International Journal of e-Collaboration
Volume 19 • Issue 5

Video-Based Metric Learning Framework for Basketball Skill Assessment

Guangyu Mu, Sanya Aviation and Tourism College, China
Tingting Li, Changchun University Of Chinese Medicine, China*

ABSTRACT

Video-based human action recognition has become one of the research hotspots in the field of computer vision in recent years and has been widely used in the fields of intelligent human-computer interaction and virtual reality. However, most of the current existing methods and public datasets are constructed for human daily activities, and the assessment of basketball skills is still a challenging problem. In order to solve the above issues, in this paper, the authors propose a coarse-to-fine video-based metric learning framework for basketball skills assessment. Specifically, they first use a variety of models to jointly represent the action video, and then the optimal distance metric between videos is learned based on the representation. Finally, based on the distance metric, a query video is coarsely classified to obtain the corresponding label of video action, and then the video is finely classified to judge whether the action is standardized. The experiments on a collected dataset show that the proposed framework can better identify and assess the non-standard actions of basketball.

KEYWORDS

Action Recognition, Artificial Intelligence, Basketball Skill Assessment, Metric Learning, Video Analysis

1. INTRODUCTION

With the development of Internet technology and the popularity of video acquisition equipment, video has become the main carrier of information. At present, the amount of video data is growing explosively, hence, how to analyze and understand the content of video becomes more and more important. As one of the important tasks of video understanding, human action recognition has become the research hotpot of computer vision. Action recognition is to learn the appearance and motion information contained in the video by modeling the spatial-temporal information of the pre segmented time-domain sequence, so as to establish the mapping relationship between the video content and the action category, so that the computer can effectively be competent for the task of video understanding. Action recognition has broad application prospects, such as action analysis, intelligent monitoring, human-computer interaction, video information retrieval and so on. However, most of the current action recognition methods are designed for human daily activities, and the action

DOI: 10.4018/IJeC.316875 *Corresponding Author

recognition for basketball is still a challenging problem. Therefore, it is of great practical significance to introduce video based action recognition methods into the field of basketball.

Basketball originated in the United States in 1891 and officially became an Olympic sport in 1936. It is a competition between the players of the two teams on the playing field, and the final result is the score obtained according to the basket. Basketball is a sport that takes into account individual ability and team cooperation. Individual technical level and team tactical level are very important for the game, which requires coaches to formulate detailed training plans in the process of player training. Up to now, in the field of sports, coaches mainly rely on observing athletes' performance on the spot to develop appropriate training plans for athletes, which has high requirements for coaches' professional quality. Unfortunately, with the increasing enthusiasm for sports in recent years, the number of professional coaches has become difficult to meet the needs of sports enthusiasts. Therefore, it is more and more urgent to study the artificial intelligence algorithms that can replace coaches' work.

The basic actions of basketball include dribbling, shooting, etc. Dribbling is the most basic action in basketball, and shooting is a necessary skill for scoring, and the accuracy of basic actions has a great impact on the score of the game. The result of professional basketball players' shooting score is related to the angle and strength of shooting action, so the practice of shooting action can improve the technical level of players. At present, the training of players in basketball is mainly aimed at basic actions, and the traditional training manner is that the coach observes the shooting action of the players, judges the standardization of the action according to his own experience, and then guides the players. However, because this manner relies on the coach's intuitive sense of judgment and lacks corresponding evaluation, it cannot give players a judgment standard. And during the training, players will analyze the hand feeling of their own shooting action, resulting in some judgment errors, which are inconsistent with the requirements of the standard action. The long-term training of non-standard actions will not only have a certain impact on the shooting results, but also have a certain sports injury to the players themselves (Zhu 2017). Therefore, the research on the assessment algorithm of action specification based on video can help players find the gap with the standard action, and improve the training according to the shortcomings of their own actions, so as to improve the intuitiveness of training and the rapidity of feedback. At the same time, the standardization of training actions can also protect the sports health of basketball players.

In view of the above problems, in this paper, a video-based basketball skill assessment framework is proposed. Specifically, because most of the current existing methods and public datasets are constructed for human daily activities, we first collect a new basketball skill assessment dataset, which contains basketball action recognition part and basketball skill assessment part. Then, we use a variety of models to jointly represent the action videos, and the optimal distance metric between different videos is learned based on the representation. Finally, based on the learned feature representation and distance metric, a query video is coarsely classified to obtain the corresponding label of video action, and then the video is finely classified to judge whether the it is standardized action (Zhu 2022). A large number of experiments demonstrate that our proposed framework can effectively classify the basketball actions and assess the basketball skills.

