Hindawi
Discrete Dynamics in Nature and Society
Volume 2023, Article ID 9808134, 1 page
https://doi.org/10.1155/2023/9808134

Retraction

Retracted: Research on Human Motion Analysis in Moving Scene Based on Timing Detection and Video Description Algorithm

Discrete Dynamics in Nature and Society

Received 22 August 2023; Accepted 22 August 2023; Published 23 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Q. Shen and S. Ye, "Research on Human Motion Analysis in Moving Scene Based on Timing Detection and Video Description Algorithm," *Discrete Dynamics in Nature and Society*, vol. 2021, Article ID 4320846, 10 pages, 2021.

Hindawi
Discrete Dynamics in Nature and Society
Volume 2021, Article ID 4320846, 10 pages
https://doi.org/10.1155/2021/4320846

Research Article

Research on Human Motion Analysis in Moving Scene Based on Timing Detection and Video Description Algorithm

Quanping Shen[1] and Songzhong Ye[2]

[1]*Minjiang Normal College, Fuzhou 350018, Fujian, China*
[2]*Sports Industry Development Research Center, Fujian Jiangxia University, Fuzhou 350108, Fujian, China*

Correspondence should be addressed to Songzhong Ye; ysz97@fjjxu.edu.cn

Received 3 November 2021; Revised 27 November 2021; Accepted 1 December 2021; Published 16 December 2021

Academic Editor: Gengxin Sun

Technical movement analysis requires specialized domain knowledge and processing a large amount of data, and the advantages of AI in processing data can improve the efficiency of data analysis. In this paper, we propose a feature pyramid network-based temporal action detection (FPN-TAD) algorithm, which is used to solve the problem that the action proposal module has a low recall rate for small-scale temporal target action regions in the current video temporal action detection algorithm research. This paper is divided into three parts. The first part is an overview of the algorithm; the second part elaborates the network structure and the working principle of the FPN-TAD algorithm; and the third part gives the experimental results and analysis of the algorithm.

1. Introduction

In recent years, with the continuous development and application of computer technology and artificial intelligence technology, vision-based human motion analysis technology has been rapidly developed and widely paid attention to. At present, vision-based human motion analysis is still a challenging topic in computer vision, mainly involving several disciplines such as pattern recognition, image processing, and virtual reality, and has a wide range of application prospects in human-computer interaction, intelligent monitoring, rehabilitation therapy, sports training, and other fields [1, 2]. Computer vision already has a variety of applications in sports training and other related fields, with prominent tasks including sports type recognition and activity recognition, tracking of athletes and analysis of other objects of interest in videos, and estimation of human pose [3].

The most central problem in motion analysis is human pose estimation, which is an important research task in the field of computer vision. The task of human pose estimation is to identify the human body and locate the position of the joints of human body parts (e.g., eyes, nose, shoulders, and wrists) through computer image processing algorithms, and to connect the positions of the joints to form the human skeleton according to the structure of the human body [4]. Applications of human posture estimation include human behavior understanding, human body reidentification, human-computer interaction, health monitoring, and motion capture [5–7].

At present, artificial intelligence technology has penetrated deep into sports training programs, and a lot of intelligent hardware and software have been created in sports training. The continuous penetration of artificial intelligence technology in sports training has promoted the development of sports [8]. Professional guidance coupled with AI-assisted training makes the training process of athletes more scientific. Through AI devices, the potential of athletes can be explored, their strength discovered, and their growth promoted. By collecting a large amount of data from the athletes' training process and using computer technology to visualize and analyze it, athletes can have a comprehensive grasp of their growth process.

The shortage of professional physical education teachers makes it difficult to teach all students in school, it is difficult to achieve continuous tracking of each

student's learning and training progress, and it is difficult to tailor the teaching to each student [9]. The demand for efficient physical education and parents' anxiety about children's competition for higher education have led to the creation of social K12 physical education training services, but some institutions use small class teaching and one-on-one VIP training as gimmicks. Use sparring as a gimmick to create a tense atmosphere, and exaggerate the training effect. At the beginning of 2019, "rope skipping training classes" and "high-priced rope skipping classes" once appeared on a large scale, which attracted the attention of mainstream media, and the mainstream media made special exposures to this undesirable phenomenon. [10]. Exploring the deep-seated reasons for the "black heart class" incident of rope skipping, we can know that a small rope skipping touches the needs of education and teaching mechanism reform on one end and touches the ardent expectations of millions of parents on the other, and it is also related to the health of a country's young people [11]. How to break the current dilemma of physical education in primary and secondary schools is not only a concern in the field of physical education teacher training, but also a concern in the field of artificial intelligence and big data. It also provides an opportunity for the research and promotion of artificial intelligence and big data in the direction of intelligent physical education [12, 13].

Traditional midterm examination rope skipping monitoring equipment is bulky and expensive, the cost and effort of manual statistics are too great, it is difficult to achieve accurate guidance for each student, and students' training is not efficient [14]. Automatically and quickly analyze whether the actions in the rope skipping process meet the standards and give the correct guidance and training plan, which is the key to improving the performance of the rope skipping test [15, 16]. In the previous design of movement analysis, it canno't effectively analyze the complex environment for In the design of previous action analysis, it cannot effectively analyze the complex environment, some only can realize the judgment of which action has been done, and there is no analysis of the standardization of the action or not.

Most current vision-based human behavior recognition algorithms have problems such as high complexity, inability to handle online videos, and harsh deployment conditions [17, 18].

