Wang L & Wang H. (2025) FOOTBALL PASSING ACTION RECOGNITION METHOD BASED ON DEEP LEARNING. Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte vol. 25 (100) pp. 164-179.
DOI: https://doi.org/10.15366/rimcafd2025.100.011

ORIGINAL

FOOTBALL PASSING ACTION RECOGNITION METHOD BASED ON DEEP LEARNING

Lanfeng Wang [1], Haoyu Wang [2*]

[1]Sports College, Guizhou City Vocational College, Guiyang, 550025, China
[2]Sports Science and Technology College, Guangzhou Institute of Applied Science and Technology, 526072, China
E-mail: why020059@163. com

Recibido 14 de junio de 2024 **Received** June 14, 2024
Aceptado 14 de diciembre de **2024 Accepted** December 14, 2024

ABSTRACT

Video identification and analysis is one of the important research contents in the field of computer vision. Among them, the fine-grained action recognition of the video is a more refined and challenging recognition task. The main challenges are the few available fine-grained action recognition datasets that limit the progress of research in this field; fine-grained action-recognition is designed to distinguish subclasses in a large action classification, which are more subtle, usually only by small local differences. Existing fine-grained recognition tasks generally use target detection, attention mechanism and other related methods to find and use the local regional information in the image. However, most of these methods are used for image recognition tasks, so they lack the utilization of timing information for video. This paper uses dual-based method to study fine-grained video action recognition. A fine-grained football video dataset, Football is presented. It consists of live videos of multiple football matches. Initially, we categorized three broad movement types: dribbling, passing, and shooting. Subsequently, these were broken down into a more detailed set of 26 specific movements. All of the experiments presented in this paper will be implemented on this dataset. All methods in this paper will complete related experiments on the Football football dataset and the MPII cooking dataset. In the process of various network optimization, these methods achieve improved results and outperformed the current mainstream methods, which verifies the effectiveness of our methods.

KEYWORDS: Fine-Grained Video Action Recognition; Football Data Set; Two-Stream Network.

1. INTRODUCTION

Video identification and analysis is one of the important research contents in the field of computer vision. With the development of technology and the popularity of the Internet. Deep learning networks are developing rapidly (Fujishima et al., 2019). Identifying actions in a video has also become a task of high research value. At present, human action recognition technology has been widely used in intelligent video surveillance, virtual reality, smart home, sports analysis and many other aspects. Compared to image classification, the visual tonic action recognition task is more challenging. The video is composed of multiple frames of continuous images. It is dynamic. In the process of shooting, the light of the scene, the change of photography Angle, and the occlusion of moving objects will all bring difficulty to the video action recognition. And, because the video is shot for a long time. So, the content contained in the video is extremely rich. This also requires the video action recognition method to grasp the key information in the video. Ignore irrelevant, redundant information. In addition, the video also contains a temporal dimension, which can provide some useful clues to the identification process. But it also makes video recognition more difficult than image recognition. Beside Video action recognition requires more data. Early video datasets such as: contain several fixed actions. It is relatively easy to identify. Later, video datasets with more actions were released. They consisted of a more diverse variety of internet videos. The more diverse large datasets were released recently. It is composed of hundreds of thousands of videos. In total, hundreds of types of actions are included. Release of these datasets. The research and development of action recognition task provide great help. In video action recognition, fine-grained action recognition is a more refined recognition task. for instance. Coarse-grained motion recognition can distinguish between coarse-grained physical activities such as running, soccer, swimming, and skiing. But you can't accurately identify the dunk and layup movements. Fine-grained action recognition is proposed based on such situations. In contrast to the coarse-grained action recognition of the human population vs. Fine-grained action recognition aims to distinguish between subclasses within a large action classification where inter-class differences are more subtle. Differ distinction can only base on local differences. Therefore, fine-grained video action recognition is even more challenging.

