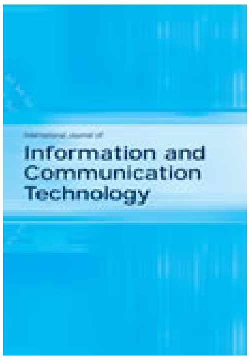
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



Information and Communication Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Enhancement of accuracy and analytical efficiency of gymnastics video using convolutional neural network and semantic analysis approach

Xiangyang Cai, Lijing Xu, Shijun Wang

DOI: 10.1504/IJICT.2025.10072456

Article History:

Received:	24 May 2025
Last revised:	16 June 2025
Accepted:	17 June 2025
Published online:	30 July 2025

Enhancement of accuracy and analytical efficiency of gymnastics video using convolutional neural network and semantic analysis approach

Xiangyang Cai

Hebi Institute of Engineering and Technology,
Henan Polytechnic University,
Hebi 458030, China
Email: qixianxiangyang@163.com

Lijing Xu*

College of Art and Design,
Zhengzhou University of Industrial Technology,
Xinzheng 451100, China
Email: xulijing@zzuit.edu.cn
*Corresponding author

Shijun Wang

School of Physical Education,
Zhengzhou Technical College,
Zhengzhou 450000, China
Email: wangshijun20080808@163.com

Abstract: Gymnastics competition's conventional video analysis approach mostly depends on shallow image processing technology, which is challenging to sufficiently capture the subtleties of participants' motions and semantic information in the competition. This work suggests a semantic analysis approach of gymnastics competition film based on domain knowledge and depth features to compensate this deficit. First, one builds a knowledge base based on action classification in addition to the domain knowledge of gymnastics competition. Second, the temporal and spatial traits of video and the dynamic performance of athletes are extracted using convolutional neural network (CNN) in combination with long-term and short-term memory network (LSTM). In the analysis of gymnastics competition video, the experimental findings reveal that this approach produces better recognition accuracy and analytical efficiency than the conventional one.

Keywords: physical gymnastics competition; video analysis; deep learning; CNN; LSTM

Reference to this paper should be made as follows: Cai, X., Xu, L. and Wang, S. (2025) 'Enhancement of accuracy and analytical efficiency of gymnastics video using convolutional neural network and semantic analysis approach', *Int. J. Information and Communication Technology*, Vol. 26, No. 30, pp.24–42.

Biographical notes: Xiangyang Cai received his Master's degree from Xinyang Normal University in 2024. He is currently an Associate Professor at Hebi Institute of Engineering and Technology, Henan Polytechnic University. His research interests include intelligent sports engineering, physical training, and sports physiology.

Lijing Xu received her Master's degree from Henan University in 2012. She is currently an Associate Professor at Zhengzhou University of Industrial Technology. Her research interests include product design, intelligent sports engineering, and sports marketing.

Shijun Wang received his Master's degree from Henan University in 2009. He is currently an Associate Professor at Zhengzhou Technical College. His research interests include intelligent sports engineering, physical training, and sports physiology.

1 Introduction

With the rapid development of artificial intelligence and deep learning technology, the application of video analysis in the field of sports has gradually become a research hotspot. The core task of sports video analysis is to extract valuable information from video data for evaluating athlete performance, match results, and providing auxiliary decision-making for referees (Brown and Cox, 2009). Physical manipulation is a comprehensive sport that highly relies on skills and performance, requiring precise evaluation of the quality, coordination, and completion of each movement. Due to the wide variety of movements involved in gymnastics competitions, many of which are highly complex and dynamic, traditional manual scoring methods are not only time-consuming and laborious, but also susceptible to subjective factors, resulting in poor consistency and accuracy of scoring. Therefore, how to use computer vision and deep learning technology to automatically analyze and score gymnastics competition videos has become an important and challenging research topic (Chen, 2024).

Particularly in terms of action identification and semantic understanding, the video study of gymnastics events has several technical difficulties. First of all, gymnastics entail not only the synchronisation of several body parts but also difficult motions including rotations, jumps, stretches, is The length and variety of these motions differ, hence the analysis technique must be able to faithfully record the spatiotemporal aspects in the video (Reily et al., 2017). Second, gymnastics contests include rigorous grading rules that take movement complexity and artistry into account in addition to precision of the motions. This makes it impossible for basic movement recognition techniques to totally satisfy the needs. Video analysis is further challenging because of often complicated backgrounds, lighting changes, and simultaneous performance of several athletes in competitive recordings.

The video analysis research of gymnastics competitions involves multiple disciplinary fields. Computer vision, deep learning, sports science, and artificial intelligence are among the several fields of study encompassing video analysis of gymnastics events. Particularly in the domains of action recognition and video semantic analysis, this branch of research has made some significant developments. Works pertaining to this research are included below.

Aimed at character identification and classification from videos, video action recognition is a classic challenge in computer vision. Early action detection techniques performed well in static scenes but had limited efficacy in complex video sequences. They depended on manually extracting features including histogram of oriented gradients (HOG) (Kim et al., 2022), histogram of optical flow (HOF) (Feichtenhofer et al., 2022), local binary pattern (LBP) (Ojala et al., 1996). Action recognition performance has been much enhanced in recent years by the extensive use of deep learning techniques – especially convolutional neural networks (CNN) in image and video processing. By combining spatial and temporal data, Tong et al. (2022) two stream networks set the stage for video action recognition.

Video content analysis has extensively benefited from deep learning, particularly with regard to CNN and RNN. Based on 3D CNNs, Shoaib et al. (2023) introduced a video action recognition system that effectively caught the spatial and temporal elements of videos. Furthermore, offering fresh approaches for video analysis are the benefits of LSTM in processing time series data. Good results were obtained by Gan et al. (2022) modelling action sequences in films using LSTM.

