Hindawi
Advances in Mathematical Physics
Volume 2022, Article ID 9965764, 10 pages
https://doi.org/10.1155/2022/9965764

*Research Article*

# Motion Feature Retrieval in Basketball Match Video Based on Multisource Motion Feature Fusion

**Biao Ma and Minghui Ji**

*School of Physical Education, Huainan Normal University, Huainan 232038, China*

Correspondence should be addressed to Minghui Ji; jiminghui@hnnu.edu.cn

Received 19 November 2021; Accepted 17 December 2021; Published 11 January 2022

Academic Editor: Miaochao Chen

Copyright © 2022 Biao Ma and Minghui Ji. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Both the human body and its motion are three-dimensional information, while the traditional feature description method of two-person interaction based on RGB video has a low degree of discrimination due to the lack of depth information. According to the respective advantages and complementary characteristics of RGB video and depth video, a retrieval algorithm based on multisource motion feature fusion is proposed. Firstly, the algorithm uses the combination of spatiotemporal interest points and word bag model to represent the features of RGB video. Then, the directional gradient histogram is used to represent the feature of the depth video frame. The statistical features of key frames are introduced to represent the histogram features of depth video. Finally, the multifeature image fusion algorithm is used to fuse the two video features. The experimental results show that multisource feature fusion can greatly improve the retrieval accuracy of motion features.

## 1. Introduction

The traditional training method is for coaches to make a training plan according to their own training theory and training experience, combined with the skill level of basketball players. The subjectivity of this training mode is very strong, and coaches need to spend a lot of time to analyze the sports characteristics of athletes. And it is difficult to objectively evaluate the training effect of athletes [1, 2]. The core of modern sports training is precision and efficiency. If coaches can accurately control the sports characteristics of athletes, the training effect can be greatly improved. Therefore, collecting and analyzing the sports data of basketball players and searching the sports characteristics are of great significance to improve the scientifically of the coaches' training plan and the training effect of the athletes, and it is a new research direction [3, 4].

The information represented by the motion video has diversified characteristics, and the user can search the video according to the diversified features in the video [5]. One is to search for similar or identical video clips in the video library by submitting short videos [6]. The other is to search the video tags by entering keywords [7]. The traditional video search method is mainly based on the keywords marked by video content manually [8, 9]. First of all, the main meaning of the video content is annotated by manual text. Then, form keywords to describe the video content. Finally, the video is retrieved according to the keywords. This method of manual annotation is easy to implement, and the query effect is better in some cases. However, due to the lack of depth information, the discrimination of its feature description is low [10].

(1) Annotated video keywords lack objectivity, and different annotators may have different opinions on the same video. Therefore, there will be subjective thoughts of the annotators, leading to a certain deviation in the keywords of video annotation

(2) The labeled video keywords are not comprehensive. When users want to obtain a person's video resources, manual annotation cannot guarantee whether all the information in the video has been

marked. Therefore, when users search from the video database according to the keywords manually marked, they cannot find the corresponding video resources

(3) The current video data surge, but the manual cannot continue to annotate for a long time and easy to fatigue. Therefore, manual labeling is slow

Therefore, in order to solve the above problems, according to the respective advantages and complementary characteristics of RGB images of video and depth video, the researchers propose a multisource motion feature fusion retrieval algorithm. Firstly, the algorithm uses the combination of spatiotemporal interest points and word bag model to represent the features of RGB video. Then, the directional gradient histogram is used to represent the depth video frame, and the key frame statistical feature is introduced to represent the depth video histogram feature. Finally, the multifeature image fusion algorithm is used to fuse the two kinds of video features to realize the motion feature retrieval.

## 2. Related Works

With the development of the network and self-media, the presentation level of video files increases exponentially. In the face of a large number of video data, how to retrieve interesting videos from these video libraries quickly and effectively has become a difficult problem in today's information age [11]. Video data has the characteristics of large amount of data and high dimension; so, it needs to consume a lot of memory and search time in the process of retrieval [12]. In the process of video processing based on multifeature fusion, key frame extraction is one of the key steps. Because the amount of video data is huge and complicated, and it takes a lot of time, the efficiency of video retrieval cannot meet the needs of users. Therefore, in order to compress the amount of data and reduce the computational complexity of matching features, redundant frames need to be discarded in the shot. Finally, one or more frames which can express the main content of the video are selected as key frames to improve the efficiency of video retrieval. In recent years, there have been many key frame extraction algorithms, but their emphasis is different.

Huang and Wang [13] proposed to extract key frames based on shot boundary. This method can directly extract video frames as key frames without calculation and is convenient and fast. The main purpose of this method is to extract the first frame and the last frame of the video shot as the key frame or to sample the video frame as the key frame according to a certain time interval. However, the extracted video frame is random and cannot correctly and completely express the information in the lens.

Wang et al. [14] proposed a method based on motion analysis to extract key frames. This method calculates the optical flow and analyses the motion of the object inside the lens and takes the frame of the minimum motion as the key frame. Although this method is sensitive to the motion of the object, it is impossible to judge the primary and secondary of the moving object in the video frame. And it is not sensitive to the features of the internal objects in the video; so, it is difficult to extract and identify the main target features in the video frame.

Gui and Lu [15] proposed a key frame extraction algorithm for foreground moving target feature extraction based on the background modeling algorithm. This method uses the SIFT algorithm to compare the similarity between adjacent frames and the average value of segment similarity to determine the key frame. However, the amount of calculation of this method is too large, and when the amount of data is huge and the video definition is high, the efficiency of video retrieval will be limited by the running speed of the machine.