2. Research Results at Home and Abroad

2.1. Current Status of Action Recognition Based on Deep Learning

Over the last few years. With the development of deep convolutional neural networks, the study of fine-grained action recognition has entered a new stage. Many two-dimensional convolutional neural network methods have been applied to action recognition. Among them, the proposal of double-flow (Two-

stream) network provides a great help for the study of action recognition. Two-stream networks are divided into temporal and spatial flow networks. Spatial flow network is an image classification architecture. It acts on a single-frame image. The time flow network acts on the dense light flow of multiple frames. Finally, the two networks need to be fused in the later stage. In an effort to capture extensive temporal dynamics, researchers have introduced the Temporal Segmentation Network (TSN). This framework encompasses dual pathways: one for temporal dynamics and another for spatial dynamics. Distinct from the dual-stream architecture, the TSN utilizes the aggregated continuous optical flow fields as inputs, rather than relying on individual frames or combined frame and optical flow sequences. Following the acquisition of individual prediction outcomes for each video segment, a subsequent integration step is necessary. The integrated outcomes from this process then serve as the conclusive video predictions (Quan et al., 2023). As we all know, for video action recognition. It is very important to learn the global description of the video time information changes. LSTM can use storage units to store, modify, and visit internal states to obtain a long range of time relationships. Based on the method. Scholars have proposed to use the feature pooling network and LSTM for video action recognition, respectively (Dang et al., 2020). Because the two-dimensional approach lacks information on the time dimension. Many other methods have proposed 3 D network structure, 3 D convolutional neural network extra 1 D representation is the time dimension. Such three-dimensional structures collectively consider both temporal and spatial dimensions. Enables the network to directly learn the spatiotemporal features. The C3D is a good feature extractor. It converts 2-dimensional convolution and pooling into 3-dimensional structures to construct a 3-dimensional network experimentally. The authors conclude that the convolution kernel of 333 is the best architecture. It is composed of multiple P3D Blocks. Three structures were designed: P3D-A, P3D-B and P3DC. Scholars expand the two-dimensional convolutional neural network model into three-dimensional. Increase the time dimension in all convolutional kernels as well as in the pooling layer. at the same time. The authors found that the use of shuangliu is still valuable (Chenyang et al., 2022). So, you can train the network on both images and optical streams, and then fuse the two streams at a later stage. In addition, in order to further improve the recognition performance (Khan & Ammar Taqvi, 2023; Sung et al., 2015).

2.2. Current Status of Fine-Grained Action Recognition Based on Deep Learning

In contrast to the coarse-grained actions. The SNR for fine-grained actions is very small. Usually, characteristic information with discrimination can only be obtained in tiny local regions. therefore. The key to the success of the fine-grained action recognition algorithm is: how to find the useful local area information in the video. Many earlier fine-grained methods rely on artificial bounding box annotation (Wang, 2023). bounding boxes can help exclude

redundant irrelevant information. But because manual labeling takes too much effort. In recent years, a major research trend in the fine-grained action method is to rely solely on category labels for classification. Due to the advances in deep learning research, information is annotated without the aid of artificial bounding boxes. Can also get a better identification performance (Reddy et al., 2021; Shaath et al., 2022). For instance. Part-based R-CNN. uses a bottom-up candidate region algorithm proposed by Zhang et al. The Pose Normalized CNN method proposed by BransonS et al. A pose alignment operation was performed on the images using a prototype. Both methods require the use of additional artificial bounding box annotation information (Wang, 2023). This information is generally very difficult to obtain (Jiao et al., 2021; Papernot et al., 2018). The following methods no longer need other manual annotation, but only use category labels to complete the identification task. Therefore, the researchers proposed only category labels for classification. Xiao et al proposed a two-level attention algorithm. The algorithm achieves attention from two different levels: object (Object-Level) and local (Part-Level). Obtain different levels of feature information. Lin et al. proposed to use Bilinear CNN to better represent deep convolutional features. Fu et al. proposed that RA-CNN. uses APN network (Attention Proposal Network) to locate key local regions, further improving the identification performance through the combination of classification network and APN network (Aceto et al., 2019; Rong et al., 2023). These methods all show that the combination of global and local information is the key to improve the performance of fine-grained identification. The method of this paper will also focus on the extraction of local features and by incorporating some other techniques. Constantly improve the identification efficiency (Sharma et al., 2023).

3. Building of an Action Recognition Model Based on Deep Learning

3.1. Convolutional Neural Network

Convolutional neural network structure generally includes the following components: (1) Convolutional layer. A convolutional layer is composed of various convolutional filters designed to capture distinct features from the input dataset. The dimensions of each filter dictate the extent of the receptive field it covers. During the application of convolutional processing, these operations are performed in succession across the image data, generating an output matrix that is subsequently passed to the subsequent layer (Fan, 2020). (2) pooling layer. The pooling layer was first proposed by LeNet, which is called down sampling (Sub-sample) and named by pooling (Pooling) since Alex Net. After the pooling layer receives the features extracted by the convolutional layer, they are selected and filtered. (3) Activate the function layer. In order that neural networks can solve more complex problems, some nonlinear functions can be introduced as excitation functions. (4) fully connected layer. In the CNN structures, the fully connected layer acts as a classifier. Figure 1 shows a simple

structural diagram of a linear fully connected neural network.

