IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 September 2023, accepted 14 October 2023, date of publication 18 October 2023, date of current version 26 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325401

RESEARCH ARTICLE

Feature Fusion for Dual-Stream Cooperative Action Recognition

DONG CHEN[1,2], MENGTAO WU[2], TAO ZHANG[2], AND CHUANQI LI[1,2]
[1]College of Computer Science and Engineering, Guangxi Normal University, Guilin 541004, China
[2]College of Physics and Electronic Engineering, Nanning Normal University, Nanning 530001, China

Corresponding author: Chuanqi Li (lcq@mailbox.gxnu.edu.cn)

This work was supported in part by the Innovation Project of Guangxi Graduate Education under Grant JGY2021164.

ABSTRACT Currently, the primary methods for action recognition involve RGB-based approaches, pose-based approaches (e.g., skeleton coordinates), and multi-stream fusion methods. In this paper, we propose a novel action recognition framework based on both RGB images and motion pose images to enhance the accuracy of action recognition in videos. As a single feature representation fail to effectively capture motion trends and image variation information, it cannot accurately reflect expected action judgments in real-world scenarios. Therefore, we utilize the appearance features of video frames and the motion variation features of the subject, aiming to cooperate the action itself with appearance information for precise action recognition. We construct video representations based on local spatiotemporal features and global features, and utilize the ResNet backbone network and Temporal Shift Module (TSM) to extract action representations from multi-stream information. Driven by the motion features, the fusion of multi-stream information achieves effective expression of motion features. Experimental results on public datasets demonstrate the effectiveness of our proposed method. It achieves competitive performance compared to state-of-the-art techniques while maintaining a less complex and more interpretable model. Overall, our approach demonstrates superior effectiveness.

INDEX TERMS Cooperative action recognition, feature fusion, spatial-temporal information.

I. INTRODUCTION

Action recognition is a fundamental task within the domain of video understanding, aiming to comprehend and categorize actions performed by individuals in videos. In recent years, the field has witnessed notable progress due to the rapid advancements in deep learning techniques for computer vision. Specifically, several deep learning models and approaches relying on image-based representations have been proposed and optimized. These developments in image-based methods have not only pushed the boundaries of video understanding but also substantially improved action recognition leveraging deep learning models.

Among deep learning approaches for video action recognition, the methods focusing on learning video timing information initially achieved the best results in action recognition tasks [1], [2], [3]. The Dual Stream Network introduces a new framework for video understanding, which represents and learns separately from time and space, reaching new heights in video understanding. With the advent of the 3D convolution method [4], [5], [6], a fresh perspective on video comprehension has emerged, eliminating the need to separately model time and space. In RGB-based methods, 3D models [4], [7], [8] identify temporal information between images while simultaneously extracting features from individual frames to facilitate video comprehension. These approaches effectively preserve the comprehensive information encoded in videos and achieve commendable performance. However, the utilization of complex models and the abundance of redundant information significantly increase the computational cost of inference.

To address these challenges, the utilization of dual-stream models or spatiotemporal convolutional networks, which leverage both RGB frames and optical flow images, has yielded superior results compared to 3D models based methodologies. However, a significant drawback of dual-stream networks lies in the necessity of introducing supplementary preprocessing steps for optical flow computation. Moreover, this computation introduces noise that adversely affects model performance, extends training times,

The associate editor coordinating the review of this manuscript and approving it for publication was Tianhua Xu.

and escalates computational complexity. Recently, certain approaches [9], [10], [11], [12] have garnered research attention by utilizing pseudo-images or video transformation techniques [13], [14] to construct video action representations have attracted researchers' attention. These methods substantially reduce the generation of redundant information while achieving exceptional results.

Numerous studies have demonstrated that the incorporation of optical flow information in dual-stream networks can more precisely represent motion information. Recent works [15], [16], [17], [18] explored the replacement of optical flow with human skeleton joint coordinates, encoding pose motion information in various manners. These methods substantially reduce the cost and computational overhead associated with optical flow extraction, while concurrently achieving superior results. Although these motion information encoding methods in skeleton-based recognition approaches slightly lag behind state-of-the-art graph convolutional neural network methods [19], [20], [21], the fusion of information from distinct modalities, such as RGB, optical flow, and skeleton, can mutually complement each other. Hence, the effective integration of these modalities can enhance the accuracy of action recognition. In some cases, information fusion techniques yield comparable or even superior results compared to graph convolutional methods, while also exhibiting enhanced robustness, rendering them more suitable for practical video understanding requirements.

In this paper, we propose a novel approach termed FFDC(Feature Fusion for Dual-Stream Cooperative) that combines image and pose data. For the RGB image stream, we employ a video compression technique to retain partial appearance information, whereas in the pose stream, we eliminate out the background interference and solely focus on the human actions by extracting skeletal coordinates using a human pose estimation method. This approach enables the model to concentrate on spatial pose variations without being influenced by the background information.

Furthermore, we introduce a color variation scheme in the pose representation, enabling the model to capture the dynamic changes in poses. By combining these two streams of information, we effectively preserve the background context while highlighting the primary actions. The experimental results demonstrate the efficacy of this approach.

