Automatic Sports Video Classification Using CNN-LSTM Approach*

Benoughidene Abdelhalim[1,*,†], Titouna Faiza[2,†]

[1]Department of computer science, LaSTIC Laboratory, University of Batna2, Algeria
[2]Department of computer science, LaSTIC Laboratory, University of Batna2, Algeria

Abstract

With the exponential growth and availability of sports video data, the need for video analysis has become crucial. Sports video classification is one of the most challenging problems among computer vision researchers. This paper proposes a novel method for sports video classification based on a deep learning approach using the pre-trained model Inception V3 combined with long short-term memory (LSTM). The pre-trained model Inception V3 extracts the low, middle, and high spatial features. Additionally, we will get the temporal features by feeding LSTM with spatial features. Then, we trained our InceptionV3-LSTM model based on spatial-temporal features by feeding it to the LSTM to classify the video sports into specific categories. The experiments are conducted on the UCF sports dataset. The experiments performed showed that our proposed model has obtained much more encouraging experimental results than the others.

Keywords

Transfer learning, Inception V3, Long short-term memory (LSTM), Deep learning, Sport video classification

1. Introduction

As the volume of sports video data increases, video classification becomes essential for understanding and organizing this data. Video classification is an important visual task in computer vision and has been used in a variety of applications, including motion recognition [1] and scene classification [2]. Video classification enables users to easily access and understand sports videos and provides insight into the game or sport [3].The aim of sports video classification is to automatically classify sports videos according to the sporting events they contain, which have spatial and motion features [4].

Deep learning algorithms have many applications, such as object recognition from images, search engines, and speech recognition. However, sports video classification is a relatively new field. With the development of deep learning techniques, this field has attracted the attention of researchers looking for new challenges. A video is composed of a set of sequential images. Each image contains information about the spatial content, while the temporal sequence of images contains information about the motion. In order to represent a video in a comprehensive and informative way, it is necessary to capture both spatial and temporal information and combine them [5].

To take account of the two spatio-temporal aspects of video, we use a hybrid architecture consisting of convolutional layers to handle spatial information and recurrent layers to handle temporal information. The two-stream convolutional neural network (CNN) and the recurrent neural network (RNN), which handle both spatio-temporal information, proved to be better than their single-stream counterparts [6].

In this paper, we focus on the classification of sports videos from the UCF dataset. We will use a transfer learning method to pre-train a CNN model (Inception V3) to extract low, middle, and high spatial features. Then, we will use Long Short-Term Memory (LSTM) layers to extract temporal features to classify sports videos into specific categories. The proposed model has achieved impressive performance on this dataset compared with some existing methods. Contributions to this work include:

1. A new deep neural network model based on LSTM is proposed to classify video frames into specific classes;
2. In this model, three spatial feature classes are extracted by a pre-trained CNN model (Inception V3) instead of hand-created features. This decision was taken to avoid system complexity and improve performance. The aim of this work is to prove that CNN features can outperform handcrafted ones;
3. In addition, we combine spatial information from the CNN model (Inception V3) with temporal information from the LSTM model to improve model performance and decision-making.

6th International Hybrid Conference On Informatics And Applied Mathematics, December 6-7, 2023 Guelma, Algeria
*Corresponding author.
†These authors contributed equally.
benouhalim@gmail.com (B. Abdelhalim); ftitouna@yahoo.fr (T. Faiza)
0000-0001-8969-2119 (B. Abdelhalim); 0000-0003-1970-2937 (T. Faiza)

This method is effective because it provides better-quality information by combining different sources rather than using them separately.

The paper consists of four sections. Section 2 discusses related work on sports video classification. Section 3 describes the proposed methodology. Section 4 presents our results. Section 5 discusses concluding remarks.

2. Related Work

Researchers are currently looking to develop more accurate and reliable algorithms to improve the classification of sports videos. The classification of sports videos is a new field that is developing rapidly with the use of deep learning techniques. Features can be extracted from text, audio, and visual information [7]. In this work, we focus solely on visual information, as vision is the primary means by which humans perceive information [8].