The traditional key frame extraction method has been unable to meet the needs of users. At present, the common method is to use the technology of extracting key frames to establish the index and then carry on the video retrieval. The key frame extraction methods are key frame extraction based on color feature [16] and key frame extraction based on video content [17]. However, in the above methods, the former method of extracting key frames will cause some redundancy, while the latter is not effective in the case of large amount of video and numerous contents. At present, the popular video retrieval methods, such as key frame extraction and video retrieval based on deep learning, video summary generation algorithm based on $k$-means clustering, and all extract key frames based on global features, can only consider video frames as a whole. However, due to the large amount of video data and the high redundancy of video adjacent frames, the efficiency of video retrieval is reduced.

In order to improve the quality and efficiency of video key frame extraction, a video key frame extraction algorithm based on optimal distance clustering and feature fusion expression is proposed in reference [18]. The algorithm solves the dependence of unsupervised clustering on threshold, takes into account the changes of moving objects and abnormal environment in the video, and has good performance and adaptability. However, the algorithm takes the average features of all frames in the video to represent the video features and ignores the features of different parts of the characters when taking the characters as the key research object.

Because the video is rich in content and feature information, using a single feature to retrieve video will be limited by the diversity of video types. In order to solve this problem, a key frame extraction method based on multifeatures is proposed in reference [19]. In this method, color features, wavelet statistical features, and SIFT local features are used to calculate the comprehensive similarity matrix between video frames. Then, the shot frames are grouped by an improved spectral clustering algorithm, and the central frame of each group is selected as the key frame. The number of key frames is estimated by calculating the minimum of clustering instability. However, this method takes the center frame of the lens as the key frame. There will be some errors for the gradient lens. When there are characters, the recognition of the characters in the middle frame of the video sequence may be relatively poor, which has a great impact on video retrieval.

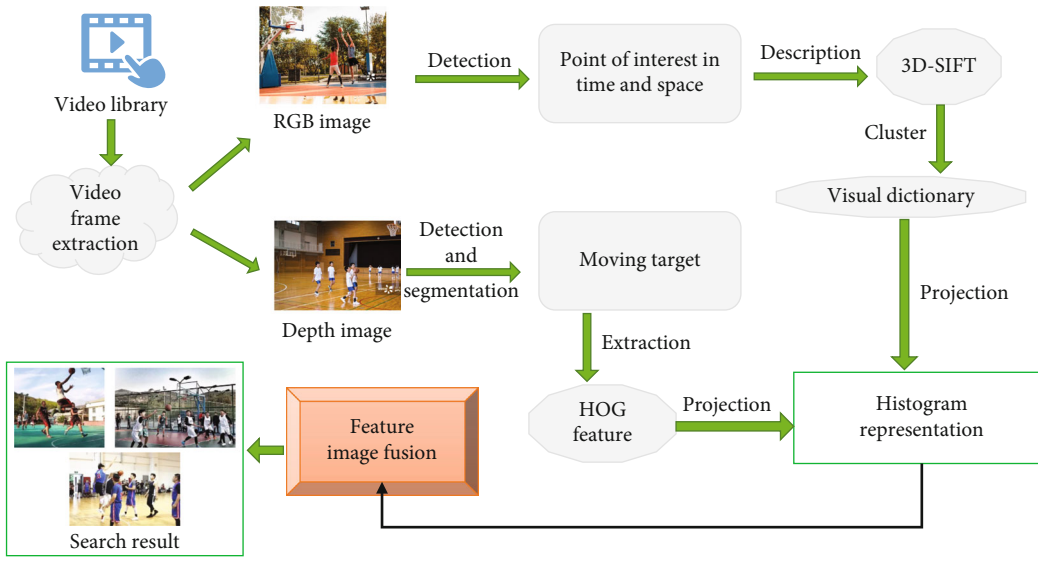



Figure 1: Algorithm structure block diagram.

Literature [20] combines image entropy and density clustering to extract further features from key frames in gesture video, so that the recognition efficiency can be improved. In addition, a feature fusion strategy is proposed to further improve the feature representation so as to improve the recognition performance. In order to overcome the great change of face quality in video stream and to improve the processing speed of facial recognition system, a key frame extraction (KFE) engine with graphics processing unit acceleration based on convolution neural network is proposed in reference [21, 22]. The purpose of this paper is to extract the key frames of high-quality human face correctly and quickly. The KFE engine based on CNN can greatly reduce the total face processing time in video recognition and improve the recognition accuracy of the back-end of face recognition. And the KFE engine proposed by this method is suitable for different facial recognition backbends. An effective key frame extraction technique is proposed in reference [23]. This method detects key frames effectively by extracting the unified local binary pattern of video frames. The distance between the unified local binary patterns of consecutive frames is calculated and compared with the threshold to extract key frames.

According to the above literature, in the process of key frame-based extraction, although the motion features in the video can be retrieved, there are the following three problems:

(1) The detection is not accurate. The phenomenon that the detected image is not detected or the detected image is a nonathlete motion process, resulting in errors caused by the fact that the extracted features do not contain motion features

(2) In the mass video, there is a large amount of calculation for comparing each frame when retrieving video, and the retrieval speed is limited to some extent

In order to solve these problems, a retrieval method of multisource feature fusion is proposed in this paper. The algorithm block diagram is shown in Figure 1. The algorithm combines RGB video features with depth video features. Finally, the multifeature image fusion algorithm is used to fuse the two kinds of video features to realize the motion feature retrieval.

The feature utilization of single feature is not as high as that of multifeature. Therefore, RGB image is used for segmentation effect is also very good, but multisource features work better than RGB images alone. Therefore, RGB images are first used for detection in this paper, and then directional gradient histogram is used for feature representation of depth video frames.

## 3. Retrieval Algorithm Based on Multisource Feature Fusion

*3.1. RGB Video Feature Representation.* In the video of human interactive behavior, spatiotemporal points of interest can correctly locate the regions with obvious motion in the video sequence with less information and have strong robustness to environmental changes and local occlusion [24, 25]. The algorithm based on spatiotemporal interest point representation behavior is widely used in the field of human behavior recognition; so, this paper uses the combination of spatiotemporal interest point and visual word bag (VWB) model [26] to represent the feature of RGB video.

