


Research Article

Automatic Description Method for Sports Videos Based on Economic Management Effectiveness Visualization Techniques

Dongming Shao¹ and Jing Han² 

¹Department of Sports, South China Agricultural University, Guangzhou, Guangdong 510521, China

²Guangdong University of Finance Athletic Department, Guangzhou, Guangdong 510521, China

Correspondence should be addressed to Jing Han; 26-070@gduf.edu.cn

Received 27 April 2022; Revised 1 July 2022; Accepted 6 July 2022; Published 8 August 2022

Academic Editor: Wen-Tsao Pan

Copyright © 2022 Dongming Shao and Jing Han. 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.

The automatic description (AD) of sports videos is a fundamental task for archiving the content of broadcasters, as well as understanding video scenes, and economic management effectiveness visualization techniques are key to the classification of sports videos. In this paper, a freestyle gymnastics video is used as an example to study the automatic video description by observing the set of movements of an athlete in a freestyle gymnastics video to generate the terminology of the movements performed by that athlete. The technique used in this paper to visualize the effectiveness of economic management is the long and short-term memory (LSTM) network model, which is used to learn the mapping relationship between word sequences and video frame sequences. Attention mechanisms (AM) are also introduced to highlight the importance of keyframes that determine freestyle gymnastics movements. The study is carried out by building a dataset of free gymnastics (FG) breakdown movements from professional events and applying a planned sampling method. Experimental results show that the method can improve the accuracy of an automatic free gymnastics video (FGV) description. The proposed method has a wide range of applications in sports analysis and instruction.

1. Introduction

In the 21st century, along with the rapid development of Internet technology, video, a common form of multimedia data, has gradually become one of the important components of multimedia data [1]. In people's daily lives, huge amounts of video data are generated, of which automatic video description allows for the effective management of these video resources [2]. With the in-depth research on automatic video description, automatic video analysis based on human movement has made significant progress in areas such as intelligent life assistance, advanced human-computer interaction, and content-based video retrieval, and is gradually receiving close attention.

There is a difficult problem in sports video analysis research. Namely, it is difficult for low-level video features to accurately reflect the needs of the human body, and the use of single features is difficult to meet the rapid growth of

available video data. As a popular sport, the research on AD of FG has made considerable achievements. Among them, the AD of FGV not only integrates theoretical knowledge of machine learning [3] but also involves several disciplines such as pattern recognition and video analysis, and we need to conduct deeper research on it.

There are many issues that have not been adequately addressed in current research on the problem of AD in freestyle gymnastics videos. In terms of practical applications, the study of the AD of FG videos has an enormous application value. Among other things, in FG movements, we need a quick identification of the various types of movements of the athletes. This paper aims to achieve high recognition accuracy in automatic video-based human movement understanding, and even real-time movement recognition and commentary. For nonexperts, if ADs can be achieved, it will not only enhance the viewing experience but also facilitate their understanding and learning of the sport of FG.

Kojima et al. [4] took an alternative perspective by studying human activity through the theory of behavioral concepts and thus described human behavior. Guadarrama et al. [5] combined the semantic hierarchy theory with semantic relations between multiple fragments. Rohrbach et al. [6] described features through mathematical modelling by studying human activity under conditional random fields. Xu et al. [7] proposed a combination of a deep video model and a joint embedding model as a kind of framework that allowed for the study of relationships between videos and words. The above research methods were limited by some syntactic structures [8], making the research results deviate from everyday descriptions.

With the continuous development of deep neural networks (NN) [9] and the emergence of many large-scale datasets in image recognition [10], many approaches to semantic representation have changed dramatically. Hochreiter et al. [11] proposed LSTM, which can effectively solve the gradient disappearance problem of Recurrent Neural Network (RNN) species. Gers et al. [12] proposed an oblivion gate mechanism, in which Graves et al. [13] improved the LSTM and proposed a bidirectional LSTM (BLSTM) NN, which has been widely used.

Venugopalan et al. [11] used a convolutional neural network (CNN) to feature extract all frames in a video and fed them into an LSTM to decode and generate text. Venugopalan [14] proposed S2VT with an LSTM in both the front and back segments. Shetty et al. [15] trained a variety of models on different kinds of features, using an evaluation network to assess and generate a description of the video by generating correlations between sentences and video features. Jin et al. [16] used multiple features and fused features to represent the video.

