Research Article

Semantic Extraction of Basketball Game Video Combining Domain Knowledge and In-Depth Features

Yufeng Du, Quan Zhao, and Xiaochun Lu

1Ministry of Sports, Xi'an Academy of Fine Arts, Xi'an 710065, Shaanxi, China
2School of Sciences, Xi'an Technological University, Xi'an 710021, Shaanxi, China

Correspondence should be addressed to Quan Zhao; zhaoquan@xatu.edu.cn

Received 13 July 2021; Accepted 15 August 2021; Published 6 September 2021

Abstract

The team sports game video features complex background, fast target movement, and mutual occlusion between targets, which poses great challenges to multiperson collaborative video analysis. This paper proposes a video semantic extraction method that integrates domain knowledge and in-depth features, which can be applied to the analysis of a multiperson collaborative basketball game video, where the semantic event is modeled as an adversarial relationship between two teams of players. We first designed a scheme that combines a dual-stream network and learnable spatiotemporal feature aggregation, which can be used for end-to-end training of video semantic extraction to bridge the gap between low-level features and high-level semantic events. Then, an algorithm based on the knowledge from different video sources is proposed to extract the action semantics. The algorithm gathers local convolutional features in the entire space-time range, which can be used to track the ball/shooter/hoop to realize automatic semantic extraction of basketball game videos. Experiments show that the scheme proposed in this paper can effectively identify the four categories of short, medium, long, free throw, and scoring events and the semantics of athletes’ actions based on the video footage of the basketball game.

1. Introduction

In recent years, the spectacle and attention of various sports events have been increasing, and a large number of sports events are broadcasted and shared on the Internet in the form of videos, and sports videos have become an efficient and irreplaceable force for information dissemination on the Internet [1]. With the explosive growth of sport video data on the Internet platform, how to manage this information scientifically has become a major challenge in the current era of big data.

The traditional method based on the manual annotation to categorize and integrate the events occurring in a video not only requires a lot of human resources but also has a high error rate because manual annotation is easily influenced by human subjective factors and cannot meet the various needs of different people [2]. So a reasonable semantic analysis of video content has important theoretical significance and broad application prospects. There are multiplayer collaborative confrontation videos in sports such as basketball, where sports videos have difficulties such as serious obscuration, fast movement speed, and complex action changes [3].

On the other hand, the current sports video dataset is quite large, containing tens of thousands of videos and hundreds of classes, and there is an urgent need to address how to use them effectively. Moreover, these classes may be specific to certain domains, such as sport, and the dataset may contain noisy labels [4]. Another key open question is as follows: what is an appropriate spatiotemporal representation pattern of the video? Recent video representations for action recognition are mainly based on two different CNN architectures: (1) 3D spatiotemporal convolution, which may learn complex spatiotemporal dependencies but is difficult to scale in terms of recognition performance; (2) dual-stream architecture, which decomposes the video into motion and appearance streams and trains separate CNNs for each stream, eventually fusing the outputs [5]. While
both approaches have made rapid progress, both stream structures typically outperform spatiotemporal convolution because they can easily exploit the new ultra-deep structure and pretrained models for still image classification.

However, the dual-stream structure largely ignores the long-term temporal structure of the video and essentially learns classifiers that operate on single frames or short blocks of a few (up to 10) frames [6], potentially forcing consistency in classification scores over different segments of the video [7]. During testing, $T$ uniformly sampled frames are independently classified, and the classification scores are averaged to obtain a final prediction. This approach raises the issue of whether such temporal averaging can model the complex spatiotemporal structure of human behavior, which is exaggerated when multiple action classes share the same subactions. For example, consider the intricate combination of “basketball shooting” shown in Figure 1. Since there are only a few consecutive video frames, it can be easily confused with other actions, such as “running,” “dribbling,” “jumping,” and “throwing.” “Using postfusion or averaging is not the right best solution, because it requires that frames belonging to the same subaction are assigned to multiple classes. What we need is a worldwide feature descriptor for the video that collects evidence about the appearance of the scene and the movement of the person throughout the video without having to uniquely assign each frame to a single action.

In order to address this problem, this paper analyzes basketball game videos in multiplayer collaborative sports videos, studies the semantic event classification problem of basketball game videos, and aims to propose a deep learning-based event analysis method for basketball sports videos to realize the automatic analysis of basketball videos, so as to assist coaches in tactics formulation, assist players in action analysis, and assist video viewers in rapid search for interesting video segments, and promote the rapid development of basketball and even sports. We develop an end-to-end trainable video-level representation that aggregates convolutional scripts to different parts of the imaged scene and the entire time span of the video.

