper, at least 10 or more fish individuals are needed, which can better reflect the group behavior of the fish shoals. However, the number of individual fish in the fish group should not be too large, otherwise the large group behavior will be decomposed into many small group decentralized behaviors due to the lack of environmental disturbance and some other reasons. The number of individual fish that can reflect the behavior of fish shoals is related to the species, living habits and living environment of fish. When the individual size of fish increases, the number of fish increases, and the imaging field of view increases, if the corresponding relationship between the pixel size of single camera and the imaging field of view meets the requirements of object-image resolution, the single camera can be used for shooting and monitoring. If the corresponding field size of single camera does not meet the requirements of object-image resolution, multi-camera imaging is required. Multi-camera uses synchronous trigger signal to shoot synchronously.

For the method proposed in this paper, the requirement of the background is to be able to highlight the fish target and easily extract the fish target from the background. Although in this paper, CNN is used for image classification, instead of traditional feature extraction in the early stage, but from the intermediate data visualization of CNN, it can be seen that the CNN for image classification is based on more edge information. Therefore, the ideal background of this method is that the background color (or gray level) is uniform and it is prominent different from that of the fish target. That is to say, the edge information of the target rather than the edge of the background pattern is expected to be reflected for the image analysis. When the background is easily confused with the target, the result of neural network training will be affected, so that the rate of false judgment will increase, or even classification failure.

Among the existing fish behavior detection models, some are based on spatial distribution characteristics, some are based on motion characteristics. But most of them are based on a certain feature to detect a certain behavior state. Here are some typical fish behavior detection models.

Reference [9] takes rainbow trout as the research object, presents an algorithm developed to calculate two indexes characterizing fish shoals' behavior. The first index is fish group dispersion index, which is calculated by summing the perimeters of each group of individuals. A group is a single black area created by the outlines of overlapping fish, so it is expected that the index would be high when shoal is spread and low when it is aggregated. The second index is fish group activity index, which is estimated by subtracting from an image any dark areas that overlap with dark areas in the previous image in a chronological order. The remaining area is proportional to the fish movement in a given time interval and gives therefore a good estimation of the swimming activity of the shoal. When swimming activity is important, it is expected therefore a high index. Behavioral variations of the shoal were estimated before and after food distribution in one test, and before and after a four hours confinement stress in

a second test. Data resulting from simulations indicate that the two indexes are sensitive to the simulated changes in the cohesion or the swimming speed of the group. Thus, indexes faithfully translated true values in simulations, with a minimum of 94% of the total variation in true values explained by indexes.

Reference [29] describes a computer vision-based method for measuring the feeding activity of an Atlantic salmon (Salmo salar) shoal. Feeding activity analysis was based on the intensity summation of the difference frame due to the motions of the fish. An overlap coefficient was defined to calibrate the error of calculation caused by the overlaps among fish bodies in images. Based on these data, a computer vision-based feeding activity index (CVFAI) was determined for measuring the feeding activity of the fish in arbitrary given duration. To assess the reliability of CVFAI, a manual observation feeding activity index (MOFAI) was determined by scoring each kind of feeding behavior in the same recordings. The CVFAI and MOFAI presented a linear relationship at a correlation coefficient of 0.9195.

Reference [12] proposes a modified kinetic energy model of the whole shoal instead tracking individuals inside a shoal. First, with the reflective regions of the water surface taken into account, two alternative ways were presented to extract the necessary spatial behavioral characteristics of the shoal. Then, optical flow, entropy and statistics were adopted to measure the dispersion status (dispersion of individuals in space) and the changing magnitude of the motion (behavioral characteristics) of the shoal. Finally, the modified kinetic energy model was obtained by combining the dispersion measurement with the changing magnitude of the behavioral characteristics. Referring to the human observation and the evacuation of the gastro-intestine contents-based long experiment of Nile tilapia (Oreochromis niloticus) in RAS, the proposed model shows good performance in detection of the emergent gathering and scattering behaviors with $97.20 \pm 1.23\%$ success rate at least and $0.61 \pm 0.08\%$ missing report rate at most.

Reference [30] aiming at the local unusual behaviors of fish school, proposes a method based on the modified motion influence map and recurrent neural network (RNN). First, the motion characteristics of the whole fish school were extracted using particle advection scheme. Secondly, the modified motion influence map representing the interaction characteristics within fish school was constructed. Then, on the basis of the constructed motion influence map, constructs a spatiotemporal vector, and sent to RNN to realize the recognition of the local unusual behaviors of fish school. Through the test on behavior dataset consisting of three typically local unusual behaviors, the performance of the presented method was verified with accuracy of 98.91%, 91.67% and 89.89% of detection, localization and recognition, respectively.

