International Journal of Computational Intelligence Systems (2023) 16:116
https://doi.org/10.1007/s44196-023-00292-9

RESEARCH ARTICLE

Integrating Vision Transformer-Based Bilinear Pooling and Attention Network Fusion of RGB and Skeleton Features for Human Action Recognition

Yaohui Sun[1] · Weiyao Xu[2] · Xiaoyi Yu[2] · Ju Gao[2] · Ting Xia[2]

Received: 19 April 2023 / Accepted: 21 June 2023

Abstract
In this paper, we propose VT-BPAN, a novel approach that combines the capabilities of Vision Transformer (VT), bilinear pooling, and attention network fusion for effective human action recognition (HAR). The proposed methodology significantly enhances the accuracy of activity recognition through the following advancements: (1) The introduction of an effective two-stream feature pooling and fusion mechanism that combines RGB frames and skeleton data to augment the spatial–temporal feature representation. (2) The development of a spatial lightweight vision transformer that mitigates computational costs. The evaluation of this framework encompasses three widely employed video action datasets, demonstrating that the proposed approach achieves performance on par with state-of-the-art methods.

Keywords Human action recognition · Multi-modal · Self-attention · Feature fusion

Abbreviations

HAR	Human action recognition
RNN	Recurrent neural network
CNN	Convolutional neural network
GNN	Graph neural network
MHA	Multi-head attention
FFN	Feed-forward networks

1 Introduction

HAR simply means identifying human actions, plays an important part in many engineering practices, for example, surveillance systems that monitor various daily human behaviors [1] and automatic navigation systems for the safe operation of drivers [2]. In addition, it is also important for many other related fields, including assistive robotics, human–computer interaction [3] and video retrieval [4], etc.

In the early times, a majority of researches concentrated on using grayscale video or RGB as an input to HAR [5], since they were popular and easy for accessibility. However, the drawbacks of RGB images, such as sensitivity to light and high noise, limited the applicability of HAR. Many works using multi-modal data have appeared in recent years [6–13], which are primarily driven by the fact that a variety of accurate and economical sensors have been developed, as well as the new generation of RGB-D cameras. Different modalities, such as skeleton, audio, depth, event stream, radar, etc., are employed for HAR. Depending on the application scenario, the RGB-D cameras can capture depth images, state information of the skeleton, and other state information. Multi-modality of HAR has significant advantages [14, 15]. Zhao et al. [14] proposed a two-stream network consisting of both recurrent neural network (RNN) and convolutional neural network (CNN) for processing RGB and skeleton data

Y. Sun, W. Xu, X.Yu, J.Gao, and T. Xia have contributed equally to this work.

✉ Weiyao Xu
bingbing2016@uzz.edu.cn

✉ Ting Xia
xiayuxue121@126.com

Yaohui Sun
mwg199391@163.com

Xiaoyi Yu
yuriiaac@hotmail.com

Ju Gao
jugao@hku.hk

[1] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

[2] School of Opto-Electronic Engineering, Zaozhuang University, Zaozhuang 277160, China

Published online: 20 July 2023

independently. Song et al. [15] presented a continuous deep CNN learning framework consisting of two skeleton-guided streams, and the network was adopted to extract features from RGB as well as optical streams.

Transformer [16] is considered as a novel deep learning model, since it was proposed, and has recently led the field of machine learning due to its powerful capabilities and promising future. Chen et al. [17] proposed a Multi-Modal Video Transformer (MM-ViT) to realize video action recognition. Currently, there are only a few works [18, 19] use Transformer in low-level vision, so further research is needed. As the video is sequential, Transformer is intrinsically suitable to video tasks [20, 21], and its performance begins to be comparable to traditional CNNs and RNNs.

Although the aforementioned HAR method based on multi-modal fusion has achieved satisfactory performance on selected benchmark datasets, however, the objective of effective modal fusion remains challenging. More specifically, there are at least threefold challenges. First, how to obtain greater action context information from multi-modal data, and how to capture more abundant feature information through the attention mechanism integration model. Second, in videos, most features are extracted from the full frame, which includes a large amount of background noise, while the object on which the action occurs is easily ignored. Third, most existing multi-modal methods have complex structures that require high computational costs. Therefore, it is necessary to solve the multi-modal HAR problem efficiently.

Inspired by the aforementioned works, the main contributions of this paper are given below.

- Different from [35], we investigate two stream complementary techniques to improve the fusion accuracy, mainly including a spatial lightweight vision Transformer block and attention module, which can effectively implement fusion of multi-modal data and perform end-to-end training.
- An effective data preprocessing method is employed for RGB video and skeleton sequences, which can help the network capture more abundant feature information, so that human behavior can be captured more accurately.
- The semantic relations between various models can be enhanced by the proposed VT-BPAN module. In addition, a spatial lightweight improvement of vision Transformer that is capable of reducing computational costs is proposed.
- The proposed VT-BPAN module provides significant improvements over existing research results in terms of action recognition through experimental analysis of multi-modal datasets.

The rest of this paper is organized as follows. Section 2 provides an outline of the related work. Section 3 describes the proposed VT-BPAN in detail. Then, in Sect. 4, its experimental implementation setup is presented as well as the experimental results and discussion. Finally, Sect. 5 concludes the paper.

2 Related Works

2.1 Multi-modal Human Action Recognition Based on RGB Video and Skeleton

RGB modalities are generally captured by RGB cameras. The two-stream 2D CNN framework generally consists of two 2D CNN branches for HAR. The framework regarded as a classical two-stream framework was proposed by Simonyan and Zisserman [22], which consisted of a spatial network and a temporal network. Wang et al. [24] fused the classification scores of videos that were split into three segments, and processed each video into three segments with a two-stream network. A long-term recursive convolutional network (LRCN) consisting of 2D CNNs for extracting RGB features at the frame level was presented in [6], and then, an LSTM was employed to generate individual action labels.

Skeleton data can naturally carry joint position informations by coordinates. Therefore, compared to RGB, the information is of a higher level. In addition, a much smaller computation is required and are capable of being robust. The human skeleton structure can represented as a graph, where each of the vertices is viewed as a human joint and the human skeleton is viewed as the connection between the joints. In recent years, a number of graph convolution networks (GNN) as well as GCN-based HAR methods [25, 26] have been introduced. Yan et al. [25] used GCN for skeleton-based HAR by introducing Spatial–Temporal GCN (ST-GCN), and both spatial and temporal modalities from skeleton data could be learned. The two-stream Adaptive GCN (2 S-AGCN) was introduced in [26]. Moreover, Chi et al. [27] introduced InfoGCN, it includes an information bottleneck for learning abundant actions, and a GCN based on attention to deduce the skeleton topology of contextual relevance.