In our model selection, we opt for the ResNet architecture due to its advantageous cross-layer connectivity within residual blocks. This connectivity facilitates the more direct propagation of gradients to earlier layers, effectively addressing the gradient vanishing problem. Consequently, ResNet enables the training of very deep convolutional neural networks, leading to strong performance across various computer vision tasks. As a result, ResNet can be pretrained as a pretrained model on large scale datasets and then fine-tuned or migrated to learn on different tasks. This capability significantly simplifies the development of high-performance models, even when working with limited datasets. In our approach, we specifically utilize the ResNet50 [22] as the backbone network. To capture temporal features, we introduce the Temporal Shift Module (TSM) [23] for extracting inter-frame variations. Additionally, FFDC leverages the pre-trained weights of ResNet50 on the ImageNet dataset. This approach not only expedites the model's convergence but also enhances its recognition performance.

As FFDC takes regular RGB images as input, it can better leverage the advancements in state-of-the-art network models for images. Moreover, the pose features can accommodate different forms of skeletal information, and the fusion of dual streams can easily incorporate other types of multi-stream information. To validate the effectiveness of FFDC, we conducted tests on the HMDB51 and JHMDB datasets, and the results demonstrated that the proposed method outperformed the baseline model, achieving superior performance.

Our contributions can be summarized as follows:

1. We design a novel dual-stream information fusion architecture that integrates information from multiple streams. This architecture combines video RGB frame data with subject motion information, employing a motion-driven fusion strategy. This design of the dual-stream structure effectively improves the robustness and recognition accuracy of the model.

2. We proposed a novel approach to skeleton design that maps spatial information about a character onto an RGB image and implicitly encodes information about movement changes through colour. Unlike heatmaps, which are affected by local joint points, this method emphasises the action as a whole and achieves the best performance among skeleton-based representations, effectively improving the accuracy of action recognition.

3. We propose a representation of compressed video information that eliminates redundant information while maintaining accuracy, thereby reducing the running cost and running time of the model and improving inference speed.

II. RELATED WORKS

A. RGB-BASED ACTION RECOGNITION

RGB-based action recognition methods are widely employed for the preprocessing of video frame sequences. However, distinct strategies are adopted for different models, including LSTM (RNN), 2DCNN, and 3DCNN. Considering the presence of temporal dynamics within videos, A proposed approach [24] focus on sequence models. This approach decomposes video sequences into multiple sub-sequences of varying lengths, utilizing them as inputs to LSTM in order to capture long-term temporal information. A novel technique [19] termed as Spatio-Temporal LSTM (ST-LSTM) is introduced for 3D human action recognition. ST-LSTM is utilized to model the spatio-temporal information present in video frame sequences, while a mechanism referred to as the ''Trust Gate'' is incorporated to dynamically learn the weights of each LSTM unit. This facilitates enhanced feature capturing within video sequences. Moreover, a dual-stream model combines image frames and optical flow, along with

an improved Inception network, extending 2D convolutions to learn temporal information through 3D convolutions. These 3D models [4], [7] have demonstrated exceptional performance on diverse datasets. Another approach [25] suggests the utilization of two branches of networks with different speeds, where the slower network processes spatial information in the video, while the faster network solely focuses on temporal information. This approach enables more accurate action recognition by effectively handling video sequence information. However, the aforementioned methods necessitate the processing of every frame in the video, leading to increased computational costs and overhead due to the substantial number of images and model parameters involved. Additionally, these methods introduce redundant information, which not only increases the inference cost but also impacts the accuracy of the model.

Recent studies [26], [27], [28] have proposed methods that extract temporal information while compressing video features, thereby reducing model inference costs and eliminating redundant information. For instance, [29] acknowledges that redundant information in videos can overshadow the ''true'' signal, compressing videos as inputs, taking into account the inter-frame correlation. The motion vectors in compressed videos allow the model to focus on the motion information in the video RGB images. Reference [30] firstly introduces the concept of dynamic images, extracting motion relationships between consecutive frames and representing them as a fused two-dimensional image using channel fusion. Dynamic images contain both spatial and temporal information, rendering them suitable for action recognition in video sequences. However, the stacking of multiple images introduces some noise and errors, which compromise the inherent information of the images and pose challenges in capturing long-term temporal dynamics. Certain studies [31], [32] aim to capture the underlying structural information of video sequences to acquire temporal information. This entails learning both the short-term actions of the subjects and capturing the long-term temporal dynamics within videos, thereby encapsulating the latent temporal structure of video sequences.

Our approach enables the fusion of a limited number of video frames at the channel level, accomplishing the dual objective of filtering redundant information and preserving the image's integral structure. Additionally, employing time-shift operations on the fused image serves the same purpose of addressing both short-term and long-term temporal dynamics.

B. SKELETON POSE-BASED ACTION RECOGNITION

Graph convolutional networks (GCNs) are a natural fit for skeleton data, considering their ability to encode skeletal movements in graph format, thereby fully exploiting the motion information within the skeleton. The ST-GCN [19] model represents a graph network model that integrates spatial and temporal information in graph convolutions, and it has exhibited remarkable success in skeleton-based action recognition. Nonetheless, the robustness and scalability of graph convolutions necessitate further exploration.

