Wang L et al. (2025) RESEARCH ON ACTION DECOMPOSITION METHOD OF FOOTBALL VIDEO IMAGE BASED ON DEEP LEARNING. Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte vol. 25 (100) pp. 313-329.
DOI: https://doi.org/10.15366/rimcafd2025.100.020

ORIGINAL

RESEARCH ON ACTION DECOMPOSITION METHOD OF FOOTBALL VIDEO IMAGE BASED ON DEEP LEARNING

Haoyu Wang [1], Lanfeng Wang [2*], Litong Wang [3], Xinran Wang [4]

[1]Sports Science and Technology College, Guangzhou Institute of Applied Science and Technology, 526072, China
[2]Sports College, Guizhou City Vocational College, Guiyang, 550025, China
[3]Gdansk University of Physical Education and Sport, Gdansk, 80-336, Poland
[4]Shandong Culture and Tourism Investment Group Financial Leasing Co., Ltd, 250000, China
E-mail: laviniawang128@163.com

Recibido 22 de junio de 2024 **Received** June 22, 2024
Aceptado 22 de diciembre de **2024 Accepted** December 22, 2024

ABSTRACT

In order to better master football skills, a football video image action decomposition method based on deep learning is proposed. Combined with the principle of deep learning, the action characteristics of football video image are collected, the denoising algorithm of football video action image is optimized, and the action decomposition steps of football video image are simplified. Finally, the experiment proves that the football video image action decomposition method based on deep learning has high practicability in the process of practical application, fully meet the research requirements.

KEYWORDS: Deep Learning; Football Video Image; Action Decomposition.

1. INTRODUCTION

Soccer video image action decomposition based on deep learning is a basic problem of image action decomposition(Jain et al., 2020). This technology holds extensive potential for application and is poised to be utilized across various domains, including video monitoring, video processing, robotic systems, and advanced interactive human-computer interfaces. The primary objective of object tracking in video sequences is to capture the path and dynamic characteristics of moving entities, including their location and dimensions. The algorithm for tracking motion in football video footage primarily encompasses two fundamental components: perceptual modeling and kinematic modeling.

These two problems are studied respectively(Acevedo et al., 2021).

In the aspect of apparent modeling, the main methodology adopted in this paper is the joint modeling of foreground (i.e. tracked object) and background, and a soccer video image action decomposition method based on deep learning is proposed respectively. In the aspect of action decomposition of football video image, taking football tracking in football video as the background, an object tracking decomposition method based on graph model is proposed to improve the accuracy of action decomposition of football video image.

2. Action Decomposition Method of Football Video Image

2.1 Football Video Image Action Acquisition

The deep learning method selects the action features of football video image based on a general criterion to improve the classification performance of the selected feature subset. Football video image action decomposition based on deep learning separates the main interfaces in the video and splits the field programmable gate columns(Sharma et al., 2021). Through the construction of different functions, functions of different types of football video image actions can be constructed in the core source program, and their functions are described, as shown in Table 1.

Table 1: Core Coding Function

NUMBER	FUNCTION NAME	FUNCTION
1	Y300*macroblock*cache*load	Macroblock encoding acquisition
2	Y300*macroblock*Forward prediction	Forward prediction
3	Y300*macroblock *Backward prediction	Backward prediction
4	Y300*macroblock *write*cavlc	Entropy coding acquisition
5	Y300*macroblock *cache*save	Current coded information storage
6	Y300*reference*update	Reconstruction coding
7	Y300*encode	Coding layering

Depth is one of the most widely used methods in color tracking. However, color is very sensitive to the change of illumination(Jenner et al., 2021). Therefore, if the illumination changes greatly in the application environment, other features (such as edge, texture, etc.) should be used for tracking.

Of course, the combination of various features is also meaningful to improve the performance of the tracking algorithm(Kong et al., 2021). Therefore, the classification system of football video object tracking method based on deep learning is shown in Figure 1:

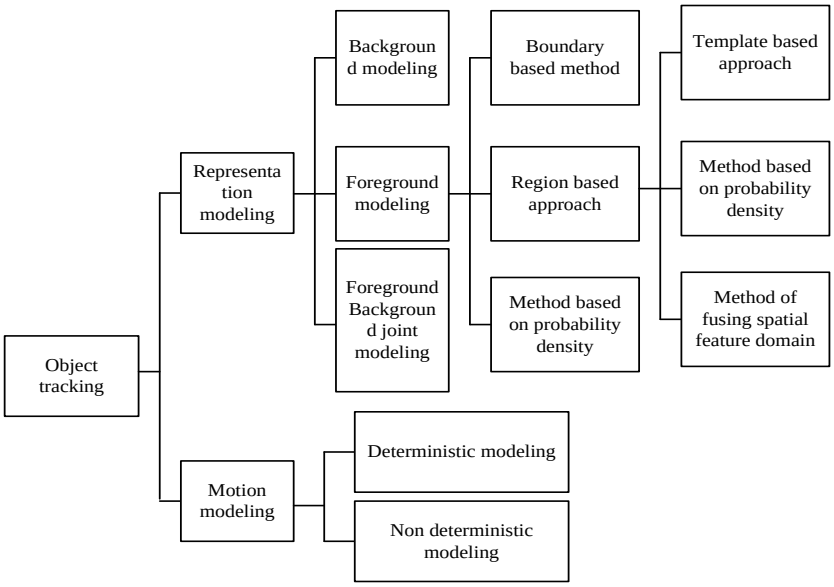

Figure 1: Classification of Soccer Video Object Tracking Method Based on Deep Learning

Football video image action acquisition adopts B / S mode based on SOA architecture, which is composed of browser, server and network transmission module. The structure is shown in the figure 2.