Transformers have shown significant potential in sequence data processing. Therefore, a great number of methods [28, 29, 31] proposed the application of Transformers on skeleton sequences with a focus on spatial–temporal modeling. In [31], a spatio-temporal tuple transformer (STTFormer) framework was proposed. Plizzari et al. [29] presented a structure of Spatial–Temporal Transformer Network (ST-TR) in which the inter-frame motion dynamics and interactions of intra-frame joint were considered to be learned by spatial and temporal self-attention modules.

2.2 Multi-modal Data Fusion Methods

There are several works investigating deep learning architectures that fuse features of RGB and HAR skeletons. Zolfaghari [32] illustrated a 3DCNN to deal with raw RGB images pose, and motion. A Markov chain model was used to fuse the streams to perform action classification. In [7], a unit-level spatial–temporal LSTM was presented, enabling efficient fusion of features in LSTM units. Li et al. [34] presented a 2 S model including R(2+1)D networks, ST-GCN networks, and guidance blocks with enhanced information related to the action in the video. Then, the model was classified with the use of score fusion. Based on a 3D CNN network, Das et al. [51] introduced an approach using RGB video as input and designed a pose-guided spatial–temporal attention-based network. Xu et al. [35] introduced a BPAN model spatial–temporal a bilinear pooling and attention fusion approach, which implemented feature fusion effectively.

3 Proposed Model

A novel VT-BPAN model of multi-modal HAR is proposed for feature fusion. Specifically, first, the two deep learning frameworks are employed to extract features separately, and then fuse the features by the VT-BPAN module. The overall architecture of the network is depicted in Fig. 1. Moreover, the proposed network is evaluated in an end-to-end way. A feature fusion approach is specified in this section, as well as a data preprocessing module and a feature extraction technique. Different from previous method given in Sect. 2, first, it utilizes a Vision Transformer architecture, which has shown promising results in various computer vision tasks, instead of traditional convolutional neural networks (CNNs) used in the previous approaches. This allows the model to capture long-range dependencies and contextual information more effectively. Additionally, the proposed method incorporates bilinear pooling, which fuses features from different modalities (RGB and skeleton) by computing outer products, enabling multi-modal interactions. Furthermore, attention network fusion is employed to dynamically weight the importance of different modality-specific features, enhancing the discriminative power of the model. These novel components in the proposed method enable better integration and exploitation of multi-modal information, leading to improved human action recognition performance.

3.1 Preprocessing Module

The RGB video is first cropped. The large amount of noise and background are generally important factors that affect the task of RGB video classification, and can also lead to computational memory consumption. Thus, focusing video detection tasks on the human body in RGB video is required, this paper adopts the technique in [35] to crop the human action in its input RGB image via pose mapping.

For the skeleton sequence, to better describe the spatial and temporal sequence of the skeleton, we adopt temporal differences and relative coordinates. As shown in Fig. 3, based on the distance among all joints in each frame and that of the joint at the center, the relative coordinates x_r can be obtained, and x_t denotes the time difference, and is computed below

$$x_t = x\,[t+1] - x\,[t]\,, \tag{1}$$

where $x[t]$ represents data at frame t, and the data of final input is concatenated and joined by x, x_r, and x_t. Based on 2 S-AGCN [26], with an origin channel C of 3, denoting the 3D coordinates at each joint. We use the preprocessing module to extract more information features. The input channel C is added to 9 by applying this module.

For 2 S-AGCN, a spatial GCN block was proposed, which can be calculated by the following equation:

$$f_{\text{out}} = \sum_{k}^{K_v} W_k f_{in}\,(A_k + B_k + C_k), \tag{2}$$

where K_v refers to kernel dimension, A_K refers to the adjacency matrix, B_k is similar to that of M_k in ST-GCN, and C_k represents the learning sample graph.

3.2 Feature Extraction Module

The recognition network model proposed in this paper is composed of two streams, which mainly include RGB data and skeleton data processing, for extracting temporal and spatial features, respectively. Next, the network is specifically described.

In general, R(2+1)D network [36] and 2 S-AGCN network [9] are used for feature extraction, respectively. Specifically, for RGB video stream, pretraining on Kinetics-400 [37] is required. A sketch of the R(2+1)D recognition block is depicted in Fig. 2. The convolution network splits the calculation into two-dimensional convolution in spatial and one-dimensional convolution in temporal terms. As the input, it regards the video data with the size of $3 \times T \times 112 \times 112$, where 3 denotes the amount of RGB channels, 112 corresponds to the image height and width, as well as T corresponds to the length of the sequence. For 3D skeleton stream. The 2 S-AGCN network [9] is adopted. First, the skeleton processing data are obtained, as shown in Fig. 3. Then, it takes the skeleton sequence with the size of $T \times C \times V \times M$ as the input, where T represents the frame in the sequence, C represents the channel amount, V represents the joints, and

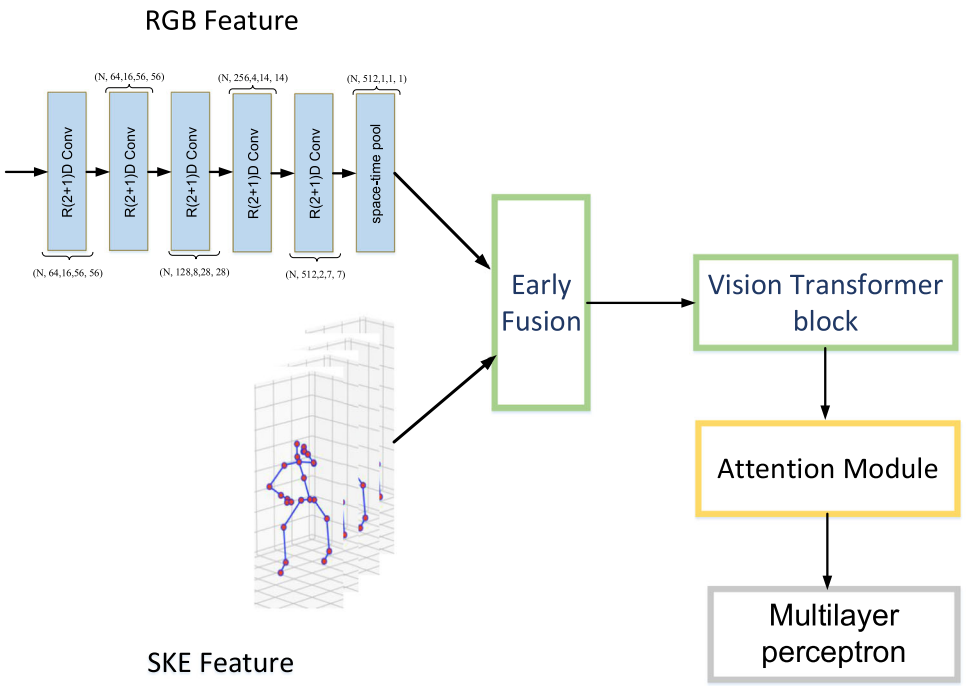

Fig. 1 An overall architecture of the network, two data preprocessing methods are applied to extract features, and the VT-BPAN block is performed to fuse the features

