Human Action Recognition Based on Three-Stream Network with Frame Sequence Features

Ruifeng Huang[1,2], Chong Chen[1,2*], Rui Cheng[1,2], Ying Zhang[1], Jiabing Zhu[3]
[1] Anhui Jianzhu University, Hefei, 230601, China
[2] Anhui International Joint Research Center for Ancient Architecture
Intellisencing and Multi-Dimensional Modeling, Hefei, 230601, China
[3] Huainan Normal University, Huainan, 232038, China
huangruifeng1997@foxmail.com, shchshch@ustc.edu.cn, {cheng_rui96, awm_obz1998, zjb3617}@163.com

2022 7th International Conference on Image, Vision and Computing (ICIVC) | 978-1-6654-6734-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIVC55077.2022.9887162

***Abstract*—In the field of human action recognition (HAR), two-stream models have been widely employed. In recent years, traditional two-stream network models have disregarded the interframe sequence characteristics of video, resulting in a decrease in model robustness when local sequence information and long-term motion information interact. In light of this, a novel three-stream neural network is proposed by combining the long-term and short-term characteristics of a frame sequence with spatio-temporal information. Initially, the optical flow sequence image frames and RGB image frames in the video are extracted, the optical flow motion information and image space information in the video is obtained, the corresponding time network and space network are entered, and the spatial information is entered into the sequence feature processing network; the three networks are then pretrained. At the conclusion of training, the operation of feature extraction is executed, the features are incorporated with the parallel fusion algorithm by adding weights, and the behavior categories are classified using Multi-Layer Perception. Experimental results on the UCF11, UCF50, and HMDB51 datasets demonstrate that our model effectively integrates the spatial-temporal and frame-sequence information of human actions, resulting in a significant improvement in recognition accuracy. Its classification accuracy on the three datasets was 99.17%, 97.40%, and 96.88%, respectively, significantly enhancing the generalization capability and validity of conventional two-stream or three-stream models.**

***Keywords*—Human Action Recognition, Frame Sequence Features, UCF11, UCF50, HMDB51.**

I. Introduction

Human Action Recognition (HAR) research has evolved in tandem with the discipline of computer science and contemporary techniques of deep learning over time. Similarly, it is one of the most promising research directions in machine learning and pattern recognition. HAR has numerous applications in the fields of intelligent video surveillance, video editing of sporting events, unmanned driving, video retrieval, and behavioral rehabilitation detection of injured individuals. However, HAR remains a very difficult video classification task due to the variability of human movement behavior, the complexity of environmental information, and the potential loss of some behavioral information. Generally, HAR is performed by extracting the spatial information provided by the image frames and the temporal information derived from the inter-frame difference information processing from a video consisting of a series of image frames. However, this approach loses the motion behavior information carried by the image

To obtain accurate recognition results and a more robust HAR model system, it is necessary to extract more useful and pertinent information. This paper presents a solution for designing a model consisting of two CNN networks and one LSTM network for the simultaneous extraction of spatial-temporal features and frame sequence features from video frames. The proposed method has been tested on three datasets, UCF11[1], UCF50[2], and HMDB51[3], all of which achieve more accurate classification results to demonstrate its efficacy.

In Section II of this paper, we will briefly review the work related to human action recognition (HAR), discussing both traditional human action recognition methods and current mainstream deep learning methods. Section III describes in detail the three-stream network approach based on frame sequence information. In Section IV, we test and evaluate the model presented in this paper on three publicly available datasets, UCF11, UCF50, and HMDB51, and analyze the results. The paper concludes with a summary.

II. Related works

With the advancement of artificial intelligence technology and machine learning, there are two primary types of recognition approaches in the field of Human Action Recognition (HAR) : traditional recognition methods based on manually extracted feature characteristics and action recognition methods based on deep learning feature extraction.

A. Traditional Method in HAR

Conventional action recognition methods use observation and analytical methods to parse the motion information in an image by manually extracting effective and appropriate human kinematic features to represent human actions. There are two primary types of algorithms: feature extraction methods based on the geometric or motion information of the entire human body, and feature descriptors that use spatio-temporal interest point methods to extract local motion information of the human body. Motion Energy Images (MEI) and Motion History Images (MHI) proposed by Bobic et al. [4] represent human motion information contained in an image

978-1-6654-6734-6/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: International Institute of Information Technology Bangalore. Downloaded on November 03,2024 at 09:19:47 UTC from IEEE Xplore. Restrictions apply.

sequence by stacking silhouettes in space-time using an encoding method. Spatial-Temporal Interest Points (STIPs) were proposed by Laptev et al. [5] as local description operators for behavioral features, and Harris corner detectors [6] and 3D-Harris corner detectors [7] were used to extract and construct spatio-temporal interest points. Due to the complex design of traditional manual feature extraction methods, their low universality, the difficulty in extracting deep features, and the impact of ambient lighting and image noise interference from environmental scenes on the model's generalization ability, deep learning-related methods are gradually becoming a proven solution for the aforementioned issues.