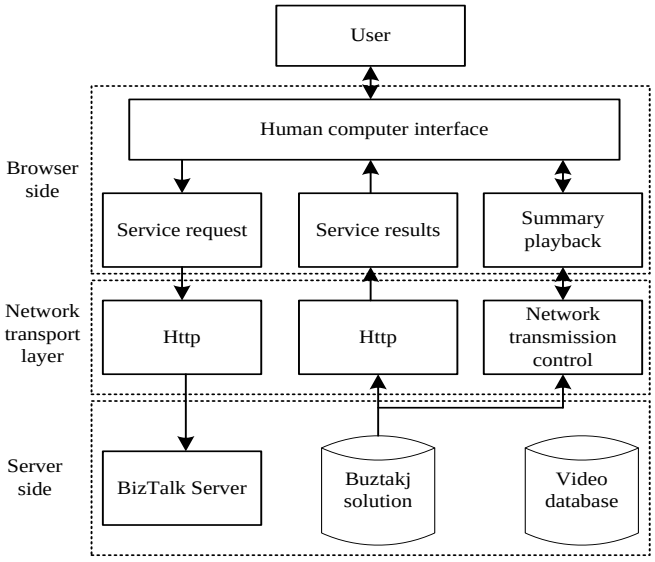

Figure 2: Structure of Football Video Image Action Acquisition

For the analysis of football match footage, it is essential to employ video capture devices to record the match's video stream into memory. Given the substantial size of video files, the consideration of storage capacity is crucial during the acquisition process(Kim et al., 2021). In cases where the in-game script data is incomplete due to the rapid tempo of play, an auxiliary mechanism is provided for post-game scene reconstruction, facilitating the real-time integration of script narratives with video content(H. Li et al., 2020). Figure 3

illustrates the organizational structure for the collection of video along with technical and tactical data.

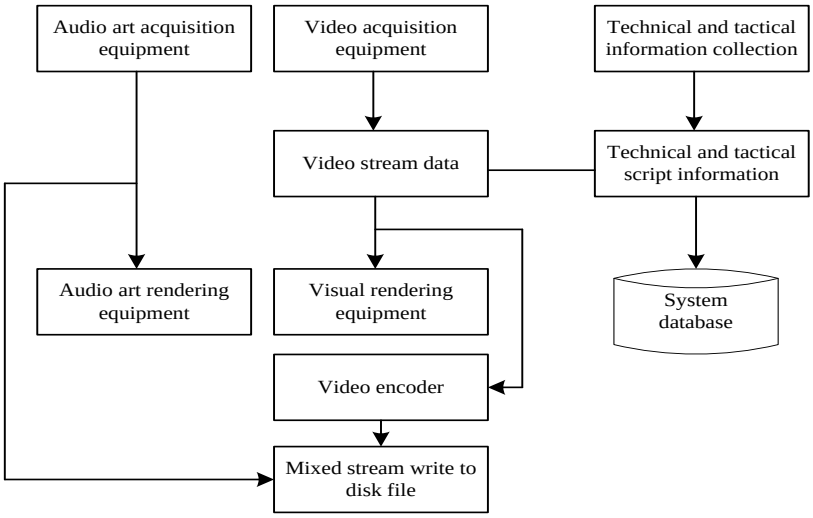

Figure 3: Structure Diagram of Video Action Data Acquisition

In order to facilitate the coaches to carry out visual analysis and statistics on the techniques and tactics of the players in the process of competition after the game(H. Li et al., 2020).

2.2 Denoising of Football Video Action Image

According to the discussion in the key technology research, for the football game video, that is to collect the players' technical and tactical information. Automatic completion and intelligent prompt depend on the support of basic skill base and team member skill base(Zebhi et al., 2020). The module structure is shown in Figure 4:

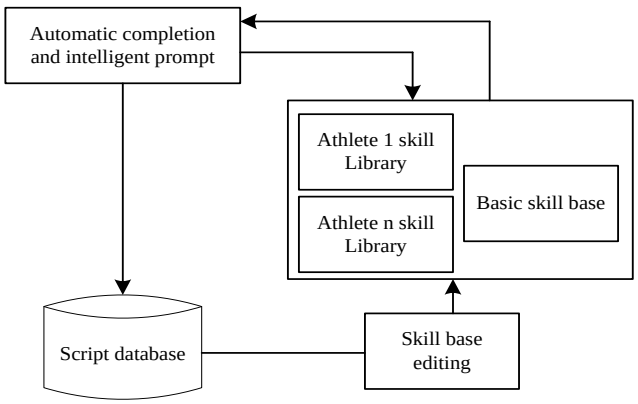

Figure 4: Automatic Completion Model of Video Action Image Data

As shown in the figure, during the collection of competition technical and tactical information, the automatic supplement and intelligent prompt module gives input prompt and supplement according to the athlete skill information

base and basic skill information base, and the collected script information is stored in the database(Feng, 2020). In order to sort the features, a criterion is needed to measure the distinguishing power of the features to the foreground and background. For a good criterion, the features with strong distinguishing power should be in the front, while the features with weak distinguishing power should be in the back(Hua et al., 2021). In the field of image, extracting the features of pictures for feature matching can get the association between different pictures. Euclidean distance is the distance between the points of two eigenvectors in Euclidean space. The calculation formula is:

$$d_{xy} = K - \sqrt{\sum_{k=1}^{K}(x_k - y_k)^2} \quad (1)$$

Where x_k and y_k represent eigenvectors and K represents the dimension of eigenvectors. Euclidean distance intuitively represents the distance between two features. The greater the distance, the smaller the degree of feature correlation. Euclidean distance focuses on the absolute value of each dimension of feature vector. If there are no other constraints on the feature itself, the value of Euclidean distance often fluctuates in a large range. In order to achieve this goal(Shamsolmoali et al., 2020). Bayesian error rate is a good choice. Theoretically, the features with small Bayesian error rate have strong discrimination, while the features with large Bayesian error rate have weak discrimination. For the pattern classification problem with two classes a and M2, the Bayesian error rate is defined as

$$P(\text{ error }) = \int_{R_2} p\left(x \mid \omega_1\right) P(\omega_1) dx + \int_{R_1} p\left(x \mid \omega_2\right) P(\omega_2) dx \quad (2)$$

In the formula, for I = 1,2, the a priori probability and likelihood functions dxrepresented by $P(\omega_1)$ and $p(x \mid \omega_1)$respectively correspond to the regions of $p(x \mid \omega_1)P(\omega_1)$ > $p(x \mid \omega_2)P(\omega_2)$ and $p(x \mid \omega_2)P(\omega_2)$ > K , then the formula can be further written

$$P(\text{ error }) = K \int^{\int \{p(x|\omega_1)P(\omega_1), p(x|\omega_2)P(\omega_2)\}} min \quad (3)$$

By using histogram to approximate the likelihood function, the discrete value of Bayesian error rate of feature f ∈ f in t frame is：

$$e_f^t = \sum_{i=1}^{m\sum\left(\boldsymbol{p}_f^t(i)P(O), \boldsymbol{q}_f^t(i)P(B)\right)} min \quad (4)$$

There are two common methods to calculate the similarity between ks-ok and pcn-ok, and there will be one method to calculate the similarity between the two joints in this paper. The present study computes the OKs scores to assess the accuracy of athletes' postures, with a score above 0.5 indicating

accurate detection. Detection of athletes and estimation of their postures fall under the category of detection tasks(Wang et al., 2020). In line with standard detection task evaluations, this research adopts three prevalent metrics: the maximum recall rate, Average Precision (AP) score, and the Precision-Recall (PR) curve. In the context of detection outcomes, four possible outcomes exist in relation to predicted and actual values: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). TP signifies the correct identification of a positive sample, FP the incorrect identification of a negative sample as positive, FN the incorrect identification of a positive sample as negative, and TN the correct identification of a negative sample. In the PR curve, 'P' denotes precision and 'R' denotes recall. The formula for calculation is presented below

$$P = \frac{TP}{TP+FP} \tag{5}$$

Suppose the matrix formula of the measured athlete characteristic points is m_{xy}, The moving image feature point matrix to be predicted is m_{xy}. If the feature points of each moving image $\begin{bmatrix} t_{xy} \\ e_{xy} \end{bmatrix}, i = 1,\ldots J, c = 1,\ldots,Q$ is the measurement matrix is m_{xy}, where J represents the number of moving image frames,Q is the ultimate goal of the matrix is to obtain the three-dimensional structure of each frame of moving image. If the three-dimensional shape of an athlete's joint is a weighted linear set of shape bases, there are:

$$F = \sum_{i=1}^{N} \delta_{xy}\lambda_x i \tag{6}$$

Whereδ_{xy} represents the weighting coefficient,λ_x represents the shape base,N represents the number of shape bases. If $N = 1$,δ_{xy} Then the state of the key area under the weak perspective projection model is:

$$m_{xy} = \begin{bmatrix} t_{x1},\ldots,t_{xQ} \\ e_{x1},\ldots,e_{xQ} \end{bmatrix} = \overline{w}\left(\sum_{i=1}^{N} \delta_{xy}\lambda_x i\right) + \overline{D}_x k_v^D \tag{7}$$

Where $\overline{w}$ represents the first two rows of the rotation matrix, $\overline{D}_x$ represents the first two elements of the translation vector. In order to optimize the operation process of the algorithm, the coordinates of each moving feature image should be changed to ensure that the origin of the moving image coordinate system is in the center of the image point, and then the translation vector can be filtered(Jiang & Tsai, 2021). The process is shown in the formula:

$$\begin{cases} \overline{t}_{xy} = t_{xy} - \frac{1}{Q}\sum_{y=1}^{Q} t_{xy} \\ \overline{e}_{xy} = e_{xy} - \frac{1}{Q}\sum_{y=1}^{Q} e_{xy} \end{cases} \tag{8}$$

By further processing the formula, we can get:

$$e_{xy} = \begin{bmatrix} \overline{t}_{x1}, \ldots, \overline{t}_{xQ} \\ \overline{e}_{x1}, \ldots, \overline{e}_{xQ} \end{bmatrix} = \overline{w}\left(\sum_{i=1}^{N} \delta_{xy} \lambda_x i\right) \quad (9)$$