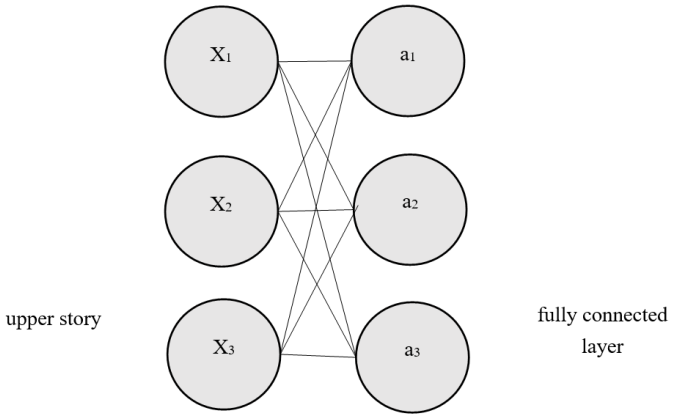

Figure 1: Structural Diagram of the Simple Linear Fully Connected Neural Network

The operational relationship is as the formula:

$$a_1 = W_{11} * x_1 + W_{12} * x_2 + W_{13} * x_3 + b_1$$

$$a_2 = W_{21} * x_1 + W_{22} * x_2 + W_{23} * x_3 + b_2$$

$$a_3 = W_{31} * x_1 + W_{32} * x_2 + W_{33} * x_3 + b_3$$

among, x_1, x_2, x_3For input, a_1, a_2, a_3For output.

3.2. Recurrent Neural Network

As shown in Figure 2, a simple RNN structure consists of the input, hidden, and output layers.

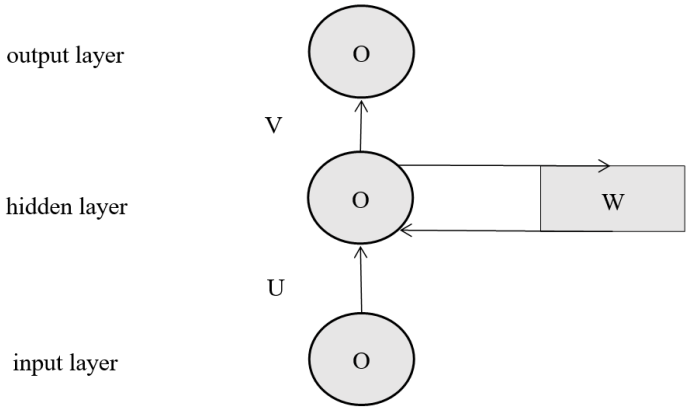

Figure 2: Simple RNN Structure

The problem of gradient explosion is easy when using recurrent neural networks. The researchers found that using a "gating mechanism" could effectively solve the problem. Long, short-term memory network (LSTM) is one

of the ways to use this mechanism. The "gating mechanism" of LSTM can selectively allow some information through to add or delete the information in the neuronal state. Also, it uses a Sigmoid layer called the "forgetting gate" to get information that needs to be forgotten.

3.3. Network Structure Algorithm

Double-stream (Two-stream) includes temporal and spatial flows, with RGB images and optical flow images, respectively. In order to obtain the optical flow image, the optical flow method is needed to calculate. Here is the basic principle of the optical flow method. First, a basic assumption is required to be constructed: (1) the brightness will not change with the movement of objects; and (2) the movement of objects between adjacent frames of the video is "small movement". Then, establish the basic constraint equation:

$$I(x, y, t) = I(x + dx, y + dy, t + dt)$$

3.3.1. NTS-Net

In fine-grained videos, informative areas are often hidden in complex backgrounds and irrelevant things. To extract these valid features, we referred to the NTS-Net to extract discriminative features in fine-grained actions by mapping to local regions. When using NTS-Net, it is not necessary to use fine-grained bounding box annotation information, which is located to regions with information in a self-supervised way. Finally, as shown in Figure 3, we fused NTS-Net into a two-stream network for action recognition from RGB images and optical flow images, respectively (Liu et al., 2021).