Different from the existing models, this paper aiming at the detection of whole behavior states of fish shoals (there are six states for the experiments of this paper) proposes a new model based on CNN model and spatiotemporal information fusion which combining the spatial information of fish distribution with the time information reflected in the movement energy. There is no feature extraction from spatial image and adopting it directly, and the L-K optical flow method is used to extract the optical flow energy map. The image of fish spatial distribution and optical flow energy map are spliced up and down to form a new image reflecting the spatiotemporal information. Then sent the spatiotemporal fusion image into a CNN for training, learning and testing to realize the recognition of different behavior states of fish shoals. Shown from the results of experiments of this paper, different behavior states of fish shoals can be recognized and classified by using this model with the total accuracy of 82.5%. Maybe the recognition accuracy is not very high at present stage, but it can show that the model is effective, and it can be improved by increasing the quantity and quality of image samples and the structure of CNN in the follow-up work. And through the work of this paper, we also further explore the following conclusions.

(1) Under different pressure environments, the group behavior states of fish shoals are quite different, which as shown in Figure 7 of this paper, six states of fish shoals can be distinguished either from the spatial distribution image or from the optical flow energy map. By combining the information of space and time (motion), behavior states of fish shoals can be distinguished and the mapping relationship with the pressure sources can be established.

(2) As shown by the training error curve in Figure 11 and experiments data in table 7 of this paper, by using CNN and training samples proposed by this paper, different behavior states of fish shoals can be recognized and classified effectively.

(3) Seen from the experiments data in table 5 and table 6 of this paper, for CNN, when using the small batch strategy to train the network, the Batch_size selects the value of less than 1 / 4 of the number of training image samples and a integer power of 2, the network training will obtain relatively high efficiency, which with less training time and higher test accuracy. In the early stage, with the increase of the number of Epoch, the error of neural network to the training samples will gradually decrease, and the classification accuracy of the test samples will gradually increase; when the network reaches a stable state, with the number of Epoch will increasing, although the error of the training samples will still decrease, but the classification accuracy of the test samples will not increase, or even decline, which indicates that the network has reached the over fitting state.

(4) This paper constructs a simple CNN composed of one convolution layer providing 20 convolution kernels with 9×9 size, one pooling layer adopting 2×2 average pooling strategy, and one full connected neural network with one hidden layer of 100 neurons. Based on the detection result of the test image samples shown as data in table 7 and the analysis below table 7 of this paper, it is found that the CNN has a higher discrimination power to the image edge feature

(the feature of spatial distribution) than the image gray-value feature (the feature of energy intensity).

(5) In this paper, visualization of the intermedia data of CNN is studied. Seen from the data visualization results shown as in Figure 14 and Figure 15, it is found that for the region information of individual fish body, or the optical flow energy intensity block, the network is convoluted to a negative value which is lower than the gray-value of background information; for the edge information of individual fish and the corner points of some optical flow energy intensity block, the network is convoluted to a positive value which is higher than the gray-value of background information. When adopting ReLU function as the neuron activation function, that is, discarding the negative value of the region information, and only sending the highlighted edge information to the full connected network for training, the network has the best stability and the highest discrimination accuracy of test samples. This is just in accordance with the conclusion of the fourth point. The CNN has a higher priority for edge features.

(6) In order to increase the influence of region information on neural network discrimination, the activation function of neuron is tested by Log-Sigmoid and Tan-Sigmoid respectively in this paper. Seen from the experiment results shown as in Figure 16, Figure 17, Figure 18 and Figure 19, it is found from the data visualization that the information sent to the full connected network by using Log-Sigmoid and Tan-Sigmoid functions includes the negative information that represents the content of the target region, but the network is either stagnant in training, or constantly oscillating, which is completely in a failure state. For the case studied in this paper, using the neuron activation function of ReLU function has the most effective experimental result.

In the follow-up research work, we can continue to carry out the following problems.

(1) The effectiveness of deep learning neural network is closely related to the quantity and quality of training samples. Network training needs a large number of sample data as support, but many public data sets do not have video data for variety behavior states of fish shoals. Therefore, in the research of fish shoals' behavior recognition and classification through deep learning, we need to collect more data of fish shoals' behavior states in various scenes as training samples to enhance the practicability of the algorithm.

