ISSN 2616-5775 Vol. 6, Issue 13: 67-75, DOI: 10.25236/AJCIS.2023.061311

Automatic Description Method of Floor Exercise Video Based on Three-dimensional Convolutional Network and Multi-label Classification

Feng He[1,a,*], **Wenda Gao**[2,b], **Yue You**[3,c]

[1]College of Continuing Education, Civil Aviation Flight University of China, Guanghan, China
[2]Department of Atmospheric Sciences, Civil Aviation Flight University of China, Guanghan, China
[3]Department of Atmospheric Sciences, Civil Aviation Flight University of China, Guanghan, China
[a]CAFUC_HF@outlook.com, [b]CAFUC_GWD@outlook.com, [c]Y13361135839@outlook.com
[*]Corresponding author

Abstract: *As people's attention to health and sports increases, the amount of data and audience of sports video is also growing rapidly. Based on this, the automatic description method of floor exercise video has attracted the attention of scientific researchers and industry. The research focus of this paper is on the automatic description of floors video, that is, to generate professional nouns by observing the movements of athletes in the video. This research has a wide range of application value, involving sports analysis, automatic interpretation and sports guidance. This paper specifically studies the automatic understanding of human movements in floor exercise videos, and combines the knowledge of computer vision and deep learning to realize the intelligent labeling and representation of specific human movements in video sequences.An automatic description method of floor exercise video based on three-dimensional convolutional network and multi-label classification is proposed. The floor exercise action is composed of multiple decomposition actions. In the work of this paper, a classifier of single decomposition action is constructed, and the automatic description problem of floor exercise action is transformed into a multi-label classification problem. Since the two-dimensional convolutional neural network loses time information when extracting features, this paper uses a three-dimensional convolutional network to extract spatio-temporal features of the video. Through multiple binary classifications, the goal of multi-label classification is achieved. In order to verify the effectiveness of the method, the classification results are randomly combined into a sentence, which is compared with the results of the automatic description method.*

Keywords: *Floor exercise; Three-dimensional convolutional network; Multi-label classification; Automatic description*

1. Background and problems

In recent years, the research on automatic description of sports video content has gradually become a hot spot[1]. With the rapid growth of sports video data and audience, as well as its potential application value, scientific researchers and industry have paid extensive attention to it[2]. However, in the study of sports video, there are relatively few studies in other fields except for football, badminton and other ball games. This paper chooses floors video as the research object, because floors plays a fundamental role in the development of other sports, which is difficult and representative[3]. The project has far-reaching practical significance in human-computer interaction, sports training assistance and other sports research and promotion.

The research goal of this paper is to realize the automatic description of the floors video, that is, to generate the professional name of the action by observing the complete set of actions of the athletes in the video.Traditional methods rely on manual interpretation, but this requires a high degree of professional knowledge of the narrator. The non-professional audience 's understanding of the game depends on the commentator 's commentary, but if there is an error in the commentary, it will affect the effect and perception of the game[4]. Therefore, it is necessary to use pattern recognition technology combined with natural language processing method to realize the automatic description of floors video.

This paper adopts the method of deep learning, based on C3D network architecture, combined with computer vision and deep learning knowledge[5], using deep convolutional neural network and support vector machine classifier (SVM) to realize the intelligent recognition and action description of specific human motion in floors video[6].Specifically, deep convolutional neural networks are used to mine deeper features in the temporal and spatial dimensions of the video. At the same time, a multi-label classification method based on support vector machine classifier is adopted to ensure the accuracy of classification.In addition, in order to solve the problem of unbalanced data, the multi-label classification problem is transformed into multiple two-class classifiers, which effectively eliminates the impact of data imbalance. Overall, this method has high accuracy and accuracy in the automatic description of floor exercise video.

2. Method of this article

2.1 The composition of network framework

The network framework used in this paper is shown in Figure 1 below. It mainly includes two parts : extracting multi-label video features of floor exercise based on three-dimensional convolution network, and using SVM for multi-label classification to transform the classification results into the real description of floor exercise[7].