Deep learning methods have recently been shown to be able to automatically extract complex features from images. Building on this, many researchers have worked on the two latest deep learning techniques, convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to extract spatial and temporal features from sports videos [4]. In [9], the authors presented a transfer learning algorithm for DNN-based video classification. The algorithm is designed as a video classification system based on a convolutional neural network. Podgorelec et al, in [10] used a transfer learning approach to fine-tune a pre-trained CNN model for a sports image classification task. They developed a classification method using CNN transfer learning with Hyper Parameter Optimization (HPO). Russo et al, in [11] proposed a model that combines deep learning and transfer learning. They combined VGGNET16, RNN, and GRU functions with transfer learning for the final classification. Chen et al, in [12] investigated simple sports classification in sports videos by detecting motion poses in video frames. For example, they can detect running, jumping, translation, and zooming. Qiu et al, in [13] showed that principal components can be used to reduce the dimensions of visual and audio video features. A time series of motion features can be used to characterize motion event classifications in soccer sports videos.

3. Proposed Methodology

In this work, we aim to develop a model for sport video classification tasks that starts with two basic steps: the feature extraction process generated by the pre-trained model InceptionV3 and the training process by LSTM to predict action as presented in Figure 1. We utilized an LSTM network, which is equipped with mechanisms known as gates. These gates are responsible for managing the cell state, thereby controlling the long-term and short-term memory of the network. They play a crucial role in determining what information should be retained or discarded.

Additionally, we employed one fully connected (FC) layer, also referred to as a dense layer, which consists of 10 neurons. The output from this layer is then processed through a softmax activation function for the final prediction. This entire approach was implemented using the Keras API. Indeed, deep learning methods have proven to be effective and perform well on tasks involving the comparison of video classifications.

3.1. Features Extraction Process

Classifying actions from videos requires extracting a set of features that are anticipated to contain the data necessary to distinguish between various actions. There are three types of features: the first is low-level features like corners, edges, and simple textures; the second is middle-level features like complex textures and shapes; and the third is high-level features like objects or parts of objects. The feature extraction technique is explained below.

3.1.1. Transfer learning via feature extraction [Low, Middle, high]:

Conventional approaches to video classification have used frame-based features to generate a representation for the videos. In this work, various types of features, such as low, middle, and high, have been used to capture various types of information. These features are obtained with the help of a transfer learning strategy. Transfer learning is an effective strategy for training networks on a small data set. The network is pre-trained on a large dataset, such as ImageNet, and then reused to train a new task. This offers a significant advantage over training the network from scratch, as it requires less time and less data. Transfer learning can save both time and computing costs [14]. There are many pre-trained models on the ImageNet dataset, such as AlexNet, VGG16, ResNet and InceptionV3. These models can be used to extract features from the data or fine-tune them to perform a new task.

In this work, we used transfer learning to extract features. Convolutional Neural Networks (CNN) automatically learn features from images at a hierarchical level.

3.1.2. Extracting features generated by InceptionV3:

Inception V3, a widely recognized convolutional neural network, is frequently utilized in tasks involving image classification and feature extraction. It demonstrates