In this paper, we take a basketball game video in a collaborative multiplayer sports video as an example for semantic event analysis of sports video. The global motion features, group motion features, and individual pose features of the video are extracted separately to realize the eventual analysis of basketball game video with multi-feature fusion. Firstly, basketball video events are classified by combining sports domain knowledge, and secondly, global and group motion patterns of players in basketball games are expressed, in which the spatiotemporal extension of NetVLAD [8] aggregation layer in deep learning [9] is used, and NetVLAD can be used well in still images for instance-level recognition tasks. For our extension of NetVLAD to videos, we also address the following issues. First, aggregation of frame-level features across time into video-level representations selects the best method for aggregation at different levels of the network from output probabilities to separate layers of convolutional descriptors [10] and shows that aggregation picks through the last layer of convolutional descriptors to perform best. Second, different strategies for aggregating features from spatial and temporal streams for best-combining signals from different streams in a multistream architecture show some surprising results as the best results can be obtained by aggregating spatial and temporal streams into their separate individual video-level representations. We support our study with quantitative experimental results while providing intuition for obtaining the results.

The contributions of this paper are as follows:

1. We first designed a scheme that combines dual-stream network and learnable spatiotemporal feature aggregation, which can be used for end-to-end training of video semantic extraction to bridge the gap between low-level features and high-level semantic events.

2. Then, an algorithm based on the knowledge from different video sources is proposed to extract the action semantics. The algorithm gathers local convolutional features in the entire space-time range, which can be used to track the ball/shooter/hoop to realize automatic semantic extraction of basketball game videos.

3. Experiments show that the scheme proposed in this paper can effectively identify the four categories of short, medium, long, free throw, and scoring events and the semantics of athletes’ actions based on the video footage of the basketball game.

2. Related Work

The purpose of basketball sports video analysis is to extract key information from video content for play summaries, computer-aided recommendations, and content insertion, among others. In recent years, the content oriented to basketball sports video analysis mainly includes player detection, action recognition, and event detection [11].
2.1 Player Detection. The task of player detection is to automatically detect the key target of the basketball video and give its location or area in the form of a candidate box search, as shown in Figure 1. In basketball sports videos, the foreground target plays a crucial role in the event scene, and the location and behavior of this target are critical in high-level semantic event analysis or sports game and tactics analysis. For example, the work in [12] uses low-level feature information while combining domain knowledge of basketball to identify specific events and implement segmentation of a given video into wide-angle and close-up shots. The study in [13] proposes a hybrid approach of manual observation and machine learning to integrate statistics in a logical rule model for highlight clip detection, where clips of player breaks are cleverly utilized to aid in event recognition. The work in [14] is based on the motion characteristics of basketball to achieve content understanding, first based on a specific region enhancement method to achieve the segmentation of the ball in the video and then based on matching key points to achieve the reconstruction of the 3D scene to establish the motion trajectory of basketball. The work in [15] focused on analyzing free-throw events in basketball videos by detecting the principal locations of free throws and calculating the hit rate of free throws by analyzing the state of players. The methods mentioned above are based on low-level information, including visual information from human observation, etc. Traditional methods are limited to matching and classifying in the low-level feature space of video frames, which is not only inefficient but also of far lower quality and far less differentiation of intraclass features.

2.2 Action Recognition. Action recognition in video means recognizing the motion of the person in the video. Extracting features of human motion from video sequences is also a crucial part of video analysis. In the telephoto case, only the trajectory of the motion target as a whole is generally extracted for analysis, but in a close-up activity like sports video, a specific expression of the target’s limbs or torso needs to be extracted from the video sequence frames, i.e., action recognition [16].

Traditional action recognition methods have static features based on shape and contour, and dynamic features based on optical flow and motion information. [17] are to ensure action recognition by doing shape matching of key video frames extracted from the video with the saved original action. The study in [18] uses fuzzy optical flow features to calculate the optical flow field of the target in the video and achieves online action recognition by template matching. The study in [19] proposed expressing the action recognition of the target by the features of key point motion history. The features extracted by traditional methods are manually designed features, and thus, the performance of action recognition is not high.

In recent years, with the development of deep learning, some new solutions have emerged in the field of basketball video. The work in [20] developed a novel recursive convolutional neural network for large-scale visual learning that is not specific to basketball video analysis but includes basketball videos in the dataset. The work in [21] analyzed a basketball game with a single shooter in life based on deep learning. The method extracts the spatial features of sampled frames as well as the motion information in the video by VGG-16 [22] and subsequently achieves the fusion of spatial and temporal features by a layer called the aggregation layer pooling in this paper, and the whole network is end-to-end structured. The work in [23] implemented basketball tracking using deep learning, which is a method to determine whether a 3-point shot is successful or not by RNN, while the model can predict the trajectory of the basketball very well. The work in [24] implemented NBA tactical analysis based on deep learning, firstly by locating the players as well as the relative positions of the ball, and secondly by implementing tactical classification based on convolutional networks. Extraction of key player regions often causes greater interference due to the existence of certain occlusions between players, and the algorithm has a complex framework and high computational complexity. The player occlusion problem is very serious during basketball games, which tends to reduce the ability to track key players, and coupled with the complex background of basketball videos, this will lead to serious target confusion while the computational effort will increase dramatically [25, 26].