B. Deep Learning Method in HAR

Deep learning methods, which differ from traditional methods, have better performance in recognition accuracy and model robustness because they reduce human labor costs and have greater extraction and expression capabilities for behavioral features due to the rapid development of deep learning in recent years. Deep learning methods can effectively extract and characterize the temporal and spatial information carried by the video; consequently, many methods extract and process both temporal and spatial information, which is an important perspective that is frequently considered when using deep learning methods to solve behavioral recognition problems.

1) Two-Stream Method: Karpathy et al. [8] proposed dividing a video into successive RGB image frames, stitching them together, and feeding them into a convolution neural network (CNN). Nonetheless, it is challenging for a single neural network to extract all action information from a video. Therefore, Simonyan et al. [9] proposed Two-Stream Convolutional Networks (TCN), which successfully extracts the spatial information of motion as well as the temporal information on the time stream, and based on this, they proposed a multi-feature fusion classification method that effectively improves recognition accuracy. Ji et al. [10] proposed a 3D convolutional neural network (3D-CNN) technique that improves model performance by stacking split frame sequence images and convolving them in 3D to simultaneously obtain spatio-temporal features. Wang et al.[11] proposed a network that combines 3D-CNN and LSTM, which significantly improves the representation of temporal information over the original 3D-CNN.

2) Three-Sream Method: Shi et al. [12] proposed a sequential Deep Trajectory Descriptor (sDTD) for 2D images to transform them into a triple-stream network in order to enhance the accuracy of the model based on the double-stream network model. In order to improve the accuracy of the model based on the two-stream network model, the sDTD is added to a two-stream network to transform it into a three-stream network. However, the aforementioned methods have limitations, such as excessive preprocessed data and ineffective sequence information extraction from the video itself.

III. Proposed Method

A. Three Stream Architecture

Figure 1 illustrates the framework we proposed. The process begins by separating the original video into RGB and optical flow image frames. The RGB image frames are processed by a sequential network to extract the frame sequence information features from the video, and the optical flow frame images are fed to a temporal network to extract the temporal features from the video. After feature extraction is complete, the spatial, temporal, and sequential features are combined to represent behavior feature information. Eventually, a multilayer perceptron is utilized to classify various actions.

B. Video Preprocessing

The RGB frame image will invoke the FFmpeg package in OpenCV to perform video frame splitting processing. If the original video length is 10 seconds, the 10-second 24fps video is split into 240 frames and stored in the corresponding folders. Next, the labeled order and the corresponding file information and category tags are extracted and put into the csv file used to store the data information, allowing the model to read and load the pre-processed data based on the csv file information.

We utilized the dense optical flow algorithm Farnebäck [13] to extract the optical flow image data from the video.

C. Spatial and Temporal Network

For feature extraction in the temporal and spatial networks, two nearly identical 2D convolutional neural network models are used. The only distinction between the two models is the number of spatial network input channels. The 2D CNN model proposed by this method consists of a series of linearly stacked layers, with each layer receiving the output of the previous layer as its input. The proposed sequential model of 2D CNN does not support nonlinear topologies, shared layers, or the presence of multiple inputs (or outputs) per layer. As a result, the information flow transfer strategy of this 2D CNN model is relatively straightforward, as well as simple to interpret and model. Fig. 3 illustrates the architecture of our sequential 2D CNN spatio-temporal network model that is used to extract deep features in time and space and classify the input data. It is composed of five 2D convolutional layers, three 2D max-pooling layers, and two batch normalization layers. In addition, the neural network component has two deactivation layers and uses ReLU activation functions between each layer to improve the deep network model's performance. The use of batch normalization and deactivation layers in the model is very practical and effective in limiting and reducing the overfitting of the model on the training data, which improves the model's performance on test samples and increases its applicability. The relevant parameter settings within each layer of this model, which are based on the model's continuous optimization and eventual adoption during training and testing, are briefly described in Fig. 3.

[14] provides detailed mathematical descriptions of back propagation, loss functions, weight initialization, and operations for numerous CNN layers, including the layers shown

Authorized licensed use limited to: International Institute of Information Technology Bangalore. Downloaded on November 03,2024 at 09:19:47 UTC from IEEE Xplore. Restrictions apply.