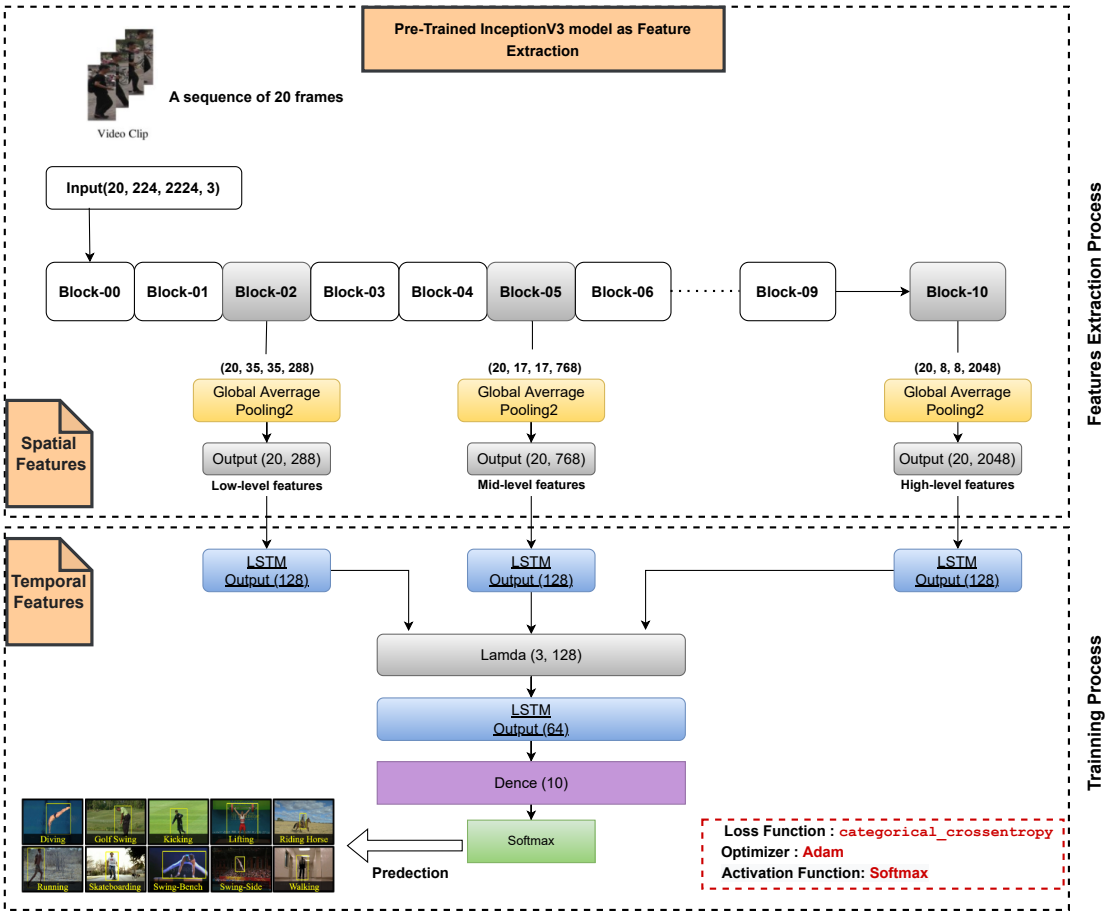

Figure 1: Proposed Method.

high accuracy, achieving over 75% - 78% on the ImageNet dataset, and is known for its speed and precision. The architecture of InceptionV3, equipped with multiple 'Inception' modules, is designed to capture intricate patterns and hierarchies in images, making it particularly suitable for sports video classification, where data often exhibits complex spatial hierarchies and patterns. However, the selection of a model can be influenced by the specific requirements of the task and the characteristics of the data. Therefore, it could be advantageous to compare the performance of InceptionV3 with other pre-trained models in our future work.

In this work, we propose an approach based on transfer learning that uses InceptionV3 to extract features from frames. The selection of specific blocks from the InceptionV3 model for feature extraction is typically based on the type of information these blocks can capture and their impact on the overall performance of the model.

The first layers of any neural network are basically responsible for identifying low-level features; the later convolutional layers identify middle-level features; and the last convolutional layers identify high-level features, which are usually very specific to the task they are trained for. In the InceptionV3 architecture, lower layers (such as block-02) are usually responsible for capturing low-level features such as edges and textures. Mid-level layers (such as block-05) can capture more complex features like shapes, and higher layers (such as block-10) can capture high-level semantic information, such as the presence of specific objects in the image. (see Figure 2)

1. **The low-level features:** Were extracted from the block02, with 288 dimensions.
2. **The middle-level features:** Were extracted from the block05, with 768 dimensions.
3. **The high-level features:** Were extracted from the last block10, with 2048 dimensions.