With its greater freedom, varied movements, and the ability to perform a complete set of moves in a set time, FG is one of the most aesthetically pleasing sports in competitive gymnastics and is the quintessential representative of competitive gymnastics. The study of automatic video descriptions of FG is of great relevance. For nonexperts, ADs would not only improve the viewing experience but also make it easier for them to understand and learn the sport of FG. In this paper, automatic video description is performed by extracting the athletes' set movements from FGVs. The LSTM network visualization techniques are used to learn the mapping relationship between word sequences and video frame sequences. An AM is also introduced to highlight the importance of the keyframes that determine the FG movements. Experiments are conducted on MSVD data and self-built datasets, using planned sampling to eliminate the differences between the training decoder and the prediction decoder.

2. Techniques for Visualizing the Effectiveness of Economic Management

2.1. Background and Issues. In recent years, the study of AD of sports video content has gradually become a hot topic, with the rapid growth of sports video data volume and audience groups. Apart from football and badminton, which

are typical representatives of ball sports, other areas of sports video research are less involved. This paper takes FGV as the object of study because it plays a fundamental role in other sports. FG has the greatest degree of freedom and difficulty among competitive gymnastics and is highly representative.

By FGV comprehension, we mean understanding a given video of FG. The terminology for the set performed by the athlete is generated by observing the set in the video, such as the method, direction, and angle of the body flip. The traditional method relies on manual commentary, and many important competitions require real-time commentary by the commentator, which demands a high level of expertise. Nonspecialists understand the competition primarily through the point of view of the commentator, and any errors in the commentary will reduce the viewing experience of these people. It is essential to use pattern recognition techniques combined with natural language processing to achieve ADs of FGVs.

2.2. Algorithmic Framework Structure. The framework of this paper is shown in Figure 1 and uses economic management effectiveness visualization techniques to analyze freestyle gymnastics video features. That is, an LSTM network [17] is used to express the mapping between the features studied and the words, to enable the description of the language. With the development of deep neural nets, many larger datasets have emerged, such as Sport-1M. In the study of this paper, the data used contain videos of FG from the Olympic Games and the National Games, and their decomposed movements are studied as a dataset.

In the FGVs, the decomposed movements mainly include the direction of the flip, the number of rotations, and the body posture of the athletes. The video frames containing the key movements of FG are defined as keyframes, and the keyframes with high discriminative power are extracted to improve the accuracy of the video description. The discriminative power of the video frames is calculated through an AM. In this paper, the AM [18] is integrated into the existing video description network to maximize the accuracy of video description by calculating the weights between different video frames.

The basic framework shown in Figure 1 begins with the construction of a free-form gymnastics decomposition movement dataset. The AD of sports videos starts with a CNN for feature extraction. For text data that have been annotated, the corresponding dictionary needs to be proposed to extract the corresponding features. Randomly selected text data are trained in the model until the model's effect stabilizes. The remaining text data are used as the test set, resulting in automated descriptions of freestyle videos.

2.3. Feature Extraction. In the AD of the freestyle video studied in the text, the data type contains not only video data but also text data. We achieve a more accurate description of this video by using CNNs to extract video features and natural language text processing for text features.

The NN model is robust and the training cost of the model is small and the classification accuracy is high. The

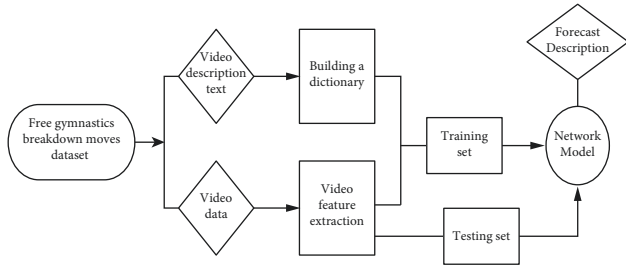


FIGURE 1: Basic framework.

CNN model has simple operating conditions and does not require much hardware for the device, and the specific morphology of the features is not considered at all when using feature extraction [19]. In this section, three different types of 2D CNN, Visual Geometry Group (VGG) [20], ResNet [21], and DenseNet [22], are used to perform feature learning on videos, respectively, and the VGG network structures are shown in Figure 2. The VGG network structure to extract feature representations of freestyle gymnastics videos has a significantly lower error rate. ResNet can use the original signal directly into the deeper layers of the neural network, speeding up the training efficiency of the network, while DenseNet builds on and improves ResNet. The feature mapping generated by DenseNet will also be used as input to all subsequent layers, ensuring that the information is passed on, and thus avoiding gradient disappearance.