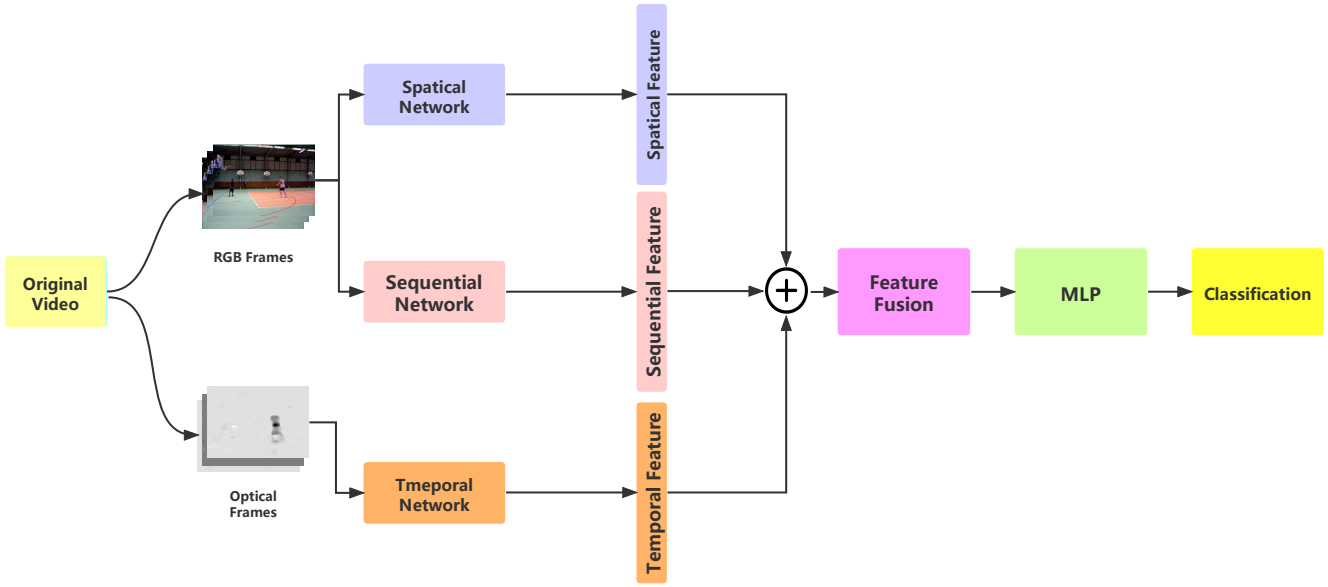

Fig. 1. Three-Stream Network Architecture

in Fig. 3. We do not implement any uncommon or complex variants of the operations within the layers; rather, we employ common and frequently used optimizer functions and a more novel and efficient activation function.

D. Sequential Network

In conventional action recognition frameworks, 2D CNNs and 3D CNNs are typically used to learn the spatial characteristics and spatiotemporal motion characteristics of behaviors. This paper takes a different approach than convolution to obtain the sequence information of video frames represented by the motion itself of the frame images in order to better extract the sequence feature information contained in the behavioral video frame images themselves. Recurrent Neural Networks (RNN), which have become popular due to their properties in natural language processing for extracting semantic information from sequential information processing of linguistic text, are typically used for processing sequential information in behavior classification and other learning-related tasks.

To better extract sequence motion information from the sequence features of frame pictures as well as the before and after behavioral action sequence information of the relative positions of the motion behaviors, the long-short-term memory network (LSTM) as a special RNN network can be used for before and after spatio-temporal correlation modeling within the action recognition framework. Our network structure thus adopts the standard LSTM model for processing and feature extraction of sequence data.

Generally, given the input of a sequence as x_p (where $p \in P$, P is the range of time step intervals for the sequence input), the LSTM unit recursively calculates and outputs h_p based on the following equations:

$$
\begin{aligned}
f_p &= \sigma\left(W_i \cdot [h_{p-1}, x_p] + b_f\right) \\
i_p &= \sigma\left(W_i \cdot [h_{p-1}, x_p] + b_i\right) \\
\tilde{C}_p &= \tanh\left(W_C \cdot [h_{p-1}, x_p] + b_C\right) \\
C_p &= f_p * C_{p-1} + i_p * \tilde{C}_p \\
o_p &= \sigma\left(W_o\left[h_{p-1}, x_p\right] + b_o\right) \\
h_p &= o_p * \tanh\left(C_p\right)
\end{aligned}
$$

x_p and h_p are the input and hidden states of the LSTM unit in the graph at sequence time p. i_p, f_p, C_p, $\tilde{C}_p$ and o_p are the states of the input gate, the forget gate, the storage unit, and the selected storage unit and output gate at time p, respectively. W_i, W_C, and W_o are the weight matrices of their corresponding states, and b_f , b_i, b_C , and b_o are the bias constants of their corresponding states, respectively. σ is the sigmoid function. In contrast to more complex model structures with bi-directional information extraction, we use a single time-oriented LSTM unit, the classical LSTM network, which can better accomplish the modeling task of motion timing information extraction with less computational effort and excellent modeling robustness. The experimental results and justification for this concept and method will be better analyzed and interpreted in Section IV's experimental analysis.