The following part of this paper is organized as follows: the related works are reviewed in Section 2; the architecture of the proposed metric learning framework is proposed in Section 3; the experiments are provided in Section 4; Section 5 is the conclusion.

2. RELATED WORKS AND ANALYSIS

Human action recognition is an important part of computer vision. At present, there are more and more researches on human action recognition all over the world. With the deepening of research, it has also produced huge economic benefits for the society. Up to now, the existing human action recognition methods can be grouped into two categories: traditional action recognition methods and deep learning based action recognition methods.

The traditional action recognition methods mainly include the spatial-temporal features based methods and the spatial features based methods.

Spatial-temporal features based methods can be further divided into template matching based and statistical model based action recognition methods.

The template matching based action recognition methods mainly use the reference template and the target template for matching, and then use the similarity between them for classification (Zhu 2021). Finally, the category with the highest similarity with the reference template is the correct category. The template matching based methods often use feature vectors such as contour, silhouette and edge for similarity comparison when perform action recognition, and also use Mahalanobis distance when using distance measurement. For instance, Bobick and Davis (Bobick 2001) used the view based method for action recognition, and proposed two concepts: motion energy image (MEI) and motion history image (MHI). Finally, the MEI and MHI invariant matrix features were matched with the template for action recognition. Through analysis, Meng and Pears (Meng 2009) discovered the limitations of MHI time template and proposed motion history histogram (MHH), which can obtain more action information than MHI. And experimental results showed that the new features can effectively make up for the shortcomings of MHI and significantly improve the accuracy of human action recognition. In conclusion, the template matching based methods have the advantages of simplicity, intuitiveness and less computation, but their robustness is poor.

The statistical model-based action recognition methods regard human action as a continuous state sequence, and use the probability transition relationship between states for modeling. Their modeling performances are more powerful than the template matching based methods. For example, Wang et al. (Wang 2014) first studied the recognition method of dynamic behavior by Markov model under depth camera, combined with the character that depth camera has rich information, improved one-dimensional HMM to multi-dimensional, and finally proposed to use the combination of multi-dimensional Discrete Hidden Markov model and multi-dimensional continuous hidden Markov model for recognition, and achieved good results. Park and Aggarwal (Park 2003) proposed a human body recognition method using hierarchical Bayesian network. This method used low-level estimation for parts of the body, high-level estimation for the whole body, and finally used dynamic Bayesian network for classification and recognition. Pavlovie et al. (Pavlovie 2000) proposed a new switching linear dynamic system (SLDS) based on dynamic Bayes, and applied it to graphic motion analysis, and achieved good results.

The spatial features based method is also called the image set based method, that is, video is regarded as an image set composed of unordered frames, and action recognition is only based on spatial features. Its characteristic is to use a certain modeling method to model the video as a whole, and then carry out the subsequent action recognition task based on the modeled feature representation. For example, Wang et al. (Wang 2018) formulated each set as a Gaussian mixture model, and used K-L divergence to measure the similarity between the distributions. Wei et al. (Wei 2019) used prototype subspace to model the video and used collaborative representation to learn the distance metric between different gesture actions.

Deep learning based action recognition methods. With the application of deep learning in the field of computer vision and the availability of large-scale video data sets, the research methods based on deep learning have achieved far better results in the field of action recognition than traditional methods, and become the mainstream research direction in this field. Specifically, Simonyan et al. (Simonyan 2014) proposed a dual flow network composed of spatial flow network and temporal flow network on the basis of 2D CNN, in which spatial flow network is used to model appearance features and temporal flow network is used to model action features. In order to solve the problem that the dual stream network has weak ability to model long-term structures, Ng et al. (Ng 2015) proposed using long short-term memory (LSTM) network to aggregate the bottom output of CNN of video frame sequences. However, the lack of modeling the underlying time information between video frames will cause the loss of timing information. Therefore, Wang et al. (Wang 2016) proposed the temporal segment networks (TSN), and introduced the sparse sampling strategy on the basis of the dual flow network, so that the network has the ability to