In this paper, the descriptors of the FGVs are transformed into features by using the one-hot vector coding. The words in the FG annotated text are first counted to construct a dictionary. The number of words used in the descriptions of the FG decomposition movements is not large, so no filtering of words is performed in the preprocessing.

3. AD Methods for Sports Videos

3.1. AD System for FG. We use LSTM to learn video features from this paper. Standard recurrent NNs are prone to gradient disappearance during backpropagation, making it difficult to continuously optimize the network parameters [23]. The LSTM is a special type of recurrent NN that can effectively solve this problem, especially in long-distance dependent tasks, where it outperforms the RNN. These are input gates, forgetting gates, and output gates. The gate control can be regarded as a fully connected layer in the CNN, and the LSTM stores and updates the information through these gate controls. The gate control quantifies the amount of information passing through each part of the cell by using a sigmoid function to obtain a probability value between 0 and 1. When the sigmoid function is 0, no information variables are allowed to pass at that moment, and when the sigmoid function is 1, all variables are allowed to pass at that moment. The gates for forgetting are called “forgetting gates” and gates for outputting are known as “output gates.”

The LSTM encodes a fixed-dimensional sequence of FG decompositions into feature sequences, which are then

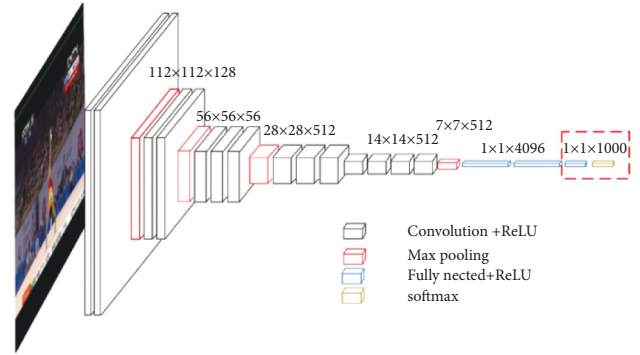


FIGURE 2: VGG16 network structure diagram.

decoded and used to generate text by the LSTM NN. First, encode the fixed dimensional FG decomposition movement feature vector $X = (x_1, \dots, x_n)$ into a feature sequence, and obtain the output $H = (h_1, \dots, h_n)$ of the corresponding hidden layer. The output of the LSTM is known to be dependent on the previous input sequence, so the feature vectors are fed into the LSTM once in sequence, and the output is a coded mapping of the sequence vectors. After the feature vector of the last frame is input, the output of the LSTM is the encoding of the sequence of frames. The LSTM in the decoding phase is fed the start character, which prompts it to begin decoding the hidden state it is subjected to a sequence of words, and the output yields a sequence of words $Y = (y_1, \dots, y_m)$ with a probability of $p(y_1, \dots, y_m | x_1, \dots, x_n)$:

$$p(y_1, \dots, y_m | x_1, \dots, x_n) = \prod_{m=1}^{t=1} p(y_t | h_{n+t-1}, y_{n-1}). \quad (1)$$

When training in the decoding phase, the log-likelihood of the predicted sentence is found under the condition that the hidden state of the frame sequence and the previously output words are known. The model is trained so that the following equation reaches its maximum value:

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{m=1}^{t=1} \log p(y_t | h_{n+t-1}, y_{n-1}; \theta), \quad (2)$$

where θ is the angle of the vector at which the maximum log-likelihood is reached and argmax indicating the maximum value. The entire training dataset is optimized using a stochastic gradient descent algorithm, which allows the LSTM to learn more appropriate implicit states. The output of the second layer LSTM z is specified by finding the most probable target word y in the vocabulary Y as shown in the following equation, where W_y indicates the weight of the output:

$$p(y | z_t) = \frac{\exp(W_y z_t)}{\sum_{y' \in Y} \exp(W_{y'} z_t)}. \quad (3)$$

3.2. Attention Mechanism. The difference in the attention allocated to different signals by the human brain when

processing signals is referred to as visual AM[24]. The area of the target on which human vision can gain focus by quickly capturing the image, in order to obtain more detailed information about the target to be focused on and to eliminate other useless information is referred to as the focus of attention [25]. In this paper, the AD of FG movements is based on the principle of the AM, which first selects the decisive video movements that can be taken, i.e., the way the athlete's body flips, the angle, and the different directions, which should be assigned more weight in order to make the AD more accurate. The introduction of this AM allows the decoder to assign weights to all feature vectors in the FGV. The structure of the model containing the AM is shown in Figure 3.