As shown in Figure 2, the two-person interactive behavior representation algorithm using the combination of spatiotemporal interest points and word bag model consists of three parts: spatiotemporal interest detection, feature description, and dictionary establishment. This clustering algorithm is nearest neighbor clustering algorithm, and the time-consuming in this algorithm is very little.

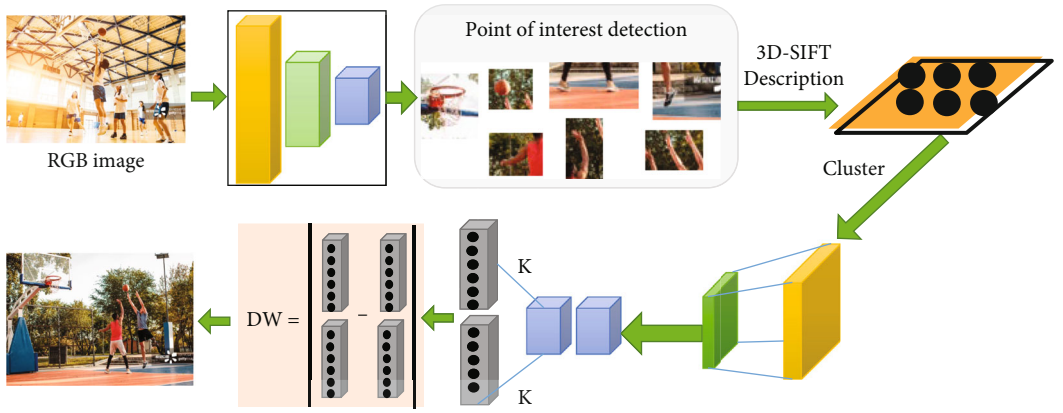



Figure 2: Generate a graphical representation of BOW descriptors.

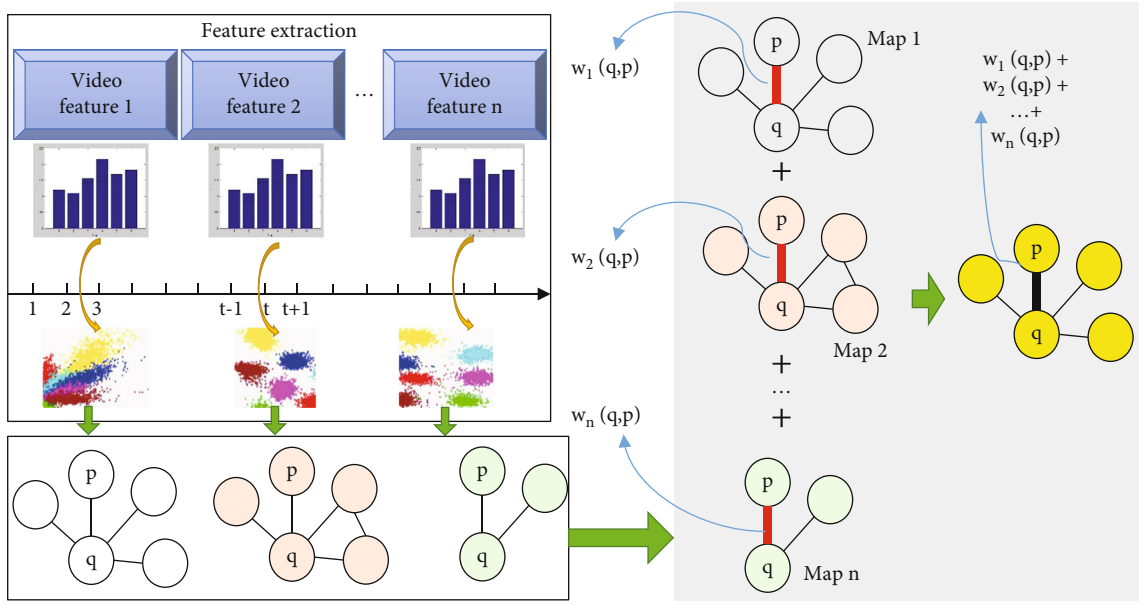



Figure 3: Multigraph fusion process.

Two separate linear filters of space and time are applied to the detector to extract rich spatiotemporal points of interest from the video sequence to fully capture the characteristics of human behavior in the video sequence. Here, it is applied to the local feature extraction of two-person interaction in color video, and its expression is as follows:

$$\mathrm{RGB} = (A \bullet \mathrm{Gau}(\mathrm{ss}) \bullet T_h)^2 + (A \bullet \mathrm{Gau}(\mathrm{ss}) \bullet T_w)^2. \quad (1)$$

Among them, Gau ($a$, $b$; ss) is a two-dimensional Gaussian smoothing kernel function, which is used for spatial domain filtering.

$$\mathrm{Gau}(a, b; \mathrm{ss}) = \frac{e^{-\left(a^2+b^2\right)/2\mathrm{ss}^2}}{2\pi \mathrm{ss}^2}. \quad (2)$$

Among them, Th and Tw are the orthogonal components of one-dimensional Gabor function, which are used for time domain filtering.

$$T_h = \cos\left(2pwt^2\right) e^{-\frac{t^2}{tt^2}},$$
$$T_w = \sin\left(2\pi wt^2\right) e^{-\frac{t^2}{tt^2}}. \quad (3)$$

Among them, ss and tt correspond to space and time scale, respectively.

The detected points of interest need to meet two conditions: the response function thr is larger than the set threshold, and the local maximum is obtained in a certain neighborhood. The selection of the threshold size can control the number of detected points of interest.