Ensure m_{xy} and m_{xy} are the motion parameter matrix and 3D structure parameter matrix can be obtained by minimizing the projection error. The process is shown in the formula:

$$min\left\|m_{xy} - m_{xy}^{'}\right\|^2 == min\sum_{x,y}\left\|m_{xy} - \left(w\sum_{i=1}^{N}\delta_{xy}\lambda_x i\right)\right\|^2 \quad (10)$$

Within this approach, the motion parameter matrix utilizes quaternion notation, structured as a 4 × G matrix, with G denoting the total count of frames in the moving image sequence. To enhance the algorithm's operational efficiency, the positional coordinates of each moving feature must be adjusted so that the moving image's coordinate system origin aligns with the centroid of the image. Subsequently, the translation vector is applied for filtering purposes. This approach leverages quaternions to represent the motion parameter matrix in a 4 × G format, where G signifies the frame count of the complete moving image sequence(Gao et al., 2020). For the football game video, relying on the football game technical and tactical description and information collection model, the process code designed and implemented in the football technical and tactical description language is used as the main input code type, and an implementation interface is reserved for the design and implementation of more coding types, such as compression code and tactical code implemented in the model(Podoprosvetov et al., 2021). There are many kinds of football technical actions. According to the statistics of athletes of various technical types, if only the three actions of service, receiving and scoring are counted, the number of action combinations that athletes may adopt in one point will reach as many as 27. The optimization framework of detection results based on time domain video frames is shown in the figure 5.

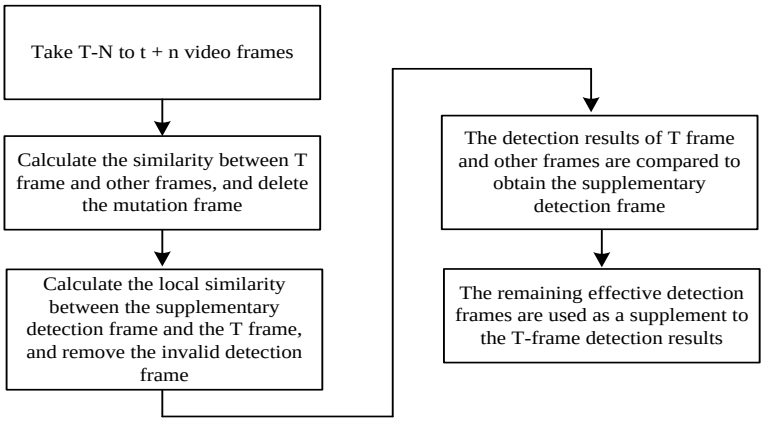

Figure 5: Detection Result Optimization Framework of Video Frame

The scheme optimizes the detection results by taking the detection results of adjacent n frames before and after t frames. Firstly, the mutation frames in adjacent frames are removed by calculating the similarity between frames, and then the detection results of different frames are compared(Sun et al., 2021). If there is a detection frame not in t frames in adjacent frames and the confidence is greater than a certain threshold, it is used as the supplementary detection frame of T frames, and then the effectiveness of the supplementary detection frame of T frames is judged by calculating the local similarity, Complete the optimization of T-frame detection results. Football techniques and tactics refer to the offensive or defensive strategies adopted by athletes in order to win the game. It consists of a series of technical actions(Climent-Perez & Florez-Revuelta, 2021). The data of the skill base is to select the most commonly used or the longest used technical and tactical actions of a player from the combination of football technical actions and techniques and tactics. The scalability of the skill base is an aspect to be considered. Here, XML document is used as the data source of the skill base(Atto et al., 2020).

2.3 Implementation of Action Decomposition in Football Video

Different from the traditional text and numerical data, video is a continuous media, which does not have structural characteristics. To manage the video data, we should first divide it into basic retrievable units, and then establish the video structure model to make it a kind of structured data, so that it can be browsed or retrieved in the way of database. For a content-based retrieval, the data model is the core, which determines the supported query type and retrieval ability(McCord et al., 2020). The structured model method divides the video data into video sequence, scene, shot and frame from top to bottom, as shown in Figure 6.

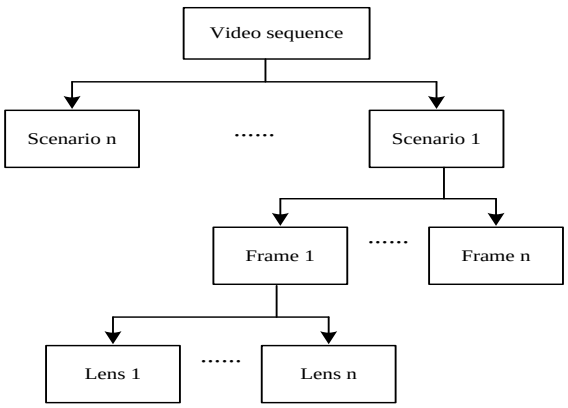

Figure 6: Frame Video Data Classification Structure

Frame is the smallest unit of video data and a still picture; Lens is the basic unit of video data. It represents the continuous action of a scene in space and time(Abdellaoui & Douik, 2020). It is determined by the beginning and end

of a camera shooting. It adopts two functions and two global arrays. The name and function are shown in Table 2. In the function, it is necessary to call the function according to the brightness and chroma signals in all HD video coding contained in the current prediction module(Kumaravel & Veni, 2019). All the obtained residual data are stored in the global data group as the input of DCT transformation.

