Multimedia Tools and Applications (2024) 83:73391–73405
https://doi.org/10.1007/s11042-023-17788-3

1232: HUMAN-CENTRIC MULTIMEDIA ANALYSIS

VT-BPAN: vision transformer-based bilinear pooling and attention network fusion of RGB and skeleton features for human action recognition

Yaohui Sun[1] · Weiyao Xu[2,3] · Xiaoyi Yu[4] · Ju Gao[3]

Received: 7 March 2023 / Revised: 8 September 2023 / Accepted: 30 November 2023 /
Published online: 11 December 2023

Abstract
Recent generation Microsoft Kinect Camera captures a series of multimodal signals that provide RGB video, depth sequences, and skeleton information, thus it becomes an option to achieve enhanced human action recognition performance by fusing different data modalities. However, most existing fusion methods simply fuse different features, which ignores the underlying semantics between different models, leading to a lack of accuracy. In addition, there exists a large amount of background noise. In this work, we propose a Vision Transformer-based Bilinear Pooling and Attention Network (VT-BPAN) fusion mechanism for human action recognition. This work improves the recognition accuracy in the following ways: 1) An effective two-stream feature pooling and fusion mechanism is proposed. The RGB frames and skeleton are fused to enhance the spatio-temporal feature representation. 2) A spatial lightweight multiscale vision Transformer is proposed, which can reduce the cost of computing. The framework is evaluated based on three widely used video action datasets, and the proposed approach performs a more comparable performance with the state-of-the-art methods.

Yaohui Sun, Weiyao Xu, Xiaoyi Yu, and Ju Gao contributed equally to this work

✉ Yaohui Sun
mwg199391@163.com; xuweiyao_2008@126.com

✉ Weiyao Xu
xuweiyao_2008@126.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 Laboratory of Advanced Information Network, Beijing University of Posts and Telecommunications, Beijing 100876, China

3 School of Opto-Electronic Engineering, Zaozhuang University, Zaozhuang 277160, China

4 School of Artificial Intelligence, Zaozhuang University, Zaozhuang 277160, China

Keywords Microsoft kinect camera · Vision transformer · Self-attention · Multi-head pooling attention · Feature fusion

1 Introduction

Human action recognition (HAR), simply explained as identifying human behavior, plays an important role in many engineering practices. The input for action recognition is a video clip. The general approach is to sample video clips into several frames, which is actually a video classification task from a presentation perspective. Technically speaking, it is mainly a spatiotemporal feature learning technique, which involves learning the temporal and spatial features in a video. Spatial features refer to what people or objects are included in the video; The temporal feature is how the people or objects in the video move. For example, automatic navigation systems [1] and AI video surveillance [2]. In addition, it is also important for many other related fields, including smart cities [3], traffic management [4], etc.

Due to the development of various precise and economical sensors, as well as the new generation of RGB D cameras. HAR adopts different modes, such as skeleton, depth, radar, etc. According to the application scenario, RGB-D cameras can capture depth images, bone state information, and other state information. Multimodal HAR has significant advantages [5, 6]. The CNN based algorithm is a relatively classic algorithm that is common in practical application scenarios. Zhao et al. [5] proposed a dual flow network composed of recurrent neural networks (RNNs) and convolutional neural networks (CNN) for independent processing of RGB and skeleton data. Song et al. [6] proposed a continuous deep CNN learning framework consisting of two skeleton guided flows, and used this network to extract features from RGB and optical flows.

Transformer [7] has been considered a new type of deep learning model since its inception. Due to its powerful functionality and broad prospects, it has recently taken a leading position in the field of machine learning. Recently, Transformer has been applied to critical computer vision tasks. At present, only a few works [8–10] use Transformers in low-level vision, and further research is needed. Due to the continuous nature of the video, Transformer is essentially suitable for video tasks [11, 12], and its performance begins to compare to traditional CNN and RNN. Most of the current enthusiasm for applying Transformer to visual tasks began with Vision Transformer (ViT) [13]. An emerging work aims to apply Transformer to visual tasks such as object detection [14], semantic segmentation [15], 3D reconstruction [16], medical image segmentation [17] etc. In this paper, we directly establish a phased model that allows channel expansion and resolution downsampling based on [13, 25]. Our goal is to connect the multi-scale feature hierarchy with the transformer model.