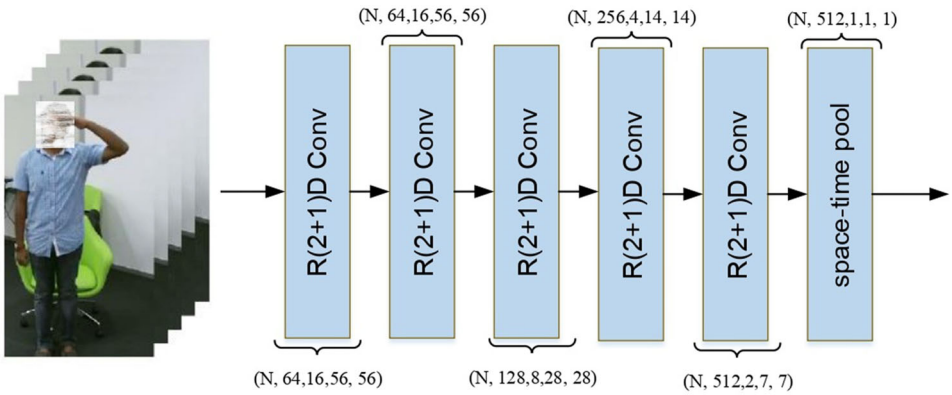

Fig. 2 An overall of skeleton preprocessing and the architecture of R(2+1)D network

M represents the skeletons within the frame. After R(2+1)D block [36] and 2 S-AGCN network as well as dimensional transformation, the features have the same dimension.

3.3 Feature Fusion Module

In this subsection, Transformer and Bilinear Pooling techniques [38, 40] have been introduced to fuse the features extracted by two preprocessing models. The proposed VT-BPAN module is depicted in Fig. 4. The module integrates the popular structure of Transformer block and gives a lightweight improvement, as shown in Fig. 5, which can efficiently fuse multi-modal for RGB-D action recognition. Denote F_{RGB} and F_{SKE} refer to the features extracted from the RGB and skeleton modules, respectively.

The specific approach to the Transformer block is described in the following two aspects.

(1) As for the vision Transformer block, each standard Transformer block is typically composed of multi-head attention layers (MHA), feed-forward networks (FFN), normalization of layers, and shortcut connections. In practice, a set of query attention functions needs to be computed simultaneously and packaged into a matrix Q, and K and V denote the key matrices and value matrices. The attention output can be computed by

$$\mathrm{Attention}\,(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{3}$$

Next, the different heads get different query matrices, key matrices, and value matrices. The input vectors are allowed to be projected into different representation subspaces, and moreover, the MHA allows the model to focus on information from different subspaces in different loca-

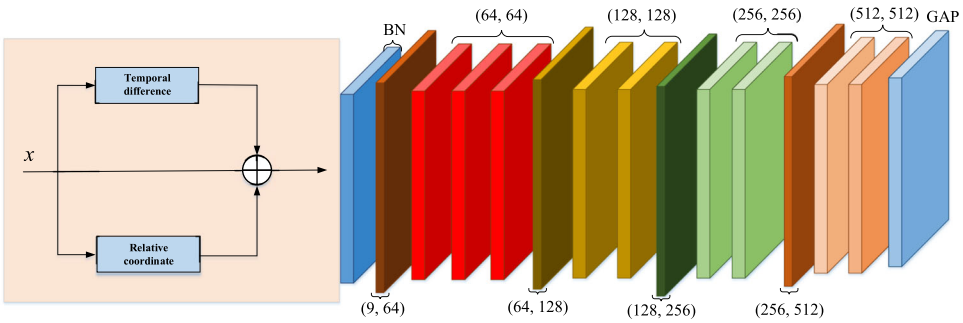

Fig. 3 An overall of skeleton preprocessing and the architecture of 2 S-AGCN with temporal difference

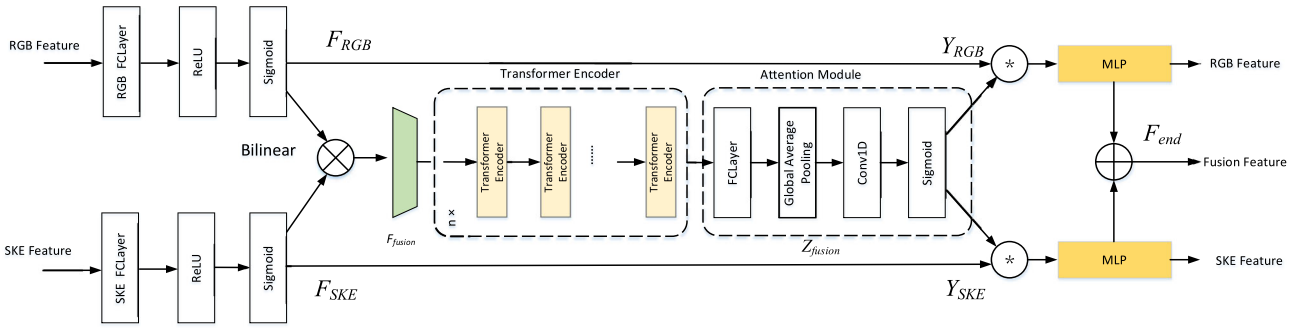

Fig. 4 An overall architecture of our VT-BPAN model. It has two streams, a skeleton sequence and an RGB frame, and the 2 S-AGCN network is for extracting skeleton features, while the R(2+1)D network is for extracting RGB features. The features have three parts, the RGB features, the skeleton features, and the eventual fusion features, and the network is used for effective feature fusion