Table 2: Function Description of Function and Global Array Name

NAME	FUNCTION DESCRIPTION
Y300*MACROBLOCK *PREDICT	Macroblock prediction
PIXEL*DIFF*WXH	Pixel resolution difference recognition
INT15*TA15*BRIGHTNESS [15*15]	Store brightness signal data
INT15*TA15*CHROMA [8*16]	Store chrominance signal data

The summary of SOA architecture is realized at the task level (i.e. completing the action acquisition of a football video image) by arranging the services(Lei et al., 2020). Therefore, the first step of video image action acquisition based on SOA architecture is to decompose the service. Service is an important feature of service-oriented architecture. It does not depend on specific implementation details. Multiple different services can be organized to form a complete solution(Ahmad et al., 2021). Therefore, how to analyze the generation process of video image action acquisition and decompose the services that can be shared by different summary generation algorithms has become a difficulty.

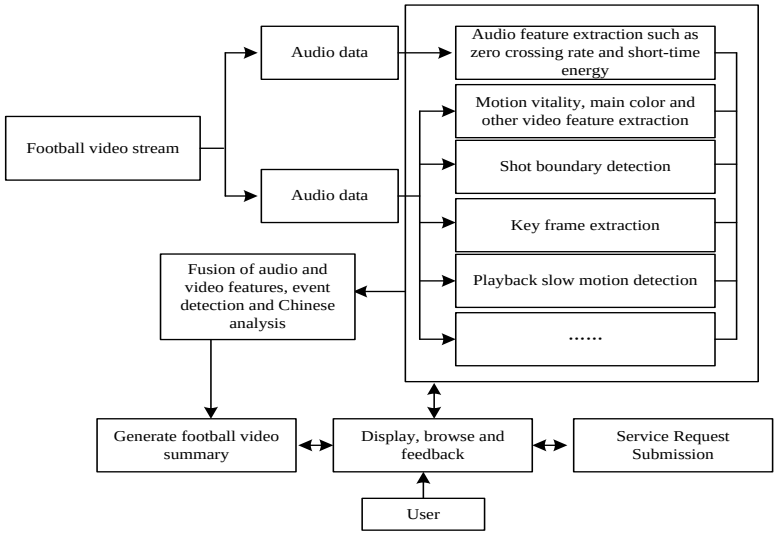

Figure 7: Key Action Image Processing Flow of Football Video

Events in a video sequence refer to a certain behavior or a series of behaviors with certain characteristic information in the sequence. Polana et al. Divided events into three categories: time structure, which has uncertain space and time range; Second, it has periodic behavior in time and limited behavior in space(Q. Li et al., 2020). The third is movement events, which are relatively

isolated actions without repetition in time and space. According to different analysis objects, video sequence processing can be divided into analysis methods based on target extraction and analysis methods without target extraction. Therefore, the flow of video sequence processing does not necessarily include the process of target detection and tracking. The specific processing flow is shown in Figure 8.

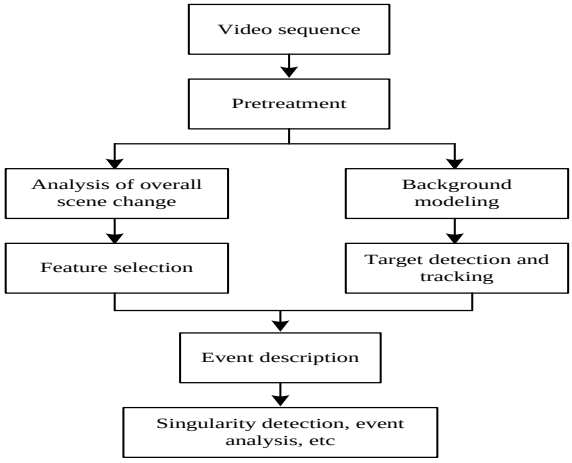

Figure 8: Football Image Video Event Analysis Process

Target detection and tracking are two different and complementary research processes. Target detection pays more attention to the information of the target itself, such as the accurate contour of the target, target attribution, etc.; while target tracking mainly focuses on the motion characteristics of the target, so as to analyze the motion changes of the target(Anvarov et al., 2020). In the process of video sequence processing, target detection can provide initialization information for target tracking, and target tracking provides space-time basic information for target detection: in turn, the result of target detection can be fed back to the process of target tracking, and the constraint conditions of target tracking can be updated to make target tracking more effective and accurate(Lejonagoitia-Garmendia et al., 2023).

3. Analysis of Experimental Results

In order to realize the detection and pose estimation of athletes in sports video, we need to process the video and run the image processing algorithm based on deep learning. Therefore, a large number of intensive operations will be generated, which has certain requirements for the software and hardware environment of the equipment. In this paper, the development environment of the is as follows) hardware environment: the CPU model is Intel Core i7-8750, and the GPU model is gtx1060 (6g) 2) software environment: the operation is windows10, and the development IDE is pychart. The main development language is python, the video processing library is opencv, the graphical interface development tool is Tkinter, and the deep learning framework is py

torch. The experimental environment used consists of three computers: a main server, a web server and a client. Its configuration is shown in the table 3.