Recent works focusing on skeleton-based approaches have proposed various encoding methods to represent skeleton information as heatmaps or matrices [9], [11], which involve aggregating these probability representations over the temporal dimension and assigning distinct colors to each frame to denote their temporal order. These heatmaps are well-suited for shallow neural networks used in action classification and demonstrate superior performance when combined with the dual-stream I3D method [10].

Another alternative approach [29], [30] involves utilizing 3D trajectories of skeleton joints for 3D action recognition. The skeleton sequences are divided into three segments, with each segment containing skeleton sequences of different frames separated by the same channel, providing spatial information of the skeleton. The devised representation method for skeleton joint sequences, specifically designed for the 3D task, incorporates motion information to capture temporal dynamics and motion modeling, enabling the model to better understand the underlying action dynamics.

To carefully factor in the global features of the joints and the posture changes of the bones across different frames, a structured representation of the action is constructed through key point information, and the multi-level features are jointly identified to improve the robustness and generalization ability of the model [31], [32]. A sequence-based view-invariant approach [33] that eliminates the impact of viewpoint variations on the spatiotemporal positions of skeleton joints. They visualize skeleton data as a series of colored images, implicitly encoding spatiotemporal information, and apply visual and motion enhancement techniques to amplify local patterns, thereby significantly improving pose robustness and discriminative features.

In the aforementioned skeleton-based action recognition methods, the skeleton serves as a source of spatial information and motion representation, effectively capturing the underlying actions. However, there might still be some loss of information in the encoding format of the skeleton. Our proposed pose design encodes skeleton data into images, fully restoring spatial information, while the color variations between frames guide the model in learning temporal dynamics. Our approach, based on a 2DCNN, readily incorporates structured representations of actions such as Cartesian coordinates, motion vectors, and others, to enhance the multidimensional representation of movements.

C. POSTURE AND APPEARANCE FUSION COOPERATIVE RECOGNITION

Research has shown that methods based on dual-stream architectures exhibit superior performance. Notably, PA3D [8] and RPAN [34] have garnered attention for their ability to learn representations of action sequences and skeleton poses by leveraging optical flow and RGB images, thereby

capturing the spatiotemporal evolution of actions. The fusion of pose and action representations is crucial for accurate video classification. Similar approaches [35], [36] have been proposed, wherein spatial streams extract information from individual frames while temporal streams capture information between consecutive frames. Nevertheless, these approaches prioritize the integration of pose and motion information to guide the learning of motion-related features. To address this limitation, a pose-driven feature integration method [37] has been introduced, employing a forget gate module to effectively combine human pose and image appearance information. This approach mitigates misclassifications arising from incomplete pose information or image backgrounds. Additionally, the method [9] proposed in for pose representation necessitates high-quality skeleton data, while the utilization of the I3D model [4] is reserved for processing RGB images. Notably, combining these two methods [10] yields significantly improved results compared to the use of a single stream alone.

In our proposed dual-stream network method, we aim to fuse RGB image and skeleton image features by leveraging posture-driven fusion. Specifically, we assign higher weights to corresponding poses in the pose stream when it performs well, facilitating accurate and efficient action recognition. Conversely, when occlusions or missing pose information occurs, the image stream is relied upon to extract motion expressions. This pose-driven fusion approach enhances performance and robustness.

III. ALGORITHM FRAMEWORK

Drawing on the design principles discussed in relevant studies on action recognition models, our objective is to develop a simple, scalable, effective, and robust action recognition model. In this paper, we introduce FFDC, a dual-stream network model that incorporates temporal modeling to capture contextual temporal information in both images and poses. Additionally, we integrate RGB stream information with the spatial pose stream to obtain a comprehensive representation of actions in videos.

Given a video, we partition it into RGB video frames, which are processed to generate appearance sequence images (denoted as A). Additionally, the spatial and motion information of individuals in the video is expressed as pose sequence images (denoted as P). These images, denoted as A and P, respectively, are inputted into separate network models to extract motion features. To facilitate feature selection, self-attention is applied to derive attention weights for the features in P. Subsequently, the two sets of features are combined to create an integrated action representation. Finally, a fully connected layer and a softmax layer are employed to output the final action label (C).

By leveraging the temporal context and fusing information from both the RGB stream and spatial pose stream, FFDC aims to capture the discriminative features for action recognition in a simplified and scalable manner, eventually achieving high effectiveness and robustness.

A. APPEARANCE STREAM

For the input of the appearance stream, we extract video clips into T video frames. For these T RGB images, we extract the temporal information of the images (refer to Fig. 1), resulting in an image sequence $A \epsilon R_T \times 3 \times H_A \times W_A$, where T, H_A, and W_A represent the number of frames, height, and width of the images, respectively. Considering the size of T, we divide the T RGB images into T/n groups of image sequences, where each group contains n images and $n \epsilon (0, T]$. Further, we then separate each group of images based on channels. Within each group, individual images are assigned different weights according to the time order, and the channel values are selected and concatenated based on these weights. This process results in a complete image with three channels, denoted as F_A, which represents the temporal information of the group of images.