In school rope skipping physical education, physical education teachers usually conduct demonstration teaching to explain the essentials in the rope skipping process. Students practice by themselves. Each student often has a different level of mastery due to different acceptance levels [19]. The research of this paper can help students to analyze where problems occur in the process of rope skipping training and what improvements need to be made. This paper introduces the artificial intelligence into the remote physical education teaching, which deeply implements the concept of "Internet + Education." The introduction of artificial intelligence into remote physical education has a positive impact on reducing the teaching pressure of physical education teachers, promoting the growth of physical education teachers, improving students' examination results, and cultivating students' joy of sports [20].

Existing vision-based algorithms for human motion behavior recognition and analysis suffer from high complexity, poor robustness, and excessive computational burden. The lack of professional staff in the direction of human motion action analysis does not allow for real-time guidance of the movement process. Therefore, the implementation of a robust and less time-consuming method for movement analysis and evaluation is important for the improvement of athletes' skills and the quality of physical education [21].

The contributions of this article are as follows.

This paper proposes a feature pyramid network-based sequential action detection algorithm, which is used to solve the problem that the action proposal module in the current video sequential action detection algorithm research has a low recall rate of small-scale sequential target action regions.

This paper proposes to use a multistage deep neural network method to design and implement an online behavior analysis algorithm based on mobile vision, which can improve the real time and correctness of sports action analysis, develop an optimized personalized training plan for individual students, and improve the efficiency of student training.

Our experiments show that the proposed method is effective, and the experimental process mainly verified the effectiveness of the two aspects of optimization proposed by FPN-TAD. The evaluation indicators of the results of the experiment include the AUC of the AR curve to measure the performance of the sequential action proposal generation and the mAP to evaluate the performance of the sequential action detection.

2. Related Work

The development of smart sports is relatively fragmented globally. Sports constitute a relatively traditional industry, and with the rapid development of IoT technology [22], the sports industry is becoming more and more closely integrated with emerging industries. Today, we can see many intelligent sports products and intelligent sports facilities. It is foreseeable that in the future, all sports such as swimming, pole vaulting, rope skipping, and cycling will be well integrated with smart hardware and smart software.

Depending on the data used in the motion analysis process, motion analysis can be divided into sensor-based human motion behavior analysis and computer vision-based human motion behavior analysis. The study in [23], in order to analyze golfers' individual swings and improve their swing techniques by using computer technology, integrated high-precision strain gauge sensors into golf clubs to obtain the whole process of players' swings and analyzed four players' swings by using linear discriminator, linear support vector machine LSVM, and KNN classification models. The results show that LSVM achieves the best results and the test accuracy is 100%. The study in [24] analyzed the nine basketball movements of walking, running, jumping, passing, catching, shooting, dribbling while running, dribbling while walking,

and dribbling while standing during playing basketball; collected the behavioral performance of leg and arm movements and the corresponding signal waveform characteristics; and used inertial sensors integrated with a three-axis gyroscope, a three-axis accelerometer, and a magnetometer fixed on the calf and arm of the subject to detect the movements of the different limbs. The data were fused with angular velocity, acceleration, and magnetic field strength by extended Kalman filtering and classified using artificial BP neural network, Bayesian network (BN), SVM, and decision tree C4.5 algorithm.

The BP network was used to classify the upper and lower extremity movements, and the final average accuracy reached 98.85%. In order to analyze and identify swimming postures, other researchers established a human swimming training state monitoring system based on wearable inertial sensors in [25] and obtained the acceleration data, gyroscope data, and magnetometer data of swimmers in backstroke, butterfly, breaststroke, and freestyle through the inertial sensor nodes bound to the waist. McGuirk also compared and analyzed the effect of different data combinations on the classification results and proved that the acceleration data can be well used for the recognition of swimming movements [26]. The study in[27] used different machine learning algorithms on HAR (Human Activity Recognition) dataset for six different daily activities. The study in [28] used a machine learning model combining the AdaBoost integrated learning algorithm and the random forest algorithm, which achieved an accuracy of 0.998 for classification.

The top-down approach is employed to separate human detection from key point detection. First, a body detector is used to find all the bodies. The study in [29] is dedicated to solving these problems and proposes to use a Symmetric Spatial Transformer Network (SSTN) to obtain the region of each person from the inaccurately detected candidate frames, in which a Single Person Pose Estimator (SPPE) is used [30]. A Spatial Transformer Network (STN) was also used to map the obtained human pose into the coordinate system [31]. A Convolution Pose Machine (CPM) network was proposed in [32], and the structure of this network was designed in stages. A two-stage network structure is used in [33] to obtain the key point location coordinates: the first stage uses Faster R-CNN [21] to detect multiple people in the image and remove unnecessary information from the border values; the second stage uses a full convolutional residual network to predict a dense heatmap and coordinate compensation for the people in each border; and finally the two types of information are fused to get precise localization of key points.

The bottom-up approach is used to first detect the key point information in the test image and then assign it to a single person. In the bottom-up approach, the combination of human articulation point connector and key point detector is mainly used in the case of multiple human key point detection; unlike the detection of a single human, it is necessary to first detect the joints of all individuals, then group these key points to the limbs of the target person by clustering, and then connect the limbs of each person. The study in [2] innovatively proposed the candidate regions of human parts; each candidate region was treated as a node in the study, and the correlation between nodes was used as the weight between graph nodes.

With the maturity of the pose estimation algorithm, more and more studies have been conducted to obtain the key point coordinate information through the pose estimation algorithm to perform the action analysis during human movement. In [5], the pose estimation algorithm in [3] was used to obtain the key point coordinate information of the player during playing tennis for pose state transformation and morphological analysis. In this study, the obtained coordinate information is used as a feature vector and combined with SVM algorithm to predict the probability of success in playing tennis. In addition, an unsupervised classification method was used to classify and perform a detailed formal analysis by comparing the appearance features that occur at each success probability, and the visualization results were fed back to the athletes for their training. Li et al. [6], in order to visualize and analyze the state during swimming, opened a new avenue of vision-based personal training for swimming by using a vision-based stance estimation system to automate this process and alleviate the overhead of extensive manual stance annotation. Human appearance features and motion features are extracted through the OpenPose [13] network models using human skeletal joint point locations. Supervised machine learning is used to identify four activity categories, namely, sitting, standing, walking, and falling.

