# Advancements, Challenges, and Future Prospects in Vision-Driven Animal Behavior Recognition Technology

Guanzhen Li

College of Information Science and Engineering, Shanxi Agriculture University, Taigu, China
20211210314@stu.sxau.edu.cn

**Abstract.** Animal behavior provides crucial insights for advancing ecological and animal husbandry research. With the advent of deep learning technologies, the accuracy and efficiency of animal behavior recognition have significantly improved. Initially reliant on manual feature extraction, the field has evolved to incorporate sophisticated deep learning models, transforming the approach to studying animal dynamics. This article reviews the current landscape of animal behavior recognition, discussing the principles, strengths, and limitations of prevalent models. It also examines the application of these models in diverse research contexts, from behavioral ecology to welfare assessments. The transition to automated systems offers nuanced behavioral insights, facilitating real-time monitoring and predictive analytics. Moreover, this paper addresses persistent challenges in animal behavior recognition, such as variability in environmental conditions and species-specific behaviors, proposing innovative solutions that leverage advancements in machine learning and computer vision. By envisioning future applications, this review underscores the potential of deep learning to revolutionize the understanding of animal behaviors, enhancing both animal welfare and ecological research. This comprehensive analysis not only charts a path forward for the field but also catalyzes new methodologies in animal behavior studies.

**Keywords:** Animal Behavior Recognition, 3D CNN, C3D.

## 1 Introduction

Animal behavior recognition leverages technological approaches to classify and analyze animal actions in specific settings, offering precise insights into their behavioral states. This understanding holds significant importance across various domains [1]. In the realm of animal husbandry, real-time monitoring of livestock health and environmental data is crucial for optimizing production efficiency. For instance, observing the aggressive tendencies in pigs, assessing lameness in cattle through their standing and walking postures, and evaluating the eating habits of horses by monitoring their interactions with feeding troughs are all vital practices. From an ecological perspective, tracking and analyzing the behavior of wild animals

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_55

provides valuable information on their activity ranges, migration patterns, and breeding habits [2]. Oceanographers, for example, use underwater acoustics to record marine life sounds, applying voiceprint recognition technology to distinguish between species and individuals, thereby aiding behavioral studies. Additionally, wildlife conservation efforts employ camouflaged cameras to monitor the movements and distribution of endangered species such as the Siberian tiger and giant panda, enabling more effective conservation strategies and ecological balance maintenance [3].

Traditionally, breeding and ecological research have relied heavily on manual observation to identify abnormal animal behaviors, a method that is inefficient, time-consuming, and highly subjective, necessitating a substantial level of expertise from the observers [4, 5]. With the advancement of deep learning technologies, these challenges in traditional ecological research and aquaculture are increasingly being addressed through computer vision. Notably, models such as Transformers and Slowfast have become prevalent in the field of animal behavior recognition. This paper begins by discussing standard theories in feature extraction and behavior classification techniques, outlining the current state of research, advantages, and limitations of animal recognition technologies, and examines the ongoing challenges in this field. The aim is to foster novel ideas and methodologies for future research [6].

## 2 Correlation Theory

### 2.1 Feature Extraction Method Anchored in Deep Learning

Feature extraction, also known as feature mapping, typically involves the generation of matrices through the convolution of an input image with a convolutional kernel. These matrices represent various features [7]. Recent advancements in deep learning have replaced manually defined feature image extractors and models with automated feature extraction layers, thus expanding the application scope of traditional machine learning models.

Unlike traditional feature extraction methods that depend on manual effort, deep learning-based methods utilizing convolutional neural networks (CNNs) can automatically identify and classify data through a multi-layer network structure and extensive input data or images. This approach offers high adaptability and robustness, making it widely used in image classification, object detection, and image segmentation tasks [8].

The evolution of deep learning can be traced back to 2012 when Alex Krizhevsky and colleagues introduced significant innovations, including deep network structures, ReLU activation functions, local response normalization, Dropout regularization, and the data-augmented AlexNet model. These contributions marked substantial progress in the field of computer vision [9]. In 2014, more profound CNN models, such as GoogLeNet and VGGNet, with enhanced feature extraction and representation learning capabilities, were proposed. These models started to be applied to behavior recognition tasks.