2.3 Event Detection. Event detection means giving the start and endpoints of semantic events in the video. Starting from when a player is ready to shoot from the free-throw line to when the ball goes in the box and land is a free throw event in a basketball game video. The study in [27] implements event detection based on CNN extended network by introducing cascaded CNN networks to express local motion information while incorporating trajectory information to achieve event analysis of surveillance videos; similar to action recognition, event analysis can be implemented based on a dual-stream CNN model, which introduces both spatial information and temporal features from a global perspective. The work in [28] tracks all players while learning the key players in basketball events based on the attention mechanism and achieves the learning of player sequence information by Bi-LSTM, the learning of global video frame sequence information by LSTM, and finally the combined expression of local and global information to achieve event detection. In [29], firstly, player tracking is achieved based on keyframes, secondly, player state changes are drawn by LSTM, and player information is fused by different kinds of team division methods to achieve event representation.

3. The Proposed Dual-Stream Network

We just try to learn a video representation, which is trained for end-to-end action recognition. To achieve this, we introduce architecture as shown in Figure 2. In detail, we sample frames from the entire video and convert aggregated features from the appearance (RGB) and motion (flow) streams into a fixed-length vector at the exclusive video level using the vocabulary “action words” [30]. This representation is subsequently passed to the classifier, which outputs a final classification score. The
parameters of the aggregation layer—a set of “action words”—are learned in a differentiated manner together with the feature extractor to accomplish the ultimate task of action classification.

In the following, we first describe the reasonable spatio-temporal aggregation layer [31]. We then discuss the possible positions of the aggregation layer in the overall architecture and strategy for combining the emergent and motion streams. Finally, we give implementation details. We more specifically describe our improvements to ActionVLAD. (1) We have improved the fusion method, as described in the introduction of the paper. (2) We changed the network structure. For example, we added CNN and FC to the network and shared their parameters based on the idea of ResNet. (3) The application scenarios of the designed model are different.

3.1. Dual-Stream Aggregation Network Model. Consider \( x_{it} \in \mathbb{R}^D \), the D-dimensional local descriptors extracted from the spatial locations \( i \in \{1 \ldots N\} \) and frames \( t \in \{1 \ldots T\} \) of the video. We want to aggregate these descriptors spatially and temporally throughout the video while preserving their information content. This is achieved by first dividing the descriptor space \( R^D \) into \( K \) cells using a vocabulary of \( K \) “action words” represented by anchor points \( \{c_k\} \) (Figure 2, note that different lines represent different aggregation states). Each video descriptor \( x_{it} \), \( T \) is then assigned to a cell and the difference between descriptors and anchor points is recorded using a residual vector \( x_{it} - c_k \). The difference vectors [32] in the whole video are then summed as

\[
V[j,k] = \sum_{t=1}^{T} \sum_{i=1}^{N} \alpha e^{-\|x_{it} - c_k\|^2} (x_{it}[j] - c_k[j]).
\]

Here, \( x_{it}[j] \) and \( c_k[j] \) are the \( j \)-th component of the descriptor vector \( x_{it} \) and anchor \( c_k \), respectively, and \( \alpha \) is a tunable hyperparameter. The first term in (1) represents the soft assignment of descriptor \( x_{it} \) to unit \( k \). The second term represents \( x_{it}[j] - c_k[j] \), the residual between the descriptor and the anchor point of unit \( k \). The two summation operators denote aggregation in time and space, respectively. The output is a matrix \( V \) where the \( k \)-th column \( V[:,k] \) represents the aggregation descriptor in the \( k \)-th cell. The columns of the matrix are then internally regularized [33], stacked, and L2 regularized [5] for a single descriptor of the entire video \( v \in \mathbb{R}^{KD} \).

Clearly known residual vectors record the differences between the extracted descriptors and the “typical actions” (or subactions) represented by the anchors. The residual vectors are then aggregated across the video by computing the sum of the residual vectors within each cell. Critically, all parameters, including the feature extractor, action words, and classifier, are learned jointly from the data in an end-to-end manner to better distinguish the target actions. This is because the spatiotemporal aggregation described in (1) is microscopic and allows backpropagation of error gradients to lower layers of the network. The aggregation used in this paper is a spatiotemporal extension of the NetVLAD [8] aggregation, while we introduce a sum across time \( T \).

3.2. Aggregation Layer. Theoretically, the spatiotemporal aggregation layers described above can be placed at any level of the network to bring the corresponding feature maps. In this section, we describe the different possible choices that will guide our experimental study.