3. Research Overview

3.1. Related Work and Motivation. Temporal action detection is a hot topic in video analysis research in recent years, and ActivityNet, an annual video understanding competition hosted by CVPR since 2016, represents the current research direction and cutting-edge level of video understanding competitions. Rakibe and Patil [12] propose a solution to detect actions in ActivityNet competition.. The BSN algorithm mainly consists of two parts: the temporal evaluation module (TEM) and the proposal evaluation module (PEM). The input of TEM is the input video features obtained from the pretrained two-stream method model TSN; the one-dimensional fully convolutional network (FCN) is used to output the input features of equal length with the probability distribution curves of "action," "start," and "end" being output with the same length as the input features; then the set of candidate proposals is obtained based on the probability distribution curves; and then the confidence score of each proposal is estimated by PEM [19, 28]. The final proposal result is obtained by removing the candidate proposals with high overlap through NMS.

However, this scheme also has two problems; one is that videos with different timing lengths are uniformly input to the TEM as fixed lengths by sampling, which makes the recall rate of multiscale action detection in long videos not high. If we can set the feature representation of the input video as $F = \{f_i\}_{i=1}^{N}$, which contains the target action as $A = \{(s_1, e_1, c_1), (s_2, e_2, c_2), \ldots\}$, and if the BSN algorithm is used, the input TEM module is sampled with the feature of $F_{\text{TEM}} = \{f_j\}_{j=1}^{L_{\text{TEM}}}$. When there exists a target action satisfying

$s_i - e_i < f_j - f_{j-1}$, it causes this action to be difficult to recall, and false detection results of (s_i, e_i') and (s_i', e_i) may occur. Second, the input features are adopted from the fully connected layer output of the TSN behavior recognition structure (before surtax), and then the input timing evaluation module extracts timing features with the one-dimensional convolution, ignoring the different effects of different channels on the results.

And, in the target detection task of images, in order to recognize objects of different sizes, it is usually necessary to construct multiscale pyramids. The simplest image pyramid is shown in Figure 1(a), where the images are made into different scales and then the corresponding multiscale features are generated at different image scales, but obviously this approach increases the time cost of the algorithm. Therefore, most target detection algorithms, such as SPPNet and Fast R-CNN, adopt this approach as shown in Figure 1(b) and use the anchor mechanism only on the last layer of feature map. SSD, on the other hand, adopts the multiscale feature fusion approach shown in Figure 1(c). There is no upsampling process, features of different scales are extracted from different layers of the network to do prediction, and this approach does not add additional computation. In contrast, FPN uses the top layer features by upsampling and the bottom layer features fusion for independent prediction at each layer, which can better achieve small target detection. Meanwhile, it has been noticed that the Baidu team has already used the Action Pyramid Network (APN) based on similar multiscale feature maps for the temporal action detection task in ActivityNet competition for video understanding and won the first place.

3.2. FPN-TAD Algorithm Research Idea. In this paper, to address the current problem of difficult detection of multiscale ground truth target actions caused by BSN of boundary-sensitive networks generating candidate proposals on fixed-size feature dimensions, we draw on the idea of FPN for prediction on multiscale feature maps and first obtain feature maps at multiple scales (corresponding to different resolutions of the same video) by FPN structure. Then, three types of probability distribution curves for target action, action start, and action end at each scale are obtained on the feature maps at different scales by drawing on the idea of BSN. Meanwhile, in order to better relate the contextual information of the video and to notice the influence of different channels, this paper adopts the output of the last pooling layer in the TSN dual-stream method feature extraction network as the input of the FPN-TAD algorithm and uses 2D timing-channel convolution instead of the 1D timing-convolution method, and the timing-channel features are modeled to obtain a better video feature representation. On the ActivityNet-1.3 and THUMOS-14 datasets, a significant performance improvement is obtained with respect to the preimprovement, reaching the current state-of-the-art level [13].

In this part of the paper, the overall framework design of the proposed FPN-based multistage proposal generation temporal action detection algorithm (FPN-TAD) is given first, followed by a detailed description of the key techniques of video feature extraction, FPN action probability evaluation,

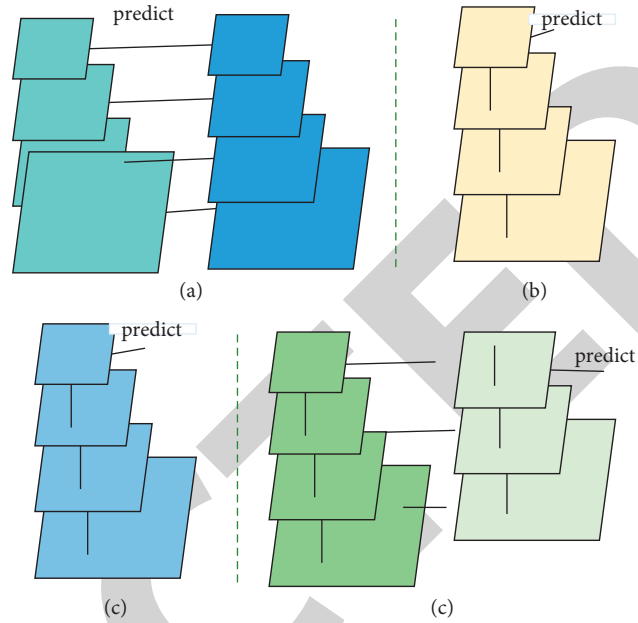

Figure 1: Structure of feature pyramid network (FPN): (a) featurized image pyramid; (b) single feature map; (c) pyramidal feature hierarchy; (d) feature pyramid network.

and candidate proposal generation, and finally a description of how to train each component of the network and the corresponding experimental parameter settings are given.