Table 3: Experimental Environment Configuration

THE SERVER	OPERATING SYSTEM	PROCESSOR	DOMINANT FREQUENCY	MEMORY
MASTER SERVER	Windows XP	Inter Xeon 8	4G	64G
DEEP LEARNING SERVER	Windows 10	AMD2800+	2G	16G
CLIENT	Windows 10	Celeron 2.6G	2.6G	32GB

The BizTalk Server is installed in the main server and is responsible for organizing and invoking the deep learning service. At the same time, the main server contains a video database, which is responsible for storing the original video stream and processing results. The in-depth learning service of extracting shot transformation rate and shot intensity is also provided by the main server. The deep learning server provides deep learning services for shot segmentation and extracting short-term audio energy. The client is responsible for submitting service requests from the browser. Functional testing is a series of functional testing to verify whether it can meet the design purpose and user needs. A complete test requires perfect test cases. This section prepares test cases for functional test and records the test results. As shown in the table, the table records the test contents, test steps, expected results and test results of the athlete detection and posture estimation function test for sports video.

Table 4: The Function Test Cases

TEST CONTENT	TESTING PROCEDURE	EXPECTED RESULTS	TEST RESULT
VIDEO SELECTION	Click file selection to load the file to be processed	Load and display video to process	Verification passed
DATA PROCESSING	Click the data processing button to process the data	A series of data processing such as framing and detection of video	Verification passed
PARAMETER SETTING	Enter the selection scheme in the text input box	If the parameters are normal, set them	Verification passed
LOADING RESULTS	After entering the scheme and threshold, click the result button	Button to draw and display the corresponding results of the parameters entered by the user	Verification passed
PLAY VIDEO	After loading the video or loading, click the play button	Play the corresponding video	Verification passed

Experiments were conducted using the football video dataset to identify the optimal set of parameters. To enhance the grid search efficiency for

parameters, the search space can be pruned by eliminating parameters that are clearly impractical, thereby reducing computational expenditure. Based on empirical data, the search parameters are defined as follows: the interval for adjacent frames (n) ranges from 0 to 10 with an increment of 1, the threshold for inter-frame similarity (s) spans from 0 to 10 with an increment of 1, the local similarity threshold (D) varies from 0 to 10 with an increment of 1, and the confidence threshold (T) for the secondary detection frame is set between 0.1 and an increment of 0.1. Upon iterative parameter optimization, the football video dataset developed in this study yields optimal outcomes with the super parameter settings as super parameter = 3S = 5, D-2, r0.3. The sequence counts and corresponding statistical outcomes are detailed in Table 5.

Table 5: Statistical Results of Encoder Sequence

RESOLVING POWER	SEQUENCE NAME	FRAME RATE（FPS）
442×133	Foreman*cif.yuv	0.70
	Silent*cif.yuv	0.73
	Container*cif*yuv	0.77
520×476	Foreman*520*476.yuv	0.10
	Container*520*476.yuv	0.09
910×768	Bigships*910*768.yuv	0.13
	Jet*910*768.yuv	0.09

Subsequently, an experimental investigation will be conducted to assess the influence of diverse parameters. By integrating the detection outcomes across several frames, it is possible to enhance the recognition recall rate; however, this approach could also result in erroneous detections. The parameter defining the interval between adjacent frames, denoted as n, significantly influences both the recall rate and the precision of the integrated outcomes. When the detection confidence threshold is set to the lowest value, the recall and accuracy of the detection results vary from 0 to 10 based on the value of n, as illustrated in Figure 9.

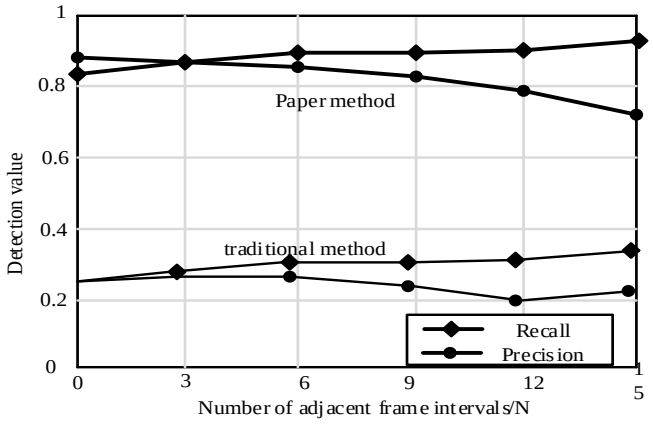

Figure 9: Influence Curve of Image Decomposition Interval N of Adjacent Video Frames

It can be seen from the figure that with the increase of adjacent frame interval n, the detection recall rate gradually increases, the rising speed is relatively stable, the detection accuracy gradually decreases, the falling speed is slow first and then fast, and the intersection of the two curves is near N= 3. The local similarity threshold D can judge whether there is a local change in the supplementary detection result. When there is a local change, the supplementary detection result cannot be added. By observing the sequence of a large number of slow shots, it is found that it has the following four modes. The zero value is removed to form a new sequence, as shown in the figure 10.