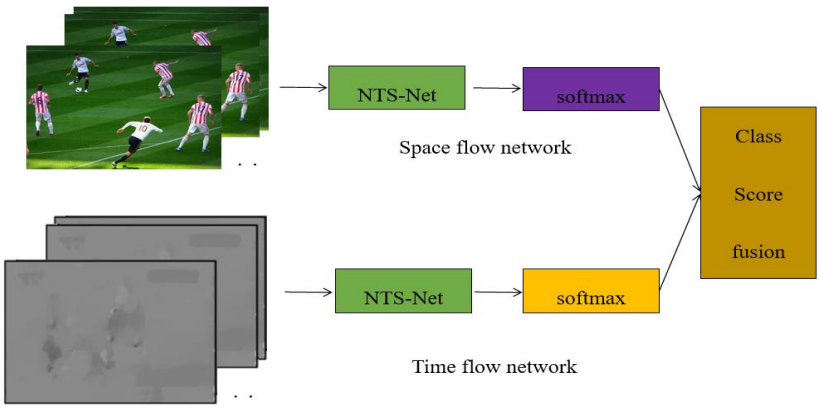

Figure 3: Two-Stream Network Structure

As shown in Figure 3, the NTS-Net contains three components: a screening network, an optimized network, and a checking network. The screening network is primarily responsible for focusing the model to regions with critical information. The optimization network is primarily responsible for evaluating the specific region obtained from the screening network and giving feedback to it. Finally, the inspection network is mainly responsible for using the

obtained effective region information for fine-grained action recognition.

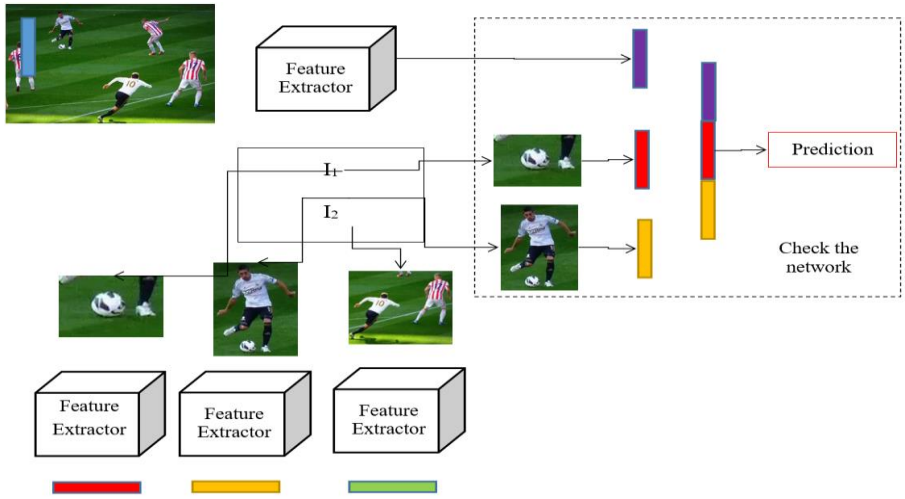

Figure 4: N T S-Net Spatial Flow Network Structure

3.3.2. Boundary Area Screening Algorithm

Filening network selection uses region generation network (RPN) to obtain local bounding box regions. Area generation network (RPN) is the core of Faster R-CNN (Nagarajan et al., 2023), which is an improved method of sliding window algorithm and a target detector. (1) Head: Generate the anchor. anchor is a pixel on an image feature map, and each anchor is centered on one anchor based on different width and scaling ratios. (2) Middle part: including classification branches and bounding box regression branches. (3) Tail: Screening for anchor. First, the transgressive anchor is removed, and then the high-repetition bounding boxes are removed using the non-maximum suppression (NMS) algorithm. Non-maximum suppression (NMS) is a method used to suppress elements that are not a maximum. From the regional generation network mentioned above, we know that multiple bounding boxes are obtained after network training, and usually we need to keep an optimal bounding box. Therefore, the NMS can be used to exclude redundant bounding boxes. The specific implementation steps are as follows:

(1) Use a function to sort all the bounding boxes, and then select the bounding boxes with the highest score; (2) Compare the remaining bounding box with the highest score boundary box and calculate the coincidence area (IOU) between the two. If the IOU exceeds a certain threshold, the boundary box will be removed and retained; (3) Repeat the above operations for the untreated bounding boxes. The specific screening process for the boundary screening algorithm presented in this chapter is as follows. First in the screening network, we used the RPN to generate a series of candidate regions containing information. Sliding-window methods are often used to predict multiple selected regions simultaneously. The size of each anchor is related to the aspect ratio as well as the scaling ratio, which is located in the center of the