Figure 2 shows the overall framework design of FPN-TAD. Overall, the FPN-TAD algorithm divides the temporal action detection task into four parts, the front video feature extraction network, the feature pyramid FPN, the temporal action proposal generation, and the action classification. The first part is the basic video feature extraction, which is mainly based on the representative structure TSN of the dual-stream method to extract the high-dimensional semantic feature representation of the input video, and it should be especially noted that the output of the dual-stream network is followed by a 2D temporal-channel convolution. The second part is the FPN module, which takes the input video feature representation and obtains a multiscale feature pyramid by 2D convolution operation and uses a top-down approach to fuse some of the small-scale feature maps as the input of the third part of the temporal action proposal module. The third part is the temporal action proposal generation module, including the temporal evaluation module (TEM), proposal evaluation module (PEM), and NMS postprocessing components. The temporal evaluation module performs one-dimensional convolutional temporal evaluation in the temporal dimension of the input multiscale feature map and generates three probability distribution curves of action region, start position, and end position, which represent the probability of the current feature corresponding to the video region as action, action start, and action end, respectively, and such action probability distribution curves are obtained for feature maps of different scales. The fourth part is the action classification module, which takes the candidate features generated in the first part according to the proposal positions obtained in the third part, uses the TSN model to

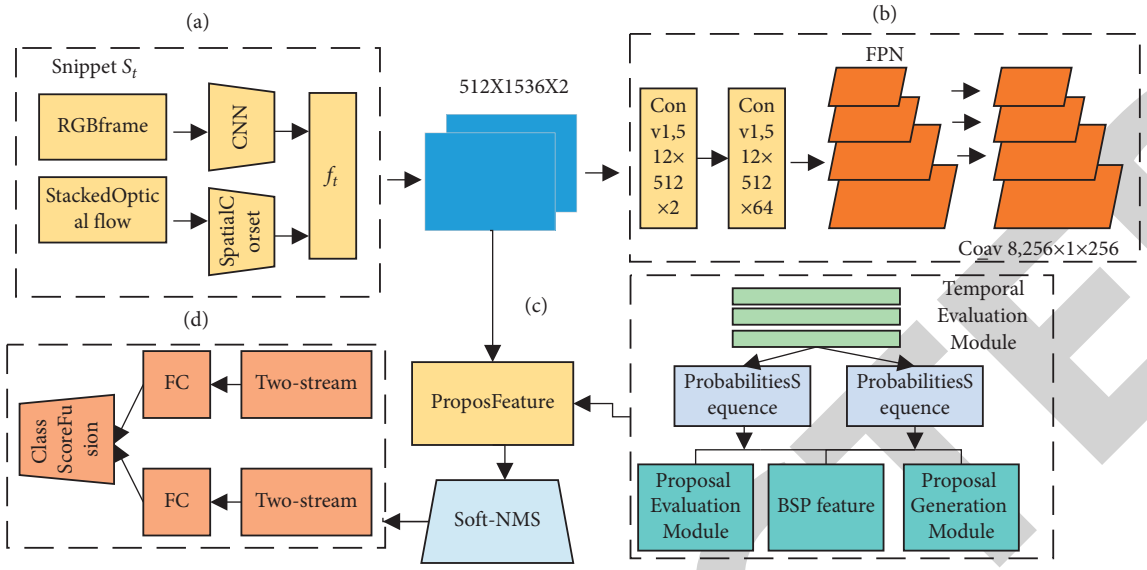

Figure 2: Overall framework of multiscale temporal action detection, FPN-TAD.

obtain the classification results of multiple snippets contained in the candidate proposals using the sing-shot classification prediction method, and finally fuses the classification results of multiple snippets in both temporal and spatial channels to finally input the action classification results of candidate proposals.

The specific process of video feature extraction is as follows: (1) Input a total of T frames of input video V, $V = \{t_i\}_1^T$, and divide the video into T/n_s snippets as the smallest unit of feature extraction (to reduce the redundancy of video stream features and reduce the computational cost), where n_s denotes the length of the snippet, and the experiment uses $n_s = 16$. (2) For each snippet, randomly select an RGB image from the sequence of n_s frames and calculate its optical flow; the result of optical flow calculation contains two optical flow images, flow_x in the x direction and flow_y in the y direction. (3) Input the RGB image input to the spatial convolutional network, the optical flow image is input to the temporal convolutional network, and the last pooling layer output of Inception-v4 is taken as the final result of video feature extraction.

The video feature extraction process, including the overall network structure of the convolutional neural network Inception-v4 and the flow diagram of the input image, is shown in Figure 3. The Inception-v4 network first completes the preprocessing of the data with the Stem module, which consists of multiple convolution and 2 times pooling, where the pooling is followed by three structures with inception modules. The role of the Reduction module between the three Inception modules is similar to that of pooling, and the same parallel structure is used to prevent the bottleneck problem. After all the convolution layers are completed, an $8 \times 8 \times 1536$ feature map (8×8 size, 1536 channels in total) is obtained, and the final 1536-dimensional image features are obtained by averaging the pooling layers. The feature extraction process of optical flow image is similar, the network structure is also Inception-v4, and the final obtained image feature dimension is also 1536. Therefore, each snippet can get 1536×2 feature representation. Considering the impact of different length videos, the obtained video features are fixed to 512 using linear interpolation. Therefore, the final feature representation of each video is $512 \times 1536 \times 2$ feature map.