The Transformer model, introduced by Vaswani et al. in 2017, captures dependencies across various points in the input sequence using the self-attention mechanism. This model has since been incorporated into behavior recognition, offering new perspectives and methodologies for multimodal data fusion and sequence modeling [10]. Due to its proficiency in handling long sequence data and modeling complex interactions within sequences, researchers have adopted Transformers to tackle action recognition challenges. Since 2017, deep learning in behavior recognition has continued to evolve, with the integration of attention mechanisms, multimodal fusion, self-supervised learning, and increasingly complex network structures. These technological advancements continually propel the development of behavior recognition, leading to ever-improving performance.

**3DCNN.** For processing three-dimensional data, a deep learning network called a 3D convolutional neural network (3D CNN) was created., which generally consists of multiple 3D convolutional layers and pooling layers as well as fully connected layers for classification tasks [11]. The convolution kernel of the 3D convolution layer is three-dimensional, with depth (D), height (H) and width (W), and can slide on the three-dimensional input data. Assuming the size of the input data, convolution kernel, output feature map are:

$$\begin{aligned} D_{\text{out}} &= \frac{D_{\text{in}} - D_k + 2P_d}{S_d} + 1 \\ H_{\text{out}} &= \frac{H_{\text{in}} - H_k + 2P_h}{S_h} + 1 \\ W_{\text{out}} &= \frac{W_{\text{in}} - W_k + 2P_w}{S_w} + 1 \end{aligned} \tag{1}$$

Among them, $P_d, P_h, P_w$ are the padding in depth, height, and width, and $S_d, S_h, S_w$ are the strides in depth, height, and width. By sliding the 3D convolution kernel across the 3D input data, the 3D convolution layer creates a 3D feature map and performing dot product operations on each position [12].

As shown in Fig 1. 3D CNN can effectively process three-dimensional data. Unlike traditional two-dimensional convolut ion, where the convolution kernel only moves in two dimensions, the convolution kernel of 3D CNN can slide along three dimensions (width, height, and time). Fig 1depicts the three-dimensional convolutional neural network's convolution process. This structure enables 3D CNN to capture the spatial and temporal features in the data, thereby better understanding video and dynamic volume data. Compared with 2D CNN, 3D CNN has the ability to process more dimensional information and therefore requires more computing resources and more training data. Its advantage is that it can capture time-related features more accurately, thus performing better when processing time series data [13].

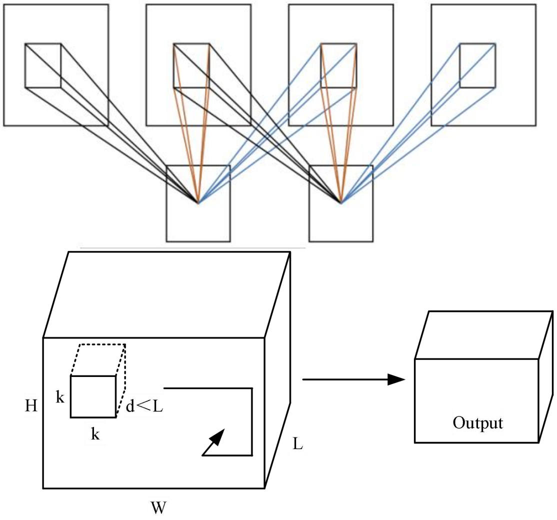

**Fig. 1.**3D CNN network structure diagram (Photo credit: Original).

**Slowfast.** SlowFast is a deep learning model for video understanding. Its main structure consists of a slow channel for processing spatial dimension information and a fast channel for processing temporal dimension information in the video. Such a structure ensures that the network can simultaneously obtain the spatial dimension information and the temporal dimension information in the video with great efficiency. The time and space information in the video can be effectively integrated through two different channels, and the task of action recognition can be completed efficiently [14].