Video analysis of gymnastics events receives quite little attention in studies. Most current studies mostly concentrate on athlete movement detection, scoring system automation, and other elements. For instance, Rangasamy et al. (2020) suggested an athlete action analysis technique including deep learning for autonomous video scoring in sporting events. By collecting important points and movement characteristics of athletes, this technique raises the automation degree of scoring. But conventional motion detection techniques are seriously challenged by the intricacy and great difficulty of physical gymnastics events.

In sports video analysis, the use of domain knowledge helps to increase model accuracy. In order to direct the training of deep models, Du et al. (2021) suggested an action recognition framework combining domain knowledge with use of action templates established by sector experts. To better grasp the technical motions in gymnastics contests, Pu and Shamir (2024) also suggested a technique to enhance athlete performance analysis by means of an expert knowledge basis.

Deep learning models have the benefit of automatically extracting features from raw data, but in complicated sports video analysis applications classical features are still rather crucial. Zhao (2023) presented a hybrid approach for motion analysis combining deep features with conventional manual elements like optical flow and edge characteristics, thereby greatly enhancing the recognition accuracy of athlete movements. Similar techniques can efficiently capture the subtleties of challenging motions in gymnastics events.

Many sports, including athletics (Woellik et al., 2014), basketball (Yao, 2021), football (Liu et al., 2020), and so on, have automated scoring systems employed in them. Still, gymnastics lacks specific automated grading systems because of its great complexity of motions and subjective performance. By using sentiment analysis methods, Jekauc et al. (2024) investigated how athlete emotions affected game performance, therefore offering a fresh angle on sports video analysis.

This work suggests a semantic analysis approach combining domain knowledge and deep features to handle these difficulties for gymnastics competition footage. This approach initially aggregates domain knowledge of gymnastics events to build a knowledge base based on action classification, therefore enabling the system to identify and recognise particular actions and approaches in contests. Second, a combination of

CNN and LSTM is used to extract spatiotemporal features from videos, and deep learning models help to analyze action sequences so strengthening semantic understanding. By means of this approach, this study may not only enhance the accuracy of video analysis in gymnastics contests but also replicate the concept of manual scoring to a wider extent, so attaining more automated and intelligent competition evaluation.

By use of creative deep learning models, this work aims to enhance the accuracy and automation of video analysis in gymnastics, so fostering the development of intelligent sports systems, and so laying the groundwork for future intelligent scoring and evaluation systems.

2 Relevant theoretical foundations

Semantic analysis of gymnastics competition films revolves mostly on how to extract useful elements from video data using suitable models and assess athletes' performance depending on domain knowledge. Several theoretical approaches and modelling strategies – including video feature extraction, deep learning models, temporal modelling, and the merging of domain knowledge – must thus be thoroughly utilised in this process. This chapter will discuss pertinent theoretical analysis including deep learning, CNN, LSTM, and domain knowledge fusion in action recognition, thereby enabling a thorough understanding of the suggested analysis approach.

2.1 Theoretical framework of anxiety

A fundamental component of video analysis, video feature extraction directly influences the success of later analysis activities such event detection, action recognition, is As a collection of temporal images, video feature extraction not only requires capturing the spatial information of each frame, but also extracting the temporal information between frames (Ali et al., 2019). In order to achieve efficient feature extraction, researchers have proposed many methods. The main theories and methods of video feature extraction will be analyzed in detail below.

A video is composed of a series of frame images, each of which contains rich spatial features that can reflect the texture, colour, shape, and other information of the image. In video analysis, spatial feature extraction is typically achieved using CNN. CNN can automatically extract hierarchical features from images through multi-layer convolution and pooling operations. Assuming the input image is *I*, after convolutional layer processing, the calculation of feature map F can be expressed as:

$$F = Conv(I, W) + b \tag{1}$$

where, *W* is the convolution kernel, *b* is the bias term, and *Con*(*I*, *W*) represents the convolution operation on the input image *I* to obtain the feature map *F*. The advantage of CNN is that it can automatically learn the spatial features of images, avoiding the complexity of manually designing features.

In addition to spatial features, temporal features in videos are also very important. The temporal characteristics reflect the motion trajectory of objects or characters in the video, which can reveal the dynamic changes of actions. One of the commonly used feature extraction methods for capturing temporal information between video frames is the optical flow method.

The optical flow method estimates the direction and velocity of motion by analyzing the displacement of pixels between video frames. The optical flow field *v* = (*u*, *v*) describes the motion of each pixel in the *x* and *y* directions in the image. The basic calculation equation for optical flow is:

$$I_x u + I_y v + I_t = 0 \tag{2}$$

where, *I*, *u* and *I*, *v* are the gradients of the image in the *x* and *y* directions, *I* is the gradient of the image over time, and *u* and *v* represent the velocity components of a point in the image in the horizontal and vertical directions, respectively. We can get the motion information of every pixel in every frame of the image by computing the optical flow, therefore enabling us to subsequently characterise the motion trajectory of objects or persons in the movie.

Motion analysis and action identification make frequent use of optical flow techniques, which may expose object motion patterns and velocity information in videos, therefore faithfully capturing the temporal aspects of videos.

In order to simultaneously consider both spatial and temporal information in videos, spatiotemporal feature extraction methods are widely used. There are two main methods for spatiotemporal feature fusion: one is to extract spatiotemporal features of videos through three-dimensional convolutional neural networks (3D-CNN), and the other is to first use CNN to extract spatial features, and then use recurrent neural networks (RNN) or LSTM to model temporal dependencies.