The PEM module outputs the temporal action proposal results, including the start and end time of the action. The proposed action classification needs to give the proposed action category according to the proposal result obtained by PEM. For a given proposal $a_i = (t_s, t_e)$, the proposed feature $f_i = (f_{s,i}, f_{t,i})$ is taken from the prior corresponding video feature extraction module, where $f_{s,i}$ and $f_{t,i}$ represent the spatial and temporal features, respectively, with length $L_{f,t} = t_e - t_s$. The loss functions for action classification are shown in the following equation:

$$L_{cls} = -\sum_{i=1}^{n_{\text{class}}} y_i \left(\widehat{y}_i - \log \sum_{j=1}^{n_{\text{class}}} e^{\vartheta y_j} \right), \tag{1}$$

where y_i denotes the proposed classification ground truth and $\widehat{y}_i$ denotes the fusion result of the classification results of all snippets on the same category i. The fusion method is used in the experiment mean function.

The training of the FPN-TAD network is divided into four stages: the first stage pretrains the predecessor video feature extraction network in the ActivityNet-1.3 dataset by the action recognition task and obtains the input snippet to obtain the 1536-dimensional feature Inception-v4 model; the second stage trains the temporal action proposal generation network, using the output of the predecessor feature extraction network as input, to obtain the multiscale temporal action probability distribution curve; the third stage trains the temporal scoring network based on the results of temporal action proposals to filter the candidate proposals; finally, the candidate proposal features are used to train the action classifier.

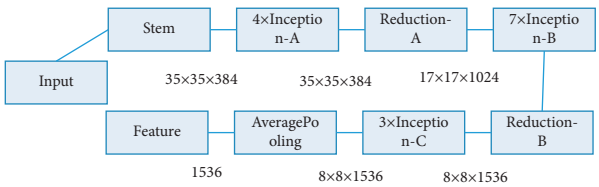

Figure 3: Inception-v4 overall structure diagram.

4. Experimental Design and Analysis of Results

The experimental scheme is designed with two validation objectives. On the one hand, the performance improvement of the proposed multistage boundary-sensitive network FPN-TAD relative to the baseline boundary-sensitive network BSN needs to be compared. On the other hand, the performance of the proposed FPN-TAD scheme needs to be compared with the existing temporal action proposal and temporal action detection schemes to verify the effectiveness of the proposed approach.

4.1. Datasets and Evaluation Criteria. The dataset of the experimental part of the paper is expanded using 2 representative datasets in the field of temporal action detection, THUMOS-14 and ActivityNet-1.3. The THUM0S-14 dataset [4] includes two tasks, behavior recognition and temporal action detection, and the number of categories of actions is 20. Most of the current papers on temporal action proposal or detection are evaluated in this dataset. ActivityNet-1.3 is a large dataset for the video understanding task, containing a total of 19994 videos, and the ratio of training, validation, and test sets is 2 : 1:1, with a total of 200 active categories. For the temporal action detection task, each video contains an average of 1.54 actions. Many recent advances in video understanding tasks, including temporal action detection, are based on ActivityNet-1.3.

The temporal action detection consists of both temporal action proposal and action recognition tasks, and the average recall (AR) calculated at different IoU thresholds is usually used as an evaluation metric in the temporal action proposal generation task. For comparison purpose, IoU = [0.5 : 0.05 : 0.95] on the ActivityNet-1.3 dataset and I10U = [0.5 : 0.05 : 0.95] on the THUMOS-14 dataset are used by convention; IoU = [0.5 : 0.05 : 1.0]. Referring to the measurement criteria of the ActivityNet 2018 competition, in order to better evaluate the relationship between recall and number of proposals, this paper calculates the average recall and average AUC, and the value range of AN is set to 0–100.

In the temporal action detection task, the common evaluation metric is mean Average Precision (mAP). The specific calculation process is to first calculate mAP over each category separately and then mAP over all categories.

4.2. Experimental Process Design. In this subsection, the experimental process of the paper is explained in detail, and the experimental results of the process are compared to verify the effectiveness of the method proposed in the paper. The baseline of the paper refers to the champion scheme of Tianwei L. et al. in ActivityNet 2018, and reproducing the scheme, the experimental process of the paper mainly verifies the effectiveness of the optimization of two aspects proposed by FPN-TAD: firstly, the introduction of FPN for timing evaluation proposal generation on multiscale timing feature maps; secondly, the application of 2D convolution on the original feature maps for joint timing-channel modeling. The evaluation metrics of the experimental results include the AUC of the AR curve, which measures the performance of temporal action proposal generation, and the mAP.

Before and after FPN improvement: the experiments first reproduce the BSN of baseline and give a temporal action detection scheme for baseline based on the results generated by this proposal in combination with the TSN behavior recognition scheme; then, in the setting of 400-dimensional features and 1D convolution after in putting FC, we directly apply FPN at multiple scales to generate candidate proposals and perform subsequent behavior recognition. Regarding the way of using FPN, the paper experiments three different schemes for multiscale temporal evaluation, where FPN-TAD1 performs a simple weighted average of the five scales of temporal feature maps obtained from the part of the feature pyramid; FPN-TAD2, on the other hand, inputs all five temporal feature maps to the temporal evaluation module for the results of fusion.