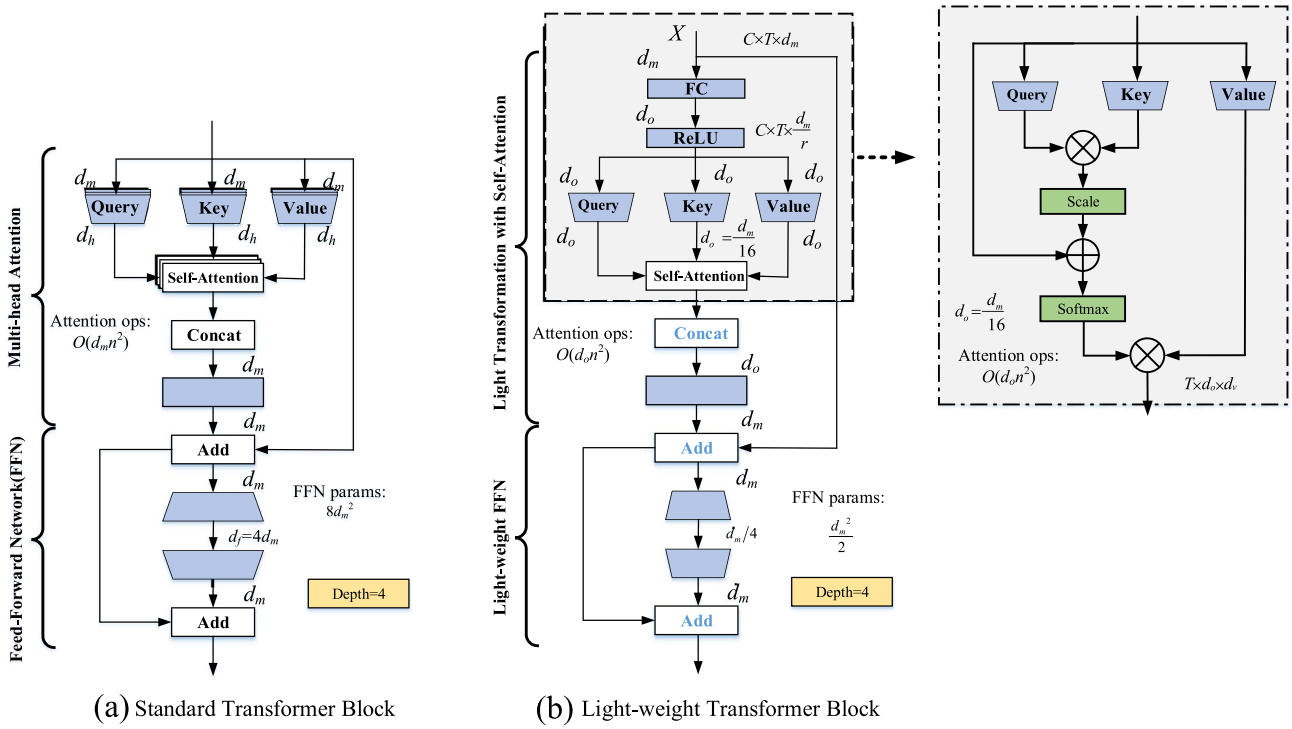

Fig. 5 **a**, **b** Block-wise comparison between the standard Transformer block and the spatial lightweight Transformer

tions. The process is shown in the following:

$$\mathrm{MHA}\,(Q, K, V) = \mathrm{Concat}\,(\mathrm{head}_1, \ldots, \mathrm{head}_h)\, W^O, \tag{4}$$

where $\mathrm{head}_i = \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$, the projections are matrices of parameters with the followings dimensions $W_i^Q \in R^{d_{\mathrm{model}} \times d_k}$, $W_i^K \in R^{d_{\mathrm{model}} \times d_k}$, $W_i^V \in R^{d_{\mathrm{model}} \times d_v}$, and $W_i^O \in R^{hd_v \times d_{\mathrm{model}}}$.

Each layer in the encoder contains a fully connected FFN, in addition to the attention layer, and can be denoted as the following function:

$$\mathrm{FFN}(x) = \max\,(0, xW_1 + \beta_1)\, W_2 + \beta_2, \tag{5}$$

where W_1 and W_2 denote weight vectors, and β_1 and β_2 denote bias vectors.

(2) The spatial lightweight transformer encoder layer is based on [29] and DelighT [30], with an improved model of the network architecture and the dimensions of the weights, as shown in Fig. 5b.

First, we consider squeezing the output feature information. The shaped input tensor X with $C \times T \times d_m$

$$\mathbf{S}_k = \sigma\left(\tilde{W}_1 ReLU\left(X\tilde{W}_2\right) + b\right), \tag{6}$$

where $\tilde{W}_1 \in R^{\frac{d_m}{r} \times d_m \times C}$ and $\tilde{W}_2 \in R^{C \times 1 \times d_m}$ denotes weight vectors. σ represents the Sigmoid activation function. b denotes the bias vector. The choice of parameter values is based on the relevant literatures [16, 29] and [30] and the source code.

Next, the matrices $X_Q \in R^{T \times C \times 1 \times d_k}$, $X_K \in R^{T \times C \times 1 \times d_k}$, and $X_V \in R^{T \times C \times 1 \times d_v}$ are acquired from rearranging the inputs. The transformation is applied to each frame separately in the T dimension. The matrices Q, K, and V are obtained by multiplying the matrices $W_Q \in R^{d_o \times 1 \times C}$, and $W_K \in R^{d_o \times 1 \times C}$ and $W_V \in R^{d_o \times 1 \times C}$, respectively, and then by conducting dimensional transformation.

Unlike the existing spatial self-attention module, we first squeeze the embedded dimension settings in each spatial lightweight Transformer encoder layer. The formula for computing the self-attention matrix is given here

$$\mathrm{Attention}\,(Q, K, V) = \mathrm{softmax} \times \left(\frac{\left(W_Q X_Q\right)\left(W_K X_K\right)^T}{\sqrt{d_k}} + \mathbf{S}_k\right)\left(W_V X_V\right). \tag{7}$$

Due to the fact that dimension of T moves in the batch, the parameters can be efficiently shared along the dimension of time and the transformations are applied separately in each frame. Through the above transformation, self-attention matrix has an output shape of $T \times d_o \times d_v$. The algorithm procedure is given in Algorithm 1.

Algorithm 1 Transformer algorithm procedures

Initialization:current period t, initial matrix Q, K, V, initial weight $W, \tilde{W}_1, \tilde{W}_2$, initial the vectors β_1, β_2 and b. Attention $(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$;
MHA: MHA $(Q, K, V) = \mathrm{Concat}\,(\mathrm{head}_1, \ldots, \mathrm{head}_h)\, W^O$;
FFN: $\mathrm{FFN}(x) = max\,(0, xW_1 + \beta_1)\, W_2 + \beta_2$;
Extrusion output feature information $\mathbf{S}_k =$
$\sigma\left(\tilde{W}_1 ReLU\left(X\tilde{W}_2\right) + b\right)$;
Attention $(Q, K, V) = softmax\left(\frac{(W_Q X_Q)(W_K X_K)^T}{\sqrt{d_k}} + \mathbf{S}_k\right)(W_V X_V)$.

Then, the output of our model can be obtained by a simple reshaping.