E. Feature Fusion

A single feature vector has a problem with low model generalization ability in terms of characterization ability of behavioral information classification, whereas a large body of literature demonstrates that the concept and method of fusing multiple features can be extremely effective at enhancing the model's reliability. We chose parallel feature fusion over maximum fusion and concatenation fusion based

Authorized licensed use limited to: International Institute of Information Technology Bangalore. Downloaded on November 03,2024 at 09:19:47 UTC from IEEE Xplore. Restrictions apply.

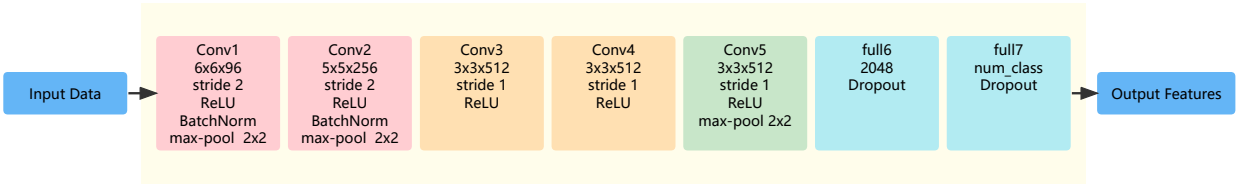

Fig. 2. Architecture of Spatial and Temporal Network

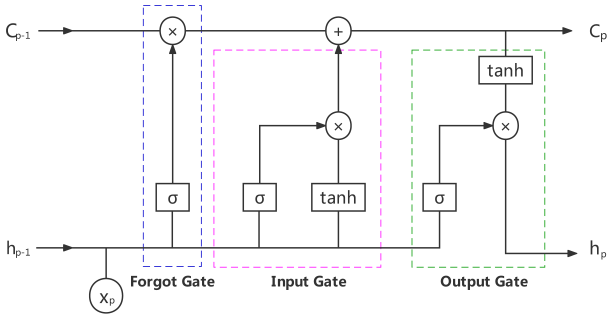

Fig. 3. Architecture of Long-Short Term Memory Unit

on experimental comparisons in this paper. Parallel feature fusion is significantly more effective than maximum fusion and concatenation fusion.

The spatial network, temporal network and sequential network are pretrained respectively, while the extracted spatial feature vector $\boldsymbol{f}_{spatial}$, temporal feature vector $\boldsymbol{f}_{temporal}$ and sequence feature vector $\boldsymbol{f}_{sequential}$ are the outputs after the last fully connected layer in the spatial domain, the time domain, and the sequential domain respectively. After parallel feature fusion, the fusion vector $\boldsymbol{f}_{fusion}$ is obtained as follows:

$$\boldsymbol{f}_{fusion} = \frac{\boldsymbol{f}_{spatial} + \boldsymbol{f}_{temporal} + \boldsymbol{f}_{sequential}}{3} \quad (1)$$

The $\boldsymbol{f}_{fusion}$ obtained after fusion is put into a Multi-Layer Perceptron (MLP) for training and testing. The recognition accuracy effect can be improved further by classifying the behavior through a layer of machine learning method on the fused feature information.

IV. Experiments

In this paper, we evaluate the proposed method on three publicly accessible datasets: UCF11, UCF50, and HMDB51. In the following subsections, we briefly describe these datasets and present the results of our experiments. We compare the classification accuracies of networks corresponding to spatial, temporal, and sequential streams with the classification accuracies of various combinations of these networks, as well as the classification accuracies of a method that fuses the three networks (the method we proposed in this paper). In the form of a confusion matrix, we also present the test performance of the proposed method on the three datasets. We conclude by comparing the performance of our method to that of other methods on the same dataset and providing our analysis and reflections.

A. Datasets and Implement Details

UCF11 dataset: This dataset was released in 2009 by the University of Central Florida (UCF) as a behavior recognition dataset. The dataset is a challenging behavioral classification dataset derived from publicly available Web video data with realistic environmental backgrounds and influences such as light masking. There are 11 mutually exclusive classification terms for dog walking, basketball shooting, cycling, diving, golfing, horseback riding, soccer playing, swinging, tennis playing, trampoline jumping, volleyball snapping, and swinging. There are 25 groups of videos under each behavior tag in all behavior categories, with each group containing at least four video clips and a resolution of 320×240.