In the research process, the problem of missing experimental data was first solved. This paper constructs a self-built floors decomposition action data set. The data set is based on floor exercises, and each action may contain multiple decomposition actions. The description of each video is transformed into a label, and the action name is divided according to the floors decomposition action name in the professional competition rules.Therefore, this paper uses three-dimensional convolutional network for feature learning, and inputs the extracted video features into the SVM classifier for classification. However, SVM is usually used for binary classification problems, so how to apply it to multi-label classification has always been the focus of research.In this paper, the method of constructing multiple binary classifiers is used to realize the multi-label classification function of SVM[8]. After completing the multi-label classification, it is necessary to find the mapping between the classification results and the natural language description of the floor exercise, and compare it with the automatic description results to complete the whole automatic description process. Through this process, the accuracy of multi-label classification can be evaluated and the true description of floor exercise can be obtained.

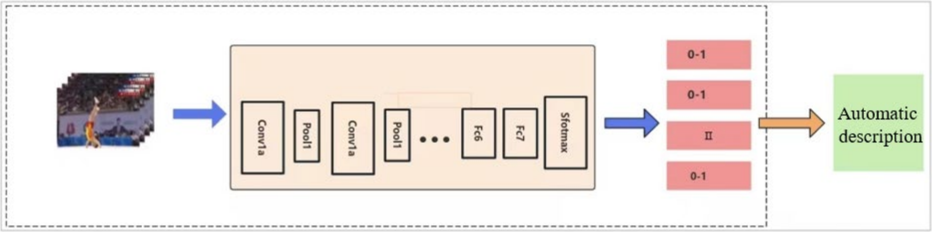

Figure 1: Network framework

2.2 Feature extraction

Compared with two-dimensional convolutional networks, three-dimensional convolutional networks can better model time information through three-dimensional convolution and three-dimensional pooling operations. But in the two-dimensional convolutional network, the process of convolution and pooling is completed in space. In a three-dimensional convolutional network, they are executed in time and space. In the previous section discussing three-dimensional convolutional networks, it was mentioned that when a two-dimensional convolutional network processes images, it produces an output image. Similarly, when operating on multiple images (considered as distinct channels), it also generates an output image. Therefore, the two-dimensional convolution network will lose the time information of the input data after each convolution operation. Only three-dimensional convolution can retain the temporal information of the input signal, thereby generating an output quantity. The same reason can be used to explain two-dimensional pooling and three-dimensional pooling. The experiment in this paper uses C3D as the feature extraction network[9].Figure 2 is the C3D network architecture.The three-dimensional convolution filters are all 3 × 3 × 3, and the step size is 1 × 1 × 1.In order to maintain the early time information, the pool1 kernel size is set to 1 × 2 × 2 and the step size is 1 × 2 × 2, and all the remaining 3D pooling layers are 2 × 2 × 2 and the step size is 2 × 2 × 2. Each fully connected layer has 4096 output units. The input of the model is to segment the video into 16 frame-length segments, with 8 frame overlaps between two consecutive segments. Each fragment is passed to the C3D network to extract fc6 activation. The activation of these fragments fc6 is averaged to form a 4096-dimensional video descriptor.