extract the global spatial-temporal features, thus effectively solving the problem that the traditional dual flow network lacks the ability to construct a long-time model. Ji et al. (Ji 2013) proposed for the first time to use 3D CNN to extract the spatial-temporal features of video, obtained the gray level, gradient and optical flow channel information between adjacent frames in the video, and generated the final fusion features through convolution and down sampling operations. On this basis, Tran et al. (Tran 2015) proposed C3D method, which used three-dimensional convolution to model spatial-temporal signals, and obtained a more compact feature representation than 2D CNN. After that, Tran et al. (Tran 2017) extended the C3D architecture to the depth residual network, and proposed the Res3D network, which maintains the consistency of network architecture parameters by changing the number of convolution layer filters, thus discussing the impact of the sampling frequency, spatial resolution and convolution type of input frame on the model performance. Kun et al. (Kun 2018) combined TSN with Res3D and proposed temporal convolutional 3D network (T-C3D). On the basis of Res3D, Wang et al. (Wang 2020) deployed an additional jump connection between adjacent residual blocks, which not only fully integrated the temporal and spatial characteristics of shallow and deep layers, but also effectively alleviated the gradient disappearance and over fitting problems that 3D CNN is prone to with the deepening of the network, further improving the performance of Res3D. Diba et al. (Diba 2017) introduced three-dimensional convolution and pooling operations into DenseNet (Huang 2017), and proposed a temporal transition layer (TTL) to build a T3D (temporal 3D convnet) network, which can extract more abundant temporal features. Carreira et al. (Carreira 2017) used 3D convolution and pooling operations to expand the inception network, proposed the inflated 3D convnet (I3D), which adopted greater spatial-temporal resolution for the input of I3D network, and proposed a new method to initialize 3D CNN.

With the continuous development of computer vision and digital image processing technology, the research on image quality assessment has attracted more and more attention. And many image quality assessment methods have been proposed. For instance, by modifying the convolutional neural network, researchers (Wang 2019) can make it suitable for solving different image quality assessment problems. Wang et al. (Wang 2019) proposed an aesthetic image reviewer model NAIR based on CNN and recurrent neural network (RNN), which can not only predict the aesthetic score, but also generate semantic assessment. Different from the traditional aesthetic classification methods, which classify aesthetics into good and bad two classes, the NIMA (Talebi 2018) method proposed by Google predicts the probability distribution of human aesthetic assessment of an image through convolution neural network. The obtained probability distribution map can more accurately understand the centralized trend of user evaluation of an image, and can more accurately guide how many people in the crowd feel good-looking. However, the current assessment methods are designed for a single image, and the video based basketball skill assessment is still a challenging problem.

3. VIDEO-BASED METRIC LEARNING FRAMEWORK FOR BASKETBALL SKILL ASSESSMENT

3.1 Data Acquisition

At present, there are many large datasets based on human daily activities, such as NTU-RGBD, Kinect-skeletons, etc. However, these datasets are not specially collected for basketball action recognition but contain many actions in daily life. And some existing basketball action recognition datasets are non-public, so we cannot directly use these datasets. In addition, many existing datasets contain not only single person behavior, but also human-human interaction behavior and human-object interaction behavior. While our collect dataset only includes the behavior of a single person. Based on the above requirements, the data we collected consists of two parts: basketball action recognition part and basketball skill assessment part. Specifically, for the basketball action recognition data part, the basketball action videos are collected according to six actions: standing dribble, walking dribble, running dribble, jump shot, free throw and layup. Finally, after processing, the basketball

action recognition dataset has a total of 1800 video sequences, each type of action contains 300 sequences, and each sequence contains dozens to hundreds of frames, the lens is unobstructed, and the background is as undisturbed as possible. For the basketball skill assessment part, 1200 of the 1800 video sequences are taken as standard action data, and the remaining 600 sequences are collected according to non-standard actions. That is, for each action category, there are 100 sequences belong to non-standard actions, and the remaining 200 sequences belong to standard actions.

The specific data collection process is as follows: select the popular basketball videos on Tik-Tok, including the highlights of NBA professional players, basketball online teaching videos, etc. All videos containing any of the six types of actions: standing dribble, walking dribble, running dribble, jump shot, free throw and layup, will be added to the pre-selected basketball technical action video set. Finally, 200 video sequences will be selected for each action category as standard actions from all these videos. For non-standard actions, select 100 video sequences for each category from the producer's video titled novice.

In order to realize the annotation of basketball skill assessment tasks, 10 professionals were recruited to score each video sequence using a 5-point system. Specifically, the standard action video sequence is scored 5 points, while other non-standard video sequences are scored respectively, with a maximum of 3 points and a minimum of 1 point. Some sample scoring examples are shown in Table 1.