In the Slowfast model, the slow flow is capable of capturing long-term temporal information in the video, while the fast flow can capture the short-term temporal information more quickly, so it has the advantage of strong spatial and temporal information capture ability. The combination of slow stream and fast stream can better adapt to video inputs of different lengths, but since the SlowFast model belongs to a dual-stream structure, it requires more computing resources and storage space, therefore, the cost of deployment and training is high, and training and adjustment also require more data to support.As shown in Fig 2.

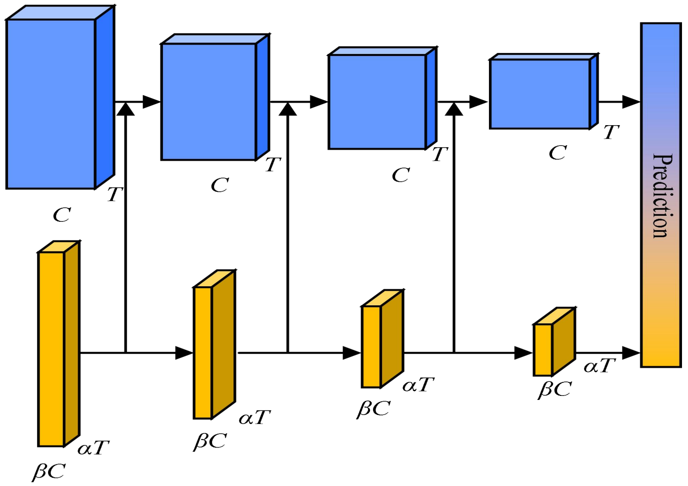

**Fig. 2.** Slowfast Model structure (Photo credit: Original).

**Transformer.** In 2017, Vaswani et al. proposed Transformer model, a deep learning model based on attention mechanism, which is used to process sequence-to-sequence. Compared with traditional recurrent neural network and long short-term memory network, Transformer does not depend on the cyclic structure, but is completely based on the attention mechanism to process the relationship between input sequence and output sequence, which is made up of two components, a decoder and an encoder, stacked on top of one another in numerous identical layers.

During recognition, the encoder maps the input sequence to a set of representations, and the decoder converts these representations into output sequences. In each layer, Transformer gathers crucial data from the input sequence using a multi-head self-attention technique and processes the position information of the sequence through position encoding. Between the encoder and the decoder, there is also a position encoder to process the positional relationship between the input sequence and the output sequence. Because there is no loop structure, Transformer can efficiently process different parts of the input sequence in parallel, and has the unique advantage of accelerating model training and reasoning.

### 2.2 Classification Methods

**Support Vector Machine.** Support vector machine, which also known as large margin classifier, is an algorithm for supervised learning applied to regression and classification tasks. The goal of SVM is to find a hyperplane that maximizes the interval between categories. Separate different categories of data by finding the optimal hyperplane in the feature space, thereby achieving the classification of new samples.

In essence, the approach determines the separation between two observations, or support vectors. The decision boundary that the SVM algorithm seeks is the boundary that maximizes its sample interval. In practice, the biggest advantage of SVM is that it can use nonlinear kernel functions to model nonlinear decision boundaries. However, SVM is a memory-intensive algorithm. Due to the significance of choosing the appropriate kernel function, it is also difficult to adjust the parameters and can not be extended to larger data sets. At present, in the industry, random forest is usually superior to support vector machine algorithm.

**C3D.** The network structure of C3D is illustrated. in Fig 3. The basic organization of its networks includes 8 convolution layers, 5 pooling layers, 2 fully connected layers and 1 Softmax output layer. In the network structure design, a 3 × 3 × 3 convolution kernel is used. To preserve temporal information and reduce computational complexity, first layer pooling kernel size is set uniquely to 1 × 2 × 2, with a 1 × 2 × 2 stride. In contrast, the other three-dimensional pooling kernels are typically set to 2 × 2 × 2 with a stride of 2 × 2 × 2. The C3D model is an improved classification model based on the three-dimensional convolutional neural network. It is mainly aimed at the classification of video information. It optimizes the selection of the size of the three-dimensional convolution kernel by changing the depth of the convolution kernel in different layers of the three-dimensional convolution network to achieve video classification tasks. After the convolutional layer, the pooling layer in the C3D model can effectively reduce the spatial dimension and time dimension of the feature map, thereby reducing the computational complexity of the model and extracting the most important behavior features. However, there is still a problem that the depth and number of convolution kernels are too large, which leads to a large amount of calculation and parameters, and the efficiency of training and reasoning is too low. At the same time, due to the limitation of network depth, the accuracy of behavior recognition is also low.As shown in Fig 3.