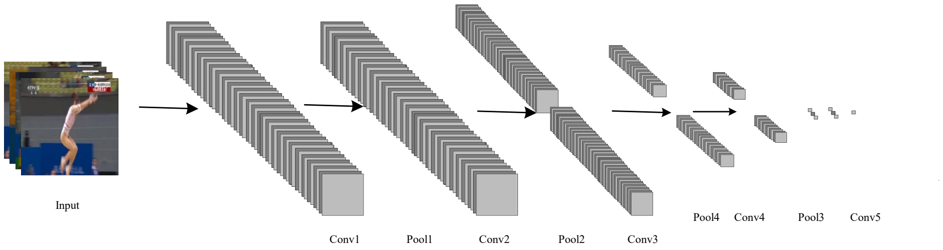

Figure 2: C3D network architecture

2.3 SVM-based video multi-classification

After obtaining the video features, this paper will establish a two-class classifier for each decomposition action to determine whether the video contains such actions[10]. In the context of establishing a binary classifier, this study utilized a SVM classifier to achieve.When using SVM classifier, the input space is transformed into a high-dimensional feature space by nonlinear transformation to obtain the optimal linear interface, so SVM is regarded as a generalized linear classifier.The two types of linear separable problem models in SVM are shown in Figure 3.In the diagram, the two categories of training samples are depicted as "*" and "_". "x1" and "x2" symbolize the two feature items of the samples. "H" represents the boundary interface known as H'. Additionally, $H_1$and H_2 respectively denote the planes parallel to the boundary interface that pass through the points nearest to the interface among the samples from the two categories. In order to ensure the minimum empirical risk in the support vector classification model, it is not only required that the optimal demarcation line can correctly separate the two types of data, but also to maximize the two types of classification interval (M in the figure). Therefore, although H' is a boundary interface that can correctly classify, it is not suitable as a decision boundary. The principle of interface selection is to make the support vector machine show better generalization ability.

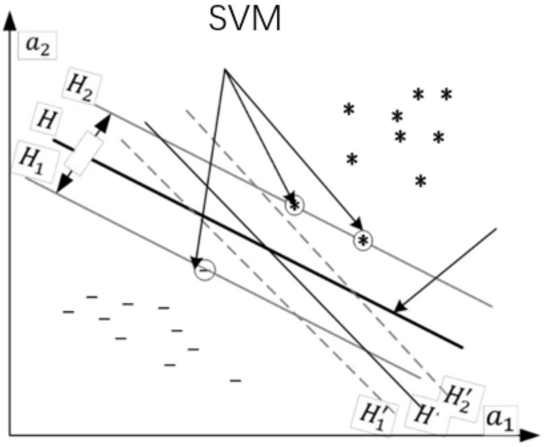

Figure 3: SVM interface diagram

Let $\{(a_1, c_1), \dots (a_N, c_N)\}$ denote a linearly separable set of samples for a binary classification problem, where$a_i \in R^d$ and d represent the dimensions.

Let $c_i \in \{-1,1\}$, $i \in [1, N]$, and w represent d-dimensional vectors for class labels, with b being a constant. From this, we can derive the linear discriminant

$$c(a) = w^T a + b \tag{1}$$

In order to obtain the maximum classification interval M, the interface needs to meet the following requirements:

$$w^T x + b \begin{cases} > \dfrac{M}{2} & for \quad y_i = 1 \\ < \dfrac{M}{2} & for \quad y_i = -1 \end{cases} \tag{2}$$

The equation (2) is normalized so that all samples can satisfy $|c(a)| \geq 1$, and the sample with the smallest distance from the interface satisfies $|c(a)| = 1$.Based on the aforementioned, it can be inferred that

$$c_i(w^T a_i + b) \geq 1 \qquad for \quad i = 1, \ldots N \tag{3}$$

$M = \frac{2}{\|W\|}$ can be derived. It can be seen that when $\| w \|$ is the smallest, the classification interval is the largest. And in order to meet the objective function to become a quadratic programming problem, so take the minimum value of $\| w \|^2$, which can be obtained :

$$\begin{aligned} & \min_{w,b} \frac{1}{2} w^T w \\ & s.t. \ c_i \left(w^T a_i + b \right) \geq 1 \quad for \quad i = 1, \ldots N \end{aligned} \tag{4}$$

According to the Lagrange method, the corresponding Larange function is obtained :

$$L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i \left(c_i (w^T a_i + b) - 1 \right) \tag{5}$$

The preceding description pertains to the case of linearly separable data. However, the actual problems often involve datasets that are linearly inseparable, known as non-linearly separable problems. As a result, the classification process inevitably encounters instances of misclassification. One of the solutions is to transform the nonlinear problem into a linear separable problem. Specifically, the kernel function is introduced to map the nonlinear separable problem of the input space to a higher-dimensional feature space. The Gaussian radial basis function kernel is employed in this paper:

$$K(a_1, a_2) = \exp(-q \| a_1 - a_2 \|^2) \tag{6}$$

The parameter q represents the width of the kernel function.