Figure 5b shows that how we integrate a spatial lightweight transformation into the Transformer block. Compared with the standard Transformer block and the lightweight Transformer block, the computational cost of calculating attention is $O\left(d_m n^2\right)$ and $O\left(d_o n^2\right)$, respectively, where $d_o < d_m$. Thus, the lightweight Transformer block has reduced the cost of calculating attention by a factor of $\frac{d_m}{d_o}$. In our experiments, we adopted $d_o = \frac{d_m}{16}$; thus, it required 16× less multiplicative addition operations compared to the Transformer structure. Note that, the advantage of dot product is that it is faster and more spatially efficient in operation, because it allows the use of highly optimized matrix multiplication codes for its implementation [16].

Consider a lightweight feed-forward network (FFN) structure, where the first layer reduces the dimension of the input from d_m to $\frac{d_m}{r}$, while the second layer expands the dimension from $\frac{d_m}{r}$ to d_m, with r being the reduction factor. The lightweight FFN thereby reduces the amount of parameters and operations in the FFN by a factor $rd_f = d_m$. In a standard Transformer, the FFN dimensions are expanded a factor of 4. In the subsequent experiments, we adopted $r = 4$. Hence, the lightweight FFN reduces the number of parameters of the FFN by 16 ×.

In this work, the compact bilinear pooling (CBP) [40] method is used for the fusion feature, and is described as follows:

$$\begin{aligned}\langle F_{\mathrm{RGB}}(\mathcal{X}), F_{\mathrm{SKE}}(\mathcal{Y})\rangle &= \sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}}\langle x_s, y_u\rangle^2 \\ &\approx \sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}}\langle \phi(x), \phi(y)\rangle \\ &\equiv \langle C(\mathcal{X}), C(\mathcal{Y})\rangle,\end{aligned} \tag{8}$$

where $C(\mathcal{X}) = \sum_{s\in\mathcal{S}}\phi(x_s)$, $C(\mathcal{Y}) = \sum_{s\in\mathcal{U}}\phi(y_u)$, and C is the channel of the feature. Define F_{fus} as the fusion feature.

Then, it can be normalized by L2 regularization

$$F_{\text{fus}} = \frac{F_{\text{fus}}}{\| F_{\text{fus}} \|_2}. \tag{9}$$

Then, the obtained F_{fus} is taken and placed into the fully connected layer and the ReLU layer. As a result, the new fusion feature $\hat{Z}_f \in \mathrm{R}^{C\times 1}$ is described by

$$\hat{Z}_f = ReLU\left(\tilde{W} \cdot F_{\text{fus}}\right), \tag{10}$$

where $\tilde{W}$ denotes the weight matrix. To get more expressive features based on attention, the network is designed similarly to ECANet [39]. Conv1d is used for computing the attention weights.

At the end of the proposed model architecture, the multilayer perceptron (MLP) is employed as a classification module, where the batch normalization layer and the ReLU function are connected sequentially. Then, a softmax function is employed to normalize the predictions to probabilistic distributions. Finally, the two features are summed to obtain an ultimate fusion feature F_{end}.

Due to the presence of two main tasks and the fact that multitask learning is based on multiple objective optimization models. We give a loss function that sums the losses of these two tasks, which can be computed as follows:

$$L_{\text{total}} = \lambda_1 L_{\text{RGB}} + \lambda_2 L_{\text{SKE}}, \tag{11}$$

where λ_1 and λ_2 represent weighting factors, respectively. L_{RGB} denotes the loss of RGB stream and L_{SKE} denotes the loss of skeleton stream.

The overall algorithm of the VT-BPAN model is given in Algorithm 2.

Algorithm 2 An overall algorithm of the VT-BPAN model

Initialization: RGB feature F_{RGB} and SKE feature F_{SKE};
Compute early F_{fusion} through CBP, $F_{fusion} \leftarrow F_{RGB}$ and F_{SKE} ;
Compute F_{fusion} through the proposed Transformer encoder;
Compute Z_{fusion} through the attention module;
Compute the loss function L_{total};
Obatin Y_{RGB} and Y_{SKE} and classified by MLP, and the two features are summed to obtain an ultimate fusion feature F_{end}.

4 Experiments

Several experiments are conducted on two public datasets to evaluate the effectiveness of the presented VT-BPAN for HAR. In addition, an extensive ablation study is conducted to investigate the performance of each module.

4.1 Datasets

(1) NTU-RGB+D Dataset [41]: The dataset is viewed as one of the most extensively used datasets available for HAR. It includes 56,880 video samples with a total of 60 action classes. In addition, it provides two classes of action categories for assessment: cross-subject (CS) and cross-view (CV). Based on both assessments, related experiments were conducted and the first ranked recognition accuracy was reported.
(2) NTU-RGB+D 120 Dataset [42]: NTU-RGB+D 120 has more action classes, with a total of 120 action classes that are classified into three major categories, including 82 daily actions, 26 interactive actions, as well as 12 health-related actions. It is composed of 114,480 RGB+D video samples coming from 106 different human participants.
(3) UTD-MHAD Dataset [43]: The dataset contained 27 actions, which are implemented by 8 human subjects, where each action is executed exactly 4 times. The dataset is left with 861 sequences after removing the three corrupted sequences. Four modalities including skeleton and RGB were available. To obtain a fair comparison, an across-subject protocol was implemented [43], with the data from subjects numbered 1, 3, 5, and 7 were employed for training, while the data from subjects numbered 2, 4, 6, and 8 were given as test data.

4.2 Implementation Details

Two NVIDIA TITAN RTX GPUs are used to perform all experiments on the PyTorch framework. For the RGB stream, the resizing of the RGB frames is adjusted to 112×112, with the video sequence length being set to 20 and the RGB model being pretrained on the Kinetics-400 dataset [37]. For the skeleton stream, with 50 being the set skeleton sequence length, the other variables are the same choices as in [26]. We select the decay of weights and rate of learning as 0.0001 and 0.01, respectively, and the model optimization is performed using stochastic gradient descent with cross entropy as the loss function of the back-propagation gradient. The learning rate is selected as 0.01.

4.3 Ablation Study

In this section, cross-topic benchmarks on the NTU-RGB+D dataset are used to validate the effectiveness of the VT-BPAN component proposed in this paper. To illustrate the performance of the model, training and testing are evaluated in each epoch. In addition, the accuracy of the model in each epoch is recorded.