In detail, we are based on the dual-stream architecture introduced by [34] on the VGG-16 network. Here, we consider only appearance streams but discuss different ways of combining appearance and motion streams with our aggregation in Section 3.1.

The two-stream model first trains a frame-level classifier that uses all frames from all videos and averages the
predictions of $T$ uniformly sampled frames at test time. Choosing this basic network (pretrained at the frame-level) as a feature generator feeds the input from different frames to our trainable aggregation layer [4, 15, 35]. But which layer’s activation should we share? We consider two main options. Consider pooling the outputs of the fully connected (fc) layers. These are represented as $1 \times 1$ spatial feature map with 4096-dimensional outputs for each T-frame of the video. In other words, we pool a 4096-dimensional descriptor from each T-frame of the video. Second, considering pooling features from the convolutional layer (we consider VAN4 3 and VAN5 3), as shown in Section 3.1, we obtain the best performance by pooling features at the highest convolutional layer (VGG-16 for conv5 and 3) [35].

3.3. Dual-Stream Aggregation Mode. The aggregation layer in this paper can be employed to pool functions across different input mode streams. In our example, we consider appearance and motion streams [36], but any number of other data streams, such as warped streams or RGB differences [37], can be spanned in pooling. There are several possible ways to combine appearance and motion streams to obtain a trainable joint representation. We will explore the most salient ones in this section and outline them in Figure 3.

A single paper aggregation layer covers appearance and motion features in tandem (concat fusion). In this case, we connect the corresponding output feature mappings from appearance and motion, essentially assuming their spatial correspondence. We place a paper aggregation layer on top of this connected feature map, as shown in Figure 3(a). This allows the codebook to be constructed using the correlation between appearance and flow features.

A single paper aggregation layer covers all appearance and motion features (early fusion). We also experimented with a single paper aggregation layer for all features from the appearance and motion streams, as shown in Figure 3(b). This encourages the model to learn a single script space $x_{ij}$ for both appearance and motion features, thus exploiting redundancy in the features.

Late Fusion. This strategy, shown in Figure 3(c), follows the standard test practice of appearance and motion weighted averaging of the last layer of features. Thus, we have two separate layers of this paper aggregation layer, one for each stream [15]. This allows both layers of this paper aggregation layer to learn a specialized representation of each input modality.

4. The Overall Framework of Integration

Domain Knowledge

The overall framework of the proposed basketball event classification method incorporating domain knowledge and deep features is illustrated in Figure 4. Our method framework is inspired by the idea of decomposing complex tasks into basic tasks and combining the work of multiple elementary channels to generate a unified network, and each part of the framework is described in detail in this subsection. First, we propose a 2-stage event classification method, which includes 5 types of coarse classification for 3 pointers, free throws, 2 pointers (including other 2 pointers and layups), dunks, and steals based on event-occ video clips with GCMP_DF_SVF information [30] and 2 types of coarse classification for layups and other 2 pointers based on pre-occ video clips with GCMP_DF_SVF information. 2 categories of coarse classification are based on the pre-occ video segment using GCMP_DF_SVF information for layups and other 2 pointers. Both event classification stages are achieved by our model to extract and express the spatial as well as sequential features of GCMPs. Secondly, we implement the prediction of event state success/failure in the video segment of the post-occ video segment based on RGB_DF_VF features [15, 21] and finally combine the results of event type classification with the event state success/failure to achieve 11 classes of event classification, which are a 3-point shot success, 3-point shot failure, free-throw success, free-throw failure, other 2-point shot success, other 2 pointers, layup success, layup failure, dunk success, dunk failure, and steal [28].

The specific process is described as follows:

1. 2-stage event classification method: firstly, we implement the event preclassification based on the depth feature basketball events in the event video segment; this stage does not consider the success and failure states only; it predicts the event type. Combined with the model described in part 3, in the event video segment, the features of other 2 pointers and layup events are relatively similar, so in this stage, we will first implement a 5-class classifier; that is, other 2 pointers and layup events are fused into one class; then, the 5 classes are 3 pointers, free throws, other 2 pointers and layups, dunks, and steals. The deep feature basketball event expression of basketball video players is extracted by optical flow graph, and then, the spatial features of basketball video are extracted by aggregation layer in this paper’s model, and the sequence feature expression of deep feature basketball events is realized by dual flow network. Next is the implementation of event subclassification within the pre-occ video segment based on depth feature basketball events. This stage requires training a 2-class classifier similar to the 5-class classifier for additional 2 pointers and layups. As further illustrated in Figures 3–6, for a video segment of an unknown class, the video sequence of the event video segment is first input to the 5-class classifier and can be output directly if the output is a 3-point, free throw,
dunk, or steal event. However, if the output is the category after the fusion of layups and other 2 pointers, the pre-event video segment of the video segment needs to be input to the 2-class classifier for further judgment, and the output of the last two classes of classifiers is the discriminative type of the event.