sliding window. Figure 3-5 shows our anchor design process for the candidate information region. We used three aspect ratios and scaling ratios. For each image of size 224,224, we used the three aspect ratios that are (1:2,3:3,3:2) and the three scaling ratios that are (24,48,96). The filtering network generates a list of all anchor information. We then rank the list from high to low according to the amount of information in this list, and select the highest top M information regions after maximum suppression (NMS). We represent the informative quantities of these several regions as $<I_1,I_2,\ldots I_M>$. ($I_1 \geq I_2 \geq \ldots \geq I_M$) Each frame in the video is processed by the above filtering network, yielding a series of information areas. Immediately after, these regions were passed into the optimization network to obtain the confidence scores, and they were ranked and expressed as $<C_1,C_2,\ldots,C_M>$ ($C_1 \geq C_2 \geq \ldots \geq C_M$)。 Finally, using the resulting two sequences of $<I_1,I_2,\ldots Iw>$ and $<C_1, C_2,\ldots,C_M>$, The screening network was optimized by judging the consistency of the two sequences.

3.3.3. Loss Function Design

The following outlines the distinct loss functions employed across the three constituent networks of the NTS-Net, along with a description of their respective optimization strategies.

(1) Screening the loss: Within the filtering network, we have designated M informative areas, denoted as $R= \{R_1, R_2, \ldots, R_M\}$ each with its informational content quantified by $I= \{I_1,I_2,\ldots,I_M\}$. Concurrently, within the refining network, the level of confidence for these M areas is denoted by $C= \{C_1, C, \ldots C_M\}$. Subsequently, the loss function for the filtering network is formulated as follows:

$$L_I(I.C) = \sum_{(i,s):C_i<C_s} f(I_s - I_i)$$

(2) Optimize the loss: We define the loss Lc of the optimized network as:

$$L_C = -\int_{i=1}^{M} logC(R_i) - logC(X)$$

Where, C is the confidence function, which represents the probability that the specified region is the true label class. The first term of the formula above represents the sum of cross-direct loss in all areas, and the second term shows cross-direct loss in the complete image.

(3) Check the loss: When the screening network obtained the most informative K regions $\{R_1,R_2,\ldots,R_x$ When, checking the network results in a fine-grained identification result = S (X, $R_1,R_2,\ldots,R_k$)。 We use the cross-direct loss as the classification loss, and its expression is as follows:

$$L_S = -logS(X, R_1, R_2, \cdots, R_K)$$

(4) Joint loss: Ultimately, we aggregate the diverse loss components to facilitate integrated training. The comprehensive loss function, which encapsulates all these elements, is formulated as follows:

$$L_{total} = L_1 + \lambda \cdot L_S + \mu \cdot L_C$$

In this experiment, we all set them to 1. We finally performed by using stochastic gradient descent methods to optimize L_{total}。

4. Football Passing Action Recognition Test Based on Deep Learning

4.1. Introduction to Dataset

Our dataset, which focuses on football, consists of recordings from various football games, amounting to a total of roughly 8 hours of footage. Captured from multiple perspectives, these videos document the movements of individual players throughout the matches. However, due to the expansive venues, presence of spectators, and other obstructions, a significant portion of the footage includes extraneous elements. During the annotation phase, we initially identified three broad categories of movements: dribbling, passing, and shooting. These were then categorized into 26 more specific actions, as detailed in Table 1. In total, our dataset encompasses 3399 detailed annotations, averaging to roughly 130 instances per action category. For every instance of an action, we have recorded the timestamps for the beginning and conclusion, along with the corresponding labels.

Table 1(a): The Basketbal1 Football Dataset

THE COARSE-GRAINED CATEGORY	ID	FINE-GRAINED CATEGORIES	SAMPLE SIZE	AP （%）	CATEGORIES THAT ARE EASILY CONFUSED
DRIBBLE	1	Behind-the-Back Dribble	119	75. 00	6,24
	2	Change the Foot Dribble	71	9. 09	7
	3	Do not Change the Foot Ball	29	0. 00	2,7
	4	The Front Unchanged to the Foot Ball	15	0. 000	3,6
	5	Cross-Leg Dribble	151	4. 35	6,7
	6	Change the Foot Dribble	717	14. 81	5,7