(1) Effect of fusion scheme: Various feature fusion schemes, including average, multiplication, sum, concatenation,

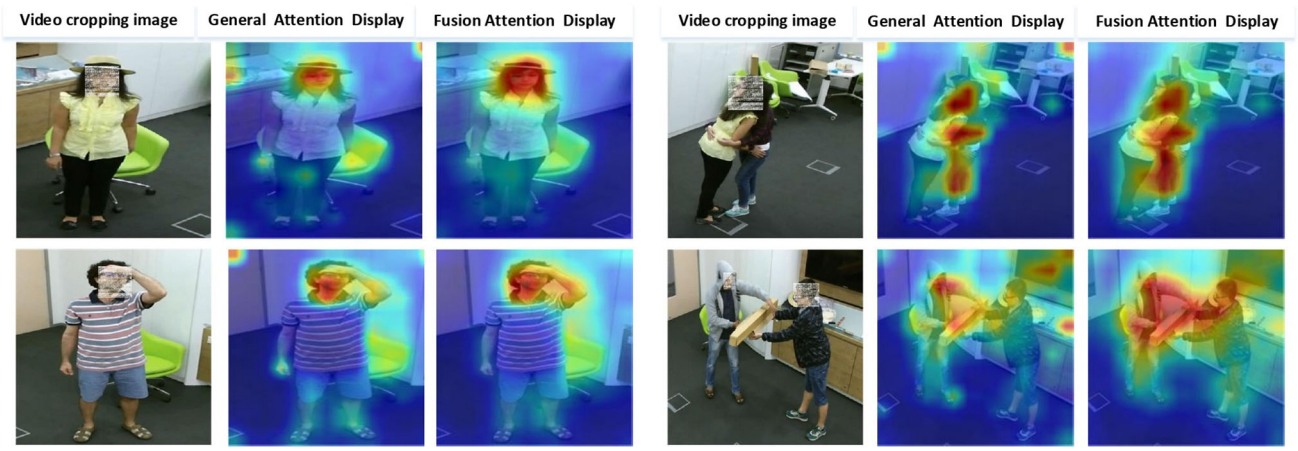

Fig. 6 Visualization results of vision Transformer fused self-attention module

Table 1 Impact of fusion scheme on performance

Methods	Accuracy
Average	93.07
Multiplication	92.20
Max	92.35
Sum	93.09
Concatenation	93.16
BPAN (Resnet 18)	94.85
VT-BPAN [ours]	**95.40**

Bold value indicates better results than other filtering methods

and maximum, are verified. Table 1 shows the results. As it can be seen in Table 1, the recognition accuracy is significantly improved by the fusion scheme and reaches 95%. Moreover, the given VT-BPAN fusion module achieves the best accuracy compared to the above model.

(2) Effect of feature extraction models: In this paper, the approach of video HAR can be briefly summarized in twofold. In one hand, a 2D+1D framework is typically used, where a 2D CNN is used for input of each frame, and is followed by a 1D module, that converges the features from each frame. The second approach alternatively applies a 3D CNN with stacked 3D convolutions to model temporal and spatial semantics in conjunction. Based on the results of [35], we compare the recognition accuracy of three convolutional neural networks [including 3D ResNet, MC3 ResNet [36], and R(2+1)D] based on ResNet-18. The results show that the selected R(2+1)D model reaches the best accuracy.

(3) Visualization analysis: For better representation of the effect of the self-attention of VT-BPAN, we further compare the only ECANet attention and the fusion self-attention produced by VT-BPAN. The results are shown in Fig. 6, it can be seen that for each input image, the learned representation of the saliency map indicates the importance of each pixel. VT-BPAN captures more meaningful pixels.

4.4 Comparisons With the State-of-the-Art

For a fair comparison, we compare with algorithms that fuse RGB features and skeleton features that appear comparable to our work. As with the original evaluation scheme, it is demonstrated the accuracy of the NTU-RGB+D, NTU-RGB+D 120, and UTD-MHAD datasets, as presented in Tables 2, 3, and 4. A comparison is made between our method and simpler fusion-based (e.g., BI-LSTM [44]) as well as attention-based methods (e.g., MMTM [47] and BPAN [35]). Compared to the NTU-RGB+D dataset, the UTD-MHAD as a smaller scale dataset still performs comparable to HAMLET [55]. Results of the evaluation indicate that attention-based fusion approaches can obtain superior results. The approach used in this paper integrates a Transformer-based bilinear pooling and an attention-based strategy. The backbone network used is Resnet 18, which is superior to the state-of-the-art results.

5 Conclusion

In this paper, a multi-modal HAR model has been proposed. For temporal and spatial feature extraction, R(2+1)D and 2 S-AGCN have been employed. Next, the VT-BPAN module that can obtain more expressive features by feature fusion and vision Transformer attention mechanism has been proposed, and in addition, a lightweight Transformer improvement has been proposed. Various fusion strategies have been compared to verify the superiority of the VT-BPAN module. In addition, we use fully connected perceptrons to get the final fusion features. End-to-end training has been performed. It has been

Table 2 Comparison of NTU-RGB+D dataset model with the state-of-the-art

Methods	Year	Multi-model	CS	CV
BI-LSTM [44]	2019	Yes	85.4	91.6
FUSION [45]	2020	Yes	91.8	94.9
MSAF [46]	2020	Yes	92.24	–
MMTM [47]	2020	Yes	91.9	95.3
VPN(I3D) [48]	2020	Yes	93.5	96.2
BPAN (Resnet 18) [35]	2021	Yes	94.85	97.4
VT-BPAN [ours(a)]	2022	Yes	**95.40**	**97.67**

Bold values indicate better results than other filtering methods

Table 3 Comparison of NTU-RGB+D 120 dataset model with the state-of-the-art

Methods	Year	Multi-model	C-Sub	C-Set
ST-LSTM [49]	2016	No	55.7	55.9
Two-streams ST-LSTM [42]	2019	Yes	61.2	63.1
separable STA [51]	2019	Yes	83.8	82.5
Verma et al. [52]	2020	Yes	76.7	77.9
VPN(I3D) [48]	2020	Yes	86.3	87.8
BPAN (Resnet 18) [35]	2021	Yes	86.6	88.1
VT-BPAN [ours(a)]	2022	Yes	**86.7**	**88.6**

Bold values indicate better results than other filtering methods

evaluated on the NTU-RGB+D dataset, the NTU-RGB+D 120 dataset, and the UTD-MHAD dataset, and has performed better than existing methods. Given the limitations of existing networks, future work will implement on fusion methods for multi-modal data and heterogeneous networks.

Table 4 Comparison of UTD-MHAD dataset model with the state-of-the-art

Methods	Year	Multi-model	Top-1 Accuracy
JDM-CNN [53]	2017	Yes	88.10
MCRL [54]	2019	Yes	93.02
HAMLET [55]	2020	Yes	95.12
BPAN (Resnet 18) [35]	2021	Yes	95.07
VT-BPAN [ours]	2022	Yes	95.10

6 Advantages, Hypothesis, and Limitations

Advantages: First, the method enables efficient extraction of spatial relationships and long-range dependencies within the RGB and skeleton features, leading to enhanced representation learning for HAR. Second, the attention network fusion mechanism effectively combines the RGB and skeleton features, leveraging the complementary information from both modalities. This fusion approach facilitates a more comprehensive understanding of human actions by capturing both appearance and motion cues. Finally, the integration of these two techniques results in improved action recognition performance, surpassing traditional methods and achieving state-of-the-art results.