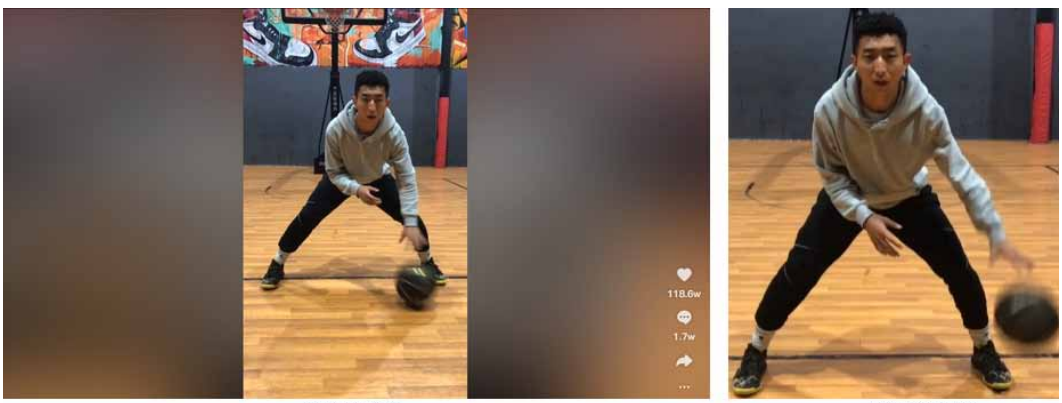
As can be seen from Table 1, since most of the example data are collected from beginners, the overall score is between 1.4-2.5.

Finally, in order to reduce the difficulty of action recognition and reduce the size of input video frames, the YoLoV5 network is used to detect the human body diagram, and the resulting frames are then resized to 300*300 pixels (A sample example is shown in Figure 1).

Table 1. Example of participants' assessment on basketball skill

ID	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Average
ID-1	2	3	3	1	2	2	3	2	3	3	2.4
ID-2	2	1	2	1	2	1	2	1	2	1	1.5
ID-3	1	1	1	2	3	2	3	1	2	2	1.8
ID-4	3	3	1	3	3	2	3	2	3	2	2.5
ID-5	2	2	1	1	1	1	1	2	2	1	1.4

Figure 1. Some sample examples of the collected basketball action recognition data



(a) Original image (b) Cropped image

3.2 Video-Based Metric Learning Framework

The keys of basketball video action recognition and basketball skill assessment lie in how to model and represent the video sequence, and how to learn distance metric based on this representation. However, the traditional single modeling based methods can only model and represent one aspect of video, resulting in limited representation ability. Based on this, this paper proposes a multiple model based distance metric learning framework. And the flow chart of the proposed framework is shown in Figure 2. Specifically, our method consists of four modules: single image feature learning module, video feature modeling and distance metric learning module, basketball action recognition module and basketball skill assessment module. The single image feature learning module mainly uses CNN network to extract the spatial feature of each frame in the video; The video feature modeling and distance metric learning module is the focus of this article, it consists of two parts: spatial feature modeling and distance metric learning part, temporal feature modeling and distance metric learning part. The first part uses subspace and Gaussian distribution to jointly model spatial features, and the second part uses the LSTM to model temporal features, then the video feature modeling and distance metric learning module maps features in different spaces to high-dimensional Hilbert space through different kernel functions, and then metric learning strategy is used to learn a more discriminating feature space, and realize feature fusion in this space. Finally, the basketball action recognition module classifies the action videos based on the learned distance metric; And the basketball skill assessment module further assesses the video based on the classification results.

3.2.1 Single Image Feature Learning Module

This module is mainly used to learn the spatial feature of each frame in the video. The CNN network here adopts ResNet-50 network, and the cross entropy loss is used. However, since the output dimension of ResNet-50 is 2048, which is not conducive to the subsequent distance metric learning, a new FC layer is added to the top of ResNet-50 to reduce the output dimension to 512.