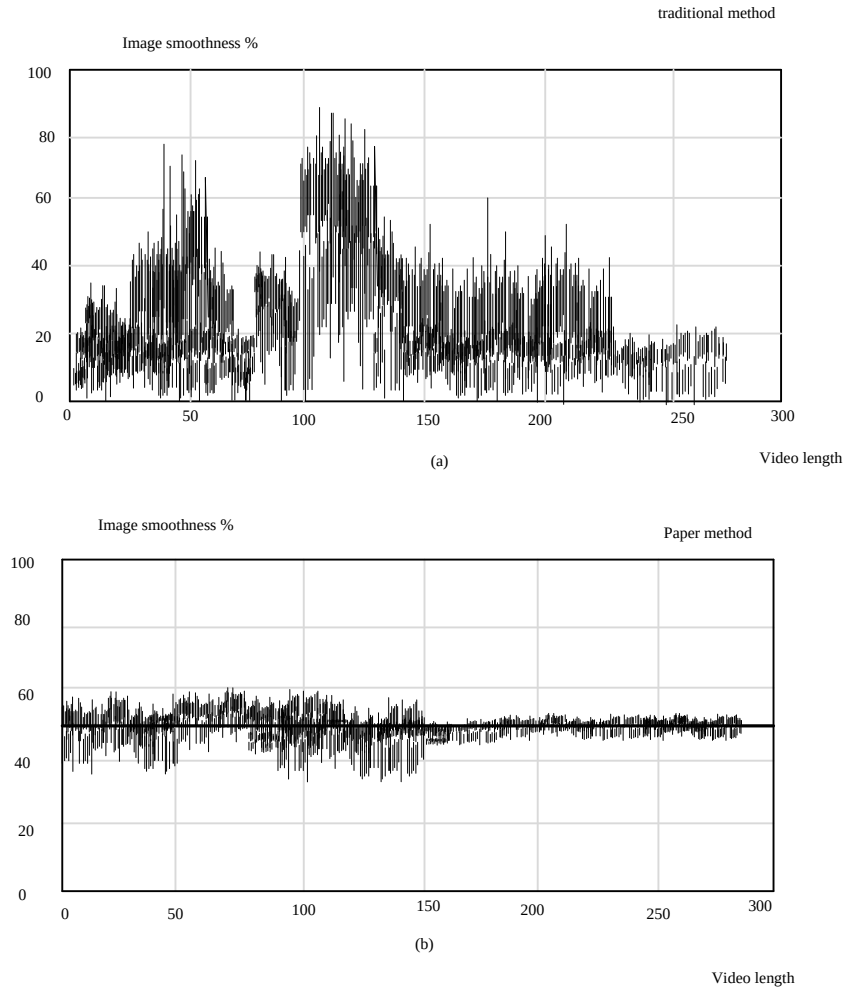

Figure 10: Video Image Motion Extraction Smoothness Contrast Detection Results

It can be seen from the figure that with the increase of local Hamming distance similarity threshold D, the detection recall rate increases and the detection accuracy decreases slightly. The itemized comparison diagram of the precision and recall of the above data is given. Based on the above comparative detection results, it is not difficult to find that compared with the traditional methods, there are significantly more decomposable images of football video in this method, and the image quality is relatively higher. The second frame

difference method has the highest precision, but its recall is less than 26%. The zero-intersection method is similar to this method in some visual aspects, but its recall is less than 44 and the precision is less than 36%. Especially in the performance of a video, the sub frame difference method and zero intersection method detect that the correct shot is zero, and the fluctuation range of the two methods is large. Generally speaking, it has good application value.

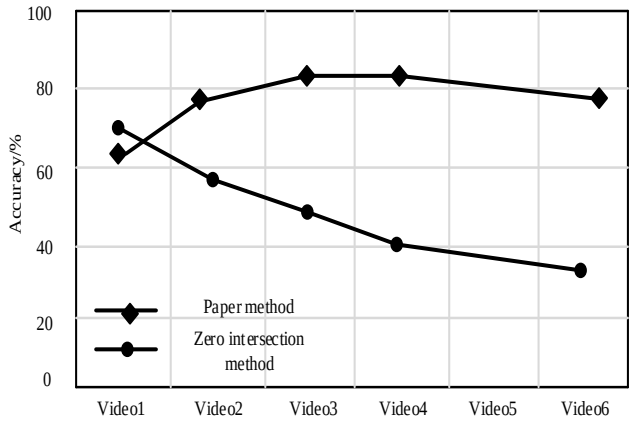

Figure 11: Accuracy Detection Results of Video Action Decomposition Processing

After introducing the local similarity threshold D, the accuracy is improved when the recall rate is basically unchanged. This proves that the algorithm proposed in this paper has high precision and high application value. The football image action decomposition function using deep school technology is built at the workflow level. Its powerful function of XII ml message processing and workflow management makes it have a good loose structure, and its maintainability, compatibility and scalability have been strengthened. On the basis of the above experiments, in order to further clarify the effect of action decomposition of this method. The overall effect of image decomposition of this method and traditional methods is compared and displayed. The specific results are shown in the following figure12,13:

Figure 12: Video Image Decomposition Results of this Method

Figure 13: Video Image Decomposition Result of Traditional Method

FIGS. 12 and 13 show that the video action decomposition picture of this method is clear and four actions can be decomposed. The traditional method of comparison can only achieve two decomposed actions, and the decomposed picture is relatively blurred.

4. Conclusions

Taking football video as the research object, based on the analysis of the existing football video image action acquisition and generation technology, a hierarchical model is established, and a B / S mode football video image action acquisition and generation is designed and implemented in the way of service-oriented architecture. In view of the shortcomings of traditional construction methods in maintainability, compatibility and scalability, various technologies for generating video image action acquisition are published as deep learning services, which are organized in the form of loose coupling to provide users with multi-layer video services. Through multi-level definition of summary results, it lays a foundation for a breakthrough in adaptive transmission in the future.

REFERENCES

Abdellaoui, M., & Douik, A. (2020). Human Action Recognition in Video Sequences Using Deep Belief Networks. *Traitement du Signal*, *37*(1).