While the method of integrating Vision Transformer-based Bilinear Pooling and Attention Network Fusion of RGB and Skeleton Features for Human Action Recognition offers several benefits, it also has certain limitations. Here are some limitations of this approach:

1. Complexity and computational requirements: The integration of Vision Transformer-based Bilinear Pooling and Attention Network Fusion introduces increased complexity and computational requirements. Vision Transformers are already computationally expensive due to their self-attention mechanisms. Combining them with bilinear pooling and attention network fusion further exacerbates the computational cost, making the method less efficient for real-time applications or systems with limited computational resources.
2. Limited interpretability: Vision Transformers are known for their black-box nature, meaning that they lack interpretability. While they excel at capturing complex patterns and relationships in visual data, understanding the underlying reasoning for their predictions becomes challenging. This limitation persists when integrating them with other techniques like bilinear pooling and attention network fusion, which may hinder the ability to analyze and interpret the model's behavior.
3. Data requirements: The success of deep learning methods heavily relies on the availability of large and labeled

datasets. Collecting and annotating diverse datasets for action recognition, especially involving multiple modalities like RGB and skeleton, can be a time-consuming and expensive process.

Acknowledgements The authors would like to thank the anonymous reviewers for their valuable comments and suggestions that helped improve the quality of this manuscript.

Author Contributions YS and WX performed the validation and wrote the manuscript. TX performed the data analysis; JG and XY performed the formal analysis.

Funding This study was funded by National Nature Science Foundation of China (No. 11974304 and No. 12175194).

Availability of data and materials The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interest This manuscript has not been partially or entirely published or presented elsewhere, and is not under consideration by other journals. The authors declare that there are no conflicts of interest.

References

1. Lin, W., Sun, M.T., Poovandran, R., Zhang, Z.: Human activity recognition for video surveillance. Proc IEEE Int Symp Circuits Syst. 2737–2740 (2008). https://doi.org/10.1109/ISCAS.2008.4542023
2. Lu, M., Hu, Y., Lu, X.: Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. Appl. Intell. **50**(4), 1100–1111 (2020). https://doi.org/10.1007/s10489-019-01603-4
3. Kuo, Y.M., Lee, J.S., Chung, P.C.: A visual Context-Awareness Based sleeping-respiration measurement system. IEEE Trans. Inf. Technol. Biomed. **14**(2), 255–265 (2010). https://doi.org/10.1109/titb.2009.2036168
4. Liu, J., Sun, C., Xu, X., et al.: A spatial and temporal features mixture model with body parts for video-based person re-identification. Appl. Intell. **49**(9), 3436–3446 (2019). https://doi.org/10.1007/s10489-019-01459-8
5. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010). https://doi.org/10.1016/j.imavis.2009.11.014
6. Donahue, J., Hen, L.A., Saenko, K.: Long-term recurrent convolutional networks for visual recognition and description. Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR): 2625–2634 (2015). https://doi.org/10.1109/CVPR.2015.7298878
7. Liu, J., Shahroudy, A., Wang, G., Duan, L.Y., Kot, A.C.: Skeletonbased online action prediction using scale selection network. IEEE Trans. Pattern Anal. Mach. Intell. **42**(6), 1453–1467 (2020). https://doi.org/10.1109/TPAMI.2019.2898954
8. Al-Janabi, S., Al-Janabi, Z.: Development of deep learning method for predicting DC power based on renewable solar energy and multi-parameters function. Neural Comput. Appl. (2023). https://doi.org/10.1007/s00521-023-08480-6
9. Wang, Y., et al.: 3DV: 3D dynamic voxel for action recognition in depth video. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR): 511–520 (2020). https://doi.org/10.1109/CVPR42600.2020.00059
10. Al-Janabi, S., Al-Barmani, Z.: Intelligent multi-level analytics of soft computing approach to predict water quality index ($IM^{12}CP$-WQI). Soft Comput. **27**, 7831–7861 (2023). https://doi.org/10.1007/s00500-023-07953-z
11. Kadhuim, Z., Al-Janabi, S.: Codon-mRNA prediction using deep optimal neurocomputing technique (DLSTM-DSN-WOA) and multivariate analysis. Results Eng. **17**, 100847 (2022). https://doi.org/10.1016/j.rineng.2022.100847
12. Al-Janabi, S., Alkaim, A., Al-Janabi, E., et al.: Intelligent forecaster of concentrations (PM2.5, PM10, NO2, CO, O3, SO2) caused air pollution (IFCsAP). Neural Comput. Appl. **33**, 14199–14229 (2021). https://doi.org/10.1007/s00521-021-06067-7
13. Wang, F., Song, Y., Zhang, J., Han., J., Huang, D.: Temporal unet: sample level human action recognition using wifi (2019). arXiv preprint arXiv:1904.11953 [Online]
14. Zhao, R., Ali, H., Vander, Smagt P.: Two-stream RNN/CNN for action recognition in 3D videos Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 4260–4267 (2017). https://doi.org/10.1109/IROS.2017.8206288
15. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: Skeleton-indexed deep multi-modal feature learning for high performance human action recognition. Proc. IEEE Int. Conf. Multimedia Expo (ICME): 1–6 (2018). https://doi.org/10.1109/ICME.2018.8486486
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need.Proc. Adv. Neural Inf. Process. Syst., 5998–6008 (2017). https://doi.org/10.48550/arXiv.1706.03762
17. Chen, J., Ho, C.M. : MM-ViT: multi-modal video transformer for compressed video action recognition. Proc. IEEE/CVF Win880 ter Conf. Appl. Comput. Vis. (WACV), 786–797 (2022). https://doi.org/10.1109/WACV51458.2022.00086
18. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Gao, W.: Pre-trained image processing transformer. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 12299–12310 (2021). https://doi.org/10.48550/arXiv.2012.00364
19. Parmar, N., et al. Image transformer (2018). arXiv preprint arXiv:1802.05751 [Online]
20. Zhou, L., et al.: End-to-end dense video captioning with masked transformer. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 8739–8748 (2018). https://doi.org/10.1109/CVPR.2018.00911
21. Zeng, Y., et al.: Learning joint spatial-temporal transformations for video inpainting. Proc. Eur. Conf. Comput. Vis. (ECCV), 528–543 (2020). arXiv:2007.10247. [Online]
22. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Proc. Adv. Neural Inf. Process. Syst. 568–576 (2014). https://doi.org/10.1002/14651858.CD001941.pub3
23. Diba, A., Sharma, V., Van Gool, L.: Deep temporal linear encoding networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit.