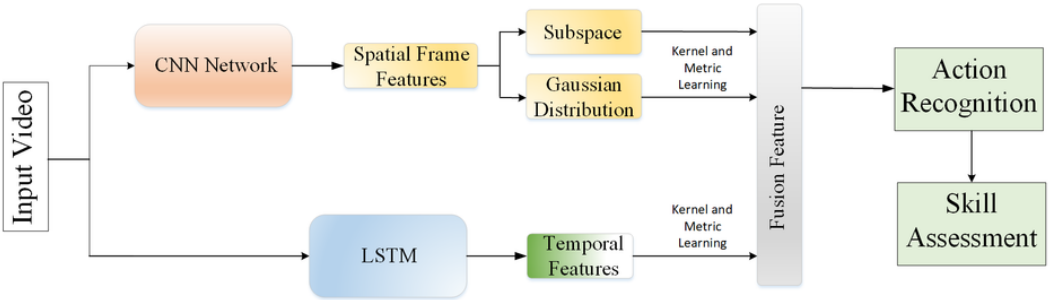
3.2.2 Video Feature Modeling and Distance Metric Learning Module

Let $X = \left\{x_1, x_2, \cdots, x_n\right\}$ be a video sequence of n frames, and $X^\theta = \left\{x_1^\theta, x_2^\theta, \cdots, x_n^\theta\right\}$ be the output of CNN network, where θ denotes the parameter of CNN network, and $x_1^\theta \in \mathbb{R}^{512}$. Then for the spatial features, the subspace and the Gaussian distribution are used to jointly modeling the video.

Subspace: The subspace is usually obtained by principal component analysis (PCA) (Gao 2019), which reduces to the eigen-decomposition of the matrix $X^\theta X^{\theta T}$, here X^θ is rewritten as $X^\theta = \left[x_1^\theta, x_2^\theta, \cdots, x_n^\theta\right]$:

$$X^\theta X^{\theta T} = P \, \Sigma \, P^T \tag{1}$$

Figure 2. Flowchart of the proposed metric learning framework



where P is the $d \times q$ orthogonal matrix $P = \left[p_1, \cdots, p_q\right]$ contains the q largest eigenvectors of $X^{\theta} X^{\theta T}$ and serves as the basis of the q -dimensional linear subspace $span\left(P\right)$, and $d = 512$.

Gaussian distribution: In video action recognition field, it is often insufficient to model the video with one single Gaussian model, because frames in this video are usually highly nonlinear and cover large data variations. Therefore, a multi-modal density mixture model, i.e., Gaussian mixture models (GMM), is utilized to represent these variations efficiently in this study. Then the estimated GMM can be written as:

$$\begin{aligned} G(x^{\theta}) &= \sum_{i=1}^{k} w_i \; g_i(x^{\theta}) \\ g_i(x^{\theta}) &= N(x^{\theta} \mid \mu_i, \Sigma_i) \end{aligned} \tag{2}$$

where $g_i\left(x^{\theta}\right)$ is a Gaussian component with prior probability w_i, mean vector μ_i, and covariance matrix Σ_i, k denotes the number of Gaussian models, and x^{θ} denotes the deep feature of a frame in this video. In addition, since GMM is sensitive to the initialization of the model, and different videos usually have varying numbers of frames, the Hierarchical Divisive Clustering (HDC) (Wang 2011) algorithm is used to generate the initialization adaptively and efficiently.

For the **temporal features**, the LSTM is used to encoder the video. Let φ denotes the parameter of LSTM network, then the video $X = \left\{x_1, x_2, \cdots, x_n\right\}$ can be encoded to a vector C^{φ}, i.e., C^{φ} contains all the temporal information of video X.

In conclusion, now we have three modeling representations for each video, i.e., the subspace, Gaussian mixture distribution and the temporal feature. However, these representations locate on heterogeneous spaces, we cannot directly fuse them. For example, the subspaces lie on Grassmann manifold, the Gaussian distributions lie on a specific Riemannian manifold, while temporal features lie on Euclidean space. To solve this problem, we want to learn three mapping functions, such that the three heterogeneous spaces are mapped to a common subspace. Specifically, we first embed these representations into high dimensional Hilbert spaces using the corresponding kernel functions. After embedding, three different transformations are learned to get the common subspace.

According to the above modeling representations, we have the following three corresponding kernel functions.

Projection kernel: The projection kernel is a generalization of the projection distance. It maps points on Grassmann manifold to RKHS. The formulation of the projection kernel is:

$$k_p\left(P_i, P_j\right) = P_i^T P_{jF}^2 \tag{3}$$

Bhattacharyya Kernel: The Bhattacharyya Distance (BD) is a widely used distance measure in statistics. For Gaussian distributions g_i and g_j, BD can be computed as follows,

$$BD\left(g_i, g_j\right) = \frac{1}{8}\left(\mu_i - \mu_j\right)^T \sum^{-1}\left(\mu_i - \mu_j\right) + \frac{1}{2}\ln\left(\frac{\det \sum}{\sqrt{\det \sum_i \det \sum_j}}\right) \tag{4}$$

where $\sum = \dfrac{\sum_i + \sum_j}{2}$.