UCF50 dataset: The UCF50 dataset is also a human action recognition dataset; the UCF released it in 2012. This dataset expands on UCF11 in terms of the number of videos and types of behaviors for which data was collected from the YouTube video website. This dataset is more difficult than the UCF11 dataset as it contains more video data and behavior categories. This dataset contains 50 behavior categories. Compared to the UCF11 dataset, which contains 1600 videos, the UCF50 dataset contains 6676 videos, with at least 100 video clips for each action, effectively increasing the complexity of behavior information. This dataset contains only videos with a resolution of 320×240.

HMDB51 dataset: This dataset is a multi-source human action dataset for complex scenes, released by Brown University in 2011. It acquires videos of real-world human behavior from publicly available films, YouTube, and Google videos. Therefore, the movie clips effectively increase the difficulty of behavior classification in this dataset, which is a challenging behavior dataset with higher requirements for model robustness and robustness. The dataset contains 51 behaviors and 6849 video clips, with at least 101 video clips corresponding to each behavior. This dataset's action categories can be roughly categorized as facial behavior and expressions, face-related action behaviors, common body movements, action behaviors interacting with objects, and action behaviors interacting with the human body. Similar to the two preceding datasets, the new dataset has a separate rate of 320×240.

The experiments were conducted using the Ubuntu 20.04 operating system and the deep learning environment PyTorch. This experiment utilizes an AMD R7 4800H CPU and an NVIDIA RTX3060 (laptop) 6 GB GPU. This model's learning

Authorized licensed use limited to: International Institute of Information Technology Bangalore. Downloaded on November 03,2024 at 09:19:47 UTC from IEEE Xplore. Restrictions apply.

rate is 0.000125 due to the impact of GPU video memory. In order to improve the generalization ability of the experimental model, a ten-fold cross-validation method is used to divide the dataset into a training set and a test set. The network is then trained and tested using the training set. The loss is computed using the cross-loss entropy function, and the Adam optimizer is employed.

B. Experimental Results and Analysis

The recognition accuracy rate is used as the evaluation standard in this paper's model evaluation, and the accuracy rate is:

$$Accuracy = \frac{TP}{Total} \times 100\% \tag{2}$$

$Total$ represents the total number of samples, while TP represents the number of correctly classified samples. The recognition precision has a more obvious and intuitive effect on the model's overall capacity for representation and generalization.

We first compare and evaluate the classification accuracy of the networks corresponding to each stream in the three-stream network, as well as the classification accuracy of various combinations of these networks for comparison, and then compare them to the classification accuracy of the three-stream fusion method proposed in this paper. In Tables I, II, and III, we present the performance results of each network, its mutual combinations, and our method on the UCF11, UCF50, and HMDB51 datasets.

TABLE I
PERFORMANCE OF DIFFERENT STREAMS ON UCF11 DATASET

Stream	Accuracy(%)
Spatial	80.37
Temporal	76.72
Sequential	82.16
Spatial+Temporal	96.71
Spatial+Sequential	97.82
Temporal+Sequential	96.43
All three stream	**99.17**

TABLE II
PERFORMANCE OF DIFFERENT STREAMS ON UCF50 DATASET

Stream	Accuracy(%)
Spatial	73.38
Temporal	69.89
Sequential	76.71
Spatial+Temporal	92.18
Spatial+Sequential	91.58
Temporal+Sequential	87.46
All three stream	**97.40**

From Tables I, II and III, it can be analyzed that the proposed three-stream network model combining temporal, spatial and sequential information significantly improves recognition classification accuracy and has significant performance advantages in multiple classification tasks, complex scenarios, and real-world scenarios compared to single network and two-stream network approaches.

TABLE III
PERFORMANCE OF DIFFERENT STREAMS ON HMDB51 DATASET

Stream	Accuracy(%)
Spatial	71.46
Temporal	66.85
Sequential	68.49
Spatial+Temporal	87.21
Spatial+Sequential	83.54
Temporal+Sequential	88.92
All three stream	**96.88**

We implemented and tested our proposed model on the UCF11, UCF50, and HMDB51 datasets, respectively, and plotted the confusion matrices as depicted in Figs. 6, 7, and 8. The confusion matrices can effectively and visually illustrate the relationship between the predicted and actual action categories of the model action classification.