The recognition of the appearance stream is based on image recognition, however, it differs from recognizing a single image as the sampled sample consists of multiple images from a video. In addition to extracting spatial information at different time points is obtained from individual frames, while temporal information between images is also extracted. To represent the temporal information, We utilize varying numbers of images, denoted as n, Detailed experimental results can be found in Table 1. In our model, we employ ResNet50 as the backbone network and introduce the Temporal Shift Module (TSM) to learn the temporal information of the images. Here, we provide a concise explanation of the principle behind the Temporal Shift Module:

Let us first establish the notion that data movement and computation can be decoupled in convolution. Consider a normal convolution operation, specifically a 1-D convolution with a kernel size of 3 as an example. Suppose the convolution has a weight $W = (w_1, w_2, w_3)$ and the input X is an infinite-length one-dimensional vector. The convolution operator $Y = Con(W, X)$ can be expressed as $Y_i = w_1 x_{i-1} + w_2 x_i + w_3 x_{i+1}$. We can separate the convolution operation into two steps: shifting and multiplicative accumulation.Firstly, we shift the input X by -1, 0, $+1$ and multiply them by w_1, w_2, w_3 respectively. Then, we sum the results to obtain Y. Formally, the shift operation is defined as follows:

$$X_i^{-1} = X_{i-1}, X_i^0 = X_i, X_i^{+1} = X_{i+1}. \tag{1}$$

The cumulative addition operation is given by:

$$Y = w_1 X^{-1} + w_2 X^0 + w_3 X^{+1} \tag{2}$$

The first shift step can be performed without any multiplication. While the second step is more computationally expensive, the Temporal Shift module integrates the multiplicative accumulation into the 2D convolution, thereby incurring no additional cost compared to the 2D CNN-based model.

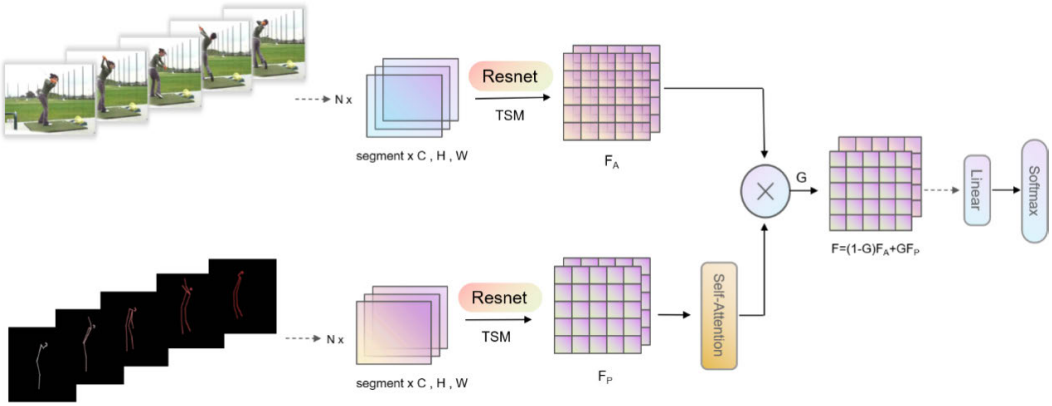


FIGURE 1. Overall framework. For a sample comprising T frames of RGB images and T frames of pose images extracted by the pose estimation module, these frames are encoded and normalized into input tensors of the model [N, $segment * C$, H, W]. Here, N denotes the batch size, and segment represents the proportion of image channel transfer. The feature tensor undergoes ResNet for extracting image features, while the TSM module is used to extract temporal information. The model features of the dual-stream network are merged using attention scores to obtain the ultimate motion representation feature, which is used for category prediction.

B. POSTURE STREAM

Traditional representations of skeleton joint sequences mainly focus on the spatial information of the joints, including their coordinates and orientations. However, these representations may not fully exploit the crucial temporal dynamics and motion patterns essential for accurate action recognition. In the processing of the pose stream, we employ a human pose estimation method to extract keypoint information from the human body, which serves as a critical preprocessing step for skeleton-based action recognition. The extraction of human skeleton or pose significantly impacts the final recognition accuracy. Considering that depth information is typically absent in video data and the quality of depth information obtained from pose estimation methods is often suboptimal, the quality of 2D poses on the two-dimensional image plane tends to be superior to 3D poses. We visualize the coordinates of each joint in each frame as a trajectory of human body motion, resulting in the pose sequence $P \epsilon R_T \times 3 \times H_P \times W_P$, where H_A and W_A represent the height and width of the image, respectively. After feature extraction, we obtain pose features F_P. To facilitate better learning, we assign distinct colors and transparency to different skeletal joints to guide the model.Experimental results have demonstrated the effectiveness of this approach(Table 2).