3D-CNN is an extension of traditional 2D CNNs, which adds temporal convolution operations on the basis of spatial convolution (Alakwaa et al., 2017). By convolving in three-dimensional space (height, width, time), 3D-CNN can capture both spatial and temporal features simultaneously. Assuming that the input of a video frame is a three-dimensional tensor V (with dimensions of *C* × *H* × *W* × *T*, where *C* is the number of colour channels, *H* and *W* are the height and width of the frame, and *T* is the number of frames in the video), after 3D convolution operation, the obtained spatiotemporal feature F can be expressed as:

$$F = Conv3D\left(V, W_{3D}\right) + b \tag{3}$$

where, W_{3D} is a three-dimensional convolution kernel, and the formula represents performing three-dimensional convolution on the input video frame to obtain the spatiotemporal feature *F*.

Through 3D convolution, 3D-CNN is able to extract comprehensive features of videos from both spatial and temporal perspectives, making it highly suitable for action recognition tasks in videos.

Another commonly used spatiotemporal feature extraction method is to combine CNN and LSTM models. The basic process of this method is as follows: first, extract the spatial features of each frame of the image through CNN; then, the extracted feature sequence is input into LSTM, which models the temporal dependencies of the features through a time gating mechanism to capture the temporal dependencies in the video. The equation is as follows:

$$h_t = LSTM\left(F_t, h_{t-1}\right) \tag{4}$$

where, F_t is the spatial feature extracted by CNN at time t, h_{t-1} is the hidden state of the previous time, and h is the hidden state of the current time. LSTM effectively captures temporal information through its inherent gating mechanism. This combination method can fully leverage the advantages of CNN in spatial feature extraction, while processing the time series information of videos through LSTM, especially suitable for tasks such as action recognition and event detection.

2.2 Convolutional neural network

Currently among the most successful image processing models, CNN is also frequently employed in video analysis to extract spatial information of every frame. Whereas the pooling layer reduces dimensionality and feature selection on the convolutional features to minimise overfitting. CNN's fundamental architecture consists in this:

$$F^l = RELU\left(Con\left(F^{l-1}, W_l\right) + b_l\right) \tag{5}$$

where, F^l is the output feature of the l^{th} layer, W_l and b_l are the convolution kernel and bias term of the l^{th} layer, ReLU is the activation function, and $Con(F^{l-1}, W^l)$ represents the convolution operation on the feature of the previous layer.

In video analysis, CNN is mainly used to extract spatial features of each frame, while temporal information between consecutive frames needs to be captured through other methods.

2.3 Long short term memory network

LSTM is a special type of RNN that can effectively process time series data. The main advantage of LSTM is that it overcomes the gradient vanishing problem that traditional RNNs encounter during long sequence training by introducing 'memory units' and 'gating mechanisms'. The basic calculation formula of LSTM is as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{6}$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{7}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{8}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{9}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{10}$$

$$h_t = o_t \cdot \tanh\left(C_t\right) \tag{11}$$

where, ft is the forget gate, i is the input gate, o_t is the output gate, $\tilde{C}_t$ is the candidate memory unit, C_t is the current memory unit, h_t is the hidden state at the current time, x_t is the current input, W is the weight matrix, and b is the bias term. Through the gating mechanism, LSTM can maintain important information in long sequences and forget irrelevant information, which gives LSTM a great advantage in processing video frame sequences.

LSTM can be applied in the video analysis of gymnastics events to model the temporal dependencies of movements, gather the temporal dynamic information of every movement, and so raise recognition accuracy.

3 Asemantic analysis method for shape gymnastics competition videos that integrates domain knowledge and deep features

This work presents a semantic analysis approach combining domain knowledge and deep features for gymnastics competition footage. The sport of physical manipulation demands great accuracy and difficulty in movement performance; so, evaluation of this sport depends mostly on standardising, coordinating, and fluency of its motions. We utilised deep learning methods and domain knowledge to build a model based on CNN and LSTM, so improving the model's capacity to recognise and interpret distinctive motions in gymnastics, so better achieving semantic analysis of gymnastics competition films. The methodological framework of this work, the building of domain knowledge, the design and implementation of deep learning models, and the approach for model fusion will be thoroughly explained in this chapter.

3.1 Domain knowledge of gymnastics competitions

As a competitive sport combining physical fitness, artistry, and expressiveness, rhythmic gymnastics (RG) not only calls for athletes to have great degrees of physical flexibility, strength, balance, and coordination, but also demands aesthetic performance when completing movements Gymnastics competition domain knowledge mostly consists on standardised movement definitions, scoring systems, technical criteria, and knowledge of other sports. By means of methodical domain knowledge, competition scoring is more objective and officials may more fairly assess athletes depending on action criteria. Covering action classification, scoring criteria, technological requirements, building of a knowledge base, this part will offer a thorough introduction to the domain knowledge required in gymnastics competitions.

Important parts of semantic analysis in competitions are the classification and standardising of gymnastics moves. Gymnastics' movements can be loosely separated into the following groups:

3.1.1 Basic actions

The basic movements are the foundation of all physical gymnastics performances, including the core movements and postures of athletes. Common basic movements include extension, such as stretching the arms, legs, or back, to showcase elegance and strength. Flexion, the movement of bending the legs or hands, is commonly used to demonstrate flexibility and control. There is also a turn, which is used to represent movements of balance and rotation control. Common examples include 360 degree rotation, jumping rotation, is These basic movements can form complex rhythmic routines through different combinations.

3.1.2 Instrument action

Shape gymnastics usually uses various instruments such as flower balls, ribbons, rings, sticks, and ropes. The movements can be divided into ball movements (such as flower balls and balls), such as throwing and receiving balls, rolling balls, and swinging balls. This type of action emphasises coordination and ball sense. Rope movements, including swinging, rotating, swinging, is, emphasise rhythm and fluency. Circular actions, such as rolling, throwing, jumping, and piercing circles, emphasise precise control and force transmission. And stick movements, including throwing and turning sticks, emphasise the combination of force and technique.