For N binary classification problems of N decomposition actions, N binary classification SVMs are constructed. Each SVM is trained in a one-to-many manner, that is, the video containing the current decomposition action is used as a positive sample, and other data is used as a negative sample.

2.4 Automatic understanding based on multi-classification

In the multi-classification automatic understanding task, this paper uses multiple two-class SVM classifiers to classify[10], and combines multiple classification results of each video into a sentence to form a description statement of floors action. This process does not involve the processing of video data, but only compares the classification results with the test description of video automatic description. Given that a majority of the video data exhibits limited categories and lacks clear logical relationships between each category, the semantic information is temporarily disregarded. Instead, the classification results are directly connected to form descriptive statements, completing the entire automated description process.

In order to carry out comparative experiments, the identified floors description sentences are compared with the correct description of the video automatic description, and the evaluation index BLEU of the video automatic description is used for comparison[11]. It is worth noting that BLEU is an evaluation index for machine translation tasks, which measures the similarity between the generated translation results and the reference translation. In this paper, it is applied to the description generation task, that is, the results of the automatic description are compared with the correct description to evaluate the quality of the generated results.

By using the BLEU index, the experimental results are compared for objective evaluation, and the similarity and quality between the description generated by the classification results and the correct description of the automatic video description are measured. This helps to evaluate the performance and accuracy of the automatic understanding model.

3. Experimental results and analysis

3.1 Experiment setting

The experiment in this paper is completed on the operating system application Ubuntu 16.04 version. The code to implement the experiment is based on the Tensorflow1.6.0 framework, and the language used is Python2.7. The network model is trained on two NVIDIA Titan 1080 graphics cards with 11 GB memory. The input video data is sampled every 5 frames. The input of the C3D feature extraction model is a 16-frame long fragment. There are 8 frames of overlap between the two consecutive fragments. The fc6 activation of these fragments is averaged to obtain a 4096-dimensional video descriptor. This paper utilizes the Libsvm classification tool[12], LIBSVM (A Library for Support Vector Machines), which is a support vector machine library developed by Professor Lin Chih-Jen in Taiwan in 2001[13]. It is primarily used for classification and requires preprocessing of the learned video features in the fc6 format, which follows the pattern of <label> <index1>:<value1> <index2>:<value2> ...

The construction of floors decomposition action data set is the basic work of floors automatic description. For the construction of the dataset, this paper collected a large number of high-standard events of professional athletes, including multiple heavyweight events of men and women such as the Olympic Games, the World Championships, and the National Games. First of all, the video of these events is preprocessed. A complete video of the floor event is completed by the participation of many athletes. During the period, it will be interspersed with wonderful moments of playback, slow-motion commentary, and judges ' scoring ranking.In the massive video, the athletes are cut as a unit, and only the athletes ' floors routines are retained[14]. The final training data contains 298 videos, and the test data contains 45 videos. The floors decomposition actions in the test set appear in the training set.Due to the absence of subtitles in sports commentary, the verbal explanation of decomposed action names by commentators is often accompanied by various distractions such as judging remarks and replay of exciting moments. These numerous interfering factors make it impossible to achieve accurate recognition through technologies like speech recognition. Therefore, a scientific, standardized, and academic approach is required to overcome these challenges.There is no way to achieve it through voice recognition and other technologies.The floors decomposition action description data set used in this paper can only be manually annotated based on real-time sports commentary. Word segmentation and word frequency statistics are performed on 298 descriptions corresponding to 298 videos in the test set. The statistical results show that a total of 48 words appear in all descriptions, and the number of occurrences of each word is shown in Table 1.It can be seen that half of the words appear less than 10 times, and the number of words that appear once and twice accounts for nearly half. Figure 4 also makes a histogram analysis of the frequency of 25 words that appear more than 10 times. It can be seen that there are still very few words that appear more than 150 times.