C. FEATURE FUSION

To obtain robust features F utilized for accurate classification by fusing the pose features F_P and appearance features F_A, we utilize the pre-trained parameters of the ResNet50 model on ImageNet [38] for our model fine-tuning. The preprocessed features from the same video are simultaneously input into distinct models, resulting in two different feature representations. In consideration of the differences between appearance and pose, feature fusion is performed within the last two layers. Cooperative feature fusion: Due to the potential presence of noise within the features obtained from the appearance stream, while the pose stream solely includes action-related information, the integration of appearance information is determined based on the reliability of the pose stream. To this end, we use attention mechanism to control the weights of pose and appearance, dynamically adjusting their contributions. By learning the pose self-attention weights, the model can automatically focus on the primary classification derived from the current pose features and dynamically integrate appearance information via weighted fusion of pose and appearance. Finally, fusion is accomplished through element-wise addition of the features from appearance and pose, generating the integrated action features $F \epsilon R_T \times C$, as follows:

$$F = GF_A + (1 - G)F_P. \tag{3}$$

This integrated feature is utilized for the final action classification. Thus, if the pose features offer adequate information for action recognition, the proposed fusion method prioritizes them, thereby circumventing the potential influence of strong contextual cues stemming from the appearance features. Alternatively, in the absence of sufficient information from the pose features, more contextual information from the appearance features is leveraged.

IV. EXPERIMENT

In order to evaluate the proposed model in this paper, we conducted experiments were conducted on the publicly available datasets JHMDB [39] and HMDB-51 [40]. All skeleton data were estimated using the Mediapipe's pose module [41] from the raw videos, following a top-down pipeline. All RGB data were directly extracted from the videos.

A. DATASETS

JHMDB: The JHMDB consists of 21 human action categories. Each video frame maintains the fixed frame rate of 30 FPS and an image resolution of 320 × 240. JHMDB

comprises videos sourced from YouTube, with each video featuring one single action.

HMDB-51: The HMDB-51 comprises 6.7k videos distributed across 51 categories, wherein each category containing includes no less than 101 videos. The videos exhibit varying lengths, ranging from 40 to 400 frames.

B. ABLATION EXPERIMENTS ON BRANCH NETWORKS

To demonstrate the effectiveness of the FFDC, we evaluated the performance of the pose stream features (F_P) and appearance stream features (F_A) was evaluated employing diverse design approaches.

For the pose image design, the pose heatmap encoding approach Potion [9] has also been frequently used in subsequent action recognition works [10], [16]. Potion has demonstrated its ability to enhance the understanding of human actions by effectively integrating spatial and temporal information. However, it exhibits limitations in modeling complex actions, primarily concentrating on heatmaps of joints, which are susceptible to the accuracy of pose estimation results. In contrast, our proposed pose coding methodology offers a representation of the entire figure's pose to the model. This approach enhances robustness, enabling accurate judgments even in scenarios involving partial occlusion. It also better preserves the comprehensive spatial information of the pose and maintains temporal consistency by strategically controlling color changes to convey timing information. Our design draws inspiration from the pose proposed in [34], wherein they propose using pose attention mechanisms are employed to encode distinct body parts using identical colors.

Here, we present two design options. The design concept for Pose I involves assigning identical colors to corresponding joints across all frames, thereby rendering the model incapable of distinguishing the mirrored relationship of poses. Each color represents a specific joint. Moreover, we enable the ResNet and TSM modules to learn temporal information. For Pose II, the design approach invovles assigning the same color to all skeletons within a unified frame, with colors varying across frames, thereby generating a gradient transition over time, as depicted in Fig.2. The experimental results of both methods are depicted in Table 1, validating the superiority of the Pose II design approach. Therefore, we proceed with this design approach in the subsequent experiments.

Various approaches have been proposed to address the issue of redundant information in videos has been addressed through various approaches, such as the direct fusion of multiple images across different channels or the compression of video information. These methods have demonstrated their effectiveness. In our study, we compare the effectiveness of different temporal modeling methods on the RGB stream. The fusion of image channels weighted by time to generate a new image with temporal information produces diverse effects on model accuracy. When a large number of image channels are fused to construct a temporal image, the original image structure may be compromised, leading to lower recognition accuracy. Conversely, stacking a small number of images may not provide sufficient temporal information to fully capture motion dynamics, thus impacting the model's judgment.

TABLE 1. Due to a significant portion of occlusion, and the subsequently loss of quality of the pose data, it remains challenging to convincingly assess this effect. To address this limitation, the experiment focuses exclusively on the JHMDB for its superior quality of human key point data to evaluate the effect of two distinct pose design methods. The reported experimental results represent the average accuracy obtained through multiple iterations of the experiments.

Methods	JHMDB
Chined(RGB+Flow +Pose)[6]	56.8
PA3D[8]	60.1
EHPI[11]	60.5
PoTion[9]	67.9
JMRN[16]	68.55
POSEI	68.98
POSEII	70.11

TABLE 2. A comparison of the effects of different models and different frame numbers T on the JHMDB and HMDB51 datasets. The models in the table refer to the ResNet50 model after pre-training on ImageNet, respectively. Additionally, We also compare the effect of Vision transform (ViT) [42] on different datasets after pre-training.