(2) The spatial-temporal data fusion used in this paper is simply splicing the spatial distribution image and the optical flow energy image in the form of up and down. In the follow-up study, variety modes for spatial-temporal data fusion can be explored. For example, different weight coefficients are used to enhance or weaken some kinds of behavior state information; the fusion mode can be variety forms of up and down, left and right, weighted superposition, and so on. The influence of different information fusion methods on the discrimination efficiency of CNN can be made a further research and discussion.

(3) In this paper, the intermediate data of CNN is visualized. The result of data visualization in this paper is consistent with the discrimination result of neural network. However, whether the data of the intermediate process of CNN has visualization significance, whether the visualization analysis of data has guidance significance for the construction of network structure, and how to adjust the network parameters according to the visualization results, all of those need further discussions in the follow-up research work.

ACKNOWLEDGMENT

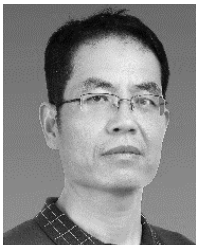
Acknowledge the support from School of Electrical and Electronic Engineering for providing them the experiment conditions.

REFERENCES

[1] G.-L. Luo *et al.*, "Application of computer vision in aquaculture," *Animal Husbandry Feed Sci.*, vol. 38, pp. 91–92, Dec. 2017.
[2] Y. Huang, J.-S. Zhang, and X.-B. Han, "Vision-based real-time monitoring on the behavior of zebrafish school," *Acta Scientiae Circumstantiae*, vol. 34, pp. 398–403, Feb. 2014.
[3] S.-J. Fu, Z.-D. Cao, and L.-Q. Zeng, *Fish Swimming*. Beijing, China: Science Press, 2014.
[4] Z. Cui, J.-F. Wu, and H. Yu, "A review of the application of computer vision technology in aquaculture," *Mar. Sci. Bull.*, vol. 20, pp. 53–66, Jan. 2018.
[5] J. He *et al.*, "Research progress on recognition and quantification of fish behavior in aquaculture based on computer vision technology," *Fishery Mod.*, vol. 46, pp. 7–14, Mar. 2019.
[6] S. Kato, M. Devadas, K. Okada, Y. Shimada, M. Ohkawa, K. Muramoto, N. Takizawa, and T. Matsukawa, "Fast and slow recovery phases of goldfish behavior after transection of the optic nerve revealed by a computer image processing system," *Neuroscience*, vol. 93, no. 3, pp. 907–914, Aug. 1999.
[7] Y. Ishibashi, H. Ekawa, H. Hirata, and H. Kumai, "Stress response and energy metabolism in various tissues of Nile tilapia Oreochromis niloticus exposed to hypoxic conditions," *Fisheries Sci.*, vol. 68, no. 6, pp. 1374–1383, Dec. 2002.
[8] T. Pinkiewicz, J. Purser, and R. Williams, "Estimating the swimming speed of salmon in aquaculture sea cages using computer vision," in *Proc. Australas. Aquaculture Conf.*, Aug. 2008, pp. 3–6.
[9] B. Sadoul, P. Evouna Mengues, N. C. Friggens, P. Prunet, and V. Colson, "A new method for measuring group behaviours of fish shoals from recorded videos taken in near aquaculture conditions," *Aquaculture*, vol. 430, pp. 179–187, Jun. 2014.
[10] X. Yu, X. Hou, H. Lu, X. Yu, L. Fan, and Y. Liu, "Anomaly detection of fish school behavior based on features statistical and optical flow methods," *Trans. Chin. Soc. Agricult. Eng.*, vol. 30, pp. 162–168, Feb. 2014.
[11] P.-R. Zhu *et al.*, "Learning-based zebrafish detection and tracking," *Comput. Appl. Softw.*, vol. 32, pp. 227–230 and 250, Sep. 2015.
[12] J. Zhao, Z. Gu, M. Shi, H. Lu, J. Li, M. Shen, Z. Ye, and S. Zhu, "Spatial behavioral characteristics and statistics-based kinetic energy modeling in special behaviors detection of a shoal of fish in a recirculating aquaculture system," *Comput. Electron. Agricult.*, vol. 127, pp. 271–280, Sep. 2016.
[13] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[14] M. Woźniak and D. Połap, "Object detection and recognition via clustered features," *Neurocomputing*, vol. 320, pp. 76–84, Dec. 2018.
[15] M. Woźniak and D. Połap, "Soft trees with neural components as image-processing technique for archeological excavations," *Pers. Ubiquitous Comput.*, pp. 1–13, Jan. 2020.
[16] D. Rathi, S. Jain, and S. Indu, "Underwater fish species classification using convolutional neural network and deep learning," in *Proc. 9th Int. Conf. Adv. Pattern Recognit. (ICAPR)*, Dec. 2017, pp. 1–6.
[17] R. Mandal, R. M. Connolly, T. A. Schlacher, and B. Stantic, "Assessing fish abundance from underwater video using deep neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–6.
[18] S. Marini, E. Fanelli, V. Sbragaglia, E. Azzurro, J. Del Rio Fernandez, and J. Aguzzi, "Tracking fish abundance by underwater image recognition," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 13748.
[19] D. A. Konovalov, A. Saleh, M. Bradley, M. Sankupellay, S. Marini, and M. Sheaves, "Underwater fish detection with weak multi-domain supervision," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 14–19.