Table 1: Word frequency statistics

Words	Times	Words	Times	Words	Times
stretched	165	half	19	Thomas	4
backward	164	handspring	16	planche	3
forward	157	handstand	14	Spring	3
two	145	arm	13	step	3
tucked	78	Straight	13	ring	2
salto	70	flexion	13	sit	2
twist	64	extension	13	spin	2
twohalf	57	vertical	12	push-up	2
onehalf	55	fast	11	flip	2
piked	47	back	10	threehalf	1
one	45	and	9	pushup	1
three	42	spring	9	deceleration	1
jump	31	change	7	straightened	1
Arab	22	Flare	5	with	1
leg	20	Straddle	5	double	1
roll	19	split	4	withdrawal	1

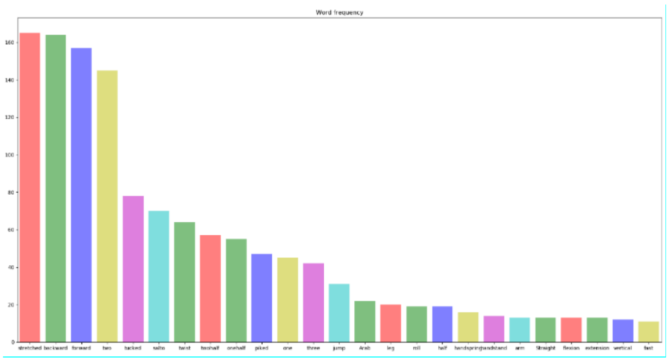

Figure 4: Shows the word frequency of more than 10 times.

3.2 Build a multi-classification dataset of floor exercise

Table 2: Number of data occurrences per category

Categories	Numbers	Categories	Numbers	Categories	Numbers
1	11	12	10	23	24
2	5	13	16	24	5
3	5	14	16	25	13
4	27	15	7	26	6
5	22	16	5	27	17
6	38	17	7	28	43
7	4	18	4	29	17
8	38	19	23	30	2
9	11	20	16	31	5
10	9	21	2		
11	9	22	7		

In order to ensure the accuracy and reliability of the comparative experiment, this paper uses the data set and the video automatic description experiment to be the same data set, and the description of the data does not change.

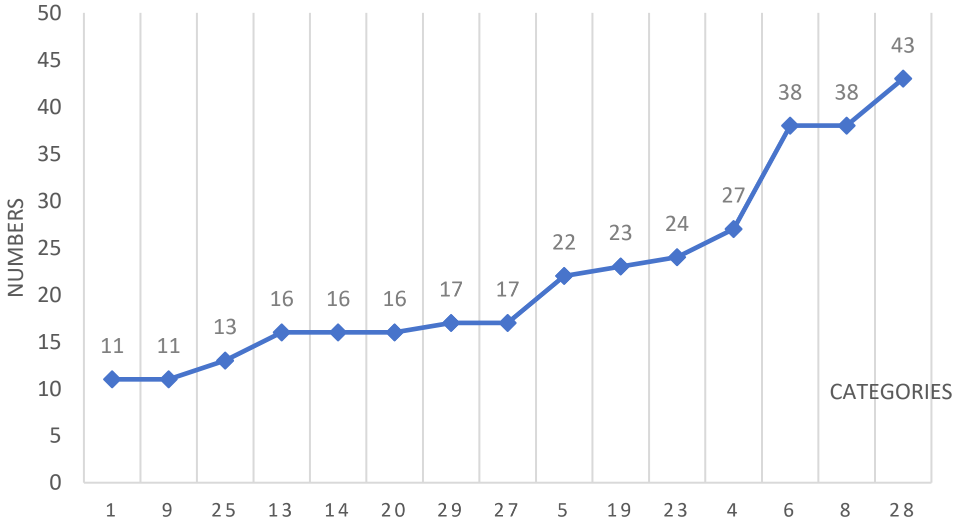

Figure 5: The number of categories with data volume greater than 10 broken line graph

In the experiment, we classify the data according to the professional movements of the floors competition to construct the data set. Each video data may be composed of more than one decomposition action, so each video data is marked as at least one category, and the category is represented by a positive integer (starting from 1), with a total of 31 categories. Each category may require multiple words to describe. The number of categories in this article is not directly related to the number of words in the video automatic description.