Methods	JHMDB	HMDB51
ResNet+TSM(T = 0)	86.56	80.50
ResNet+TSM(T = 2)	85.48	80.47
ViT(T = 2)	77.67	70.18
ViT(T = 8)	76.71	70.03
ResNet(T = 8)	61.75	50.69

The Temporal Shift Module (TSM) is a technique specifically designed for temporal modeling. By incorporating the temporal residual connections, TSM preserves the spatial structure of the original image while integrating temporal information. This approach proves highly effective in learning spatiotemporal information within the dual-stream framework. We evaluated the performance of different frames numbers(T) in our experiments, and the results are presented in Table 2. Notably, due to the JHMDB only consisting of short videos with fewer than 50 frames, while the HMDB51 comprising longer videos with hundreds of frames, our method significantly reduces the number of images, thereby reducing the model's learning time for redundant information, and the learning time for the model. However, even in relatively longer videos, the compressed frame representation achieves results comparable to the uncompressed representation.

C. COOPERATIVE FEATURE FUSION

FFDC employs feature inputs from both RGB streams and pose streams. The ResNet50 utilizes pretrained model parameters from ImageNet to dynamically fuse the appearance stream information features based on the recognition effect of pose streams. Table 3 illustrates the results obtained across different datasets. Notably, our approach outperforms and exhibits faster performance compared to the multi-stream model that utilizes 3D convolution. The recent 3D model-based work have also incorporated RGB images and pose features as inputs. For instance, PoseC3D [10] utilizes an enhanced C3D model to extract spatio-temporal features from

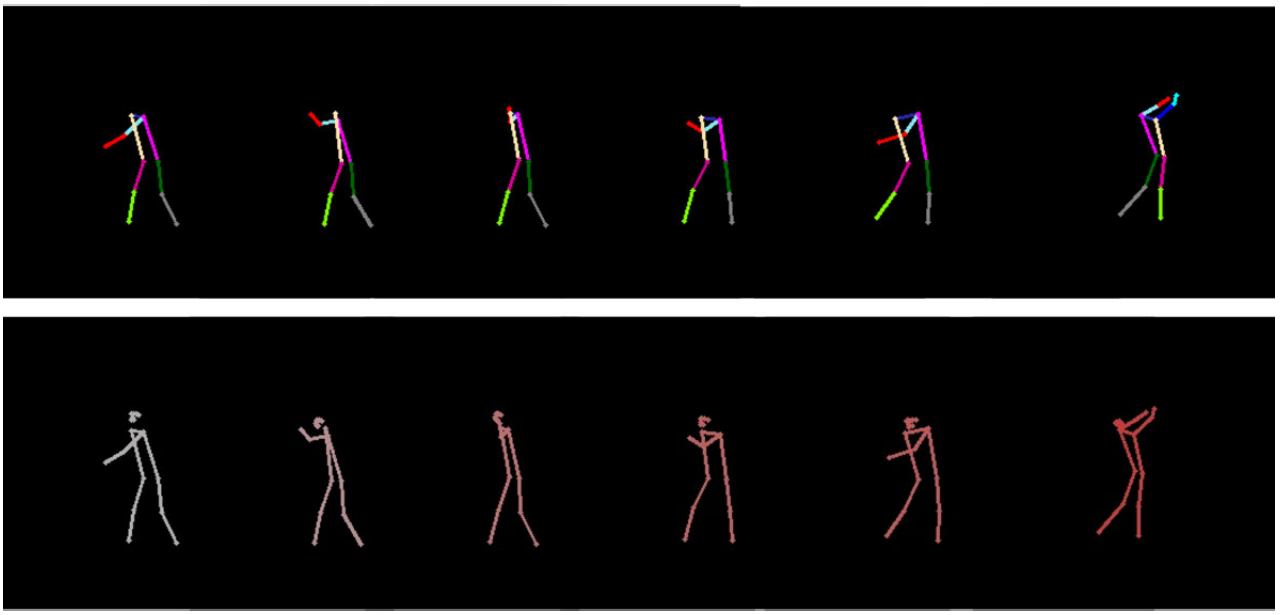


FIGURE 2. Different skeletal poses designed for model implicit learning, the upper part is Pose, each joint is represented by a different color, and the lower part is Pose, one color is used to represent one frame, and the color changes in time order before and after.

TABLE 3. Comparison of dual-stream network methods. Accuracy under JHMDB and HMDB51 datasets.

Methods	JHMDB	HMDB51
PA3D[8]	56.8	80.50
PA3D+I3D[8]	-	82.1
PA3D+RPAN[34]	86.1	-
MR-TS R-CNN[43]	71.1	-
TSN[2]	80.0	69.4
R(2+1)D[44]	-	78.7
MSNet-R50[27]	-	81.9
I3D[4]	84.1	80.7
I3D+PoTion[10]	85.5	80.9
JMRN+I3D[16]	88.3	82.5
FFDC	89.57	83.14

both RGB frames and 3D heatmaps. This method demonstrates its effectiveness on various datasets and achieves state-of-the-art. However, despite advancements in spatio-temporal feature construction and scalability, it still deals with the computational complexity and time required for reasoning with 3D models. Additionally, the model's effectiveness remains heavily relay on the quality of the heatmaps. The 2D model based approach JMRN [16] employs heatmap models to capture joint population features. It then individually extracts motion features from each joint through a shared motion encoder and performs collective inference to derive motion representations. While this method eliminates the need for pose tracking and enhances model robustness for pose recognition, it remains susceptible to errors due to misestimations of certain joints, which can propagate into the final collective inference. In comparison, FFDC not only achieves competitive performance, but also stands out for its simplicity in terms of model complexity. Furthermore, in terms of robustness, our approach places greater emphasis on overall pose information, reducing the impact of local joint misestimations and ensuring more accurate estimations.