In this paper, a dynamic weighted sum of temporal feature vectors is used, with the following equation:

$$\varphi_t(X) = \sum_n^{i=1} \alpha_i^{(t)} \chi_i, \quad (4)$$

where t denotes the moment t and x_i denotes the vector. $\sum_n^{i=1} \alpha_i^{(t)} = 1, \alpha_i^{(t)}$ is the proportion of the overall score that the output of the hidden layer at that moment matches the entire video representation vector, calculated as follows:

$$\alpha_i^{(t)} = \frac{\exp(\text{score}(x_i, h_i))}{\sum_n^{j=1} \exp(\text{score}(x_j, h_j))}, \quad (5)$$

where $\text{score}(x_i, h_i)$ denotes the fraction of the video feature vector x_i occupied by the output h_i of the i th hidden layer, the larger the fraction, the greater the attention of the input at this moment in that video, which is calculated as follows:

$$\text{score}_{x_i, h_i} = w^t \tanh(Wx_i + Uh_i + b), \quad (6)$$

where w, W, U are weight vectors and b are offsets.

4. Model Analysis

4.1. Experimental Setup. The graphics card in this paper is an NVIDIA Titan 1080 and the memory size is 11 GB. During the network training, the input data were resized to $227 * 227$, and the VGG-16 pretrained model provided in the model parameter training was performed directly on the ILSVRC-2012 image set, a subset of the ImageNet. Comparison experiments were added in order to verify the impact of the features extracted from the different 2D CNN on the description results of the freestyle gymnastics videos. We conducted experiments on the ResNet101, ResNet50, and DenseNet201 CNN to compare the results of the experiments after feature extraction and input to the model for coding and decoding.

4.2. Dataset Construction. The construction of a dataset of decomposed movements for FGs is an essential task for the AD of FGs. The experimental dataset in this paper is mainly collected from videos of professional athletes competing in professional competitions, such as the Olympic Games, World Championships, National Games, and several other heavyweight events. These collected videos are first

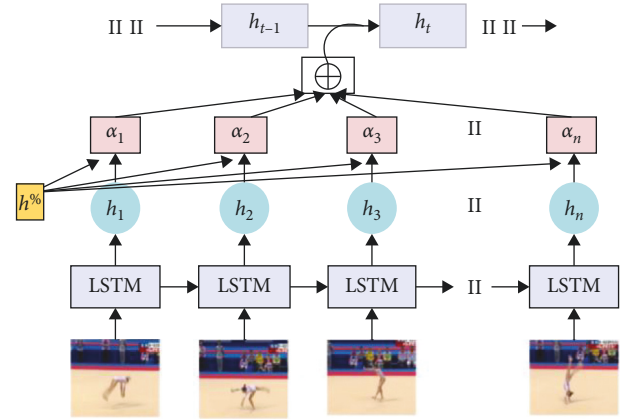


FIGURE 3: Integrated network architecture.

preprocessed, with a number of video frames per athlete being cut off to include only the athlete's FG movements. Because these videos are interspersed with highlights, replay, and slow-motion commentary, which together make up a video, these are the parts that were ignored in the AM.

Through data collection, we obtained a total of 298 training video data and 45 test video data. After pre-processing all the videos, there are still some problems. As these videos are among those obtained live, there is no real-time caption display for the narrator's words, and we address the effect of distracting factors by using speech recognition. In this paper, word frequency statistics were performed on the 298 video descriptions collected, and the results showed that a total of 48 words appeared in these descriptions, and the word frequencies of all words are shown in Table 1. The words occur less than 10 times, and nearly half of the words occur once and twice. Figure 4 also analyses the frequency of the 25 words with more than 10 occurrences. It can be seen that the number of words with more than 150 occurrences is still very small and the names of the words in Figure 4 are replaced by the first two letters.

4.3. Scheduled Sampling. In the decoder of the training phase, it is the target sample that is used as the input for the next predicted subsample. Whereas in the prediction phase the decoder takes the previous prediction result and uses it as an input for the next prediction value. This difference leads to the problem that the training and prediction scenarios are different. In prediction, if the previous word is predicted incorrectly, all subsequent ones will follow, whereas the training phase does not.