Each type of equipment has its own unique movement standards, involving the coordination between the equipment and the body, the control of the equipment, and the creative expression of movements.

3.1.3 Combination action

The movements of gymnastics are not just basic movements or instrument operations, but more importantly, the combination of movements. The combination of actions has action connections, such as the connection between different actions, which needs to demonstrate fluency and coordination. For example, transitioning from a stretching motion to a spinning or jumping motion. There are also complex physical expressions. Usually reflecting athletes’ inventiveness, bodily coordination, and musical sense, these moves also represent.

3.2 Scoring criteria for gymnastics competitions

Gymnastics events have a technical and artistic score system, both of which are judged according on distinct factors.

Mostly evaluating athletes’ precision, dexterity, and execution standards in executing motions, the technical score Among the scoring criteria are the action’s degree of difficulty: Higher scores usually go for actions with more effort, including jumping, somersaults, throwing and receiving. The difficulty level is usually determined based on the complexity of the action and the standards of execution. Action accuracy (Precision): Whether the movements are executed accurately according to technical requirements, including body posture, coordination between athletes and equipment, is And motion control: Can athletes maintain stability and elegance during the execution of movements, such as stability during rotations, jumps, and instrument control.

The art score mainly evaluates the performance ability, creativity, and overall aesthetic appeal of athletes. The scoring factors include fluency: is the connection between movements natural, and can athletes smoothly transition to the next movement without obvious interruptions. Expression: including athletes’ facial expressions, emotional expressions, and synchronisation with music. Athletes need to convey the emotions of their movements through body language. And the coordination between music and action (Musicality): the coordination between movements and background music is crucial, and athletes need to demonstrate their understanding of music through their movements and sense of rhythm.

In addition to bonus points, the competition also has deduction points. For example, exceeding the prescribed time (usually 2 minutes and 30 seconds) will result in deduction of points. Mistakes such as falling and dropping equipment can also affect the rating. Failure to complete the technique or technical action according to the action requirements will result in deduction of points.

Each gymnastics movement has detailed technical requirements, and athletes need to complete it within the prescribed movement framework. For stretching, bending, and other movements, there is usually a required range of motion to ensure elegance and strength. Athletes' body posture should meet the standards of their movements, such as maintaining appropriate angles of their limbs when performing rotations. Many movements, especially jumping and spinning, require athletes to maintain good balance and control to showcase their technical abilities. Figure 1 shows the relevant videos of the gymnastics competition.

Figure 1 The gymnastics competition (see online version for colours)

3.3 *Construction of knowledge base in the field of gymnastics*

In order to effectively conduct semantic analysis of gymnastics competition videos, it is necessary to construct a knowledge base that includes information such as action classification, scoring criteria, and technical requirements. Along with thorough descriptions, technical criteria, and difficulty ratings for every movement, this knowledge base includes classification information for all basic movements, instrument motions, and combination movements of gymnastics.

Detailed records of the scoring criteria and technical requirements for each action and competition event. Including the rating range, difficulty level, execution criteria, is for each action. In addition, it also includes possible deductions during scoring.

In order to better integrate deep learning models for analysis, an action knowledge graph can be constructed, which includes the relationships between various actions, the technical requirements of actions, and their corresponding scoring criteria. Through graph structure, semantic reasoning can be effectively carried out to improve the model's understanding ability of actions.

The application of domain knowledge in gymnastics competitions in deep learning models is crucial. In this study, we combine domain knowledge with deep learning techniques. In action recognition, utilising action classification and scoring criteria from domain knowledge bases can help deep learning models more accurately identify different actions in videos and label each action. To assist with scoring, the model can use scoring criteria based on domain knowledge to assess the performance of athletes. Through the standards in the knowledge base, the model can continuously adjust its recognition strategy for movements to more accurately match the standards of gymnastics competitions.

3.4 Design of deep learning model framework

The main goal of the semantic analysis of gymnastics competition movies is to extract efficient feature information from the source video and apply it to classify and assess movements. Simple spatial feature extraction cannot satisfy the modelling needs for temporal information. We present a deep learning model comprising CNN and LSTM to handle this problem. While LSTM models temporal data to capture temporal correlations between video frames, CNN is applied to extract spatial information from every frame of video. This model may efficiently manage action identification and semantic analysis chores in gymnastics competition films by means of this combination.

Two key components form our model framework: temporal modelling module and spatial feature extraction module. The spatial feature extraction module specifically employs CNN to process every frame of video and extract important features from the image; the temporal modelling module analyzes the time series of these spatial features and records the dependency links between video frames. To finish the action recognition and scoring process, LSTM’s output is finally translated to the action categorisation space via a fully linked layer.

The movements in gymnastics clearly show time series properties. We used LSTM networks to grasp the temporal relationships of activities in films. By means of a ‘gating’ mechanism, LSTM is a unique kind of RNN that solves the gradient vanishing issue faced by conventional RNNs in modelling lengthy time series. Particularly in the analysis of difficult sequences of gymnastics movements, LSTM can filter out extraneous data and retain vital information over a lengthy time range.

The basic structure of LSTM includes three main gates: forget gate, input gate, and output gate. They determine which information should be remembered, forgotten, and outputted at the current moment by controlling the flow of information.

Our model combines CNN with LSTM, first using CNN to extract spatial features of each frame, and then feeding these spatial features as input into the LSTM model for temporal modelling. The specific fusion steps are as follows:

- Spatial feature extraction: firstly, each frame of the video is input into the CNN, which extracts the spatial features of that frame through multi-layer convolution and pooling. The video features of each frame are represented as a vector, whose dimension is determined by the output of the convolutional layer.