(CVPR), 2329–2338 (2017). https://doi.org/10.1109/CVPR.2017.168
24. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: towards good practices for deep action recognition (2016). arXiv preprint arXiv:1608.00859. [Online]
25. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. Proc. 32nd AAAI Conf. Artif. Intell.: 1–9 (2018). https://doi.org/10.48550/arXiv.1801.07455
26. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR): 12026–12035 (2019). https://doi.org/10.48550/arXiv.1805.07694
27. Chi, H.-G, Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: representation learning for human skeleton-based action recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR), 20154-20164 (2022). https://doi.org/10.1109/CVPR52688.2022.01955
28. Zhang, Y., Wu, B., Li, W., Duan, L., Gan, C.: Stst: spatial-temporal specialized transformer for skeleton-based action recognition. Proc. 29th ACM Int. Conf. Multimedia, 3220–3228 (2021). https://doi.org/10.1145/3474085.3475473
29. Li, X., Zhang, J., Wang, S., et al.: Two-stream spatial graphormer networks for skeleton-based action recognition. IEEE Access **2022**(10), 100426–100437 (2022). https://doi.org/10.1002/14651858.CD001941.pub3
30. Mehta, S., et al.: DeLighT: deep and light-weight transformer (2021). arXiv preprint arXiv:2008.00623 [Online]
31. Qiu, H., Hou, B., Ren, B., Zhang, X.: Spatio-temporal tuples transformer for skeleton-based action recognition (2022). arXiv preprint arXiv:2201.02849 [Online]
32. Zolfaghari, M., G. Oliveira, L., Sedaghat, N., Brox, T.: Chained multistream networks exploiting pose, motion, and appearance for action classification and detection. Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2904–2913. https://doi.org/10.1109/iccv.2017.316
33. Liu, J., Li, Y., Song, S., Xing, J., Lan, C., Zeng, W.: Multi-modality multi-task recurrent neural network for online action detection. IEEE Trans. Circuits Syst. Video Technol. **29**(9), 2667–2682 (2019). https://doi.org/10.1109/TCSVT.2018.2799968
34. Li, J., Xie, X., Pan, Q., Cao, Y., Zhao, Z., Shi, G.: SGM-Net: skeletonguided multimodal network for action recognition. Pattern Recognit. Art. 107356 (2020). https://doi.org/10.1016/j.patcog.2020.107356
35. Xu, W., Wu, M., Zhao, M., Xia, T.: Fusion of skeleton and RGB features for RGB-D human action recognition. IEEE Sens. J. **21**(17), 19157–19164 (2021). https://doi.org/10.1109/JSEN.2021.3089705
36. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 6450–6459 (2018). https://doi.org/10.48550/arXiv.1711.11248
37. Zhang, Z.: Microsoft Kinect sensor and its effect. IEEE MultiMedia Mag. **19**(2), 4–10 (2012). https://doi.org/10.1109/MMUL.2012.24
38. Hu, J.F., Zheng, W.S., Pan, J., Lai, J., Zhang, J.: Deep bilinear learning for RGB-D action recognition. Proc. Eur. Conf. Comput. Vis. (ECCV), 335–351 (2018). https://doi.org/10.1007/978-3-030-01234-2_21
39. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit (CVPR), 11531–11539 (2020). https://doi.org/10.1109/CVPR42600.2020.01155
40. Gao, Y., Beijbom, O., Zhang, N., Darrell, T.: Compact bilinear pooling. Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR), 317–326 (2016). https://doi.org/10.1109/CVPR.2016.41
41. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: A large scale dataset for 3D human activity analysis. Proc. Comput. Vis. Pattern Recognit (CVPR), 1010–1019 (2016). https://doi.org/10.48550/arXiv.1604.02808
42. Liu, J., Shahroudy, A., Perez, M.L., Wang, G., Duan, L.-Y., Chichung, A.K.: NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2684–2701 (2019). https://doi.org/10.1109/TPAMI.2019.2916873
43. Chen, C., Jafari, R., Kehtarnavaz, N.: UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. Proc. IEEE Int. Conf. Image Process. (ICIP), 168–172 (2015). https://doi.org/10.1109/ICIP.2015.7350781
44. Liu, G., Qian, J., Wen, F., Zhu, X., Ying, R., Liu, P.: Action recognition based on 3D skeleton and RGB frame fusion. Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 258–264 (2019). https://doi.org/10.1109/IROS40897.2019.8967570
45. De Boissiere, A.M., Noumeir, R.: Infrared and 3D skeleton feature fusion for RGB-D action recognition. IEEE Access., 168297–168308 (2020). https://doi.org/10.1109/ACCESS.2020.3023599
46. Su, L., Hu, C., Li, G., Cao, D.: MSAF: Multimodal split attention fusion (2020). arXiv:2012.07175 [Online]
47. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: MMTM: Multimodal transfer module for CNN fusion. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 13289–13299 (2020). https://doi.org/10.48550/arXiv.1911.08670
48. Das, S., Sharma, S., Dai, R., Bremond, F., Thonnat, M.: VPN: Learning video-pose embedding for activities of daily living. Proc. Eur. Conf. Comput. Vis., 72–90 (2020). https://doi.org/10.1007/978-3-030-58545-7_5
49. Liu, J., Shahroudy, A., Dong, X., Gang, W.: Spatio-temporal LSTM with trust gates for 3D human action recognition. Proc. Eur. Conf. Comput. Vis., 816–833 (2016). https://doi.org/10.1007/978-3-319-46487-9_50
50. Liu, M., Yuan, J.: Recognizing human actions as the evolution of pose estimation maps. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 1159–1168 (2018). https://doi.org/10.1109/CVPR.2018.00127
51. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome: Real-world activities of daily living. Proc. IEEE Int. Conf. Comput. Vis.(ICCV), 833–842 (2019). https://doi.org/10.1109/ICCV.2019.00092
52. Verma, P., Sah, A., Srivastava, R.: Deep learning-based multimodal approach using RGB and skeleton sequences for human activity recognition. Multimed. Syst. **26**(6), 671–685 (2020). https://doi.org/10.1007/s00530-020-00677-2
53. Li, C., Hou, Y., Wang, P., Li, W.: Joint distance maps based action recognition with convolutional neural networks. IEEE Signal Process. Lett. **24**(5), 624–628 (2017). https://doi.org/10.1109/LSP.2017.2678539
54. Liu, T., Kong, J., Jiang, M.: RGB-D action recognition using multimodal correlative representation learning model. IEEE Sens. J. **19**(5), 1862–1872 (2019). https://doi.org/10.1109/JSEN.2018.2884443
55. Islam, M.M., Iqbal, T.: HAMLET: A hierarchical multimodal attention-based human activity recognition algorithm. Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 1–8. 406–413 (2020). https://doi.org/10.1109/IROS45743.2020.9340987