This paper modifies the model of the decoder during training by introducing a planned sampling approach. The base model will only use the true annotated data as input, the training decoder with the addition of planned sampling is to select the model's output with a probability P as the input for the next prediction and the true markers with $1-P$ as the input for the next prediction. That is, the sampling rate of P varies during the training process. In the beginning, when training is not sufficient, start by making P smaller and try to use the true description as input, and as training progresses,

TABLE 1: Word frequency statistics.

Words	Frequency	Words	Frequency	Words	Frequency
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

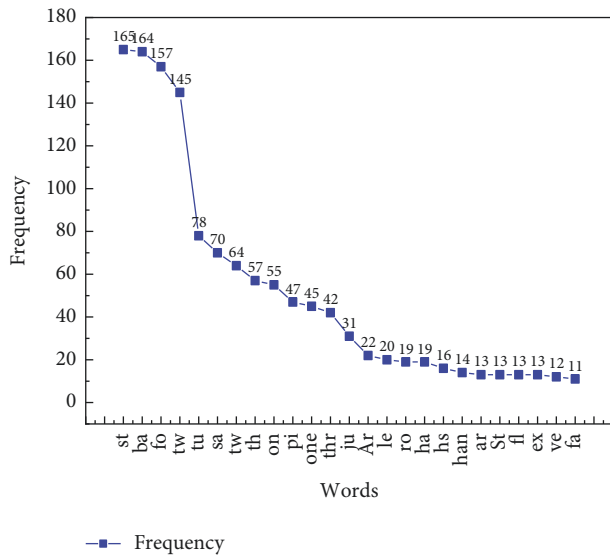


FIGURE 4: Frequency of words occurring more than 10 times.

increase P and use more of its own output as input for the next prediction. As training progresses, P gets larger and larger, and the training decoder model eventually becomes the same as the prediction decoder model. Eventually, the difference between the training and the prediction decoder is reduced by planning the sampling scheme.

4.4. *Loss Function.* The iterations of the loss function in the experiments are presented using the visualization tool TensorBoard, as shown in Figure 5, which shows the variation of the training loss value with the number of iterations for the original model and the model after the introduction of the AM. The loss values of both models gradually decrease and converge. The model with the AM has an increased rate of convergence as the time complexity increases and the starting loss value is larger.

4.5. *Evaluation Metrics and Performance Comparison.* The result of the AD of FG is the description of the decomposition of FG movements, which is a kind of

natural language, and the evaluation of the result of the description can be referred to the metrics used in natural language to evaluate the quality of machine translation results. Bleu is the closest metric to the human rating at present. Bleu is a matching principle using N -gram. N -gram is the representation of a sentence as a sequence of n consecutive words. This paper conducts experiments on two corpora, the MSVD and the self-built dataset. The experimental results are shown in Table 2.

The experimental results compared in Table 2 are the mean Blue from Bleu_1 to Bleu_4. The table shows three datasets, two of which are our own, each of which is different when tagging the video descriptions. OURS(1) is the most straightforward natural language, and since the professionalism required for the description of FG breakdown movements is high, OURS(2) is different when describing markers; the descriptive statements are adapted to the specialist terminology. From the results of the three datasets, we know the model with the AM introduced in this paper performs better on both the MSVD dataset and the self-built dataset.

The different test results for the three models are given in Figure 6. From Figure 6, we know that the MSVD has the best performance among the three models and it accounts for the largest percentage. Whereas OUS(1) is the least effective and OUS(2) is the second most effective, indicating that the effectiveness of the modified model has improved.

Table 3 shows the comparison of the experimental results of feature extraction using ResNet101, ResNet50, and DenseNet201 networks on the self-built dataset OURS (2). In Table 3, two evaluation metrics, ROUGE_L and METEOR, are also added, and in order to highlight the gaps in the experimental results more, Table 3 compares the results of the Bleu evaluation metrics in Table 2 with specific The average Bleu from Table 2 is compared and expanded into Bleu_1, Bleu_2, Bleu_3, and Bleu_4 specific results. The specific experimental results show that the DenseNet201 performs the best in all evaluation metrics compared to VGG16.