The extracted spatial features are used as inputs for LSTM, and the LSTM network models the temporal information through its gating mechanism to capture the temporal dependencies between video frames. The output of LSTM is the hidden state at the current time, which represents the feature representation after time evolution. Finally, the output of LSTM is fed into a fully connected layer for classification. By using the Softmax activation function, the model outputs the probability distribution of the action category:

$$P\left(y_t \mid X\right) = Softmax\left(W_h \cdot h_t + b_h\right) \tag{12}$$

where, W_h and b_h are trainable parameters of the fully connected layer, y_t is the action label corresponding to each time step in the video sequence, and $P(y_t|X)$ is the predicted probability of the action category given input X.

3.5 Model training and optimisation

This section will elaborate on the training and optimisation process of a deep learning model that combines CNN and LSTM. We will introduce the setting of training parameters, the training steps of the model, optimisation methods, training convergence graphs, is, to ensure that the model can effectively learn and achieve good performance in semantic analysis tasks of gymnastics competition videos.

Reasonable parameter settings are crucial in the training process of deep learning models. The following are the main training parameters and their selection criteria for our model.

Learning rate is a hyperparameter that controls the step size of model parameter updates. If the learning rate is set too high, it may cause the model to fail to converge or even diverge; If the setting is too small, the convergence speed of the model will be too slow. We have adopted an adaptive learning rate optimisation algorithm (such as Adam), so the initial learning rate is set to 10^{-3} and dynamically adjusted during the training process based on the training effect.

The batch size determines the number of data samples used each time the parameters are updated. The choice of batch size directly affects the memory usage and training speed during the training process. Generally speaking, larger batches can estimate gradients more stably, but may consume more memory. We set the batch size to 32, which means that each iteration uses 32 frames of video data for training.

The number of training rounds determines the number of times the entire training dataset is used during the training process. To ensure that the model fully learns the features in the data and avoids overfitting, we set the number of training rounds to 50. After each round of training, the validation set will be evaluated to monitor the performance changes of the model.

For multi classification problems, we use the cross entropy loss function to calculate the difference between the model output and the true labels. The cross entropy loss function is defined as:

$$L = -\sum_{t}\sum_{c=1}^{C} y_{t,c}\log\left(P\left(y_{t,c} \mid X\right)\right) \tag{13}$$

where, c is the number of action categories, $y_{t,c}$ is the true label of category c corresponding to time step t in the video sequence, and $(P(y_{t,c}|X))$ is the probability predicted by the model.

We use the Adam optimiser, which is an adaptive learning rate optimisation algorithm that can adjust based on the first and second moments of the gradient, typically resulting in faster convergence speed and better generalisation performance. The update rules for Adam optimiser are as follows:

Usually producing faster convergence speed and higher generalisation performance, the Adam optimiser – an adaptive learning rate optimisation method – can be adjusted depending on the first and second moments of the gradient. Adam's optimiser has the following update rules:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{14}$$

where, θ_t is the model parameter, η is the learning rate, $\hat{m}$ and $\hat{v}$ are estimates of the first and second moments, and ε is a small constant added to avoid zero division operations.

Before training the model, the original video data needs to be pre-processed first. Video frames need to be divided into fixed sized image blocks and then normalised so that the pixel values of each frame are between 0 and 1. In addition, to improve training effectiveness, we use data augmentation techniques, including:

1. Crop. Randomly crop different regions from the original video frames to increase data diversity.
2. Rotate and flip. Randomly rotate or horizontally flip video frames to enhance the robustness of the model.
3. Colour jitter. Fine tune the brightness and contrast of video frames to simulate their performance under different lighting conditions.

The processed video frame data will be loaded into the training in chronological order to ensure that the model can capture the dependencies of the time series.

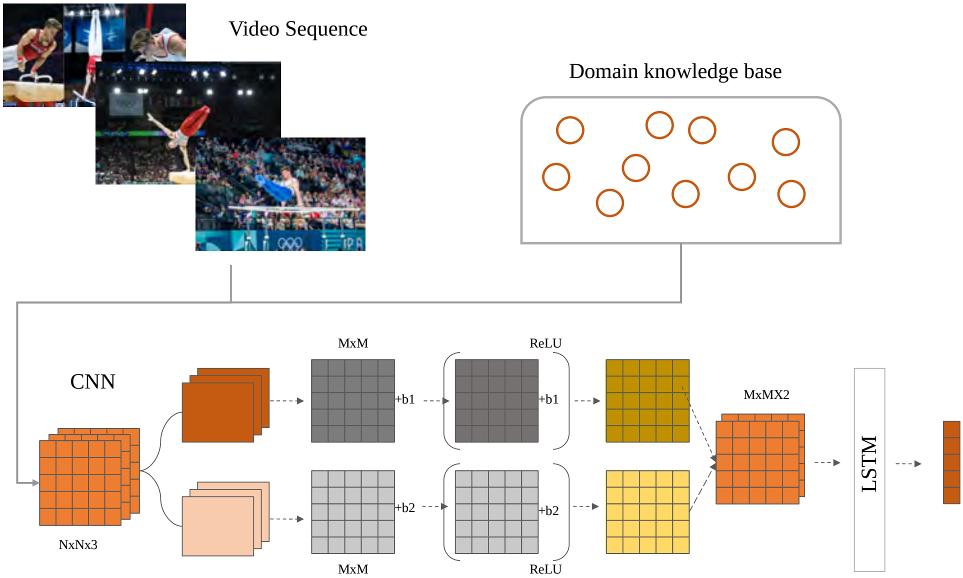
Before training begins, the parameters of the model need to be initialised. We use Xavier initialisation (also known as Glorot initialisation) to initialise the weights of convolutional and fully connected layers, with the bias term initialised to 0. Xavier initialisation can maintain a balance of gradient sizes in forward and backward propagation, reducing the phenomenon of gradient vanishing or exploding.