The rationale behind selecting these specific blocks could be that the combination of low, mid, and high-level fea-

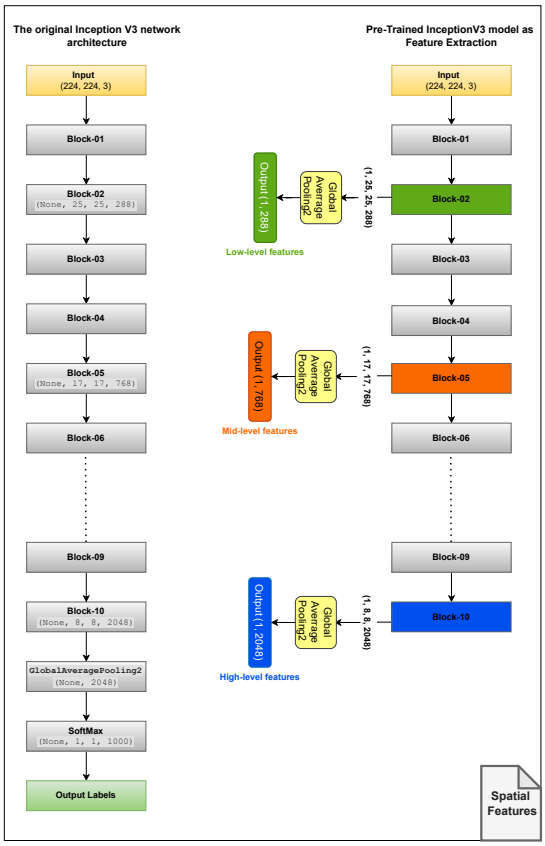

Figure 2: Transfer learning approach as a feature extractor. **Left:** The original InceptionV3 network architecture outputs probabilities for each of the 1,000 ImageNet class labels. **Right:** Removing the FC layers from InceptionV3 and returning the block-02, block-05 and block-10 as output of features

tures provides a comprehensive representation of the video content, which is beneficial for the task of sports video classification.

3.2. Training process

We train the LSTM and Fully Connected (FC) layer to perform the classification process. Whereas CNN blocks cannot be trained to extract features. This saves us less consumption of resources, parameters and computation time. After passing the data through the model, we gradually decrease the computed error using the categorical cross-entropy formula provided by the following loss function.

$$CE = -\sum_{c}^{M} y_{o,c} \log(p_{o,c}) \quad (1)$$

Where :

- CE : Loss function (categorical Cross-Entropy);
- M : The number of classes;
- Y : The ground truth;
- P : The expected probabilistic observation of class c.

Model parameters are updated using gradient descent and backpropagation error. During the learning process, the joint weights of the pre-trained model are frozen, in particular from block 0 to block 10. Finally, the LSTM output passes through a fully connected layer with softmax activation functions.

Experimental results show that our method is effective in classifying sports videos, with the model achieving 89.5% accuracy, 90% precision, 88% recall, and 88% f1 score.

4. Experimental results and discussion

4.1. UCF Sports Action Data Set

We evaluated the performance of our proposed model by testing it on the UCF sports action dataset, which contains videos of various actions from different sports Figure 3. The dataset includes human-annotated action boundaries, which are shown in yellow. UCF is considered one of the best datasets for applications that require action localization and recognition.

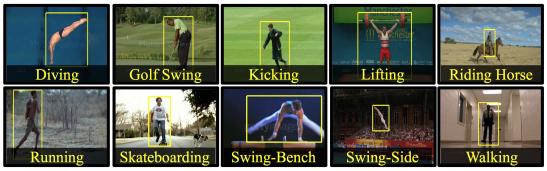

Figure 3: UCF Sports Dataset: sample frames of 10 action classes along with their bounding box annotations of the humans shown in yellow

The dataset comprises 150 videos divided into 10 categories, filmed in different environments. Since each category has a different number of videos. Figure 4 shows the distribution of the number of videos for each category. The video resolution is 720 x 480 and the frame rate is 10 fps. The total duration of the dataset is 958 seconds, and the average sequence length is 6.39 seconds. Table 1 shows the characteristics of the data set [15] [16].