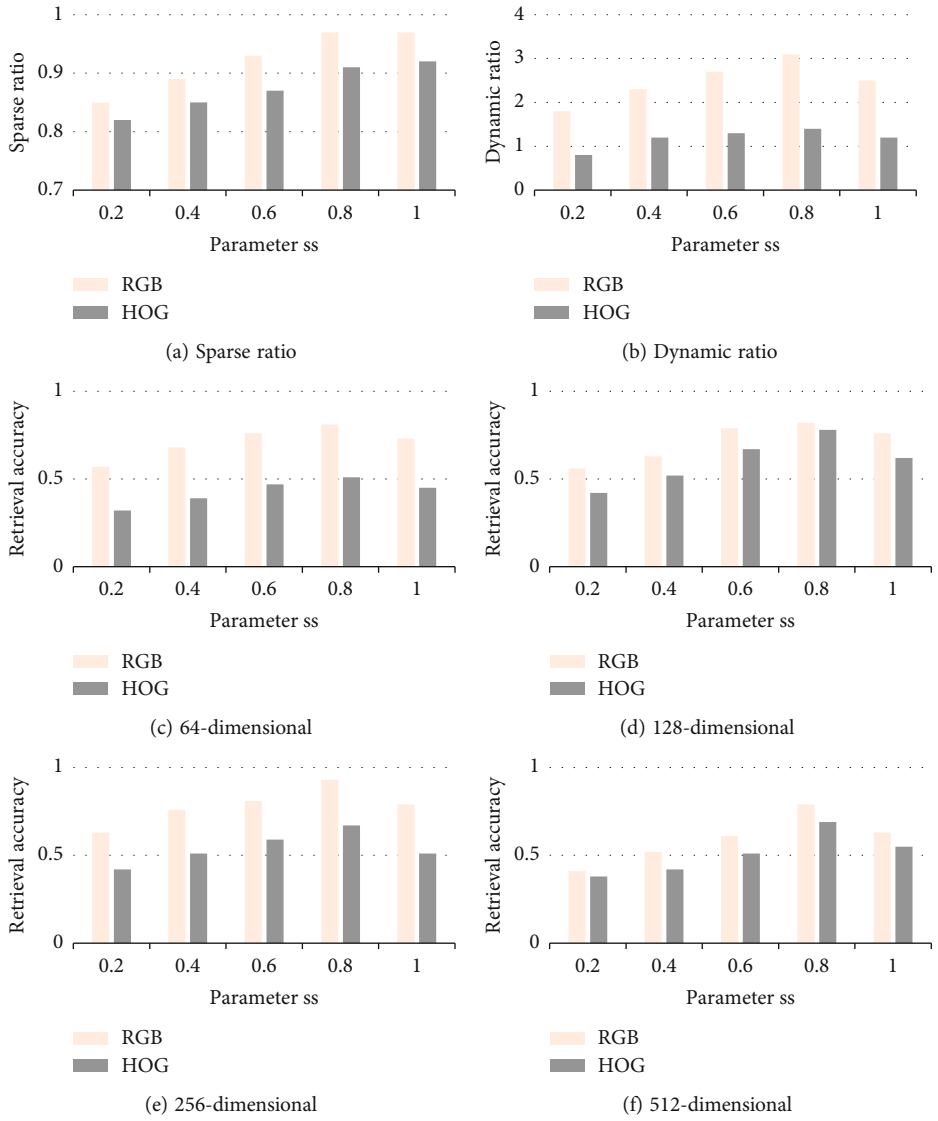



Figure 4: Experimental results under different parameters ss.

The 3D-SIFT method [27] is used to describe the points of interest, and the steps are as follows:

(1) The spatiotemporal cube is extracted from the neighborhood around the point of interest and divided into fixed-size unit subcubes

(2) The polyhedral sphere is used to calculate the spatiotemporal gradient histogram of each unit cube

(3) All the unit cube histograms are combined to form 3D-SIFT descriptors of spatiotemporal points of interest

In this paper, the cube with the size of $H \times W \times C$ pixels is divided into $N$ subcubes. $J$ faces are used to describe $k$ gradient directions; so, the feature dimension of each point is thr × sC, which is used to describe the spatiotemporal interest features of interactive behavior.

A simple and effective $K$-means clustering algorithm is used to cluster all feature vectors. First, $K$ samples are randomly selected as the clustering center in the data set. According to a certain similarity measure, all samples are divided into the classes represented by the nearest cluster center, and $K$ clusters are formed. Then, recalculate the cluster centers for the $K$ clusters and reclassify the samples according to the new clustering center. This iteration continues until the criterion function of formula (4) converges.

$$Y = \sum_{i=1} \sum_{j \in C_i} (j - \mu_i)^2, \tag{4}$$

where $Y$ is the sum of the square errors of all the research objects, $j$ is the point of space, that is, the data object, and $\mu i$ is the average value of the cluster $C_i$.