In each training iteration, the video frame data is first subjected to spatial feature extraction through a CNN layer. The spatial feature representation of each frame is fed into the LSTM layer for temporal modelling. The LSTM models the temporal dependencies through its gating mechanism and ultimately outputs the hidden state. Then, the model maps the hidden states to the category space through a fully connected layer and outputs the predicted probabilities for each category.

Calculate the difference between the predicted category and the true label using the cross-entropy loss function, and then calculate the gradient using the backpropagation algorithm. The gradient propagates back to each layer in the network through the chain rule, updating the weights and bias terms. During backpropagation, gradients propagate through multiple time steps of LSTM, and special attention should be paid to the problem of gradient vanishing or exploding. The Adam optimiser will adjust the parameters of the model and update the weights based on the calculated gradient. Update the model parameters through the Adam optimiser to ensure that each parameter is adjusted towards minimising the loss function.

We assess the validation set following every training cycle and compute the current model's performance indicators – accuracy, precision, recall, is Should the validation set performance not show appreciable improvement over numerous rounds, we will halt instruction early to prevent overfitting. Early Stopping is the name given to this approach, which helps the model to have better generalising capacity. Figure 2 shows the overall framework of the model.

Figure 2 The overall framework of the model (see online version for colours)



4 Experiment

This chapter will do a comparison analysis using a set of experiments to confirm the efficiency of the semantic analysis approach for shape gymnastics competition films suggested in this work, which combines domain knowledge and deep features. We will evaluate the proposed deep learning model with conventional action recognition techniques including HOG + *SVM (Dadi and Pillutla, 2016), investigate their performance variations in action identification and semantic analysis, and investigate the model's performance in challenging circumstances. Analyzing the experimental findings helps one confirm the benefits of the suggested approach in raising the accuracy, precision, and recall rate of video analysis in gymnastics events.

4.1 Dataset

This study made use of a video dataset of gymnastics events, comprising whole movies of several contests. Every competition lasts between five and ten minutes on video. With action labels matching each video frame, the dataset comprises athletes of various age groups and sexes, and the video material contains standard gymnastics actions including forward bending, balance movements, rotations, is All movies go through a homogeneous preparation procedure comprising video frame cropping, normalisation, and data augmentation to guarantee the fairness of the experiment and the dependability of the outcomes.

4.2 Experimental setup

All experiments were conducted on a computer platform with CUDA support, with NVIDIA GeForce RTX 3090 GPU, Intel Core i9-10900K CPU, 64GB of memory, and PyTorch 1.10.0 deep learning framework.

To verify the effectiveness of the proposed method, we selected several traditional action recognition methods for comparison.

- HOG + SVM: this method extracts the HOG features of video frames and uses support vector machines (SVM) for classification. HOG features can capture local gradient information in images, while SVM is used to map these features to the action category space.
- CNN (Chua 1997): using CNN to extract spatial features of video frames and employing fully connected layers for action classification. This method does not consider the dependency relationship of time series, and only identifies by processing a single frame image.

4.3 Experimental result

Firstly, we evaluated the action recognition accuracy of the three methods. Accuracy is a measure of the proportion of correct predictions made by a classifier among all predictions, and its calculation formula is as follows:

$$Accuracy = \frac{correct}{total} \tag{15}$$

In addition to accuracy, we also evaluate the classification performance of the model through precision and recall. Precision and recall are defined as

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{16}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{17}$$

The F1 value is the harmonic mean of accuracy and recall, commonly used to evaluate classification performance in imbalanced datasets. The calculation formula for F1 value is as follows:

$$F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \tag{18}$$

From Figure 3, it can be seen that the traditional HOG + SVM method achieves an accuracy of 82.4%, while the simple CNN method achieves an accuracy of 88.2%, which is an improvement over the traditional method. The model proposed in this study, which combines CNN and LSTM, has an accuracy of 94.6%, which is 6.4 percentage points higher than CNN. This indicates that the deep learning model incorporating temporal modelling has significant advantages in the recognition of gymnastics movements.

Figure 3 Comparative experimental results (see online version for colours)

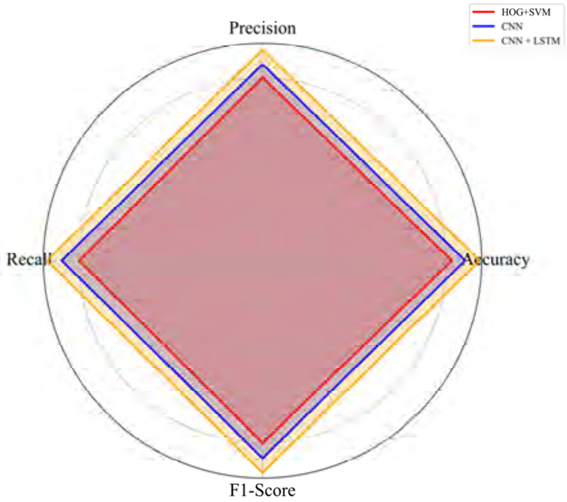
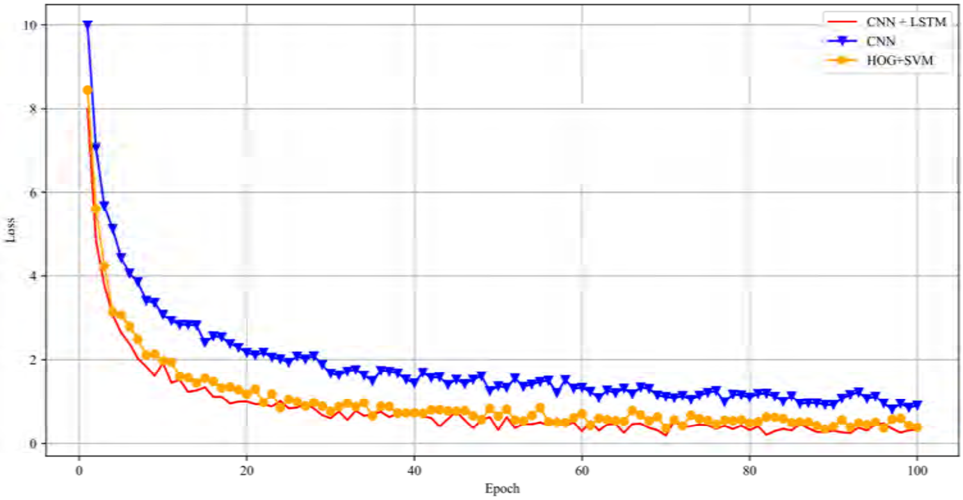