The 20 frames of each video were used for video classification, as shown in Table 2.

4.2. Evaluation Metrics

The quality of video classification models is measured using various performance metrics. Precision, accuracy,

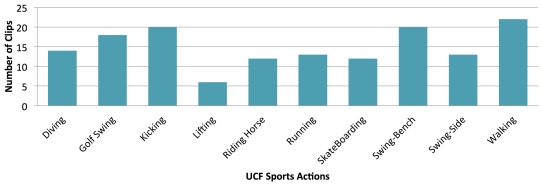

Figure 4: Number of clips per action class

Table 1
Summary of the characteristics of UCF Sports.

Actions	10	Durations	958s
Clips	150	Frame rate	10fps
Mean clip length	6.39s	Resolution	720x480
Min clip length	2.20s	Max clips	22
Max clip length	14.10s	Min clips	6

Table 2
Experimental Setup.

Parameters	Values
The maximum number of frames	n = 20
The resize of input frames	224 x 224 x 3
Learning rate	0.01
The batch_size	32
Maximum epochs	100

recall and F1 score are among the most common measures. We uses these metrics to evaluate ours model. These metrics are defined as follows:

1. **Accuracy:** It is a measure of how well a model performs across all categories. Accuracy is useful when all categories are equally important. It is calculated by dividing the number of correct predictions by the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

2. **Precision:** It is the ratio of the number of positive samples correctly classified to the total number of samples classified as positive, including incorrectly classified positive samples.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

3. **Recall:** It is the ratio of the number of positive samples that were correctly detected to the total number of positive samples. The higher the recall, the more positive samples are detected.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

4. **F1 score:** It is a measure that combines accuracy and recall into a single score. F1 score ranges from 0 to 1, with a score of 1 indicating the best model performance.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

Where (TP) is True Positive, (TN) is True Negative, (FP) is False Positive and (FN) is False Negative.
Table 3 compares the performance of our proposed model with some recent works on the UCF Sports Motion dataset.

4.3. Results and Discussion

The goal of a learning algorithm is to find a model that has a good fit, meaning that the model is not overfitting or underfitting. Our model has a good fit, as evidenced by the decrease in the training and validation losses to a point of stability with a minimum gap between the final loss values.

In Figure 5, it can be observed that the maximum training accuracy for the model was reached at epoch 100, where the validation accuracy also reached its maximum, which is 89.5%. The learning curve plot shows good agreement, which confirms the good fit of our model.

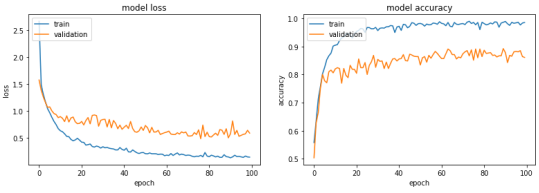

Figure 5: The accuracy and loss presentation in the training and validation phases for UCF Sports action dataset. These results present high validation accuracy 89.5%

Figure 6 shows a confusion matrix for ten actions in the UCF Sport action recognition experiment using a batch normalization layer. The diagonal elements indicate the accuracy of recognizing each action type. Each row of the confusion matrix represents the true action, and each column represents the predicted action. Our model performs well on some actions, achieving a validation accuracy of 100% for seven actions.

The classification report is shown in Figure 7 including three criteria: precision, recall, and F1-score. Our method successfully performs the classification between ten actions, as the model yields 89.5% of accuracy, 90% of precision, 88% of recall, and 88% of f1 score.

Table 3 compares our results on the UCF sports dataset with some state-of-the-art methods. Our model significantly outperformed all studies that used the handcrafted

Table 3
Comparison of the proposed model with other models evaluated over UCF sports action dataset.