D. EFFECTIVENESS OF FUSION

Our experiments demonstrate the effectiveness of pose-driven fusion in action recognition. While skeletons provide a direct representation of human behavior, the quality of the data significantly influences recognition accuracy. This conclusion is further supported by the superior performance of graph convolutional network (GCN) methods over conventional image-based methods on the NTURGB+D [44] confirms this conclusion. However, real-world video data often contains occlusions and motion jitter, which can impair the skeletal motion and subsequently affect the action recognition. GCN methods tend to exhibit a significant performance drop when skeletal information is missing.

To address these challenges, our approach goes beyond simple addition or concatenation of corresponding features during the fusion process. To achieve effective pose-driven fusion and ensure robust recognition, we pass the pose stream feature (F_P) and appearance stream feature (F_A) through an additional fully connected layer. By incorporating fusion attention parameters and concatenating the resulting features, we obtain the final motion feature representation utilized for action recognition. Our designed experiments validate the effectiveness of our proposed method.

E. EXPERIMENT ANALYSIS

The experimental environment of the algorithm is based on Nvidia 3090 series graphics card.Based on the Pytorch framework, the above experimental results were designed and implemented. The main parameter settings for the experiments were that initial learning rate is 0.0005 and every 10 epochs decrease by 0.1, dropout rate is 0.5, and segment value is 8. The optimizers elected during training is Adam. During thet raining process, the accuracy and loss change curve before and after the feature fusion is shown in Fig.3.

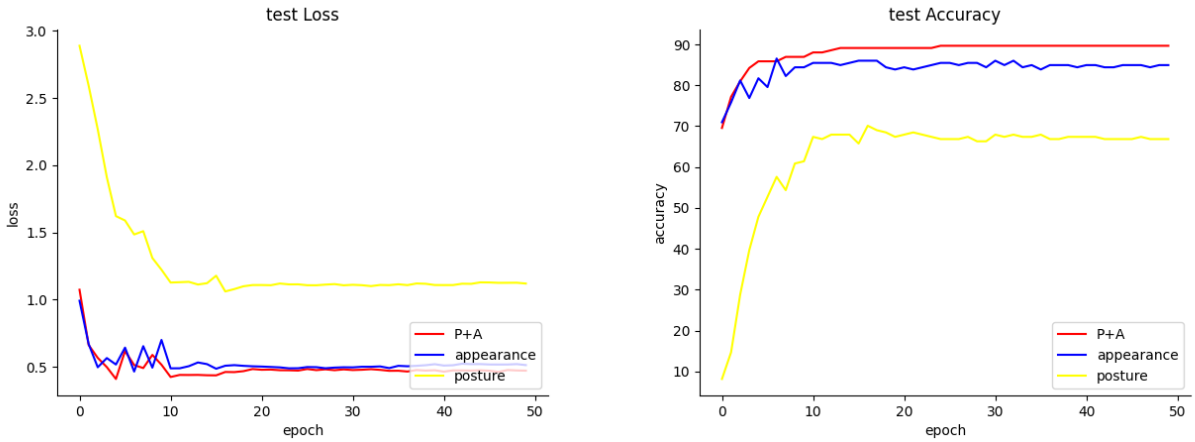


FIGURE 3. Loss and Accuracy change curve in JHMDB. The term "apperance" denotes the result of using only video RGB frames, "posture" refers to the result of using only gesture motion images, and "P+A" indicates the result of using both features.

TABLE 4. A comparative analysis of cooperative fusion results. The term "Add" denotes the direct merging of features from both streams, "concatenate" refers to the concatenation of the two feature types, and "Motion-driven" signifies the results obtained by leveraging fusion attention weights.

Methods	JHMDB	HMDB51
Motion-driven	89.57	83.14
Concatenate	88.25	82.36
Add	88.31	82.19

The curve clearly illustrates that "P+A" consistently outperforms any single stream in terms of accuracy. "P+A" stands for the methodology that incorporates the characteristics of multiple streams. This result demonstrates the effectiveness of fusing multi-stream information.

V. CONCLUSION

We design a simple and efficient multi-feature fusion method for recognising subject actions in videos. Our method constructs the spatio-temporal information of the video subject through motion images and appearance images, and discriminates the action in concert with the motion information and appearance features of the video subject. Through the application of the time shift module and channel transfer technology, we efficiently extract the spatio-temporal information representation of video. The fusion of pose and appearance features significantly enhances the recognition ability of the model, improves its accuracy and robustness. Consequently, our approach delivers competitive performance when benchmarked against state-of-the-art methods.