(2) Classification method for success and failure event states: the success and failure state prediction is deployed in the post-event video segment based on the original video frame RGB_DF_VF. Combined with the analysis of different video segments of the same event in chapter 3 on the event classification ability, we extract RGB_DF_VF features for the success and failure classification of events, except for the stealing events (here only the success and failure states are set for the pitching events, without considering stealing). Finally, the entire event state analysis is achieved by voting the results of each frame in the post-event.

The final classification of all the 11 types of events is done by fusing the analysis of stages (1) and (2) above. Based on the analysis of (1) above, the prediction vectors of 6 types of events can be obtained through the 2-stage event classification, which can be expressed as points, free throws, layups, other points, dunks, and steals, and the prediction vectors of success and failure of the event state stages can be expressed as success and failure based on the analysis of (2) above. It should be noted here that since steals cannot be determined whether their status is success or failure, the five categories of events that need to be analyzed in this paper are all shooting events, whose expression vectors are points, free throws, layups, other points, and dunks. The vectors of these events are all binary vectors, and each element of the predicted event is either “1” or “0” [20, 30].

5. Simulation Experiment and Result Analysis

Our experimental video data were selected from the video scenes captured by the main camera. First, five video sequences of basketball games from the Nth NCTU tournament were segmented into many video shots [7, 10], from which we selected about 157 video shots from these scenes.

In this thesis, a neural network model is constructed using the PyTorch framework [5] with a server operating system of Linux release (core). 7 Two Tesla V100 chips with 32 G of video memory each and a CPU frequency of 2.20 GHz are used.

5.1. Ball Detection and Tracking. The ball detection is built on a coarse to the fine method. Coarse processing is to search for possible locations of the ball in a larger spatial area, while exceptional processing is to precisely determine the shooter/ball location in a smaller search area. The position of the ball...
5.3. Shooting Detection and Tracking. We find that the edge in Figure 6. (140) with an aspect ratio equal to 0.3. Here, we use this different scales (i.e., the block length can be between 50, 60, that the image size of the shooter can be quantified into 10 the shooter. To determine the size of the shooter, we assume around the ball and search for the approximate location of the fusion architecture locates the ball, we can search the area for the shooters in the game. Using the color histogram features and luminance features to calculate this paper model score. Figure 5(a) represents the ball tracking from the initial position (dashed surrounds) of the previous frame to the best position (solid surrounds) of the current frame.

The intensity of each block in Figure 5(b) indicates a paper model score. Blocks with lower intensity indicate lower paper model scores, and the position is measured in pixels. The arrows indicate the tracking trajectory of the current search from the initial position in the previous frame to the best position of the current frame and the local maxima of this paper’s model score. We search for the brightest pixel, where the position indicates the predicted position of the ball.

5.2. Shooter Recognition. In basketball game videos, players from two different teams always wear two different colors. Here, we trained two different models of this paper model for the shooters in the game. Using the color histogram of the shooter’s clothing as a feature vector, two distinct models of this paper model are pretrained. Once the knowledge fusion architecture locates the ball, we can search the area around the ball and search for the approximate location of the shooter. To determine the size of the shooter, we assume that the image size of the shooter can be quantified into 10 different scales (i.e., the block length can be between 50, 60, ..., 140) with an aspect ratio equal to 0.3. Here, we use this paper model tracking to identify the size of the shooter. Results of the shooter identification experiments are shown in Figure 6.

5.4. Pitch Detection. The estimation of ball release timing is further complicated by the fact that players have different heights and arm lengths. Taller players may also have longer arms. To detect whether the ball was released, we compared the product of the pitcher’s distance ($h_1$) with the pitcher’s height and ratio ($r_1$). The pitcher’s distance in the vertical direction (i.e., $h_1$) can be obtained from the distance between the ball tracking and the upper boundary of the pitcher’s torso.

Here, we find a certain relationship between $h_1$ and $h_2$ (i.e., $r_1 = h_1/h_2$) when the player is about to release the ball. If the thrower’s distance is greater than the product, the thrower is releasing the ball and a shot event may occur; the relationship between $h_1$, $h_2$, and $r_1$ is shown in Figure 8. Here, ratio 1 (i.e., $r_1$) can be achieved by proper training. By training a sequence of approximately 70 ball release examples, we find that the average ratio $r_1 = 0.655$.