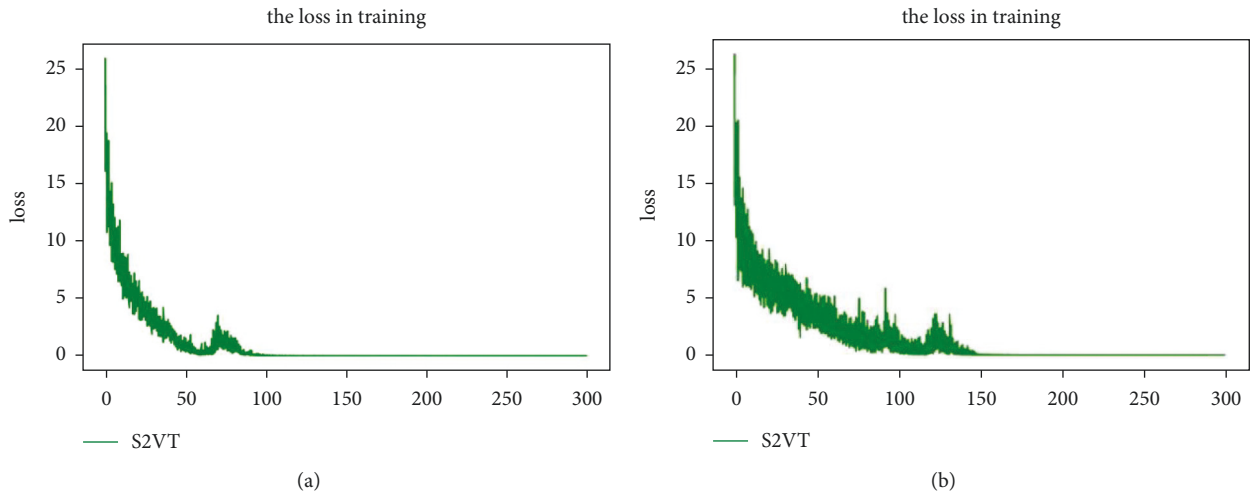


FIGURE 5: Loss function values. (a) Original model. (b) The proposed model.

TABLE 2: Analysis of experimental results of the improved model.

Category	S2VT	AMs	Planned sampling
MSVD	17.2	17.9	18.8
OURS (1)	8.7	9.3	10.4
OURS (2)	10.9	11.2	12.3

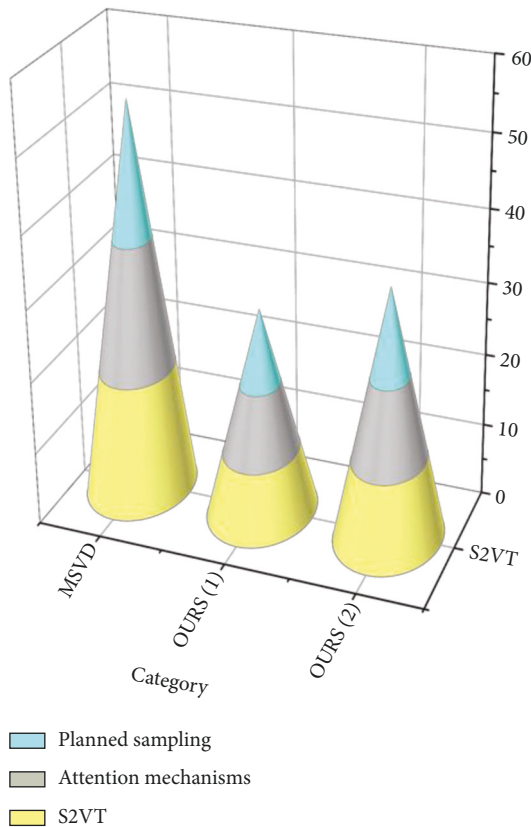


FIGURE 6: Comparison of experimental results of improved models.

4.6. Comparison of the Experimental Results of the FG Video Depiction Examples of the Different Models on the Self-Constructed Dataset. Compared to the S2VT, the results of the

model with the AM are similar for the direction test as “forward” in blue, but for the body posture test as “stretched” in red, the improved model is more specific. A comparison of the visual results of the different feature extraction networks is shown in Figure 7.

The ultimate goal of the multiclassification in this paper is to achieve AD of the video, using the improved method for testing. To ensure experimental rigour, the mean Blue from Bleu_1 to Bleu_4 is still taken here as the evaluation metric, and the freestyle description statements identified are compared with the correct descriptions labelled from the previous section. The comparison of the experimental results is shown in Table 4, and it is clear that the method of using video multilabel classification transformation for AD of freestyle gymnastics videos gives better experimental results.

The pie charts of the experimental results of the five methods are shown in Figure 8. The experimental results of the method in this paper are the best, which verifies the effectiveness of the proposed method.