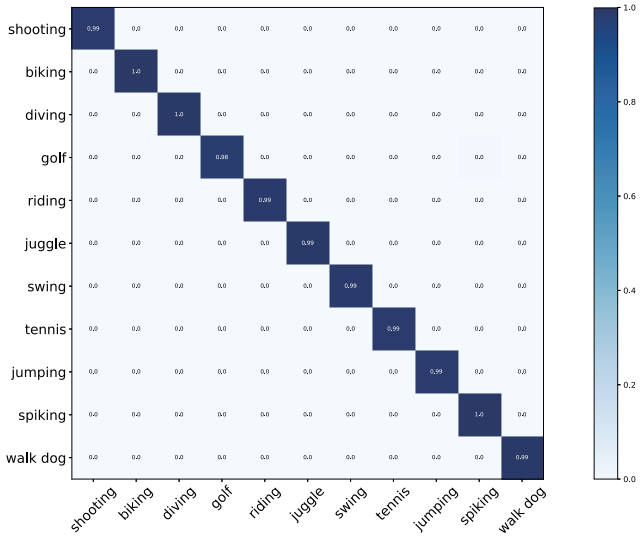

Fig. 4. Confusion Matrix for UCF11

Comparing the method presented in this paper to other 2D behavioral identification methods yields the results shown in Tables IV, V, and VI.

On benchmark datasets UCF11, UCF50, and HMDB51, our proposed methods outperform existing methods and depict their robustness. The method achieves an accuracy of 99.17% for UCF11, which is not only vastly superior to the traditional recognition methods in terms of recognition accuracy, but also vastly superior to the two-stream network represented by Two-stream LSTM (94.6%) and the multiple 3D convolutional complex network represented by MDF (97.7%). Combined Fig. 6 and Table V demonstrate performance improvements across all action behavior categories of UCF50, with a more discrete distribution of overall classification error cases and excellent overall correct classification.

Our model has a significant advantage over iDTs, a multiple-descriptor approach, in that it improves recognition

Authorized licensed use limited to: International Institute of Information Technology Bangalore. Downloaded on November 03,2024 at 09:19:47 UTC from IEEE Xplore. Restrictions apply.

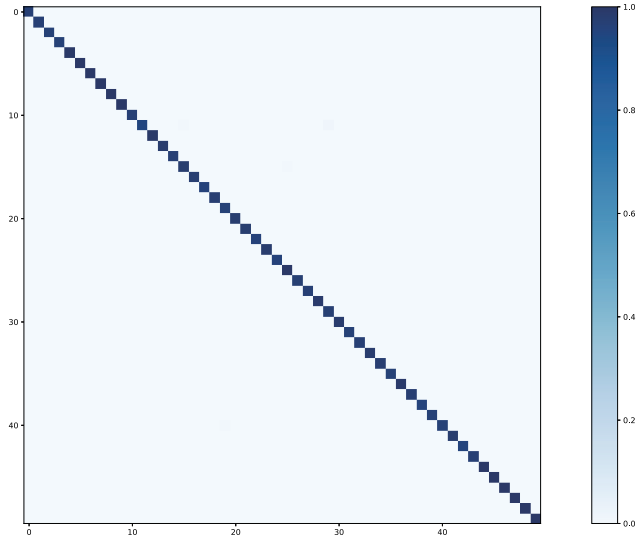

Fig. 5. Confusion Matrix for UCF50

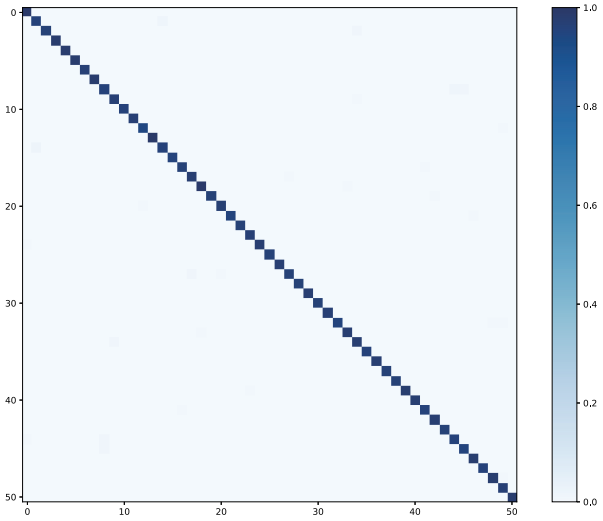

Fig. 6. Confusion Matrix for HMDB51

TABLE IV
ACCURACY (%) COMPARISON BETWEEN OUR PROPOSED METHOD WITH MODERN METHODS ON UCF11 DATASETS

Method	Accuracy(%)
OF+HOG3D [15]	72.7
ColorSTIP(HOG3D) [16]	78.6
Soft attention [17]	84.9
Sig min-Hash [18]	86.7
Naïve VR + VRTD HOG [19]	89.27
Snippets [20]	89.5
H-VLBP+ Two layer FFNN [21]	90.2
HOG+HOF+MBH+SVM [22]	91.47
Two-stream LSTM [23]	94.6
BoE [24]	96.68
MDF:MD3DCNN+4LSTM [25]	97.7
Ours	**99.17**