The HAR method based on Transformer's multimodal fusion still poses challenges in achieving the goal of effective modal fusion. More specifically, there are at least three challenges. Firstly, how to obtain greater action context information from multimodal data, and how to capture richer feature information through attention mechanism integration models. Secondly, most of the features in the video are extracted from the entire frame, which includes a large amount of background noise, and the objects that undergo actions are easily overlooked. Concurrently, it is necessary to consider advanced features that are rough but complex in space to model visual semantics. Thirdly, most existing multimodal methods have complex structures and high computational costs. Therefore, it is necessary to design lightweight channel capacity to effectively solve multimodal HAR problems.

Motivated by the above works, the main contributions of this work are as follows.

- To enhance fusion accuracy, we employ two complementary flow techniques, namely a spatial ViT architecture and an attention module designed for visual data modeling. These components enable effective fusion of multimodal data and facilitate end-to-end training. Notably, in contrast to the traditional Transformer as mentioned in reference [25], our approach utilizes Multi-head pooling attention to pool the sequence of latent tensors, offering multiple channel scales.
- An effective data preprocessing method has been adopted for RGB videos and skeleton sequences, which can help networks capture richer feature information and more accurately capture human actions.
- The proposed VT-BPAN module effectively handles spatial roughness but complex high-level features to model visual semantics. In addition, a spatial lightweight improvement of the ViT that can reduce computational costs has been proposed.Through experimental analysis of multimodal datasets, the proposed VT-BPAN module has significantly improved in action recognition compared to existing research results.

The remainder of this paper is organized as follows. An overview of the related work is given in Section 2. Section 3 describes the proposed VT-BPAN in detail, and then its experimental implementation setup and the experimental results and discussion are presented in Section 4. Finally, Section 5 concludes the paper.

2 Related works

2.1 Human action recognition integrating RGB video and skeleton data

RGB mode is typically captured using RGB cameras, whereas skeleton data naturally encodes joint positions through coordinates. Consequently, this skeleton information possesses a higher level of abstraction compared to RGB data. Additionally, it demands fewer computational resources and offers enhanced robustness. The human skeleton's structure can be portrayed as a graph, with each vertex representing a human joint, and the connections between these vertices forming the human skeletal structure. In their work, Simonyan and Zisserman [18] introduced a classic dual-flow framework encompassing spatial and temporal networks. Furthermore, they proposed a long-term recursive convolutional network (LRCN) composed of 2D cellular neural networks in [19]. This network's purpose is to extract RGB features at the frame level and subsequently generate a single action label using LSTM. Yan et al. [20] introduced the spatiotemporal GCN (ST-GCN) for human action recognition based on skeleton data. Another approach, the two-stream adaptive GCN (2S-AGCN), was presented in [21]. Moreover, Chi et al. [22] introduced InfoGCN, which incorporates information bottlenecks to facilitate the learning of complex actions.

Transformers have demonstrated significant potential in processing sequential data. In [23], Qiu et al. proposed a spatio-temporal tuple Transformer (STTFormer) framework. Plizzari et al. [24] introduced a spatiotemporal transformation network (ST-TR) structure, which considers learning inter-frame motion dynamics and intra-frame joint interactions through spatial and temporal self-attention modules.Additionally, in [25], a Multiscale Vision Transformer (MViT) was proposed for video and image recognition. The MViT combines a multi-scale feature hierarchy with Transformer architecture, incorporating several channel resolution scale stages. These stages start from the input resolution and small channel dimensions, progressively expanding channel capacity while reducing spatial resolution.

2.2 Multi-modal data fusion methods

The exploration of deep learning architectures that seamlessly integrate RGB and HAR skeleton features has garnered significant attention. Zolfagari [26] applied a 3DCNN to process both pose and motion information extracted from the original RGB images. They employed Markov chain models to fuse this flow of information for action classification. In a more recent development, Li et al. [27] introduced a two-stream model comprising R (2+1) D networks, ST-GCN networks, and guided blocks that enhance action-related information in videos. Subsequently, they employed score fusion techniques for model classification. Building upon a 3D CNN network, Das et al. [28] proposed a method utilizing RGB video as input. They devised a pose-guided spatiotemporal attention network to capture relevant information effectively. In a different approach, Xu et al. [29] introduced the BPAN model, incorporating spatiotemporal bilinear pooling and attention fusion techniques, which proficiently achieved feature fusion.