The final experimental results obtained are shown in Table 1. The AUC performance of the FPN-TAD1 approach relative to the baseline approach is improved by only 0.25, and that of FPN-TAD2 is improved by 1.21, while the FPN-TAD approach achieves the best AUC improvement of 1.45. Based on this experimental result, the following conclusions can be drawn: (1) Comparing FPN-TAD1 and FPN-TAD2 illustrates that the utilization of multiple temporal feature maps should occur after TEM (FPN-TAD2) rather than directly in the feature dimension for fusion (FPN-TAD1), mainly because small-scale information affects the resolution of large-scale information. (2) Comparing BSN-baseline and FPN-TAD well demonstrates that the FPN-based temporal action detection (FPN-TAD) algorithm has a significant performance improvement, mainly due to the way of generating proposals at different resolutions, which can improve the recall of actions in the case of a large range of variations in the target action timing length.

The fusion method of FPN based on 2D convolution: In order to better extract video features, the output (512, 1536) before full connection in TSN is taken as the input of the temporal action proposal module, and 2D convolution is used to jointly model the temporal and channel features; the experimental results obtained are shown in Table 2.

The experimental results show that the joint modeling of timing and channel is very effective, and compared to using only ID convolution to extract timing features, 2D convolution can better link contextual information while focusing on the different effects of different channel features on the results. Therefore, the final FPN-TAD implementation scheme is the 2D convolutional temporal-channel joint

Table 1: Results of the experimental procedure of FPN-TAD.

Method	AR@10	AR@100	AUC
BSN-baseline	—	74.22	66.29
FPN-TAD1	54.84	74.91	66.54
FPN-TAD2	55.59	74.30	67.48
FPN-TAD3	55.68	76.86	67.72

Table 2: Experimental results of 2D convolutional timing-channel joint modeling.

Method	AR@10	AR@100	AUC
BSN-baseline	—	74.25	66.29
FPN-TAD3	54.68	76.86	67.72
FPN-TAD	55.14	77.06	67.48

modeling FPN-TAD algorithm, and the experiments refer to the evaluation metrics of the ActivityNet 2018 video understanding competition, and Figure 4 shows the relationship between the recall rate and the number of candidate proposals under different IoU.

As shown in Figure 4, the horizontal coordinate of the AR IoU curve is the average number of candidate proposals per video used for evaluation, the vertical coordinate is the average recall rate for 200 categories, multiple dashed lines indicate different IoU thresholds, and the black solid line indicates the average recall rate of the algorithm at IoU = [0.5 : 0.05 : 0.95]. Figure 4(a) depicts the reproduction results of the BSN-baseline before improvement, and Figure 4(b) shows the experimental results of the FPN-TAD algorithm based on the feature pyramid proposed in this paper. From the experimental results, the FPN-TAD algorithm has a significant performance improvement at different IoU thresholds, which proves that the multiscale feature pyramid approach can better capture the boundary information.

After obtaining the temporal action proposal results of FPN-TAD, the mAP performance obtained for different IoU cases using the TSN action recognition framework is shown in Table 3. The experimental results show that (1) the results of temporal action proposal directly affect the results of temporal action detection and (2) the FPN-TAD algorithm can significantly improve the results of temporal action detection under different IoU requirements.

4.3. Comparison of Experimental Methods. To demonstrate the effectiveness of the algorithm, the paper experiments on the THUMOS-14 dataset to compare the FPN-TAD algorithm proposed in the paper with the current state-of-the-art algorithms, including five representative reference algorithms, namely, DAPs, SCNN, SST, TURN, and BSN.

The experimental results are shown in Figure 5. Figure 5(a) shows the AR curve, with the horizontal coordinate indicating the average number of proposals per video and the vertical coordinate indicating the average recall rate. The experimental results in Figure 5(b) show that the proposed FPN-TAD algorithm and the BSN algorithm used as a reference significantly outperform the other four baseline algorithms of DAPs, SCNN, SST, and TURN when the number of candidate proposals used for evaluation is the same; in particular, when the number of candidate proposals is less than 200, the improvement of the average recall rate is around 20%. In addition, when the number of candidate offers is in the interval of 50–150, the FPN-TAD algorithm based on multiple temporal feature maps has a significant performance improvement relative to the BSN-baseline, which fully demonstrates that the FPN-TAD algorithm can generate higher quality candidate proposals, especially when the number of candidate proposals is further increased to achieve better algorithmic results. Figure 5 shows the recall curves of different algorithms at IoU = 0.5 when the number of candidate proposals is equal to 1000, and we can see that the proposed FPN-TAD curve has a significant performance improvement compared with other algorithms; in particular, the recall rate exceeds all the compared algorithms at IoU = 0.7, which indicates that the FPN-TAD algorithm can produce candidate proposals with higher IoU and more accurate boundary localization of target actions.

The quantitative experimental results of the comparison of temporal action proposal algorithms are shown in Table 4, based on which the following conclusions can be drawn: (1) Comparing the TURN of C3D and 2-Stream, we find that the 2-Stream feature significantly outperforms C3D in the temporal action proposal task. (2) The performance advantage of the proposal generation methods based on the probability curves of temporal actions (BSN and FPN-TAD) over the fixed-size sliding window or anchor methods is obvious when the number of proposals is small. (3) The introduction of multiple scales has a significant recall improvement when the number of proposals is small, and the recall rate is close to the average of AR@50, AR@100, and AR@200, which all have a performance improvement of nearly 1%.