Table 1(b): The Basketbal1 Football Dataset

THE COARSE-GRAINED CATEGORY	ID	FINE-GRAINED CATEGORIES	SAMPLE SIZE	AP （%）	CATEGORIES THAT ARE EASILY CONFUSED
	7	High dribble	270	9. 76	6,14
PASS	8	Foot is passing the ball	16	40. 00	7,14
	9	One-foot rebound pass	78	4. 17	7,14
	10	Pass on one foot across the chest	25	0. 00	7,14,23
	11	Single-foot side pass	170	3. 92	6,9,11,14
	12	Backball with both feet	91	7. 14	7,14
	13	chest pass	535	21. 12	7,12,15,21
	14	head pass	229	7. 25	11,14
SHOOT	15	follow shot	12	25. 00	18,21
	16	penalty shot	114	97. 14	6
	17	pivot shot	65	90. 00	23
	. . .	. . . .	. . .	. . .	. . .
	26	pivot shot	65	90. 00	23

4.2. Introduction to the MPII Cooking Dataset

In addition to the Basketbal1 football dataset presented in subsection 3. 3. 1, we performed experiments on another existing fine-grained action recognition dataset. The MPII Cooking dataset is a fine-grained video dataset for recording cooking activity using a fixed camera in the kitchen. The 12 people involved in the shooting performed 65 different cooking activities, such as slicing, taking materials, pouring wine, etc.

The dataset contains 44 videos with a total length of over 8 hours. To assess the efficacy of our suggested approach and contrast it against prevailing registration techniques, we employ the mean Average Precision (mAP) as our metric of evaluation. First of all, I will introduce the following indicators: 1) True Positives (TP): Instances where the model correctly identifies positive samples as positive; 2) True Negatives (TN):

Cases where the model correctly identifies negative samples as negative; 3) False Positives (FP): Instances where the model incorrectly classifies negative samples as positive; 4) False Negatives (FN): Cases where the model incorrectly classifies positive samples as negative. As depicted in Table 2, the value 1 signifies the presence of the positive class, while 0 indicates the absence, representing the negative class.

Table 2: Evaluation Indicators

		CALCULATE		
		1	0	Amount to
REALITY	1	TP	FN	TP+FN
	0	FP	TN	FP+TN
AMOUNT TO		TP+FP	FN+TN	TP+FN-+FP+TN

The accuracy (precision) and recall (recal1) are calculated in the following formula:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FP}$$

According to this formula, the accuracy rate is basically inversely proportional to the recall rate. The high accuracy rate and the recall rate will be very low, and the recall rate will be very high. You can use the average accuracy rate (Average Precision, AP), which has the formula:

$$AP = \int_0^1 p(r)dr$$

The integral approaches the formula:

$$AP = \sum_{k=1}^{N} P(k)\Delta r(k)$$

Mean Average Precision (mAP) is the mean of the AP values for all categories.

4.3. Interpretation

Our approach involves employing a variety of standard 2D network architectures, such as LSTM, dual-stream networks, and TSN, as comparative benchmarks for our novel network design. Initially, the network underwent pre-training on the ImageNet dataset, followed by fine-tuning to accommodate our Football dataset and the MPII cooking activities dataset. (1) The Football football dataset. For the conducted experiment, a dataset comprising 2364 samples was allocated for the training phase, while 1035 samples were designated for the evaluation phase. Initially, we will conduct training and testing for the three broad action categories: dribbling, passing, and shooting. Tables 3-5 present the mean Average Precision (mAP) scores, with the final

column displaying the mAP for the three primary categories and the central column detailing the mAP for the 26 细分类别. Analysis of the data indicates that within the various action recognition techniques, TSN outperformed in both temporal flow and integrated networks, whereas our network, which leverages local features, demonstrated superior performance on RGB data. Overall, these techniques exhibit a high degree of accuracy in recognizing the three general ball movements, with an average mAP of approximately 80%. Subsequently, our focus shifts to examining the performance of various methods across the 26 detailed action categories. Table 3 illustrates that our region-based screening method outperforms other networks on RGB data, with a nearly 9% increase in mean Average Precision (mAP) compared to the top-performing TSN model in our benchmarks (27. 40% compared to 19. 06%). Nevertheless, in terms of optical flow analysis, TSN maintains its superiority. Synthesizing the findings from both coarse-grained and fine-grained recognition analyses, it is evident that methods leveraging local features exhibit superior accuracy in processing RGB images. This enhanced performance may be attributed to the fact that RGB images are capable of preserving more pertinent regional details than optical flow images.