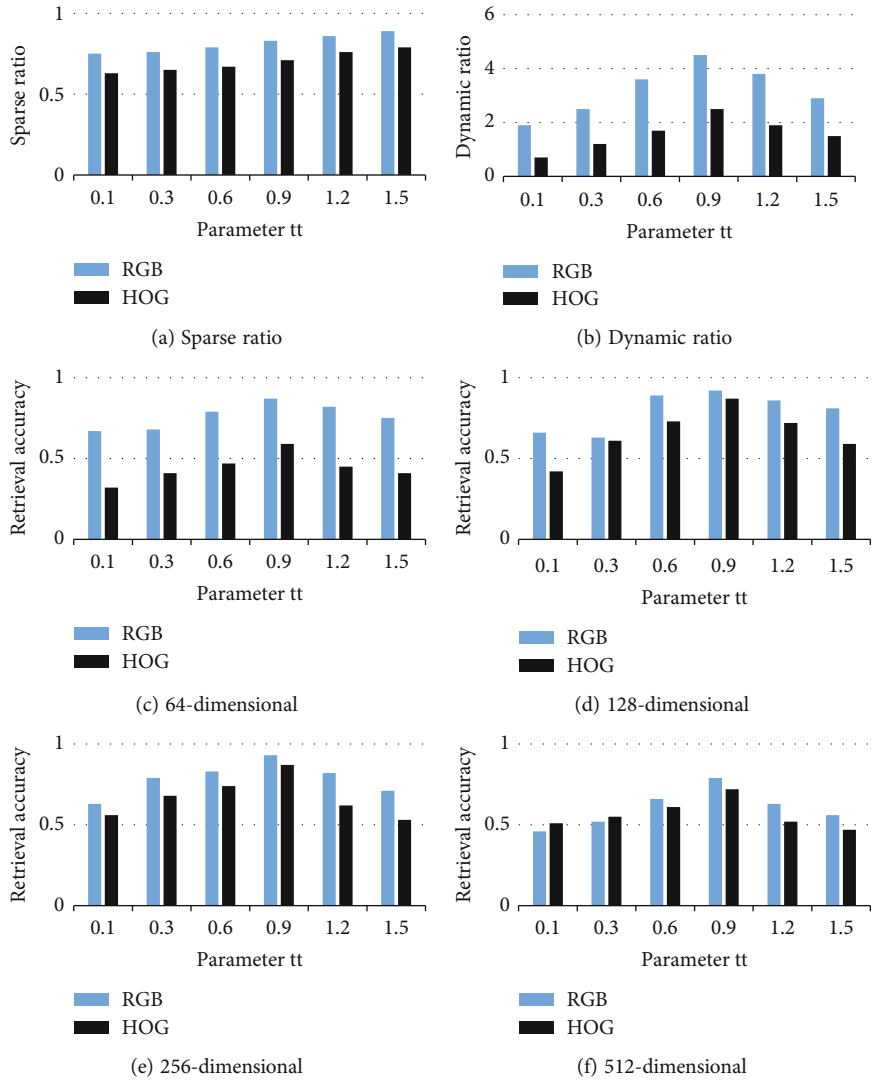



Figure 5: Experimental results under different parameters tt.

Each clustering center is regarded as a word in the dictionary, and each feature vector is represented by the word closest to it. Then, count the frequency of the words in the dictionary in the video and construct the histogram representation of the video.

*3.2. Depth Video Feature Representation.* Depth image, also known as distance image, refers to the image that takes the distance from the image collector to each point in the scene as the pixel value. When there is a certain distance between the foreground and the background of the human body, the depth image can be directly reflected by the grayscale value information. Considering that HOG [28] can better extract and represent the edge information around the human body, the HOG feature is selected for the global representation of depth video. In this paper, the frame difference method is used to detect the moving target in the range image.

The construction of HOG feature is realized by calculating and counting the gradient direction histogram of the local region of the image. The extraction operation for the gradient can not only capture the contour and texture information but also reduce the impact of illumination changes. The gradient of the Abscissa and Abscissa directions of a pixel $(i, j)$ can be expressed as

$$\begin{aligned} T_i(i,j) &= A(i+1,j) - A(i-1,j), \\ T_j(i,j) &= A(i,j+1) - A(i,j-1), \end{aligned} \tag{5}$$

where $T_i(i,j)$, $T_j(i,j)$, and $A(i,j)$ represent the horizontal gradient, vertical gradient, and pixel value at the pixel point in the input image, respectively.

The amplitude and direction of the gradient at the pixel $(i, j)$ are expressed as follows:

$$\begin{aligned} T(i,j) &= \sqrt{\left(T_i(i,j)\right) + T_j(i,j)\left)\left(T_i(i,j) - T_j(i,j)\right)\right.}, \\ \alpha(i,j) &= \cot\frac{T_i(i,j)}{T_j(i,j)}. \end{aligned} \tag{6}$$

HOG feature extraction is carried out in the region where the moving target is located. That is, the gradient image is equally divided into $j \times k$ nonoverlapping subregions, and the contribution weight of the gradient of pixels in each region to $K$ different directions is calculated, which is superimposed on all gradient directions to construct the gradient direction histogram. Finally, the $j \times k \times K$ dimensional feature vector of the moving object in each frame is obtained.

In order to effectively integrate with the feature of the point of interest, the statistical feature of key frame is used to represent the depth video. That is, the $K$-means clustering method is used to generate the key frame feature library for the HOG features of the training video. Then, according to the similarity measure function, the frequency of all the frame features in a video to be tested in the key frame feature library is counted, and the statistical histogram representation of the depth video is obtained.

*3.3. Multifeature Image Fusion Method.* In this algorithm, the input is $n$ different feature graphs $F(v, e, w) = \{F_1, F_2, \cdots\}$, and the fused graph $F' = (v, e, w)$ satisfies the following conditions:

$$v = \lim_{n \longrightarrow \infty} \sum_{d=1}^{n} v_d,$$

$$e = \lim_{n \longrightarrow \infty} \sum_{d=1}^{n} e_d, \tag{7}$$

$$w(i, j) = \lim_{n \longrightarrow \infty} \sum_{d=1}^{n} i * w_d(i, j).$$

Figure 3 intuitively shows the fusion process. First, all the nodes existing in the graph are put into the new graph. For two of the nodes $p$ and $q$, if there are edges between them in multiple graphs, the edge weight $w_i(p, q)$ is added as the edge weight of the new graph $w(p, q)$. If there is no edge between two nodes in any graph, then no edge is added in the new graph, and it can be considered that there is an edge between them with an edge weight of 0.