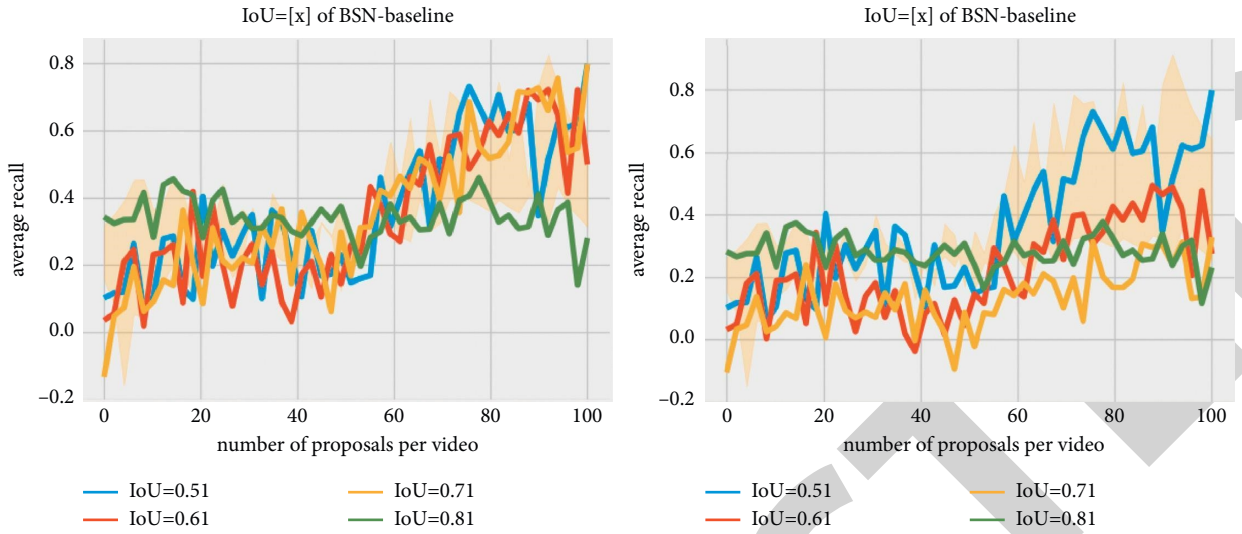

Figure 4: IoU curves before and after improvement.

Table 3: Experimental results of 2D convolutional timing-channel joint modeling.

Method	IoU = 0.5	0.75	0.95	Average
BSN-baseline	52.50	33.53	8.85	33.72
FPN-TAD	56.41	35.14	9.52	38.25

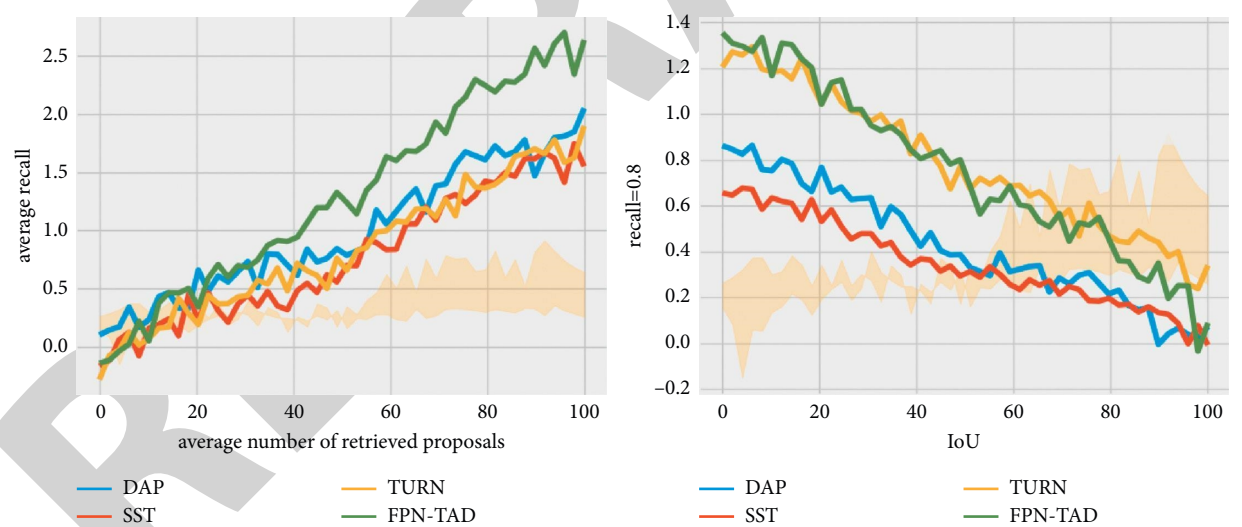

Figure 5: Comparison of FPN-TAD with mainstream algorithms.

Table 4: Comparison of AR@AN experimental results (THUMOS-14).

Feature	Method	@50	@100	@200	@500	@1000
C3D	DAPs	13.47	23.53	33.98	48.57	56.21
C3D	SCNN	17.85	26.10	37.83	51.14	58.06
C3D	SST	19.86	27.82	37.83	54.02	60.52
C3D	TURN	19.90	27.93	38.45	53.32	60.87
2-Stream	TURN	22.04	37.52	46.25	58.72	64.58
2-Stream	BSN	33.21	41.05	50.02	60.16	65.71
2-Stream	FPN-TAD	34.01	43.08	51.17	60.24	65.58

5. Conclusions

This paper details the feature pyramid network-based temporal action detection (FPN-TAD) algorithm. The main content is divided into three parts: a review of more specific related work in the direction of temporal action detection, an introduction to the proposed feature pyramid network-based temporal action detection (FPN-TAD) algorithm, and the outline of the algorithm design and working principle. The network structure of the FPN-TAD algorithm is proposed, and the main components and algorithm flow of FPN-TAD are described.

Data Availability

The datasets used in this paper are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this work.

References

[1] X. Wang, "Discussion on application of multimedia teaching in college english vocabulary teaching," *Open Journal of Modern Linguistics*, vol. 6, no. 3, pp. 177–181, 2016.

[2] M. Ahltorp, M. Skeppstedt, S. Kitajima, A. Henriksson, R. Rzepka, and K. Araki, "Expansion of medical vocabularies using distributional semantics on Japanese patient blogs," *Journal of Biomedical Semantics*, vol. 7, no. 1, p. 58, 2016.