[20] J.-Y. Xu, "Behavioral response of tilapia (Oreochromis niloticus) to acute stress monitored by computer vision," Ph.D. dissertation, Zhejiang Univ., Zhejiang Province, China, 2005.
[21] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.
[22] C. H. Antink, T. Singh, P. Singla, and M. Podgorsak, "Advanced Lucas Kanada optical flow for deformable image registration," *J. Crit. Care*, vol. 27, no. 3, p. e14, Jun. 2012.
[23] M. Woźniak, D. Połap, L. Kośmider, and T. Cłpa, "Automated fluorescence microscopy image analysis of pseudomonas Aeruginosa bacteria in alive and dead stadium," *Eng. Appl. Artif. Intell.*, vol. 67, pp. 100–110, Jan. 2018.
[24] J. Su, R. Miyazaki, T. Tamaki, and K. Kaneda, "High-resolution representation for mobile mapping data in curved regular grid model," *Sensors*, vol. 19, no. 24, p. 5373, Dec. 2019.
[25] K. Phil and W. Zou, *Deep Learning for Beginners: With MATLAB Examples*. Beijing, China: Beijing Univ. of Aeronautics and Astronautics Press, 2018.
[26] J. Bin, Z. Zi-Liang, H. Hao, Z. Yun-Peng, Z. Yong-Jian, and Q. Mei-Xia, "Automatic identification of WDMS spectra based on anti-Bayesian learning paradigm," *Spectrosc. Spectral Anal.*, vol. 39, pp. 1829–1833, Jun. 2019.
[27] K. Jia, S. Li, Y. Wen, T. Liu, and D. Tao, "Orthogonal deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 21, 2019, doi: 10.1109/TPAMI.2019.2948352.
[28] Y.-T. Fei, *Error Theory and Data Processing*. Beijing, China: Mechanical Industry Press, 2017.
[29] Z. Liu, X. Li, L. Fan, H. Lu, L. Liu, and Y. Liu, "Measuring feeding activity of fish in RAS using computer vision," *Aquacultural Eng.*, vol. 60, pp. 20–27, May 2014.
[30] J. Zhao, W. Bao, F. Zhang, S. Zhu, Y. Liu, H. Lu, M. Shen, and Z. Ye, "Modified motion influence map and recurrent neural network-based monitoring of the local unusual behaviors for fish school in intensive aquaculture," *Aquaculture*, vol. 493, pp. 165–175, Aug. 2018.

FANGFANG HAN was born in Chengde, Hebei, China, in 1978. She received the Ph.D. degree in instrument science from Tianjin University, Tianjin, China, in 2012. She is currently a Lecturer with the School of Electrical and Electronic Engineering, Tianjin University of Technology. Her research interests include computer vision, digital image processing, and artificial intelligence.



JUNCHAO ZHU was born in Kaifeng, Henan, China, in 1972. He received the Ph.D. degree in optical engineering from Tianjin University, Tianjin, China, in 2007. He is currently a Professor with the School of Electrical and Electronic Engineering, Tianjin University of Technology. His research interests include computer vision, photoelectric detection technology, and embedded measurement and control systems.

BIN LIU was born in Tianjin, China, in 1983. He received the Ph.D. degree in instrument science from Tianjin University, Tianjin, in 2010. He is currently an Associate Professor with the School of Electrical and Electronic Engineering, Tianjin University of Technology. His research interests include computer vision and photoelectric detection technology.

BAOFENG ZHANG was born in Tianshui, Gansu, China, in 1962. He received the Ph.D. degree in instrument science from Tianjin University, Tianjin, China, in 2002. He is currently a Professor with the School of Electrical and Electronic Engineering, Tianjin University of Technology. His research interests include test measurement technology, advanced testing methods, and intelligent instruments.

FUHUA XIE was born in Datong, Shanxi, China, in 1995. She is currently pursuing the degree major in control science and technology with the School of Electrical and Electronic Engineering, Tianjin University of Technology.