Figure 4 Convergence graph (see online version for colours)



The model proposed in this study, which integrates CNN and LSTM, outperforms other methods significantly in terms of accuracy (92.5%) and recall rate (93.8%), especially in terms of recall rate. This means that our method can better identify actual actions and reduce missed detections.

Through the comparison of F1 values, we can see that the method proposed in this study achieved an F1 value of 93.1%, far exceeding traditional methods and CNN methods. This further proves the superiority of the model in handling the task of recognising gymnastics movements.

The convergence of the training process can be evaluated by monitoring changes in the loss function and accuracy. The convergence diagram is shown in Figure 4.

During each round of training, the loss function (such as cross entropy loss) continuously decreases, indicating that the model is gradually learning patterns from the data. Ideally, the loss value should decrease rapidly in the early stages of training and then stabilise. The following is a typical loss function convergence graph:

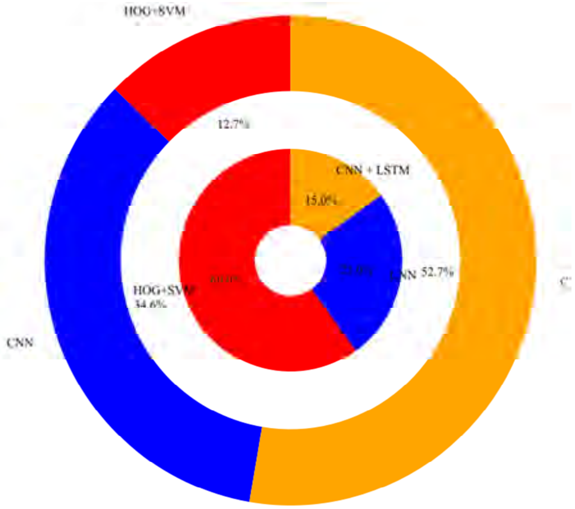
From the graph, it can be seen that the loss function decreases rapidly in the first few rounds and then tends to stabilise, indicating that the model gradually converges.

In addition to the loss function, we can also observe the accuracy changes on the training and validation sets. In the early stages of training, the model may perform well on the training set, but the accuracy on the validation set may be lower. As the training progresses, the accuracy of the validation set gradually improves and eventually stabilises. The following is a typical accuracy convergence graph:

From the graph, it can be seen that the accuracy of both the training and validation sets gradually improves and tends to stabilise after a certain number of rounds, indicating that the model has achieved good performance on both the training and validation sets.

In addition to model performance, training and inference speed are also important indicators for evaluating the effectiveness of a model. Figure 5 shows the training and inference time for different methods. From the table, it can be seen that although the method proposed in this study has increased the training time by 18.7 hours, the LSTM layer can effectively process temporal information, resulting in a significant reduction in inference time. The inference time is 0.03 seconds per frame, which is much better than the CNN method and traditional methods.

Figure 5 Training and reasoning speed (see online version for colours)



The experimental results show that the proposed CNN+LSTM method outperforms the HOG + SVM and CNN methods in all evaluation metrics, especially in terms of accuracy, recall, and F1 score. The combination of LSTM's temporal modelling capability enables this method to fully capture the temporal dependencies of movements, improving the understanding and recognition of movements in gymnastics competitions.

The model of this study performs well in complex scenes and can maintain high recognition accuracy under the interference of background noise and occlusion. Compared to traditional HOG + SVM methods, CNN methods and traditional deep learning models, the fusion of LSTM in temporal modelling can effectively reduce misidentification caused by factors such as inter action similarity, background changes, and occlusion.

5 Conclusions

This study proposes an analysis method that integrates domain knowledge and deep features to address the complexity and accuracy of action recognition in semantic analysis of gymnastics competition videos. By introducing domain knowledge of gymnastics competitions, constructing a knowledge base of standard movements and scoring criteria, and combining the spatial feature extraction and temporal modelling capabilities of deep learning models (CNN + LSTM), efficient semantic analysis of gymnastics competition videos has been achieved. This study constructed a knowledge base for classifying movements in gymnastics competitions, embedding the standard movement features and scoring criteria of gymnastics into the model to provide prior knowledge for deep learning models, thereby enhancing the model's understanding of the unique movements in gymnastics. Extracting spatial features of video frames through CNN and modelling the temporal dependence of actions using LSTM, enabling the model to capture dynamic changes in actions. Through experiments on the video dataset of gymnastics competitions, the superior performance of the proposed method in terms of accuracy, recall, F1 score, is in action recognition has been verified, especially in complex scenes such as background noise and action occlusion, showing good robustness. The experimental results show that the method proposed in this study can effectively improve the performance of semantic analysis in gymnastics competition videos, providing valuable references for video semantic analysis tasks in other fields.