5.5. Ball and Ring Contact Detection. Ball ring contact is important information for identifying shot put events. We identify ball-loop contact by calculating the ball-loop distance between the ball and the basket. Ball tracking and basket tracking steps can be used to get the position of the ball and the basket. We can find the ball ring contact when the positions of the ball’s envelope and the basket’s envelope box are very similar to each other. As shown in Figure 9, the distance between the ball and the basket is obtained by calculating the following:

1. The upper boundary of the basket and the lower boundary of the ball as the distance Y of the ball ring
2. The left/right boundary of the basket and the right/left boundary of the ball as the distance X of the ball ring

5.6. Ball Direction Recognition. The direction of the ball bouncing after the ball ring contact detection is usually unpredictable. We must establish the direction of the bouncing ball’s motion relative to the basket. First, the ball position and basket position in the image are generated using ball tracking and basket detection techniques. When the direction of motion of the bouncing ball relative to the basket is downward, we can deduce that the video shot is a scoring shot. Figure 10(a) shows the scoring shots on the left and right sides, respectively. If the direction of motion of the bouncing ball is not downward, it may move upward or sideways, as illustrated in Figure 10(b).

5.7. Lens-Based Event Recognition. This subsection enables the retrieval of video footage by semantic understanding and classification of a large amount of input basketball video footage. In this section, we illustrate how to test the performance of our system and show experimental results in some aspects. We emphasize that the recognition rate metric is based on video shot units, while the updated probability of the model in this paper is based on frame units [7, 15].
Similarly, experimental video data are five basketball games consisting of two women’s basketball games and three men’s basketball games selected from the basketball video scenario described in the previous section. For the bottom node, preprocessing and feature extraction are utilized to obtain the shooter position, ball position, and basket position in each frame.

The interpretation of basketball video shots is achieved by classifying the video shots. The root node of the shot event consists of four states, indicating four different shot event categories. The input video shots are now classified according to the maximum posterior probability of the corresponding state. Accuracy and recall are calculated to measure the performance of shot event classification and score event detection indicated in Tables 1 and 2. It is clear that the performance of the schemes in this paper is very good; for example, both recall and precision in the short shot can reach more than 92%, which is a rare performance result in video data, and these are attributed to our dual-stream network setup and the integration of video knowledge.

5.8. Network Operations at the Aggregation Layer. Here, we evaluate the different locations in the network where the layers of this paper’s aggregation layer can be inserted. Specifically, we compare the placement of this paper’s aggregation layer after the last two evolutionary layers (conv4 3 and conv5 3) and the last fully connected layer (fc7), embedding subsampled conv and fc features from a set of videos in tSNE as shown in Figure 11. In each case, we trained to the block of layers before this paper’s aggregation layer; thus, in the case of conv4, conv4 1 indicates a loss, and in the case of fc7, fc7 indicates a loss. The results are shown in Figure 11 and clearly show that the best performance is obtained by aggregating the last evolutionary layer (conv5 3). fc6 characteristics obtain similar performance to fc7.

We believe that this is due to two reasons. First, pooling on fully connected layers prevents the aggregation layers in this paper from modeling spatial information, since these layers already compress a lot of information. Second, fc7 features are more semantic, so features from different frames are already comparable to each other and do not take advantage of the modeling capabilities of the aggregated layers in this paper; i.e., they would normally fall in the same cell. To test this hypothesis, we visualize the conv5 3 and fc7 appearance features from the same frame using public tSNE [22, 30] embedding in Figure 11. These plots clearly show that fc7 features from the same video (displayed in the same color) already resemble each other (Figure 11(b)), while conv5 3 features are more diverse and can benefit from the ability of this paper to aggregate layers to capture complex distributions in the feature space (Figure 11(a)), as described in Section 3.1.
5.9. Different Fusion Effects. It can be clearly seen from Table 3 that late fusion has the best accuracy under different basketball skills. For example, in terms of basic facts, late fusion is 0.4 higher than intermediate fusion. Similarly, the other two aspects are also better than other hybrid model schemes. This is also the reason why this paper selects late fusion as the fusion method of our model.

Figure 7: (a) Search range of basket detection. (b) Detection of the basket. (c) Edge information of the object.

Figure 8: Relationship between $h_1$, $h_2$, and $r_1$. $h_1$ is the pitcher distance, $r_1$ is the ratio 1, and $h_2$ is the pitcher clothing height.
Figure 9: Ball ring distance.

(a) A bouncing ball moving sideways (b) A bouncing ball moving downward.