accuracy by nearly 5%. It is also superior to the most recent LDM+MIFS method by more than 1.5%. The method presented in this paper demonstrates a strong capacity for generalization on both small classification targets and multi-classification targets. The largest accuracy improvement is observed on HMDB51, with a 40% improvement compared to the traditional method HOG+HOF+MBH+ELM composite; and with the more popular previous methods Two-stream I3D, RepFlow-50 (R[2+1]D), and the most recent depth synthesis model DEEP-HAL with ODF+SDF improved 15.98%, 15.78%, and 9.32%.

TABLE V
ACCURACY (%) COMPARISON BETWEEN OUR PROPOSED METHOD WITH MODERN METHODS ON UCF50 DATASETS

Method	Accuracy(%)
LDM+HOG/HOF [26]	49.6
BoF+HOG [27]	67.8
MIP Encoding [28]	72.68
distance transform+entropy[29]	79.84
STFP [30]	82.1
LSSVM [31]	88.04
Yang et al. [32]	89.2
iDTs [33]	92.07
LDM+IDT [26]	92.8
BoE [24]	93.42
LDM+MIFS [26]	95.0
Ours	**97.40**

TABLE VI
ACCURACY (%) COMPARISON BETWEEN OUR PROPOSED METHOD WITH MODERN METHODS ON HMDB51 DATASETS

Method	Accuracy(%)
LDM+HOG/HOF [26]	24.4
HOG+HOF+MBH+ELM [22]	56.47
Two-Stream [9]	59.4
LDM+MIFS [26]	66.2
Two-Stream(VGG16) [34]	65.4
VidTr-L [35]	74.4
Hidden Two-Stream [36]	78.7
D3D + D3D [37]	80.5
Two-stream I3D [38]	80.9
RepFlow-50(R[2+1]D) [39]	81.1
R2+1D-BERT [40]	85.10
SCK(I3D) [41]	86.11
DEEP-HAL with ODF+SDF (I3D) [42]	87.56
Ours	**96.88**

As shown in Table VI, our proposed method achieves state-of-the-art performance on HMDB51 in multi-classified complex scenes and accounts for the impact of film clips.

We believe that the model presented in this paper can achieve such excellent results on the extremely difficult HMDB51 dataset without the extraction of frame sequence features, which effectively obtains the human actions under the influence of clips with behavioral coherence features and helps the model to become more robust.

V. CONCLUSION

In this paper, a practical and effective multi-feature fusion network model for HAR is proposed. Different features are extracted from spatial information, temporal information, and frame sequence information, respectively, in the framework

Authorized licensed use limited to: International Institute of Information Technology Bangalore. Downloaded on November 03,2024 at 09:19:47 UTC from IEEE Xplore. Restrictions apply.

developed in this paper. After extracting the three networks and their respective characteristics, the characteristics are merged in series by simple addition. This approach to feature fusion permits the effective fusion of each component's features. By using the weights of the multilayer perceptron to output classification results during the classification stage, the model can achieve more accurate classification results. Lastly, in the evaluation of three publicly accessible datasets in real-world settings, our method achieves a higher level of excellence.

ACKNOWLEDGMENT

The paper is supported by "the Doctoral Scientific Research Foundation (2020QDZ24)", "Key Research Project of Natural Science in Anhui Province (KJ2020A0471, KJ2021A0622, 202003a05020031)" and "University Synergy Innovation Program of Anhui Province (GXXT-2021-028)".

Especially, Ruifeng Huang would like to thank his girlfriend who silently help him without any interruption for 25 years.

REFERENCES

[1] J. Liu, L. Jiebo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 1996–2003.

[2] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[3] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, Conference Proceedings, pp. 2556–2563.

[4] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[5] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[6] C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151.

[7] I. Sipiran and B. Bustos, "Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes," *The Visual Computer*, vol. 27, no. 11, p. 963, 2011.

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 1725–1732.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 568–576.

[10] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[11] X. Wang, L. Gao, J. Song, and H. T. Shen, "Beyond frame-level cnn: Saliency-aware 3-d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*, vol. 24, pp. 510–514, 2017.

[12] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream cnn," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[13] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, J. Bigun and T. Gustavsson, Eds. Springer Berlin Heidelberg, Conference Proceedings, pp. 363–370.

[14] J. Teuwen and N. Moriakov, *Convolutional neural networks*, 2020, pp. 481–501.

[15] V. F. Mota, E. A. Perez, L. M. Maciel, M. B. Vieira, and P. H. Gosselin, "A tensor motion descriptor based on histograms of gradients and optical flow," *Pattern Recognition Letters*, vol. 39, pp. 85–91, 2014.