On this basis, this paper adds the shortest edge clustering algorithm. Firstly, the TTNG map is built by using the initial retrieval sequence matrix of multiple features, and the local TTNG map of each sample is merged into a large image. The edge weights between two nodes are set to the sum of edge weights in two local graphs, and $n$ TTNG graphs are obtained. Finally, the shortest edge clustering algorithm is used for the nodes in the graph to get the reordering sequence.

In this algorithm, the complexity of constructing a graph is $O(k^3 \cdot \log_2 k)$, and the total time of $M$ graphs is $O(M \cdot k^3 \cdot \log_2 k)$. The complexity of TTNG constructed graph is $O(k^2)$, and the complexity of integrating $M$ graphs is $O(M \cdot k^2)$. The complexity of shortest edge clustering algorithm is $O((n(n+1)/2)2)$, and the complexity of retrieving results from the graph is $O(K \cdot \log_2 k)$. To sum up, the time complexity of the whole algorithm is $O(M \cdot k^3 \cdot \log_2 k + (2M + 1)K^2 + (n(n+1)/2)2 + K \cdot \log_2 k)$.

Table 1: Retrieval results of different features.

| Features | | RGB | HOG |
|---|---|---|---|
| Retrieval accuracy | Near | 0.81 | 0.72 |
| | Far away | 0.56 | 0.67 |
| | Exchange | 0.78 | 0.71 |
| | Hug | 0.95 | 0.93 |
| | Boxing | 0.95 | 0.91 |
| | Push | 0.78 | 0.83 |
| | Shake hands | 0.93 | 0.88 |

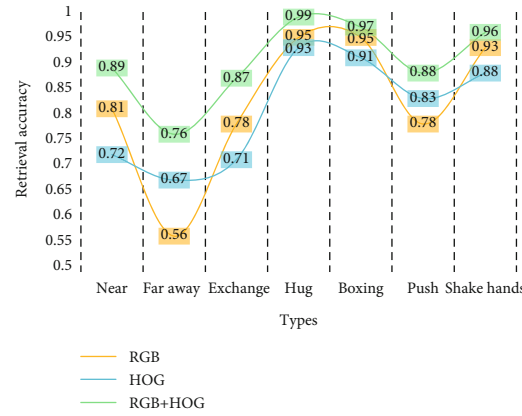



Figure 6: Comparison of single-feature and multifeature retrieval.

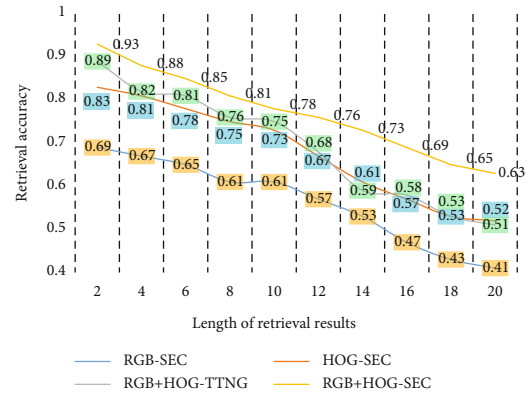
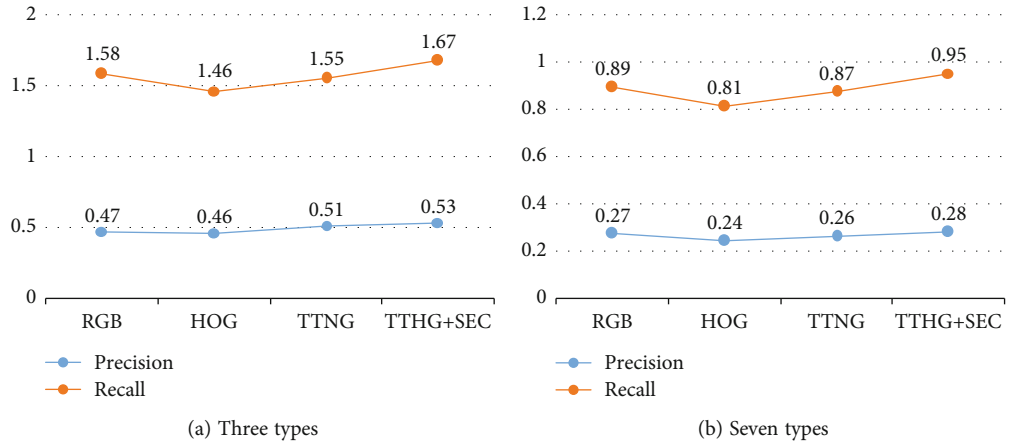



Figure 7: Retrieval results of TTNG and SEC in three kinds of data.

## 4. Experimental Results and Analysis

*4.1. Selection of Parameters.* This section compares the characteristics of the features under different parameters ss, tt, and the classification ability of the two features at the same scale. Among them, the feature learning dimension is set to [64, 128, 256, 512].

Compared with the depth video feature, the RGB feature is more than 90% sparse on average at the same feature size, and the dynamic ratio is doubled (as shown in Figures 4 and 5).




(a) Three types (b) Seven types

Figure 8: Precision and recall on different data sets.

Table 2: Confusion matrix after multifeature fusion.

| | 3-pointer | Free throw | Layup | Dunk | Snatch off | Accuracy |
|---|---|---|---|---|---|---|
| 3-pointer | 393 | 3 | 92 | 3 | 67 | 0.724 |
| Free throw | 3 | 153 | 7 | 0 | 3 | 0.961 |
| Layup | 72 | 32 | 823 | 17 | 171 | 0.747 |
| Dunk | 19 | 3 | 3 | 11 | 7 | 0.268 |
| Snatch off | 14 | 8 | 45 | 0 | 392 | 0.893 |
| Average accuracy | — | — | — | — | — | 0.718 |

The two kinds of features are applied to retrieval and recognition at the same time, and the recognition experimental results also show that the accuracy of RGB features is more improved than that of deep video features. The recognition effect of the 256-dimensional feature has been 0.93, which is much higher than that of the traditional coefficient feature.