Table 1: Shooting event classification results.

| Video snapshot | Basic facts | Correct | Miss | False alarm | Recall   | Accuracy |
|----------------|-------------|---------|------|-------------|----------|----------|
| Close-up       | 34          | 35      | 3    | 1           | 0.9142   | 0.9652   |
| Medium view    | 20          | 11      | 5    | 5           | 0.8362   | 0.7542   |
| Far view       | 13          | 13      | 2    | 3           | 0.8324   | 0.8154   |
| Penalty shot   | 17          | 17      | 2    | 6           | 0.8865   | 0.8423   |
| Total          | 84          | 76      | 12   | 15          | 0.8722   | 0.8664   |
6. Conclusions

In this paper, we propose a successful spatiotemporal video feature aggregation method for basketball video semantic extraction. Our approach is end-to-end trainable, and our method is general enough to be applied as an aggregation layer-CNN layer in this paper for future video architecture, which may be useful for related tasks such as spatiotemporal localization of human behavior in long videos. We designed a dual-stream network combined with a reasonable spatiotemporal feature aggregation scheme trying to bridge the gap between low-level features and high-level semantic events, which can fuse knowledge from different video sources to design an automatic semantic system for basketball games videos. Experiments show that the model in this paper can accurately identify different shooting items or player actions in video footage of basketball games.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest

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

Acknowledgments

The authors would like to thank ActionVLAD [35] for providing the improvement and innovation ideas of this paper.

References

[1] L. Havasi and Z. Szlavik, “Using location and motion statistics for the localization of moving objects in multiple camera surveillance videos,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, IEEE, Barcelona, Spain, 2010.
[2] H. Kerr, B. Bakken, and G. House, “Future directions in sports-related concussion management,” Clinics in Sports Medicine, vol. 40, no. 1, pp. 199–211, 2021.
[3] J. A. Blázquez López, S. García Martínez, A. Ferriz Valero, and J. Olaya Cuartero, “Cuantificación de la carga de entrenamiento y competición: análisis comparativo por posiciones en un equipo de la Liga Española de Baloncesto Oro [Quantification of training and competition load: comparative analysis by position in a Spanish Golden Basket],” Retos, vol. 42, pp. 882–890, 2021.
[4] U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger, “X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words,” IEEE Transactions on Medical Imaging, vol. 30, no. 3, pp. 733–746, 2011.
[5] R. Li, Q. Liu, J. Gui et al., “Indoor relocalization in challenging environments with dual-stream convolutional neural networks,” IEEE Transactions on Automation Science and Engineering, vol. 15, pp. 651–662, 2017.
[6] Z. Zhao, D. A. Klindt, A. Maia Chagas et al., “The temporal structure of the inner retina at a single glance,” Scientific Reports, vol. 10, no. 1, p. 4399, 2020.

Table 2: Score event detection results.

| Snapshot     | Basic facts | Correct | Miss | False alarm | Recall | Accuracy |
|--------------|-------------|---------|------|-------------|--------|----------|
| Score        | 35          | 35      | 3    | 3           | 0.8842 | 0.9152   |
| No score     | 52          | 41      | 4    | 4           | 0.9142 | 0.9242   |
| Total        | 87          | 76      | 7    | 7           | 0.9234 | 0.9258   |

Table 3: Accuracy of different hybrid models in different basketball skills.

|                  | Basic facts | Correct | Miss |
|------------------|-------------|---------|------|
| Early fusion     | 0.8652      | 0.7512  | 0.8743|
| Intermediate fusion | 0.8832  | 0.7486  | 0.8859|
| Late fusion      | 0.9231      | 0.8159  | 0.8959|

Figure 11: Embedding subsampled conv and fc features from a set of videos in tSNE. (a) conv5. (b) fc7.
[7] A. Chadha, A. Abbas, and Y. Andreopoulos, “Video classification with cnns: using the codec as a spatio-temporal activity sensor,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 2, pp. 475–485, 2017.

[8] R. Arandjelovic, P. Gronat, A. Torii et al., “NetVLAD: CNN architecture for weakly supervised place recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5297–5307, Las Vegas, NV, USA, June 2016.

[9] A. Kamilaris and F. X. Prenafeta-Boldú, “Deep learning in agriculture: a survey,” Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.

[10] L. Wang, Y. Qiao, and X. Tang, “Action recognition with trajectory-pooled deep-convolutional descriptors,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 4305–4314, Boston, MA, USA, June 2015.

[11] C. M. Chen and L. H. Chen, “Novel framework for sports video analysis: a basketball case study,” in Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), pp. 961–965, IEEE, Paris, France, October 2014.

[12] G. Nagarajan and R. I. Minu, “Multimodal fuzzy ontology creation and knowledge information retrieval,” Proceedings of the International Conference on Soft Computing Systems, Springer, New Delhi, India, pp. 697–706, 2016.

[13] D. W. Tjondronegoro and Y.-P. P. Chen, “Knowledge-discounted event detection in sports video,” IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 40, no. 5, pp. 1009–1024, 2010.

[14] Z. Gao, P. Wang, H. Wang, M. Xu, and W. Li, “A review of dynamic maps for 3D human motion recognition using ConvNets and its improvement,” Neural Processing Letters, vol. 52, no. 2, pp. 1501–1515, 2020.