Then, by exponentiating the BD, the Bhattacharyya kernel for Gaussian distributions can be defined as:

$$K_{BD}\left(g_i, g_j\right) = \exp\left(-\frac{BD\left(g_i, g_j\right)}{2t^2}\right) \tag{5}$$

3.2.2.1 Objective Function

After defining the above kernel functions, we want to learn three mapping functions f_1, f_2 and f_3, such that the three heterogeneous spaces are mapped to a common subspace. Then, fuse the learned information in the common Euclidean subspace using the following feature fusion strategy:

$$z = \left[f_1(\cdot); f_2(\cdot); f_3(\cdot)\right] \tag{6}$$

In detail, the subspace (point on a Grassmann manifold) is firstly mapped to a Hilbert space H_p by the function $\phi_P : M_p \to \mathrm{H}_p$; the Gaussian distribution (points on a specific Riemannian manifold) is first mapped to the Hilbert space H_g by the function $\phi_g : M_g \to \mathrm{H}_g$; the temporal features are first mapped to the Hilbert space H_E by the RBF kernel function $\phi_E : \mathbb{R}^d \to \mathrm{H}_E$. Then three transformation functions t_1, t_2 and t_3 are learned from the mapped data to get the common subspace. Consequently, the final mappings are $f_1 = t_1 \circ \phi_P$, $f_2 = t_2 \circ \phi_g$, $f_e = t_e \circ \phi_E$.

Finally, our objective function can be formulated as:

$$\begin{aligned}\max \quad & \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{\mathrm{cov}\left(f_i(\cdot), f_j(\cdot)\right)}{\sqrt{\mathrm{var}\left(f_i(\cdot)\right)\bullet\mathrm{var}\left(f_j(\cdot)\right)}} - \sum_{t,k=1}^{m} y_{tk}\left\|f_1(P_t) - f_1(P_k)\right\|_2^2 \\ & - \sum_{t,k=1}^{m} y_{tk}\left\|f_2(G_t) - f_2(G_k)\right\|_2^2 - \sum_{t,k=1}^{m} y_{tk}\left\|f_3(C_t^{\varphi}) - f_3(C_k^{\varphi})\right\|_2^2\end{aligned} \tag{7}$$

where the first term is the inter-modeling constraint term, which is used to maximize the correlation between different modeling features. The last three terms are the intro-modeling constraints terms, they hope that within the same model, the more congealed the samples of the same kind, the better. Here, m denotes the number of videos, $y_{tk} = 1$ if the video pair $\left(t, k\right)$ come from the same category, and $y_{tk} = 0$ if the video pair come from different categories.

3.2.3 Basketball Action Recognition Module

Through the above steps, each video sequence can now be represented using fusion features z, if the number of training videos is m_t, then all training videos can be expressed as: $\left[z_1, z_2, \cdots, z_{m_t}\right]$. When a new test sample $\hat{z}$ is available, the nearest neighbor classifier can be used for effective classification.

3.2.4 Basketball Skill Assessment Module

After obtaining the action category of the test video, you can bring it into this part for skill assessment. Since there are 6 action categories, we need to train 6 multi-class classifiers in this part. In addition, because the

number of training samples in each category is very limited, for each category, a 5-layer fully connected neural network is used to build the assessment classifiers. And the number of output neurons of each model is 5, which respectively represents the probability that the input video belongs to scoring 1,2,3,4,5.

Borrowing idea from NIMA, the EMD-based loss is used as our loss function:

$$EMD\left(p,\hat{p}\right)=\left(\frac{1}{N}\sum_{k=1}^{10}\left|\sum_{i=1}^{k}p_{s_i}-\sum_{i=1}^{k}\hat{p}_{s_i}\right|^{r}\right)^{1/r} \tag{8}$$

where p and $\hat{p}$ are the ground truth and estimated probability respectively.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The effectiveness of the proposed framework on basketball action recognition task and basketball skill assessment task is verified on our collected dataset. For the basketball action recognition task, 80% data, i.e., 240 video sequences per category, are used as the training data and 20% are used as the testing data. For the basketball skill assessment task, in each category, 120 standard actions and 80 non-standard actions are selected as the training set, and the remaining 40 (standard action) and 20 (non-standard action) video sequences are selected as the testing set.

For the action recognition task, the classification accuracy is used as the performance criteria. While for the skill assessment task, the linear correlation coefficient (LCC), spearman's rank correlation coefficient (SRCC) and the EMD value are used as our evaluation criteria.