The number of occurrences of each category is shown in Table 2. It can be seen that half of the categories appear less than 10 times, and only a few categories appear particularly frequently. Figure 5 analyzes the frequency of 16 categories that appear more than 10 times, and only the number of occurrences of individual categories is particularly high.

3.3 Experimental results and analysis

(1) Multi-classification evaluation index and experimental results analysis

Accuracy is the most common performance metric in classification models. It is suitable for binary classification models and can also be used for multi-classification models. The calculation of accuracy is also relatively simple. Assuming the classification model is g, with the test set D containing N data, the formula for calculating accuracy is:

$$A = \frac{1}{N}\sum_{i=1}^{N}(f(a_i) = label_i) \tag{7}$$

In Table 3, the classification accuracy of 31 categories is listed respectively. From the data results, there are some differences in the accuracy of each category, but this difference is not significantly related to the number of data under this category.

Table 3: The classification accuracy of each category

Categories	Accuracy(%)	Categories	Accuracy(%)	Categories	Accuracy(%)
result1	75.00	result12	75.00	result23	70.00
result2	75.00	result13	68.33	result24	71.67
result3	66.67	result14	68.33	result25	70.00
result4	68.33	result15	75.00	result26	66.67
result5	70.00	result16	66.67	result27	73.33
result6	65.00	result17	73.33	result28	70.00
result7	71.67	result18	70.00	result29	68.33
result8	71.67	result19	70.00	result30	73.33
result9	71.67	result20	70.00	result31	65.00
result10	73.33	result21	68.33		
result11	73.33	result22	78.33		

(2) Automatic understanding of evaluation indicators

The ultimate goal of the multi-classification in this paper is to realize the automatic description of the video. The evaluation index used in this paper is the same as the automatic description of the video. The average BLEU of Bleu _ 1 to Bleu _ 4 is still taken as the evaluation index, and the recognized floors description statement is compared with the correct description marked by traditional feature extraction network methods such as VGG16, ResNet101, ResNet50, DenseNet20, etc. The experimental results are shown in Tables 4. It can be clearly seen that the method of automatic description of floors video by using the transformation of video multi-label classification is used in this section. The experimental results are outstanding.

Table 4: The experimental results of this method are compared with those of three feature extraction network methods.

	$\overline{\text{B}}$leu
Ours	41.25
VGG16	12.30
ResNet101	17.85
ResNet50	17.88
DenseNet201	18.75

(3) Case analysis

Figure 6 is a representative frame of a video in the test data of the self-built dataset. Due to the limitation of the number of pages, only one instance is given as a reference. The instance of video automatic description is the same video as the experimental instance. The floors automatic description of different models on the self-built data set is compared with the experimental results of the multi-label classification method in this paper. Compared with the original model S2VT, the model obtained after introducing the attention mechanism has similar results in the direction of the blue font ' forward ' test, but in the body posture such as the ' stretched ' of the red font, the improved model will be more specific. In the classification problem, this video contains two actions. Although the classification results only identify a correct classification ' forward stretched twist three ', the category contains four correct description words, which improves the accuracy of the description.

Figure 6: The representative frame of a video in the self-built dataset test data.

Accurate description: forward stretched twist three forward stretched twist one

SV2T: twohalf forward onehalf forward

Introducing attention mechanism: onehalf forward one stretched forward

Scheduling: two stretched <PAD> stretched <PAD>

Proposed method in this paper: forward stretched twist three backward piked two

4. Summary of this article and future prospects

4.1 Research summary

This paper proposes an automatic description method of floors based on support vector machine multi-label classification. The characteristics of the research object 's floors video data are not only that there are a large number of key frames, but also that the definition of its high-level semantic things is relatively fixed, and the floors movements have certain arrangement rules. Therefore, the video automatic description problem is transformed into a video classification problem. The important technical support to promote the transformation of this problem is the existing high-precision video classification model.In order to retain the time signal of the video, this paper uses the C3D feature extractor to input the extracted feature vectors into multiple binary SVM classifiers to complete the task of multi-label classification. In order to verify the feasibility of problem transformation, the classification results are mapped into natural language descriptions. The final experimental results show that the transformed method has greatly improved the automatic description of video.