The experimental results of the different AD models of FG on the self-constructed dataset are compared with those of the classification method in this paper. Compared to the original model S2VT, the results of the model with the AMs are similar for the direction test as “forward” in blue, but for the body posture test as “stretched” in red, the improved model is more specific. In the classification problem, the video contains two actions, and although only one correct category is identified, “forward stretched twist three,” this category contains four correctly described words, thus improving the accuracy of the description.

Although the improved model improves the accuracy of the description, the AD method for FGV based on the LSTM networks only applies a two-dimensional CNN model for

TABLE 3: Comparison of experimental results for different feature extraction networks.

Category	Belu_1	Belu_2	Belu_3	Belu_4	Belu	ROUGE_L	METEOR
VGG16	21.64	16.42	10.50	0.64	12.30	40.67	14.42
ResNet101	30.63	25.27	14.83	0.67	18.75	41.94	15.66
ResNet50	30.74	25.32	14.80	0.67	17.88	41.94	15.64
DenseNet201	31.17	25.88	17.19	0.74	18.75	43.26	16.14

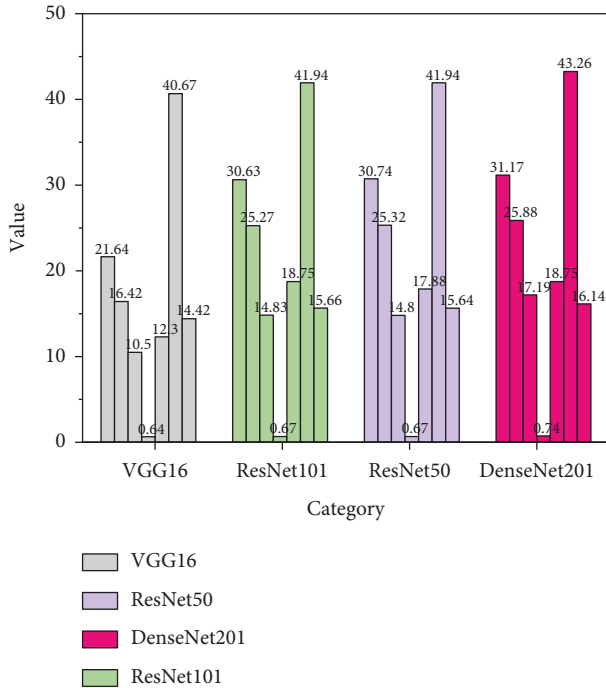


FIGURE 7: Comparison of visual results for different feature extraction networks.

TABLE 4: Comparison of experimental results.

Category	Bleu mean value
Methodology of this paper	41.25
VGG16	12.30
ResNet101	17.85
ResNet50	17.88
DenseNet201	18.75

feature extraction. This increases the risk of gradient loss due to the temporal loss of information in the video data. In the future, three-dimensional CNN could be used for feature extraction and the network could be improved by fusing multimodal video features. In addition, the introduction of the AM could be further improved by aiming to be able to introduce several attention modules at the same time to highlight the importance of the keyframes in decision-making. Attention should be paid to the speed and efficiency of the operation and to the improvement of the algorithm to improve the accuracy of the description.

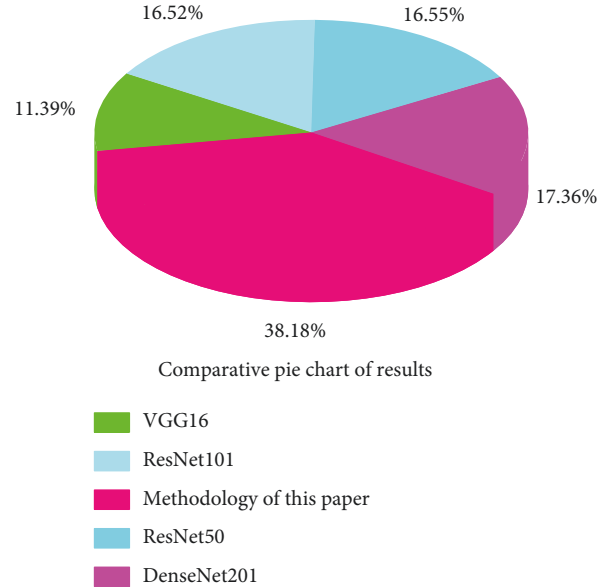


FIGURE 8: Pie chart of experimental results.