*4.2. Retrieval Results of Different Features.* It can be seen from Table 1 that the recognition effect of "Push" and "Far Away" interactive action of the feature representation used in RGB video is worse than that of HOG feature in depth video, while the recognition effect of "Hug" is better. For two-person interactive action recognition, the depth image carries more action information; so, the recognition rate based on the depth image is much higher than that of the RGB video.

According to Figure 6, we can draw the following conclusions: multifeature retrieval has higher retrieval efficiency than single-feature retrieval, and the recall rate of retrieval is improved. The use of a single feature for retrieval is always subject to many limitations, and the retrieval efficiency is not ideal. The comprehensive use of multifeature retrieval can achieve the effect of complementary advantages. For each single feature, the optimal weight distribution is obtained to improve the retrieval efficiency.

*4.3. Feature Image Fusion Result.* The video retrieval framework proposed in this paper mainly includes three parts: feature extraction, reordering, and extracting reordering from the graph. First of all, the features of the video samples are extracted, and the RGB features and depth features are obtained, respectively. The TTNG mapping method is used to fuse the two, and the shortest edge clustering algorithm proposed in this paper is implemented. Finally, the MFR algorithm is used to extract the rearrangement sequence from the graph and return the retrieval results.

The retrieval results of the first 20 samples of the multifeature fusion SEC algorithm in the three types of data sets are shown in Figure 7, in which the rearrangement results of TTNG and SEC algorithms are the optimal results selected after many iterations. As can be seen from the figure, the fusion effect of the TTNG algorithm on the two features used in this paper is not very good, because the two features are extracted using the same network and do not meet the conditions that are independent of each other. SEC algorithm has a further improvement on the basis of TTNG and finally can get a better rearrangement effect.

Figure 8 shows the accuracy of the first 20 retrieval results of the proposed video retrieval framework on three types of data sets and seven types of data sets, respectively. On the three types of data sets, the accuracy of the first 20 search results is 52.88%, which is 5.76% higher than that of RGB features and 13.84% higher than that of HOG features. On the seven types of data sets, the accuracy of the first 20 search results is 28.75%, which is 5.78% higher than the RGB feature and 17.88% higher than the HOG feature.

Next, we will continue to test the event classification after multifeature fusion based on the optimal training model, and the results are shown in Table 2. As can be seen from Table 2, the average classification accuracy of the five types of events after multifeature fusion can reach 71.856%.

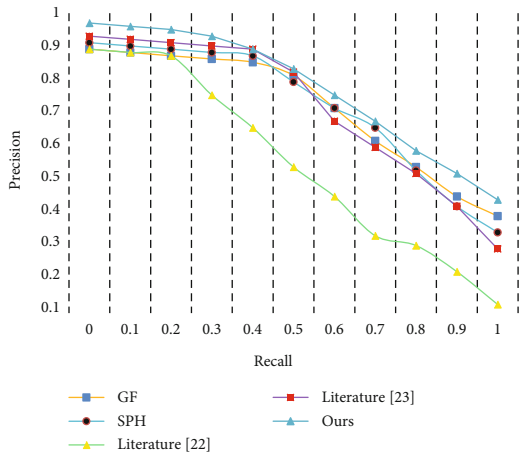



Figure 9: Precision and recall performance.

*4.4. Retrieval Performance of Different Algorithms.* In order to prove the performance of the proposed method, the proposed method is compared with the global feature-based method (GF) [29], spectral hash (SPH) [30], literature [22], and literature [23], respectively. In addition, in order to prove the mutual assistance between different features in the proposed method, this paper also uses a single level of features to carry out experiments. The experimental results are shown in Figure 9. Figure 9 shows the comparison of the PR curve of each method, and the PR curve is one of the most commonly used evaluation indexes. From the curve of Figure 9, we can also see that the method proposed in this paper is obviously better than other comparison methods.

## 5. Conclusion

Multiperson collaborative sports video analysis is not only one of the most important research directions in the field of computer vision but also a very challenging task. How to help viewers quickly identify interesting clips from a large number of videos has become particularly important. In order to solve this problem, this paper takes the video of basketball game as the research object to analyze the event. According to the characteristics of RGB image and depth image, a retrieval algorithm based on the fusion of RGB feature and depth feature is proposed. The algorithm makes full use of the respective advantages and complementary characteristics of the two kinds of video information and adopts a feature representation method which is suitable for the two kinds of video. Compared with the single feature representation method, the retrieval performance of multiple feature representation methods is up to 71.856%. Compared with the existing methods, the method proposed in this paper can make full use of the complementarity between different levels of features, make up for the defects of low-level artificially defined features in semantic expression, and finally achieve more efficient retrieval performance. Experiments in this paper also prove the effectiveness of the proposed method. This paper makes a preliminary exploration on multifeature fusion, and then we can consider integrating more features to further improve the retrieval efficiency and achieve the purpose of practical application.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the Anhui Provincial Practical Education Base Project: Talent Training Practice and Education Platform of Huainan Normal University Sports Industry (the project has no number).

## References

[1] I. A. Heikura, A. L. T. Uusitalo, T. Stellingwerff, D. Bergland, A. A. Mero, and L. M. Burke, "Low energy availability is difficult to assess but outcomes have large impact on bone injury rates in elite distance Athletes," *International Journal of Sport Nutrition and Exercise Metabolism*, vol. 28, no. 4, pp. 403–411, 2018.

[2] T. Josefsson, A. Ivarsson, H. Gustafsson et al., "Effects of mindfulness-acceptance-commitment (MAC) on sport-specific dispositional mindfulness, emotion regulation, and self-rated athletic performance in a multiple-sport population: an RCT Study," *Mindfulness*, vol. 10, no. 8, pp. 1518–1529, 2019.