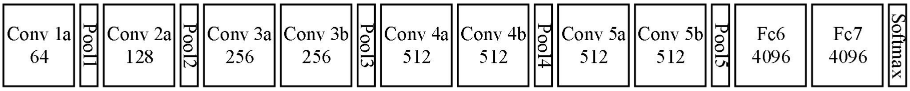

**Fig. 3.** C3D network structure diagram (Photo credit: Original).

# 3 Typical Method

## 3.1 Research Status of Behavior Recognition for Livestock

Li et al. divided the lactation behavior into three sub-behaviors: before sucking, sucking and ending sucking who introduced the SlowFast deep neural network model and embraced a construction with two streams. To achieve preliminary recognition of sow lactation behavior by extracting spatiotemporal features. Compared with other similar methods, The SlowFast model generates the most effectively when it comes to classifying fine-grained actions corresponding to sow lactation. In addition, when mixed with HMM, the fine-grained sow lactation recognition task can be effectively realized. Sequence consistency placement accuracy and behavior transition time can be as high as 90.51% and 87.05%.

Zhai Mingxin et al. established a data set by taking photos of dairy goats and annotating the pictures. After training on the C3 D and SlowFast models, the accuracy rates reached 70.38 % and 83.31 %.

Jie Liao et al. demonstrated a two-stream parallel network framework that utilizes MLMC as the model's input and combines CNN and Transformer. To categorize typical pig sounds, a convolutional neural network with four layers of convolution was applied. Transformer sequence encoding and CNN spatial feature representation were both utilized in the study, which has good robustness and generalization ability. The performance changes little for different input features. The accuracy, AUC and recall rate of domestic pig sound recognition reached 96.05%, 98.37% and 90.52% respectively.

### 3.2 Research Status of Behavior Recognition for Wild Mammals

Zhong Weifeng et al. constructed a video dataset containing nine types of macaque behaviors (MBVD-9), and based on this dataset, proposed a Transformer-enhanced SlowFast network for macaque behavior recognition (TAS-MBR). The average classification accuracy of macaque behavior reached 94.53%, effectively combining the Transformer model and the Slowfast model.

### 3.3 Current Status of Research on Insect Behavior Recognition

In order to control the migratory agricultural pest Spodoptera frugiperda, Yang LÜChun et al divided the female ovarian image of this pest into five forms, which is the color and texture of the ovaries, the length of the fallopian tubes, the number of eggs in the fallopian tubes, and the diameter of the eggs. The number of eggs in the ovaries was identified through the YOLOv4 algorithm. Finally, the five postures were classified using the SVM vector machine, and an information technology platform for accurately predicting the reproductive dynamics of adult insects was established. This automated way is capable of accurately predicting the population's reproductive dynamics of Spodoptera exigua, with an accuracy rate of 91% in determining the development stage of the female ovaries.

Zhou Tongzhou et al. trained three behaviors on C3D, I3D and X3D networks respectively for three problems of Drosophila rumination: behavior recognition, ruminant liquid extraction and trajectory tracking. Ingenious application of the concept of human behavior recognition to insect behavior recognition. The researchers applied the I3 D network to the recognition of fruit fly spit behavior, and the accuracy rate reached 96.3 %. Yolov5 + DeepSort was utilised concurrently to track the direction of the fruit fly spit.

And the accuracy rate of fruit fly detection was stable at 99.8 %.

Dong Tianyu first established a 1200-minute scientific record video of the citrus fruit fly, and divided the behavior of this insect into eight categories. A three-dimensional convolutional neural network with multi-scale feature fusion based on C3 D network is proposed. An additional channel is added to extract the shallow feature map for behavior classification, and the fusion of shallow feature information makes the utilization rate of the feature map significantly increased. The head combing, forefoot combing, midfoot combing, hindfoot combing, midfoot combing, wing combing, static rest and moving were identified, and the overall recognition accuracy reached 93.6 %.