Although this study has achieved certain results in semantic analysis of gymnastics competition videos, there are still some aspects that can be further optimised and expanded.

Future research could consider combining visual information with other modalities such as audio and sensor data to further enhance the model's understanding of complex game scenes.

The current knowledge base is constructed through manual organisation, and in the future, automated knowledge extraction technology can be used to expand the scale and coverage of the knowledge base, and improve the model's generalisation ability to other sports projects.

Although the model has met the basic requirements in terms of inference time, higher real-time performance may be needed in actual competitions. In the future, efforts can be made to combine lightweight models such as MobileNet to further reduce computational costs.

Acknowledgements

This work is supported by the research achievement of the Science and Technology Research Project of Henan Province (No. 242102321136).

Declarations

All authors declare that they have no conflicts of interest.

References

Alakwaa, W., Nassef, M. and Badr, A. (2017) 'Lung cancer detection and classification with 3D convolutional neural network (3D-CNN)', International *Journal of Advanced Computer Science and Applications*, Vol. 8, No. 8.

Ali, H., Sharif, M., Yasmin, M., Rehmani, M.H. and Riaz, F. (2019) 'A survey of feature extraction and fusion of deep learning for detection of abnormalities in video endoscopy of gastrointestinal-tract', *Artificial Intelligence Review*, Vol. 53, No. 4, pp.2635–2707.

Brown, D. and Cox, A.J. (2009) 'Innovative uses of video analysis', *The Physics Teacher*, Vol. 47, No. 3, pp.145–150.

Chen, Y. (2024) '3D Convolutional neural networks based movement evaluation system for gymnasts in computer vision applications', *Journal of Electrical Systems*, Vol. 20, No. 3s, pp.880–898.

Chua, L.O. (1997) 'CNN: a vision of complexity', *International Journal of Bifurcation and Chaos*, Vol. 07, No. 10, pp.2219–2425.

Dadi, H.S. and Mohan Pillutla, G.K. (2016) 'Improved face recognition rate using HOG features and SVM Classifier', *IOSR Journal of Electronics and Communication Engineering*, Vol. 11, No.4, pp.34–44.

Du, Y., Zhao, Q. and Lu, X. (2021) 'Semantic extraction of basketball game video combining domain knowledge and in-depth features', *Scientific Programming*, Vol. 2021, pp.1–12.

Feichtenhofer, C., Li, Y. and He, K. (2022) 'Masked autoencoders as spatiotemporal learners', *Advances in Neural Information Processing Systems*, Vol. 35, pp.35946–35958.

Gan, Z., Li, L., Li, C., Wang, L., Liu, Z. and Gao, J. (2022) 'Vision-language pre-training: basics, recent advances, and future trends', *Foundations and Trends® in Computer Graphics and Vision*, Vol. 14, Nos. 3–4, pp.163–352.

Jekauc, D., Burkart, D., Fritsch, J., Hesenius, M., Meyer, O., Sarfraz, S. and Stiefelhagen, R. (2024) 'Recognizing affective states from the expressive behavior of tennis players using convolutional neural networks', *Knowledge-Based Systems*, Vol. 295, p.111856.

Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E. and Ganslandt, T. (2022) 'Transfer learning for medical image classification: a literature review', *BMC Medical Imaging*, Vol. 22, No. 1, pp.69.

Liu, G., Luo, Y., Schulte, O. and Kharrat, T. (2020) 'Deep soccer analytics: learning an action-value function for evaluating soccer players', *Data Mining and Knowledge Discovery*, Vol. 34, No. 5, pp.1531–1559.

Ojala, T., Pietikäinen, M. and Harwood, D. (1996) 'A comparative study of texture measures with classification based on featured distributions', *Pattern Recognition*, Vol. 29, No. 1, pp.51–59.

Pu, L. and Shamir, R. (2024) '4CAC: 4-class classifier of metagenome contigs using machine learning and assembly graphs', *Nucleic Acids Research*, Vol. 52, No. 19, p.e94.

Rangasamy, K., As'ari, M.A., Rahmad, N.A., Ghazali, N.F. and Ismail, S. (2020) 'Deep learning in sport video analysis: a review', *Telecommunication Computing Electronics and Control*, Vol. 18, No. 4, p.1926.

Reily, B., Zhang, H. and Hoff, W. (2017) 'Real-time gymnast detection and performance analysis with a portable 3D camera', *Computer Vision and Image Understanding*, Vol. 159, pp.154–163.

Shoaib, M., Shah, B., Ei-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., Gechev, T., Hussain, T. and Ali, F. (2023) 'An advanced deep learning models-based plant disease detection: a review of recent research', *Frontiers in Plant Science*, Vol. 14, pp.1158933–1158933.

Tong, Z., Song, Y., Wang, J. et al. (2022) 'Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training', *Advances in Neural Information Processing Systems*, Vol. 35, pp.10078–10093.

Wang, H., Guo, X., Song, K., Sun, M., Shao, Y., Xue, S., Zhang, H. and Zhang, T. (2025) 'GO-MAE: Self-supervised pre-training via masked autoencoder for OCT image classification of gynecology', *Neural Networks*, Vol. 181, p.106817.

Woellik, H., Mueller, A. and Herriger, J. (2014) 'Permanent RFID timing system in a track and field athletic stadium for training and analysing purposes', *Procedia Engineering*, Vol. 72, pp.202–207.

Yao, P. (2021) '[Retracted] real-time analysis of basketball sports data based on deep learning', Complexity, Vol. 2021, No. 1.

Zhao, L. (2023) 'A hybrid deep learning-based intelligent system for sports action recognition via visual knowledge discovery', *IEEE Access*, Vol. 11, pp.46541–46549.