[15] X.-B. Fu, S.-L. Yue, and D.-Y. Pan, “Camera-based basketball scoring detection using convolutional neural network,” International Journal of Automation and Computing, vol. 18, no. 2, pp. 266–276, 2021.

[16] B. Li and X. Xu, “Application of artificial intelligence in basketball sport,” Journal of Education, Health and Sport, vol. 11, no. 7, pp. 54–67, 2021.

[17] R. A. Minhas, A. Javed, A. Irtaza, M. T. Mahmood, and Y. B. Joo, “Shot classification of field sports videos using AlexNet Convolutional Neural Network,” Applied Sciences, vol. 9, no. 3, p. 483, 2019.

[18] M. Rafiq, G. Rafiq, R. Ayeman, G. S. Choi, and S.-I. Jin, “Scene classification for sports video summarization using transfer learning,” Sensors, vol. 20, no. 6, p. 1702, 2020.

[19] S. Su, J. Pyo Hong, J. Shi et al., “Predicting behaviors of basketball players from first person videos,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1501–1510, Honolulu, HI, USA, July 2017.

[20] K. Lu, J. Chen, J. J. Little, and H. He, “Lightweight convolutional neural networks for player detection and classification,” Computer Vision and Image Understanding, vol. 172, pp. 77–87, 2018.

[21] Y. Yang, “Research on sports action training method based on generative confrontation network model and artificial intelligence,” Journal of Intelligent and Fuzzy Systems, vol. 4, pp. 1–11, 2021.

[22] N. A. Rahmad, N. A. J. Sufri, N. H. Muzamil, and M. A. As’ari, “Badminton player detection using faster region convolutional neural network,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 14, no. 3, pp. 1330–1335, 2019.

[23] W. Liu, C. C. Yan, J. Liu, and H. Ma, “Deep learning based basketball video analysis for intelligent arena application,” Multimedia Tools and Applications, vol. 76, no. 23, pp. 24983–25001, 2017.

[24] J. Zemgulys, V. Raudonis, R. Maskeliūnas et al., “Recognition of basketball referee signals from real-time videos,” Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 3, pp. 979–991, 2020.

[25] D. Acuna, “Towards real-time detection and tracking of basketball players using deep neural networks,” in Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 4–9, Long Beach, CA, USA, January 2017.

[26] G. Junjun, “Basketball action recognition based on FPGA and particle image,” Microprocessors and Microsystems, vol. 80, Article ID 103334, 2021.

[27] L. Wu, Z. Yang, Q. Wang et al., “Fusing motion patterns and key visual information for semantic event recognition in basketball videos,” Neurocomputing, vol. 413, pp. 217–229, 2020.

[28] X. Fu, K. Zhang, C. Wang, and C. Fan, “Multiple player tracking in basketball court videos,” Journal of Real-Time Image Processing, vol. 17, no. 6, pp. 1811–1828, 2020.

[29] L. Shen and W. Wang, “Analysis of technical features in basketball video based on deep learning algorithm,” Signal Processing: Image Communication, vol. 83, Article ID 115786, 2020.

[30] L. Wu, Z. Yang, J. He et al., “Ontology-based global and collective motion patterns for event classification in basketball videos,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 7, pp. 2178–2190, 2019.

[31] L. Liu, “Objects detection toward complicated high remote basketball sports by leveraging deep CNN architecture,” Future Generation Computer Systems, vol. 119, pp. 31–36, 2021.

[32] J. Quiroga, H. Carrillo, E. Maldonado et al., “As seen on tv: automatic basketball video production using Gaussian-based actionness and game states recognition,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 894–895, Seattle, WA, USA, June 2020.

[33] B. Mahaseni, E. R. M. Faizal, and R. G. Raj, “Spotting football events using two-stream convolutional neural network and dilated recurrent neural network,” IEEE Access, vol. 9, Article ID 61929, 2021.

[34] K. Rangasamy, M. A. As’ari, N. A. Rahmad, and N. F. Ghazali, “Hockey activity recognition using pre-trained deep learning model,” ICT Express, vol. 6, no. 3, pp. 170–174, 2020.

[35] R. Girdhar, D. Ramanan, A. Gupta et al., “Actionvlad: learning spatio-temporal aggregation for action classification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 971–980, Honolulu, HI, USA, July 2017.

[36] V. Choutas, P. Weinzaepfel, J. Revaud et al., “Potion: pose motion representation for action recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7024–7033, Salt Lake City, UT, USA, June 2018.

[37] M. Trinidad-Fernández, D. Beckwée, A. Cuesta-Vargas et al., “Differences in movement limitations in different low back pain severity in functional tests using an RGB-D camera,” Journal of Biomechanics, vol. 116, Article ID 110212, 2021.