Acevedo, J., Boden, A. L., Greif, D. N., Emerson, C. P., Ruiz, J. T., Jose, J., Feigenbaum, L. A., & Kaplan, L. D. (2021). Distal medial collateral ligament grade III injuries in collegiate football players: operative management, rehabilitation, and return to play. *Journal of athletic training*, *56*(6), 565-571.

Ahmad, T., Jin, L., Lin, L., & Tang, G. (2021). Skeleton-based action recognition using sparse spatio-temporal GCN with edge effective resistance. *Neurocomputing*, *423*, 389-398.

Anvarov, F., Kim, D. H., & Song, B. C. (2020). Action recognition using deep 3D

CNNs with sequential feature aggregation and attention. *Electronics*, *9*(1), 147.

Atto, A. M., Benoit, A., & Lambert, P. (2020). Timed-image based deep learning for action recognition in video sequences. *Pattern Recognition*, *104*, 107353.

Climent-Perez, P., & Florez-Revuelta, F. (2021). Improved action recognition with separable spatio-temporal attention using alternative skeletal and video pre-processing. *Sensors*, *21*(3), 1005.

Feng, Y. (2020). Mobile terminal video image fuzzy feature extraction simulation based on SURF virtual reality technology. *Ieee Access*, *8*, 156740-156751.

Gao, L., Li, T., Song, J., Zhao, Z., & Shen, H. T. (2020). Play and rewind: Context-aware video temporal action proposals. *Pattern Recognition*, *107*, 107477.

Hua, M., Gao, M., & Zhong, Z. (2021). Scn: dilated silhouette convolutional network for video action recognition. *Computer aided geometric design*, *85*, 101965.

Jain, D. K., Zhang, Z., & Huang, K. (2020). Multi angle optimal pattern-based deep learning for automatic facial expression recognition. *Pattern Recognition Letters*, *139*, 157-165.

Jenner, S., Belski, R., Devlin, B., Coutts, A., Kempton, T., & Forsyth, A. (2021). A qualitative investigation of factors influencing the dietary intakes of professional Australian football players. *International Journal of Environmental Research and Public Health*, *18*(8), 4205.

Jiang, H., & Tsai, S.-B. (2021). An empirical study on sports combination training action recognition based on SMO algorithm optimization model and artificial intelligence. *Mathematical Problems in Engineering*, *2021*(1), 7217383.

Kim, D. H., Anvarov, F., Lee, J. M., & Song, B. C. (2021). Metric-based attention feature learning for video action recognition. *Ieee Access*, *9*, 39218-39228.

Kong, Y., Wang, Y., & Li, A. (2021). Spatiotemporal saliency representation learning for video action recognition. *IEEE Transactions on Multimedia*, *24*, 1515-1528.

Kumaravel, S., & Veni, S. (2019). A unified model of video-based human action categorization using Chaotic Quantum Swarm Intelligence on Intuitionistic fuzzy 3D Convolution Neural Network. *Intelligent Decision Technologies*, *13*(4), 507-521.

Lei, Q., Zhang, H.-B., Du, J.-X., Hsiao, T.-C., & Chen, C.-C. (2020). Learning effective skeletal representations on RGB video for fine-grained human action quality assessment. *Electronics*, *9*(4), 568.

Lejonagoitia-Garmendia, M., Gustran-Iglesias, I., Gil, S., Ortuondo, J., & Sarasola-Ruiz, L. (2023). Foot injuries in sport climbers: footwear and other associated factors. *Revista multidisciplinar de las Ciencias del*

Deporte, *23*(90).

Li, H., Cryer, S., Acharya, L., & Raymond, J. (2020). Video and image classification using atomisation spray image patterns and deep learning. *Biosystems Engineering*, *200*, 13-22.

Li, Q., Yang, W., Chen, X., Yuan, T., & Wang, Y. (2020). Temporal segment connection network for action recognition. *Ieee Access*, *8*, 179118-179127.

McCord, A., Cocks, B., Barreiros, A. R., & Bizo, L. A. (2020). Short video game play improves executive function in the oldest old living in residential care. *Computers in Human Behavior*, *108*, 106337.

Podoprosvetov, A., Alisejchik, A., & Orlov, I. (2021). Comparison of action recognition from video and IMUs. *Procedia Computer Science*, *186*, 242-249.

Shamsolmoali, P., Celebi, M. E., & Wang, R. (2020). Advances in deep learning for real-time image and video reconstruction and processing. *Journal of Real-Time Image Processing*, *17*, 1883-1884.

Sharma, V., Gupta, M., Kumar, A., & Mishra, D. (2021). EduNet: a new video dataset for understanding human activity in the classroom environment. *Sensors*, *21*(17), 5699.

Sun, C., Song, H., Wu, X., Jia, Y., & Luo, J. (2021). Exploiting informative video segments for temporal action localization. *IEEE Transactions on Multimedia*, *24*, 274-287.

Wang, T., Chen, Y., Lin, Z., Zhu, A., Li, Y., Snoussi, H., & Wang, H. (2020). Recapnet: Action proposal generation mimicking human cognitive process. *IEEE transactions on cybernetics*, *51*(12), 6017-6028.

Zebhi, S., Al-Modarresi, S. M. T., & Abootalebi, V. (2020). Converting video classification problem to image classification with global descriptors and pre-trained network. *IET Computer Vision*, *14*(8), 614-624.