5. Conclusion

In this paper, the method of the video AD and related technical theories were introduced, and the method AD of FGV based on the LSTM networks was described in detail. Firstly, the automatic video description method was taken as the entry point, and the current status of AD video methods and sports video research studies were reviewed. From the perspective of economic management effectiveness visualization techniques, the relevant concepts and development history were introduced, and the structure of three important types of NNs was described with emphasis on the structural dissection of typical network models, respectively. The paper introduced the integration of AMs into existing video description networks, weighing the importance of video frames by the means of weight values. In the course of the experiments to improve the model's computational accuracy, application schemes were employed to reduce the discrepancy between the training decoding model and the predicted decoding model before. Finally, experiments were conducted on multiple feature extraction network structures of VGG16, ResNet101, ResNet50, and DenseNet201, through which the feasibility of the improved method was

verified. As the results of this paper were obtained in an experimental setting, there should be more extraneous factors interfering in the practical application, and the model will be improved to make it applicable in a realistic setting.

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 do not have no conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] S. Gupta, S. Hamilton, and R. Hamilton, "Marketing insights from multimedia data: text, image, audio, and video," *Journal of Marketing Research*, vol. 58, no. 6, pp. 1025–1033, 2021.
- [2] "The trustees of columbia university in the city of New York; patent issued for video description system and method," *Computer Weekly News*, vol. 86, no. 3, pp. 54–59, 2013.
- [3] K. B. Dasari and N. Devarakonda, "Detection of different DDoS attacks using machine learning classification algorithms," *Ingénierie des Systèmes d'Information*, vol. 26, no. 5, pp. 461–468, 2021.
- [4] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 171–184, 2002.
- [5] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, and T. Darrell, "Youtube2text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 2712–2719, Sydney, NSW, Australia, December 2013.
- [6] M. Rohrbach, W. Qiu, I. Titov, M. Pinkal, and B. Schiele, "Translating video content to natural language descriptions," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 433–440, Sydney, NSW, Australia, December 2013.
- [7] R. Xu, C. Xiong, W. Chen, and J. Corso, "Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified framework," *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [8] Y. Pan, T. Yao, H. Li, and M. Tao, "Video captioning with transferred semantic attributes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6504–6512, Honolulu, HI, USA, June 2017.
- [9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep CNN," *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.
- [10] O. Russakovsky, J. Deng, H. Su et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [11] S. Venugopalan, H. Xu, J. Donahue, R. Marcus, M. Raymond, and S. Kate, "Translating videos to natural language using deep recurrent NNs," 2014, <https://arxiv.org/abs/1412.4729>.
- [12] F. A. Gers, J. Schmidhuber, and F. J. N. C. Cummins, *Learning to Forget: Continual Prediction with LSTM*, vol. 12, no. 10, pp. 2451–2471, 1999.
- [13] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.
- [14] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE international conference on computer vision*, pp. 4534–4542, Santiago, Chile, December 2015.
- [15] R. Shetty and J. Laaksonen, "Frame-and segment-level features and candidate pool evaluation for video caption generation," *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1073–1076, 2016.
- [16] Q. Jin, J. Chen, S. Chen, and X. Yifan, "Describing videos using multi-modal fusion," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1087–1091, New York, NY, United States, October 2016.
- [17] M. Sundermeyer, H. Ney, and R. Schluter, "From feedforward to recurrent LSTM neural networks for language modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 517–529, 2015.
- [18] L. Yao, A. Torabi, K. Cho et al., "Describing videos by exploiting temporal structure," *Proceedings of the IEEE international conference on computer vision*, pp. 4507–4515, 2015.
- [19] A. Sharif Razavian, H. Azizpour, J. Sullivan, and C. Stefan, "CNN features off-the-shelf: an astounding baseline for recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, San Juan, PR, USA, June 2014.
- [20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, <https://arxiv.org/abs/1409.1556>.
- [21] K. He, X. Zhang, S. Ren, and S. Jian, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, San Juan, PR, USA, June 2016.
- [22] G. Huang, Z. Liu, L. Van Der Maaten, and Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.
- [23] H. Zhao, S. Sun, and B. Jin, "Sequential fault diagnosis based on LSTM neural network," *IEEE Access*, vol. 6, pp. 12929–12939, 2018.
- [24] L. Han, Y. Zhao, H. Lv, Y. Zhang, H. Liu, and G. Bi, "Remote sensing image denoising based on deep and shallow feature fusion and attention mechanism," *Remote Sensing*, vol. 14, no. 5, p. 1243, 2022.
- [25] L. Wang, H. Zhang, and G. Yuan, "Big data and deep learning-based video classification model for sports," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 1140611, 2021.