Table 3: Experimental Results of the Football Datasets

METHOD	26 SUBCLASS MAP (%)	THREE MAJOR CATEGORIES OF MAP (%)
CNN+LSTM	18. 26	72. 52
BILINEAR CNN	21. 63	76. 96
TWOSTREAM-RGB	17. 49	70. 03
TSN-RGB	19. 06	72. 71
OURS-RGB	27. 40	78. 10
TWOSTREAM-F1OW	23. 03	77. 38
TSN-F1OW	26. 39	86. 77
OURS-F1OW	19. 89	76. 50
TWOSTREAM-FUSI ON	24. 01	77. 86
TSN-FUSION	26. 70	87. 13
OURS-FUSION	27. 23	80. 44

(2) The MPII cooking dataset. In this experiment, there were 3886 instances for training and 1723 instances for testing. As can be seen from Table 4, our proposed region-based screening method is the best for RGB images, obtaining 33. 79% mAP, and it is 4. 56% higher than the spatial flow network of dual-flow structure and TSN, respectively (33. 79% vs. 29. 23%) and 2. 0% (33. 79% vs. 31. 79%). For optical flow images, the TSN network achieved the best results, with 33. 10% mAP obtained. After merging the two networks, the TSN fused best with a final mAP of 33. 98%, while our region-based screening network performed almost consistently with the TSN fusion, only 0. 05% lower.

Table 4: Experimental Results of the M P I I Cooking Dataset

METHOD	MAP (%)
CNN+LSTM	30. 34
BILINEAR CNN	31. 46
TWOSTREAM-RGB	29. 23
TSN-RGB	31. 79
OURS-RGB	33. 79
TWOSTREAM-F1OW	32. 16
TSN-F1OW	33. 10
OURS-F1OW	30. 16
TWOSTREAM-FUSION	32. 19
TSN-FUSION	33. 98
OURS-FUSION	33. 93

From the experimental results of the above two datasets, TSN works better on optical flow, while the region-based screening method works better on RGB, so we can consider integrating the two methods in space and time in order to achieve better results. Underpinned by this hypothesis, we conducted subsequent experiments employing linear weighted averaging to determine the combined outcomes for the networks. The spatial flow network was assigned a weight of 0. 4, while the temporal flow network received a weight of 0. 6. Table 5 presents a comparative analysis of our outcomes against those of the two preceding methodologies. It is evident that the integrated approach enhanced the performance across both video datasets; the mean Average Precision (mAP) for the Football dataset rose from 27. 40% to 29. 78%, and for the MPII cooking dataset, it increased from 33. 98% to 34. 78%.

Table 5: Experimental Results of the M P I I Cooking Dataset

DATA SET	METHOD	MAP (%)
FOOTBALL DATASET	TSN-RGB	19. 06
	TSN-F1ow	26. 39
	TSN-Fusion	26. 70
	Ours-RGB	27. 40
	Ours-F1ow	19. 89
	Ours-Fu sion	27. 23
	Ours-RGB+TSN-F1ow	29. 78
THE MPII COOKING DATASET	TSN-RGB	31. 79
	TSN-Flow	33. 10
	TSN-Fusion	33. 98
	Ours-RGB	33. 79
	Ours-F1ow	30. 16
	Ours-Fusion	33. 93
	Ours-RGB+TSN-F1ow	34. 78

However, when evaluating the aggregate outcomes, it is evident that the precision of fine-grained video action recognition remains relatively low. Specifically, for the Football dataset, the recognition accuracy for the 26 fine-grained actions is significantly lower than that for the three broader action categories, indicating the complexity of fine-grained recognition tasks. Table 5's fifth column presents the Average Precision (AP) values for each action class following the fine-grained identification using the aforementioned fusion network. The table reveals substantial variations in AP within the same action class. For instance, within the shooting class, the AP for free throws is as high as 97. 14%, whereas for short-stop jumpers, it is a mere 18. 52%. The final column identifies the most commonly mistaken subcategories during the identification process for each subclass. It is observed that fine-grained subclasses often confuse one another within the same broader class. As previously noted, this confusion arises due to the subtle differences between fine-grained actions. Consequently, effectively extracting distinctive features from complex backgrounds is crucial for accurately differentiating between similar actions.