[3] R. Ji, "Research on basketball shooting action based on image feature extraction and machine Learning," *IEEE Access*, vol. 8, pp. 138743–138751, 2020.

[4] L. Zhao and W. Chen, "Detection and recognition of human body posture in motion based on sensor technology," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 15, no. 5, pp. 766–770, 2020.

[5] J. Spitz, P. Moors, J. Wagemans, and W. F. Helsen, "The impact of video speed on the decision-making process of sports officials," *Cognitive Research: Principles and Implications*, vol. 3, no. 1, pp. 1–10, 2018.

[6] I. Errekagorri, J. Castellano, I. Echeazarra, and C. Lago-Peñas, "The effects of the video assistant referee system (VAR) on the playing time, technical-tactical and physical performance in elite soccer," *International Journal of Performance Analysis in Sport*, vol. 20, no. 5, pp. 808–817, 2020.

[7] N. A. Rahmad, M. A. As'ari, N. F. Ghazali, N. Shahar, and N. A. J. Sufri, "A survey of video based action recognition in Sports," *Science*, vol. 11, no. 3, pp. 987–993, 2018.

[8] H. C. Shih, "A survey of content-aware video analysis for Sports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1212–1231, 2018.

[9] M. Qi, Y. Wang, A. Li, and J. Luo, "Sports video captioning via attentive motion representation and group relationship modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2617–2633, 2020.

[10] W. Song, M. Xu, and Y. Dolma, "Design and implementation of beach sports big data analysis system based on computer Technology," *Journal of Coastal Research*, vol. 94, no. sp1, pp. 327–331, 2019.

[11] W. Yang, "Analysis of sports image detection technology based on machine learning," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, 2019.

[12] K. M. Shah and R. M. Makwana, "Shot boundary detection using logarithmic intensity histogram: an application for video retrieval," *International Journal of Applied Engineering Research*, vol. 12, no. 22, pp. 12616–12624, 2017.

[13] C. Huang and H. Wang, "A novel key-frames selection framework for comprehensive video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 577–589, 2020.

[14] P. Wang, H. Liu, L. Wang, and R. X. Gao, "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration," *CIRP Annals*, vol. 67, no. 1, pp. 17–20, 2018.

[15] J. Gui and Z. M. Lu, "A fast key frame extraction method for the surveillance video based on the moving targets," *The Journal of Intelligent Information Hiding and Multimedia Signal Processing*, vol. 10, no. 2, pp. 250–262, 2019.

[16] K. Kumar, D. D. Shrimankar, and N. Singh, "Eratosthenes sieve based key-frame extraction technique for event summarization in videos," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 7383–7404, 2018.

[17] X. Li, B. Zhao, X. Lu et al., "Key frame extraction in the summary space," *IEEE Transactions on Cybernetics*, vol. 48, no. 6, pp. 1923–1934, 2018.

[18] H. Tang, H. Liu, W. Xiao, and N. Sebe, "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion," *Neurocomputing*, vol. 331, pp. 424–433, 2019.

[19] H. Yan, M. Xie, P. Wang, Y. Zhang, and C. Luo, "Kernel-correlated filtering target tracking algorithm based on multi-features fusion," *IEEE Access*, vol. 7, pp. 96079–96084, 2019.

[20] Y. Sun, P. Li, Z. Jiang, and S. Hu, "Feature fusion and clustering for key frame extraction," *Mathematical Biosciences and Engineering*, vol. 18, no. 6, pp. 9294–9311, 2021.

[21] H. Tu, "A 3D model retrieval method based on multi-feature fusion," *International Journal of Information and Communication Technology*, vol. 15, no. 2, pp. 121–131, 2019.

[22] X. Qi, C. Liu, and S. Schuckers, "CNN based key frame extraction for face in video recognition," in *2018 IEEE 4th international conference on identity, security, and behavior analysis (ISBA)*, pp. 1–8, Singapore, 2018.

[23] Q. Zhang, J. Cao, Y. Zhang, S. Zhang, Q. Zhang, and D. Yu, "FPGA implementation of quantized convolutional neural networks," in *2019 IEEE 19th International Conference on Communication Technology (ICCT)*, pp. 1605–1610, Xi'an, China, 2019.

[24] Y. Li, R. Xia, Q. Huang, W. Xie, and X. Li, "Survey of spatio-temporal interest point detection algorithms in Video," *IEEE Access*, vol. 5, pp. 10323–10331, 2017.

[25] C. Wu, X. Ye, F. Ren, and Q. du, "Check-in behaviour and spatio-temporal vibrancy: An exploratory analysis in Shenzhen, China," *Cities*, vol. 77, pp. 104–116, 2018.

[26] A. Olaode and G. Naghdy, "Adaptive bag-of-visual word modelling using stacked-autoencoder and particle swarm optimisation for the unsupervised categorisation of images," *IET Image Processing*, vol. 14, no. 9, pp. 1769–1776, 2020.

[27] J. Yang, J. Huang, Z. Jiang et al., "3D SIFT aided path independent digital volume correlation and its GPU acceleration," *Optics and Lasers in Engineering*, vol. 136, article 106323, 2021.

[28] H. S. Dadi and G. K. Mohan Pillutla, "Improved face recognition rate using HOG features and SVM Classifier," *IOSR Journal of Electronics and Communication Engineering*, vol. 11, no. 4, pp. 34–44, 2016.

[29] L. Du, J. Tan, H. Yang et al., "Ssf-dan: separated semantic feature based domain adaptation network for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 982–991, Seoul, Korea (South), 2019.

[30] Q. Liu, G. Liu, L. Li, X. T. Yuan, M. Wang, and W. Liu, "Reversed spectral hashing," *Ieee Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2441–2449, 2018.