[3] E. Lee and M. Park, "Student presentation as a means of learning English for upper intermediate to advanced level students," *Journal of Pan-Pacific Association of Applied Linguistics*, vol. 12, no. 1, pp. 47–60, 2008.

[4] L. Su, J. N. Liu, L. F. Ren, and F. Zhang, "An object classification approach based on randomized visual vocabulary and clustering aggregation," *Applied Mechanics and Materials*, vol. 433-435, pp. 778–782, 2013.

[5] L. Wang, C. Zhang, Q. Chen et al., "A communication strategy of proactive nodes based on loop theorem in wireless sensor networks," in *Proceedings of the 2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP)*, pp. 160–167, IEEE, 2018.

[6] H. Li, D. Zeng, L. Chen, Q. Chen, M. Wang, and C. Zhang, "Immune multipath reliable transmission with fault tolerance in wireless sensor networks," in *Proceedings of the International Conference on Bio-Inspired Computing: Theories and Applications*, pp. 513–517, Xi'an, China, October 2016.

[7] Z.-Y. Lu and S.-H. Fan, "'A mixed-method examination of adopting focus-on-form TBLT for children's English vocabulary learning," *English Language Teaching*, vol. 14, no. 2, p. 37, 2021.

[8] Y. Luo, M. Wang, Z. Le, and H. Zhang, "An improved kNN text categorization algorithm based on cluster distribution," *Journal of Computational Information Systems*, vol. 8, no. 3, pp. 1255–1263, 2012.

[9] N. Pazyura, N. Nychkalo, J. Wang, L. Lukianova, and N. Muranova, "Use of task-based approach in teaching vocabulary to business English learners at university," *Advanced Education*, vol. 16, no. 16, pp. 98–103, 2020.

[10] A. Irawan, A. Wilson, and S. Sutrisno, "The implementation of duolingo mobile application in English vocabulary learning," *Scope: Journal of English Language Teaching*, vol. 5, no. 1, p. 08, 2020.

[11] A. Jain, D. K. Tayal, and S. Yadav, "Retrieving web search results using Max–Max soft clustering for Hindi query," *International Journal of System Assurance Engineering & Management*, vol. 7, no. 1 Supplement, pp. 1–12, 2016.

[12] R. S. Rakibe and B. D. Patil, "Background subtraction algorithm based human motion detection," *International Journal of scientific and research publications*, vol. 3, no. 5, pp. 2250–3153, 2013.

[13] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 1, pp. 13–24, 2009.

[14] J. K. Aggarwal and Q. Cai, "Human motion analysis: a review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[15] D. Wu, C. Zhang, L. Ji, R. Ran, H. Wu, and Y. Xu, "Forest fire recognition based on feature extraction from multi-view images," *Traitement du Signal*, vol. 38, no. 3, pp. 775–783, 2021.

[16] J. Shotton, A. Fitzgibbon, M. Cook et al., "Real-time human pose recognition in parts from single depth images," in *Proceedings of the CVPR 2011*, pp. 1297–1304, Colorado Springs, CO, USA, June 2011.

[17] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: an experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1442–1468, 2013.

[18] J. J. Tompson, A. Jain, Y. LeCun, and C. Bergler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1799–1807, 2014.

[19] A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, 2010.

[20] J. Shotton, R. Girshick, A. Fitzgibbon et al., "Efficient human pose estimation from single depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2821–2840, 2012.

[21] T. Xie, C. Zhang, Z. Zhang, and K. Yang, "Utilizing active sensor nodes in smart environments for optimal communication coverage," *IEEE Access*, vol. 7, pp. 11338–11348, 2018.

[22] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 755–762, San Francisco, CA, USA, June 2010.

[23] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1995–2006, 2013.

[24] Z. Zhang, C. Zhang, M. Li, and T. Xie, "Target positioning based on particle centroid drift in large-scale WSNs," *IEEE Access*, vol. 8, pp. 127709–127719, 2020.

[25] R. T. Collins, A. J. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 745-746, 2000.

[26] C. McGuirk, *A Multi-View Video Based Deep Learning Approach for Human Movement Analysis*, Université d'Ottawa/ University of Ottawa, Ottawa, Canada, 2021.

[27] Y. Liu, "Human motion image detection and tracking method based on Gaussian mixture model and CAMSHIFT,"

Microprocessors and Microsystems, vol. 82, Article ID 103843, 2021.

[28] C. Zhang, T. Xie, K. Yang et al., "Positioning optimisation based on particle quality prediction in wireless sensor networks," *IET Networks*, vol. 8, no. 2, pp. 107–113, 2019.

[29] A. Miao and F. Liu, "Application of human motion recognition technology in extreme learning machine[J]," *International Journal of Advanced Robotic Systems*, vol. 18, no. 1, Article ID 1729881420983219, 2021.

[30] H. Zheng, D. Liu, and Y. Liu, "Design and research on automatic recognition system of sports dance movement based on computer vision and parallel computing," *Microprocessors and Microsystems*, vol. 80, Article ID 103648, 2021.

[31] S. Yue, "Human motion tracking and positioning for augmented reality," *Journal of Real-Time Image Processing*, vol. 18, no. 2, pp. 357–368, 2021.

[32] C. H. Cao, Y. N. Tang, and D. Y. Huang, "IIBE: an improved identity-based encryption algorithm for WSN security," *Security and Communication Networks*, vol. 2021, Article ID 8527068, 8 pages, 2021.

[33] R. M. Lemos Baptista, "Human motion analysis using 3D skeleton representation in the context of real-world applications: from home-based rehabilitation to sensing in the wild," Doctoral Thesis, University of Luxembourg, Esch-sur-Alzette, Luxembourg, 2021.