4.1 Performance on Action Recognition Task

First, we want to verify the effect of the number of video frames on action recognition task. We randomly select 5, 20, 30, 50 frames from each video, and the average classification accuracies are reported in Table 2.

It can be seen from this table that when the size of the input video frame is same, the setting of the number of frames has a certain impact on the recognition performance. Specifically, when the number of frames is too small, for example, only one frame is used, our method will degenerate into single image based action recognition method. In this way, the actions recognized only based on one frame image have no temporal information, and the performance of continuous action recognition is not ideal. As the number of frames increases, the recognition accuracy increases. However, we also note that the higher the number of frames, the higher the training amount of the model, and the longer the time required. Therefore, 30 frames per video is an appropriate choice for our task.

Then, we randomly select 30 frames per video, and compare the classification accuracy of different methods on action recognition task. The comparison results on our collected dataset are shown in Table 3. From Table 3, we can observe that in all cases, our proposed framework achieves the best results, this demonstrates that our proposed framework can effectively deal with basketball action recognition task. Specifically, our proposed framework achieves the average classification accuracy 90.6%, which is higher than other methods, especially compare with Deep LSTM, our framework

Table 2. Average classification results of our proposed framework on our collect dataset with different frame size (%)

Frames	Accuracies
5	80.5
20	85.6
30	90.6
50	93.5

achieves a 3.9% improvement, which means that our multi-model fusion strategy is effective. We also observed that the first three actions are more difficult to recognize in all methods.

4.2 Performance on Basketball Skill Assessment Task

We also carried out experiments on the basketball skill assessment task to verify the effect of the method proposed in this paper. Like the above experiments, we also verify the impact of different frame numbers on the assessment task. The experimental results are shown in Table 4. As shown in Table 4, Our method is sensitive to the number of frames in the video, and when the number of video frames is greater than 30, our method has strong correlation with the ground truth, hence, in what follows, it is acceptable to select 30 frames per video.

Then, we will compare the comparison experiments with some state-of-the-art assessment methods, such as NIMA, A-Lamp CNN (Ma 2017). We also compare our method with some baselines, such as ResNet50 + our assessment network, LSTM + our assessment network. The comparison results on our collected dataset are shown in Table 5.

From Table 5, it can be seen that our proposed method is superior to NIMA and A-Lamp CNN method in all attributes, and in most cases, our baselines achieve the comparable experimental results, we guess this is because NIMA and A-Lamp CNN are both single image based methods, which are difficult to get the video level features. And the comparison among our proposed method and the two baselines demonstrate that our multi-model fusion strategy can capture the complementary discriminant features.

Table 3. Average classification results of different methods on our collect dataset (%)

Methods	Action 1	Action 2	Action 3	Action 4	Action 5	Action 6	Average
C3D	75.0	83.3	83.3	93.3	93.3	96.7	87.6
I3D	81.7	83.3	83.3	93.3	95.0	95.0	88.6
Deep LSTM	73.3	83.3	83.3	91.7	91.7	96.7	86.7
Our	83.3	85.0	86.7	95.0	93.3	100	90.6

Table 4. Average performance of the proposed framework in predicting the score of basketball skill on our collect dataset with different frame size (%)

Frames	LCC	SRCC	EMD
5	0.49	0.50	0.11
20	0.54	0.55	0.07
30	0.65	0.63	0.05
50	0.67	0.65	0.03

Table 5. Performance of the proposed method in predicting the score of basketball skill compared to the state-of-the-art

Methods	LCC	SRCC	EMD
ResNet50 + Our	0.61	0.59	0.10
LSTM + Our	0.60	0.58	0.08
A-Lamp CNN	0.56	0.57	0.11
NIMA(Inception-v2)	0.60	0.59	0.08
Our	0.65	0.63	0.05

5. CONCLUSION

In this paper, aiming at the problem of basketball skill assessment, we first construct a new dataset which contains video action recognition and basketball skill assessment two parts. Then, a new metric learning framework is proposed, which uses a variety of models to jointly represent the action video, and then the optimal distance metric between videos is learned based on the representation. Finally, based on the distance metric, a query video is coarsely classified to obtain the corresponding label of video action, and then the video is finely classified to judge whether the action is standardized. The experimental results on our collected dataset show the superiority to some state-of-the-art methods. Our future research includes continuing to expand the basketball skill assessment dataset and exploring more intelligent basketball skill assessment methods.