Indeed, while our method embodies scalability through multi-feature fusion, it also grapples with the challenge of redundant information. For example, motion images contain vital information concentrated in a small image portion, but the model processes a significant amount of irrelevant data. One possible approach for improvement is to represent motion images in a more precise form. Moreover, 2D images inherently lack 3D information. Therefore, integrating auxiliary elements such as motion patterns into a multi feature fusion framework to improve the motion representation of bone posture in a single data stream is another way for future improvement. In summary, there is still considerable potential for improvement in our research.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[2] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016.

[3] J. Zhang, F. Shen, X. Xu, and H. T. Shen, "Cooperative cross-stream network for discriminative action representation," 2019, *arXiv:1908.10136*.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[5] P. Lee, T. Kim, M. Shim, D. Wee, and H. Byun, "Decomposed cross-modal distillation for RGB-based temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2373–2383.

[6] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2923–2932.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[8] A. Yan, Y. Wang, Z. Li, and Y. Qiao, "PA3D: Pose-action 3D machine for video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7914–7923.

[9] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose motion representation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7024–7033.

[10] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2959–2968.

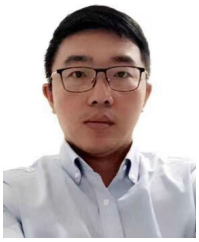
[11] D. Ludl, T. Gulde, and C. Curio, "Simple yet efficient real-time pose-based action recognition," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 581–588.

[12] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 102–106.

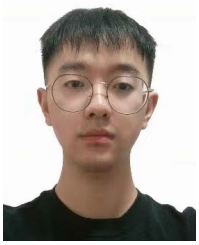
[13] J. Park, J. Lee, and K. Sohn, "Dual-path adaptation from image to video transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2203–2213.

[14] C.-Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Krähenbühl, "Compressed video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6026–6035.

[15] A. Ali, E. Pinyoanuntapong, P. Wang, and M. Dorodchi, "Skeleton-based human action recognition via convolutional neural networks (CNN)," 2023, *arXiv:2301.13360*.
[16] A. Shah, S. Mishra, A. Bansal, J.-C. Chen, R. Chellappa, and A. Shrivastava, "Pose and joint-aware action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 141–151.
[17] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 597–600.
[18] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2405–2415, Aug. 2018.
[19] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–9.
[20] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 180–189.
[21] Z. Chen, S. Li, B. Yang, Q. Li, and H. Liu, "Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1113–1122.
[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[23] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7082–7092.
[24] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.
[25] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6201–6210.
[26] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3034–3042.
[27] H. Kwon, M. Kim, S. Kwak, and M. Cho, "MotionSqueeze: Neural motion feature learning for video understanding," in *Proc. 16th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 345–362.
[28] B. Fernando, E. Gavves, J. Oramas M., A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 773–787, Apr. 2017.
[29] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4570–4579.
[30] C. Caetano, J. Sena, F. Brémond, J. A. D. Santos, and W. R. Schwartz, "SkeleMotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition," in *Proc. 16th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2019, pp. 1–8.
[31] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proc. ACM Multimedia Asia*, Dec. 2019, pp. 1–6.
[32] R. Hachiuma, F. Sato, and T. Sekii, "Unified keypoint-based action recognition framework via structured keypoint pooling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22962–22971.
[33] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.
[34] W. Du, Y. Wang, and Y. Qiao, "RPAN: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3745–3754.
[35] B. Su, P. Zhang, M. Sun, and M. Sheng, "Direction-guided two-stream convolutional neural networks for skeleton-based action recognition," *Soft Comput.*, vol. 27, no. 16, pp. 11833–11842, Aug. 2023.
[36] G. Moon, H. Kwon, K. M. Lee, and M. Cho, "IntegralAction: Pose-driven feature integration for robust human action recognition in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3334–3343.
[37] B. Hosseini, R. Montagne, and B. Hammer, "Deep-aligned convolutional neural network for skeleton-based action recognition and segmentation," *Data Sci. Eng.*, vol. 5, no. 2, pp. 126–139, Jun. 2020.
[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[39] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.
[40] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.
[41] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," 2020, *arXiv:2006.10204*.
[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[43] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Proc. 14th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 744–759.
[44] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.



DONG CHEN received the M.S. degree from Harbin Engineering University, China, in 2009. He is currently pursuing the Ph.D. degree with Guangxi Normal University. He is currently an Associate Professor with Nanning Normal University, China. His main research interests include artificial intelligence and deep learning.



MENGTAO WU was born in Shangrao, Jiangxi. He received the B.E. degree in computer science from the Jiangxi University of Traditional Chinese Medicine, Nanchang, in 2020. He is currently pursuing the M.S. degree with Nanning Normal University. His research interests include machine learning and computer vision.

TAO ZHANG was born in Yuncheng, Shanxi. He received the M.S. degree from the College of Physics and Electronic Engineering, Nanning Normal University, Nanning, in 2023. His research interests include machine learning and video understanding.

CHUANQI LI received the M.S. degree from the Institute of Plasma Physics, Chinese Academy of Sciences, China, in 1991, and the Ph.D. degree from Southeast University, China, in 2004. He is currently a Professor with Nanning Normal University, China. His research interests include optical communication and artificial intelligence.