Feature extraction method	Methods	Accuracy %	Year
Handcrafted features	Rodriguez et al. [15]	69.2	2008
	Yeffet et Wolf. [17]	79.2	2009
	klaser et al. [18]	86.7	2010
	Wang et al. [19]	88.2	2011
	Yu et al. [20]	81.07	2014
Learned features	Oliveira Silva et al. [21]	78.46	2017
	Zare et al. [22]	82.14	2019
	Jaouedi et al. [23]	89.01	2020
	Ours	89.5	2023

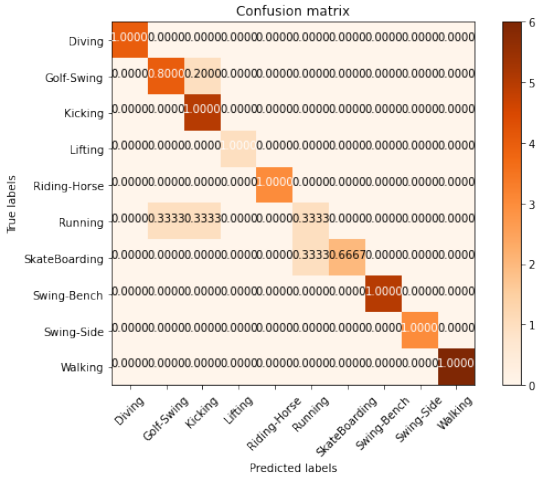

Figure 6: The confusion matrix presentation of ten action in the UCF Sport action dataset

	precision	recall	f1-score	support
Diving	1.00	1.00	1.00	4
Golf-Swing	0.80	0.80	0.80	5
Kicking	0.71	1.00	0.83	5
Lifting	1.00	1.00	1.00	1
Riding-Horse	1.00	1.00	1.00	3
Running	0.50	0.33	0.40	3
Skate-Boarding	1.00	0.67	0.80	3
Swing-Bench	1.00	1.00	1.00	5
Swing-Side	1.00	1.00	1.00	3
Walking	1.00	1.00	1.00	6
accuracy			0.89	38
macro avg	0.90	0.88	0.88	38
weighted avg	0.90	0.89	0.89	38

Figure 7: Classification report

features method, but for studies that used our same approach to extracting features, that is, the learned features, our model achieves competitive results compared to [21], [22] and [23] in terms of accuracy.

5. Conclusion and future work

This study proposes a new approach to identifying human action recognition in sports. The approach is based on feature extraction and uses two models: Inception V3 to extract spatial features (low, middle, and high-level) and LSTM to extract temporal features. The FC layer receives the output of the final LSTM layer, and the softmax layer predicts the type of action. One of the key benefits of the proposed method is its ability to aggregate spatial features from Inception V3 and temporal features from LSTM at each time step and in each video frame. This approach improves the model's performance and decision-making by combining information from two sources rather than using them independently. Experimental results on the UCF sports dataset show that our proposed method achieves higher classification accuracy 89.5% than state-of-the-art sports video classification methods. Our proposed method has shown promising results, but it has its limitations. One potential limitation could be the reliance on the InceptionV3 model for feature extraction, which may not be ideal for all sports videos. Also, the LSTM model, used for capturing time-related changes, might struggle with long sequences.

For future work, we suggest looking into other pre-trained models for feature extraction, exploring more advanced versions of LSTM like Gated Recurrent Units (GRU) or Transformer models, and combine more data like player stats or game context. We believe these improvements could make our method even better at classifying sports videos.

References

[1] H. Liu, N. Shu, Q. Tang, W. Zhang, Computational model based on neural network of visual cortex for human action recognition, IEEE Transactions on Neural Networks and Learning Systems 29 (2018) 1427–1440. doi:10.1109/tnnls.2017.2669522.

[2] K. G. Derpanis, M. Lecce, K. Daniilidis, R. P. Wildes,

Dynamic scene understanding: The role of orientation features in space and time in scene classification, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012. doi:10.1109/cvpr.2012.6247815.