[16] I. Everts, J. C. van Gemert, and T. Gevers, "Evaluation of color spatio-temporal interest points for human action recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1569–1580, 2014.

[17] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv [cs.LG]*.

[18] A. Gilbert and R. Bowden, "Image and video mining through online learning," *Computer Vision and Image Understanding*, vol. 158, pp. 72–84, 2017.

[19] T. P. Moreira, D. Menotti, and H. Pedrini, "Video action recognition based on visual rhythm representation," *Journal of Visual Communication and Image Representation*, vol. 71, 2020.

[20] M. Ravanbakhsh, H. Mousavi, M. Rastegari, V. Murino, and L. S. Davis, "Action recognition with image based cnn features," *arXiv [cs.CV]*.

[21] K. Kiruba, E. D. Shiloah, and R. R. C. Sunil, "Hexagonal volume local binary pattern (h-vlbp) with deep stacked autoencoder for human action recognition," *Cognitive Systems Research*, vol. 58, pp. 71–93, 2019.

[22] H. Xu, Q. Tian, Z. Wang, and J. Wu, "A joint evaluation of different dimensionality reduction techniques, fusion and learning methods for action recognition," *Neurocomputing*, vol. 214, pp. 329–339, 2016.

[23] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Conference Proceedings, pp. 177–186.

[24] S. Nazir, M. H. Yousaf, J. C. Nebel, and S. A. Velastin, "A bag of expression framework for improved human action recognition," *Pattern Recognition Letters*, vol. 103,

Authorized licensed use limited to: International Institute of Information Technology Bangalore. Downloaded on November 03,2024 at 09:19:47 UTC from IEEE Xplore. Restrictions apply.

pp. 39–45, 2018.

[25] H. Ou and J. Sun, "The multidimensional motion features of spatial depth feature maps: An effective motion information representation method for video-based action recognition," *Mathematical Problems in Engineering*, vol. 2021, p. 6670087, 2021.

[26] T. Wang, C. Liu, and L. Wang, "Action recognition by latent duration model," *Neurocomputing*, vol. 273, pp. 111–119, 2018.

[27] F. Shi, R. Laganière, and E. Petriu, "Local part model for action recognition," *Image and Vision Computing*, vol. 46, pp. 18–28, 2016.

[28] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," ser. Computer Vision - ECCV 2012. Springer Berlin Heidelberg, Conference Proceedings, pp. 256–269.

[29] P. Ramya and R. Rajeswari, "Human action recognition using distance transform and entropy based features," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 8147–8173, 2021.

[30] H. F. Hu, Z. K. Liao, and X. Xiao, "Action recognition using multiple pooling strategies of cnn features," *Neural Processing Letters*, vol. 50, no. 1, pp. 379–396, 2019.

[31] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural svm," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 23, no. 8, pp. 1422–1431, 2013.

[32] Y. H. Yang, R. S. Liu, C. Deng, and X. B. Gao, "Multi-task human action recognition via exploring super-category," *Signal Processing*, vol. 124, pp. 36–44, 2016.

[33] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.

[34] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Conference Proceedings, pp. 1933–1941.

[35] Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, H. Chen, I. Marsic, and J. Tighe, "Vidtr: Video transformer without convolutions," *arXiv [cs.CV]*, 2021.

[36] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Computer Vision - ACCV 2018*, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Springer International Publishing, Conference Proceedings, pp. 363–378.

[37] J. C. Stroud, D. A. Ross, C. Sun, J. Deng, and R. Sukthankar, "D3d: Distilled 3d networks for video action recognition," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Conference Proceedings, pp. 614–623.

[38] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Conference Proceedings, pp. 4724–4733.

[39] A. Piergiovanni and M. S. Ryoo, "Representation flow for action recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, Conference Proceedings, pp. 9937–9945.

[40] M. E. Kalfaoglu, S. Kalkan, and A. A. Alatan, "Late temporal modeling in 3d cnn architectures with bert for action recognition," ser. Computer Vision - ECCV 2020 Workshops. Springer International Publishing, Conference Proceedings, pp. 731–747.

[41] P. Koniusz, L. Wang, and A. Cherian, "Tensor representations for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 648–665, 2022.

[42] L. Wang and P. Koniusz, *Self-supervising Action Recognition by Statistical Moment and Subspace Descriptors*. Association for Computing Machinery, 2021, pp. 4324–4333.

Authorized licensed use limited to: International Institute of Information Technology Bangalore. Downloaded on November 03,2024 at 09:19:47 UTC from IEEE Xplore. Restrictions apply.