5. Conclusion

Video identification and analysis is one of the important research contents in the field of computer vision. In the action recognition task, the fine-grained action recognition of videos is a more refined recognition task. The researchers have proposed many methods to solve various problems in action recognition. Among them, the mainstream networks have dual-stream and three-dimensional-based convolutional neural network structures. The dual-flow method is used for regional screening. Since there are relatively few datasets about fine-grained videos, this paper presents a new fine-grained football video dataset Football and completes relevant experiments on this dataset. Based on the structure of the two-stream network, we mainly propose the following methods. In order to obtain informative key regions in fine-grained video segments, we propose a dual-stream method based on region screening to locate the informative regions and extract discriminative properties in fine-grained actions. However, this method only uses high-level features for prediction, and does not adopt low-level features. And considering that the attention mechanism can effectively obtain key information, we then propose the following two methods. The first establish a feature pyramid based on the attention mechanism of spatial domain. It uses multi-scale fusion technology and can use different feature scales of the image to locate to different local regions. Through this feature pyramid, the model can learn the local regional features with key information. The second is based on the attention mechanism of the channel domain, using a multi-scale channel attention module to study the interdependence between the channels to improve the network performance. Channel attention is achieved across multiple scales by varying the size of the spatial pooling. Finally, the features are fused through the

attention feature fusion module.

REFERENCES

Aceto, G., Ciuonzo, D., Montieri, A., & Pescapé, A. (2019). Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges. *IEEE Transactions on Network and Service Management*, *16*(2), 445-458.

Chenyang, D., Jianzhong, F., Bin, Q., Linyan, B., & Panpan, W. (2022). Jujube garden detection and recognition in GF-6 image using deep learning. *Bulletin of Surveying and Mapping*(3), 54.

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), 483.

Fan, T. (2020). Research and realization of video target detection system based on deep learning. *International Journal of Wavelets, Multiresolution and Information Processing*, *18*(01), 1941010.

Fujishima, M., Narimatsu, K., Irino, N., Mori, M., & Ibaraki, S. (2019). Adaptive thermal displacement compensation method based on deep learning. *CIRP journal of manufacturing science and technology*, *25*, 22-25.

Jiao, L., Zhang, R., Liu, F., Yang, S., Hou, B., Li, L., & Tang, X. (2021). New generation deep learning for video object detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(8), 3195-3215.

Khan, N., & Ammar Taqvi, S. A. (2023). Machine learning an intelligent approach in process industries: A perspective and overview. *ChemBioEng Reviews*, *10*(2), 195-221.

Liu, Q., Pang, B., Li, H., Zhang, B., Liu, Y., Lai, L., Le, W., Li, J., Xia, T., & Zhang, X. (2021). Machine learning models for predicting critical illness risk in hospitalized patients with COVID-19 pneumonia. *J Thorac Dis*, *13*(2), 1215.

Nagarajan, J., Mansourian, P., Shahid, M. A., Jaekel, A., Saini, I., Zhang, N., & Kneppers, M. (2023). Machine Learning based intrusion detection systems for connected autonomous vehicles: A survey. *Peer-to-Peer Networking and Applications*, *16*(5), 2153-2185.

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). Sok: Security and privacy in machine learning. 2018 IEEE European symposium on security and privacy (EuroS&P),

Quan, Y., Liu, A.-f., Li, C., Jin, Y.-x., Liu, Y.-e., Zhou, J., Yin, X.-x., Li, X., Jin, M., & Lv, J. (2023). Factors influencing the rate of residual stenosis in athletic patients after endovascular intervention for symptomatic carotid artery stenosis. *Revista multidisciplinar de las Ciencias del Deporte*, *23*(89).

Reddy, D. S., Rajalakshmi, P., & Mateen, M. (2021). A deep learning based approach for classification of abdominal organs using ultrasound images. *Biocybernetics and Biomedical Engineering*, *41*(2), 779-791.

Rong, J., Qi, L., Wu, H., & Chen, S. (2023). Energy Efficiency Evaluation of HVDC Converter Station Based on Deep Neural Network Model. *Applied Mathematics and Nonlinear Sciences*, *8*(2), 2003-2012.

Shaath, H., Vishnubalaji, R., Elango, R., Kardousha, A., Islam, Z., Qureshi, R., Alam, T., Kolatkar, P. R., & Alajez, N. M. (2022). Long non-coding RNA and RNA-binding protein interactions in cancer: experimental and machine learning approaches. Seminars in cancer biology,

Sharma, N., Haq, M. A., Dahiya, P. K., Marwah, B., Lalit, R., Mittal, N., & Keshta, I. (2023). Deep Learning and SVM-Based Approach for Indian Licence Plate Character Recognition. *Computers, Materials & Continua*, *74*(1).

Sung, Y.-T., Chen, J.-L., Cha, J.-H., Tseng, H.-C., Chang, T.-H., & Chang, K.-E. (2015). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior research methods*, *47*, 340-354.

Wang, Z. (2023). Intelligent recommendation of open educational resources: Building a recommendation model based on deep neural networks. *International Journal of Advanced Computer Science and Applications*, *14*(6).