# 4 Challenge Analysis

## 4.1 The Main Difficulties Existing in The Current Animal Behavior Algorithm

First of all, there is a lack of large-scale data sets in the field of animal recognition. From the perspective of research objects, researchers usually study animals of different ethnic groups. Compared with human pose estimation which have multiple large-scale labeled data sets, the wide variety of animal behavior recognition and the large differences in appearance and bones between species make it difficult to establish a data set that can contain most animal behaviors. From the perspective of animal behavior, when establishing data sets, animals often obstruct each other, leading to issues such as challenging recognition between animals, reduced image or video quality, and detection errors.

Secondly, the labeling cost of the data set is too high, and the behavior or characteristics of animals are diverse. For each sample, it may need to be labeled at multiple levels and angles, which increases the complexity and workload of labeling. For example, Zhong Weifeng et al. first divided macaques into group and monomer behaviors, and then processed the behaviors of nine macaques respectively, which undoubtedly greatly increased the workload of researchers. And because the quality of the data set annotation directly affects the training effect and performance of the model, the annotator needs to annotate the pictures and videos in the data set one by one, which is time-consuming and slow, and also puts forward higher requirements for the professional knowledge and annotation accuracy of the annotator. What is more, there may be significant differences in animal behavior and characteristics among different populations and individuals. To guarantee the algorithm's capacity for generalisation, different environmental conditions and different individual data may be required for processing.

Lastly, the models that are currently being employed in the field of animal recognition has inadequate interpretability and transferability. The fundamental reason is that most studies use self-built data sets for research, and self-built data sets are often collected under specific environmental conditions, resulting in poor generalization ability of the model in other environments. The self-built data set is usually constructed for specific tasks or specific problems, resulting in low applicability and generalization ability of the model on other tasks or other problems. From the perspective of researchers, self-built data sets may also be affected by sampling bias, uneven or incomplete sample distribution, so the performance of the model on other data sets may decline.

### 4.2 Potential Solutions

Try to build a large-scale data set. In the current field of animal recognition, there are many large data sets, such as the Animal-pose data set covering most domestic pets and livestock, mainly for the Tiger10 data set of large cats, including the zebra image data and the corresponding key point annotation zebra pose data set. However, these existing data sets are only for a single species.. The common data sets of animal posture are shown in Table 1.

The field of human behaviour recognition has achieved relatively outstanding strides, and the method of human behavior recognition can be used to participate in the field of animal behavior recognition. While continuously improving the algorithm and model, a hierarchical classification method can be used to decompose some complex actions into basic actions such as standing, walking, and forelimb bending. It can also learn from the methods in the field of human recognition. Digital enhancement technology can be used to increase the diversity of data and effectively improve the generalization ability of the model.

**Table 1.** Common data sets of animal posture.

| Dataset name | sample size | animal type |
|---|---|---|
| Animal pose | ———— | Cats, dogs, horses, cattle, sheep |
| Zebra pose | 2977 | zebra |
| Dairy goat behavior recognition dataset | 3000 | milch goat |
| MBVD-9 | 3000 | macaque |

The combination of multiple models can be used to identify animal behavior. According to what was previously discussed, Zhong Weifeng et al. implemented the Transformer model with the Slowfast model to identify macaque behaviour with more precision. Zhai Mingxin et al. first trained on a single C3D model, and only achieved an accuracy of 70.38 %. However, after combining the Slowfast model with the spatio- temporal and channel hybrid attention mechanism, the original data set can be trained to achieve an accuracy of 85.28 %.

## 5 Applicable Analysis

### 5.1 Animal Husbandry Field

Combined with sensor technology and Internet of Things technology, the development of animal recognition is conducive to the construction of intelligent breeding systems. Through real-time monitoring of animal behavior and environmental data, automatic management and remote monitoring of the breeding process can be achieved, improving breeding efficiency and reducing labor costs. For example, the application of some intelligent inspection equipment, instead of the traditional human inspection, can effectively identify the abnormal behavior of animals and respond in time, thus producing huge economic benefits. At the same time, by monitoring the activity patterns, eating habits and sleep conditions of animals, the signs of disease can be found in advance, and measures can be taken in time for prevention and treatment.