However, we also realize that there are still some shortcomings in this paper. Firstly, our method still has certain limitations in feature extraction and further improvements are required to enhance classification accuracy. Secondly, for the fine description of the floors video, our method still needs to be improved. In addition, our data set is relatively small and needs to be further expanded and enriched.

4.2 Future prospect

(1) The automatic description method of floors based on support vector machine multi-label classification also has many areas that can continue to be studied and improved. According to the network structure, feature extraction is still a very important module, and there is a lot of room for improvement. It can also integrate multi-modal video features such as optical flow and sound to improve the network.For the classifier, the SVM classifier selected in this paper is initially a two-class model. After

the amount of data increases, the calculation of multiple two-class SVM models to complete the multi-label classification task will be very large. The subsequent research can improve its algorithm and build a fast and efficient SVM multi-label classifier. Other classifiers can also be used to construct models to improve the accuracy of classification results. In the process of natural language transformation of classification results, the codec structure can also be introduced to improve the accuracy.

(2) At present, the data set capacity of the experiment is small, and the data distribution is not uniform. In the future, it is also necessary to enrich the self-built floors decomposition action video data set, and ensure that the proportion of various floors decomposition actions is balanced enough. This can better train the algorithm model and make it have better generalization ability.

References

[1] Qin Dan.Research on video content automatic classification algorithm based on multi-feature combination and SVM [D].Shanghai Jiaotong University, 2009.

[2] Li Haixia. Design and implementation of video automatic description system based on deep learning [D].University of Electronic Science and Technology of China, 2018.

[3] Zhang Lu.Comparative study on the current situation of amateur gymnastics training between China and the United States [D].Chengdu Institute of Physical Education, 2014.

[4] Xin Lijuan. Architecture and Application of Smart Sports [J]. Electronic Technology and Software Engineering, 2018 (15): 128-128.

[5] Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions[J]. International Journal of Computer Vision, 2002, 50(2): 171-184.

[6] Wang Jianhong.Research on key technologies of video-based human action recognition [D]. Southeast University, 2017.

[7] Liu Duanyang, Qiu Weijie. Multi-label classification based on weighted SVM active learning [J]. Computer Engineering, 2011, 37 (8): 3.DOI: 10.3969 / j.issn.1000-3428.2011.08.062.

[8] Zhao Pengpeng, Jiao Yang, Xian Xuefeng, etc. A multi-label active learning classification method and system based on SVM: CN201410184086.8 [P].CN201410184086.8 [2023-11-30].

[9] Qian Wenzhuo. Human action recognition technology based on MA-C3D neural network [J].Modern computer, 2021, 27 (35): 6.

[10] Jiao Chunpeng. Comparative study of multi-classification methods based on binary SVM [D].Xi 'an University of Electronic Science and Technology [2023-11-30].DOI: CNKI: CDMD: 2.1013.111129.

[11] Callison-Burch C , Osborne M , Koehn P .Re-evaluation the Role of Bleu in Machine Translation Research[C]//EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy.2006. DOI:http://dx.doi.org/.

[12] Zhang Yu. Research on the optimization design of individual complete sets of movements of Chinese women's competitive gymnastics under the international gymnastics rules in 2013-2016 [D].Hunan Normal University, 2014.

[13] Tong Zhanbei, Zhong Jianwei, Li Zhenwei, et al. Power quality disturbance classification based on VGG16 image feature extraction and SVM [J].Electricity, 2023 (7): 7-13.

[14] Mogan J N, Lee C P, Anbananthen K S M ,et al. Gait-DenseNet: A Hybrid Convolutional Neural Network for Gait Recognition[J].IAENG Internaitonal journal of computer science, 2022(2 Pt.2): 49.