REFERENCES

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3), 257–267. doi:10.1109/34.910878

Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action recognition? A new model and the kinetics dataset. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4724-4733. doi:10.1109/CVPR.2017.502

Diba, A., Fayyaz, M., & Sharma, V. (2017). *Temporal 3d convnets: new architecture and transfer learning for video classification*. https://arxiv.org/abs/1711.08200v1

Gao, X., Sun, Q., Xu, H., Wei, D., & Gao, J. (2019). Multi-model fusion metric learning for image set classification. *Knowledge-Based Systems*, *164*, 253–264. doi:10.1016/j.knosys.2018.10.043

Huang, G., Liu, Z., & Der, M. L. V. (2017). Densely connected convolutional networks. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261-2269. doi:10.1109/CVPR.2017.243

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231. doi:10.1109/TPAMI.2012.59 PMID:22392705

Kun, L., Liu, W., & Gan, C. (2018). T-C3D: Temporal convolutional 3D network for real-time action recognition. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 7138-7145.

Ma, S., Liu, J., & Chen, C. (2017). A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Meng, H., & Pears, N. (2009). Descriptive temporal template features for visual motion recognition. *Pattern Recognition Letters*, *30*(12, 12SI), 1049–1058. doi:10.1016/j.patrec.2009.03.003

Ng, J. Y., Hausknecht, M., & Vijayanarasimhan, S. (2015). Beyond short snippets: Deep networks for video classification. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, 4694-4702.

Park, S., & Aggarwal, J. K. (2003). Recognition of two-person interactions using a hierarchical Bayesian network. *Proc of ACM SIGMM International Workshop on Videl Surveillance*, 65-76. doi:10.1145/982452.982461

Pavlovic, V., Rehg, J. M., & Cham, T. J. (2000). A dynamic Bayesian network approach to tracking using learned switching dynamic models. *International Workshop on Hybrid Systems: Computation and Control*, 366-380. doi:10.1007/3-540-46430-1_31

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Proceedings of the Advances in Neural Information Processing Systems*, 568-576.

Talebi, H., & Milanfar, P. (2018). NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, *27*(8), 3998–4011. doi:10.1109/TIP.2018.2831899 PMID:29994025

Tran, D., Bourdev, L., & Fergus, R. (2015). Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the 15th IEEE International Conference on Computer Vision*, 4489-4497. doi:10.1109/ICCV.2015.510

Tran, D., Ray, J., & Shou, Z. (2017). *ConvNet architecture search for spatiotemporal feature learning*. https://arxiv.org/abs/1708.05038

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. *Lecture Notes in Computer Science*, *9912*, 20–36. doi:10.1007/978-3-319-46484-8_2

Wang, N. (2014). *Gesture recognition based on hidden Markov model with binocular camera* [Thesis]. Northeastern University.

Wang, R. (2011). Maximal linear embedding for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(9), 1776–1792. doi:10.1109/TPAMI.2011.39 PMID:21358001

Wang, W., & Wang, R. (2018). Discriminant Analysis on Riemannian Manifold of Gaussian Distributions for Face Recognition with Image Sets. *IEEE Transactions on Image Processing*, *27*(1), 151–163. PMID:28866497

Wang, W. S., Yang, S., Zhang, W. S., & Zhang, J. (2019). Neural Aesthetic Image Reviewer. *IET Computer Vision*, *13*(8), 749–758. doi:10.1049/iet-cvi.2019.0361

Wang, X., Xie, L., & Peng, L. (2020). Double residual network recognition method for falling abnormal behavior. *Journal of Frontiers of Computer Science and Technology*, *14*(9), 1580–1589.

Wei, D., Shen, X., Sun, Q., Gao, X., & Yan, W. (2019). Prototype learning and collaborative representation using Grassmann manifolds for image set classification. *Pattern Recognition*, *100*, 107123. doi:10.1016/j.patcog.2019.107123

Zhu, F., Gao, J., Xu, C., Yang, J., & Tao, D. (2017). On selecting effective patterns for fast support vector regression training. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3610–3622. PMID:28841559

Zhu, F., Gao, J., Yang, J., & Ye, N. (2022). Neighborhood linear discriminant analysis. *Pattern Recognition*, *123*, 108422. doi:10.1016/j.patcog.2021.108422

Zhu, F., Ning, Y., Chen, X., Zhao, Y., & Gang, Y. (2021). On removing potential redundant constraints for SVOR learning. *Applied Soft Computing*, *102*, 106941.