### 5.2 Ecology Field

Animal behavior recognition algorithms can provide data support for ecological research, help scientists understand the structure and function of ecosystems, explore the interaction mechanism between organisms and the environment, and provide scientific basis for ecosystem management and protection. Identifying and analyzing the behavior of wild animals can help us understand their range of activities, migration paths, breeding habits and other information, and help protectors to better formulate protection strategies, protect endangered species and maintain ecological balance.

## 6 Conclusions

Recent advancements in science and technology have significantly enhanced the prospects and research value of animal behavior recognition technology. Models originally developed for human behavior recognition, such as Transformers, Slowfast, and C3D, have been successfully adapted for animal recognition, achieving notable progress. However, due to the considerable differences in posture and behavior among species, most research in animal recognition relies on custom-built datasets to train models, which limits their generalizability and interpretability. Future research should focus on expanding dataset construction, refining algorithms, and optimizing models. Animal behavior encapsulates diverse information. From an ecological perspective, monitoring and intelligent identification of animal behavior can enhance animal welfare and provide timely access to critical information, such as migration patterns and breeding data. In the realm of animal husbandry, recognizing livestock behaviors allows farmers to take proactive measures to prevent issues like disease transmission, thereby improving production efficiency. Through ongoing technological innovation and methodological enhancements, animal behavior technology is poised to play a pivotal role in future ecological and breeding industries, continuing to advance more intelligent animal behavior monitoring systems.

## References

1. Broomé, S., et al. Going Deeper than Tracking: A Survey of Computer-Vision Based Recognition of Animal Pain and Emotions. Int. J. Comput. Vis. 131(2), 572–590 (2023).
2. Wei, C., et al. Masked Feature Prediction for Self-Supervised Visual Pre-Training. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2022).
3. Fan, H., et al. Multiscale Vision Transformers. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. (2021).
4. Yan, S., et al. Multiview Transformers for Video Recognition. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2022).

5. Li, B., et al. Recognition of Fine-Grained Sow Nursing Behavior Based on the SlowFast and Hidden Markov Models. Comput. Electron. Agric. 210, 107938 (2023).
6. Zhai, M. X. Research on Behavior Recognition Method of Dairy Goats Based on Three-Dimensional Convolutional Neural Network Northwest A&F University, (2023).
7. Liao, J., et al. Domestic Pig Sound Classification Based on Transformer CNN. Appl. Intell. 53(5), 4907–4923 (2023).
8. Zhong, W., et al. A Multi-Behavior Recognition Method for Macaques Based on Improved SlowFast Network. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi 40(2), 257–264 (2023).
9. Lü, C.-Y., et al. Accurate Recognition of the Reproductive Development Status and Prediction of Oviposition Fecundity in Spodoptera Frugiperda Based on Computer Vision. J. Integr. Agric. 22(7), 2173–2187 (2023).
10. Zhou, T., Zhan, W., & Xiong, M. A Series of Methods Incorporating Deep Learning and Computer Vision Techniques in the Study of Fruit Fly Regurgitation. Front. Plant Sci. 14, 1337467 (2024).
11. Dong, T. Y. Research on Intelligent Recognition Method of Combing Behavior of Drosophila Insects Based on Three-Dimensional Convolutional Neural Network. Yangtze University, (2023).
12. Wu, S., Wu, J., Cheng, G., et al. Research Progress on Animal Behavior Recognition Based on Pose Estimation. J. China Agric. Univ. 28(6), 22–35 (2023).
13. Guo, Y., Du, S., Qiao, Y., et al. Research and Application Progress of Deep Learning in Smart Livestock Breeding. Smart Agriculture (Chinese and English) 5(1), 52–65 (2023).
14. Li, X., Wang, L. ZeroI2V: Zero-Cost Adaptation of Pre-Trained Transformers from Image to Video. (2023)