[3] J. Zheng, X. Cao, B. Zhang, X. Zhen, X. Su, Deep ensemble machine for video classification, IEEE Transactions on Neural Networks and Learning Systems 30 (2019) 553–565. doi:10.1109/tnnls.2018.2844464.

[4] M. S. Sarma, K. Deb, P. K. Dhar, T. Koshiba, Traditional bangladeshi sports video classification using deep learning method, Applied Sciences 11 (2021) 2149. doi:10.3390/app11052149.

[5] M. A. Russo, A. Filonenko, K.-H. Jo, Sports classification in sequential frames using cnn and rnn, in: 2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT), 2018, pp. 1–3. doi:10.1109/ICT-ROBOT.2018.8549884.

[6] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[7] D. Brezeale, D. Cook, Automatic video classification: A survey of the literature, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38 (2008) 416–430. doi:10.1109/tsmcc.2008.919173.

[8] M. C. Darji, D. Mathpal, A review of video classification techniques, IRJET Journal 4 (2017).

[9] H. Guangyu, Analysis of sports video intelligent classification technology based on neural network algorithm and transfer learning, Computational Intelligence and Neuroscience 2022 (2022) 1–10. doi:10.1155/2022/7474581.

[10] V. Podgorelec, Š. Pečnik, G. Vrbančič, Classification of similar sports images using convolutional neural network with hyper-parameter optimization, Applied Sciences 10 (2020) 8494. doi:10.3390/app10238494.

[11] M. A. Russo, L. Kurnianggoro, K.-H. Jo, Classification of sports videos with combination of deep learning models and transfer learning, in: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, 2019. doi:10.1109/ecace.2019.8679371.

[12] C. Lifu, W. Hong, C. Xianliang, G. Zhenghua, J. Zhiwei, Convolutional neural network sar image target recognition based on transfer learning, Chinese Space Science Technology 38 (2018) 45–51.

[13] N. Qiu, X. Wang, P. Wang, S. Zhou, Y. Wang, Research on convolutional neural network algorithm combined with transfer learning model, Computer engineering and applications 56 (2020) 43–48.

[14] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, Proceedings of the IEEE 109 (2020) 43–76.

[15] M. D. Rodriguez, J. Ahmed, M. Shah, Action MACH a spatio-temporal maximum average correlation height filter for action recognition, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008. doi:10.1109/cvpr.2008.4587727.

[16] K. Soomro, A. R. Zamir, Action recognition in realistic sports videos, in: Computer vision in sports, Springer, 2014, pp. 181–208.

[17] L. Yeffet, L. Wolf, Local trinary patterns for human action recognition, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009. doi:10.1109/iccv.2009.5459201.

[18] A. Klaser, M. Marszalek, I. Laptev, C. Schmid, Will person detection help bag-of-features action recognition?, Research Report RR-7373, INRIA, 2010. URL: https://hal.inria.fr/inria-00514828.

[19] H. Wang, A. Klaser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in: CVPR 2011, IEEE, 2011. doi:10.1109/cvpr.2011.5995407.

[20] J. Yu, M. Jeon, W. Pedrycz, Weighted feature trajectories and concatenated bag-of-features for action recognition, Neurocomputing 131 (2014) 200–207. doi:10.1016/j.neucom.2013.10.024.

[21] V. de Oliveira Silva, F. de Barros Vidal, A. R. S. Romariz, Human action recognition based on a two-stream convolutional network classifier, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2017. doi:10.1109/icmla.2017.00-64.

[22] A. Zare, H. A. Moghaddam, A. Sharifi, Video spatiotemporal mapping for human action recognition by convolutional neural network, Pattern Analysis and Applications 23 (2019) 265–279. doi:10.1007/s10044-019-00788-1.

[23] N. Jaouedi, N. Boujnah, M. S. Bouhlel, A new hybrid deep learning model for human action recognition, Journal of King Saud University - Computer and Information Sciences 32 (2020) 447–453. doi:10.1016/j.jksuci.2019.09.004.