3 Proposed model

In this section, the data preprocessing module and feature extraction strategy are introduced, and the feature fusion strategy is explained. Two deep learning frameworks are used to extract features separately, and the features are fused through the VT-BPAN module.

3.1 Preprocessing module

In this paper, we utilize the approach detailed in [29] to direct the focus of video detection towards the human body in RGB video, as depicted in Fig. 1. We achieve this by cropping the actions of the human body in the input RGB image through pose mapping.

In addition, for skeleton sequences, in order to better describe the spatiotemporal sequence of the skeleton, we used time difference and relative coordinates. As shown in Fig. 2, the relative coordinate x_r can be obtained based on the distance between all joints in each frame and the distance between the center joints. x_t represents the time difference, which can be calculated as follows.

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

where the notation $x[t]$ signifies the data at frame t, and the final input data is concatenated and combined from x, x_r, and x_t. Building upon the framework of 2S-AGCN [21], where the original channel C is set to 3 to denote the three-dimensional coordinates of each joint, we employ preprocessing modules to extract additional information features. This module expands the input channel C to 9. Within 2S-AGCN, a spatial GCN block is introduced, which can be computed using the following equation.

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

where K_v denotes the dimension of the kernel, A_K corresponds to the adjacency matrix, B_k bears a similarity to M_k as seen in ST-GCN, and C_k represents the learned sample graph.

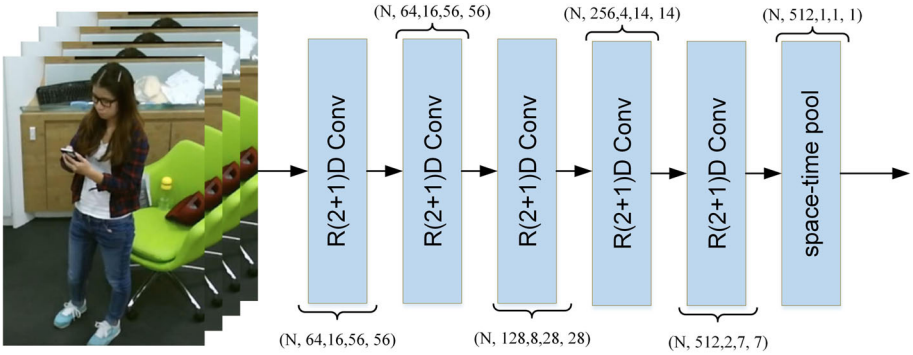


Fig. 1 An overall of skeleton preprocessing and the architecture of R(2+1)D network

3.2 Feature extraction module

In this section, the R (2+1) D network [30] and the 2S-AGCN network [21] are used for feature extraction, respectively. Specifically, for RGB video streams, pre training is required on Kinetics-400 [31]. The schematic diagram of the R (2+1) D recognition block is shown in Fig. 1. A complete 3D convolution can be more conveniently approximated through 2D and 1D convolutions, dividing spatial and temporal modeling into two separate steps.The network architecture of R(2+1)D is composed of M_i 2D convolutional filters of size $N_{i-1} \times 1 \times d \times d$ and N_i 1D time convolution filters of size $M_i \times t \times 1 \times 1$. The hyperparameter M_i determines the dimension of the intermediate subspace, where the signal is projected between spatial and temporal convolutions. As the input, it regards the video data with the size of $3 \times T \times 112 \times 112$, where 3 represents the number of RGB channels, 112 corresponds to the image height and width, as well as T corresponds to the length of the sequence.

Regarding the 3D skeleton stream, we employ the 2S-AGCN network. Initially, we acquire and process the skeleton data, as illustrated in Fig. 2. The skeleton sequence is characterized by dimensions $T \times C \times V \times M$, where T denotes the frames in the sequence, C denotes the number of channels, V represents the joints, and M represents the skeletons within each frame. Following the application of R(2+1)D blocks, the 2S-AGCN network, and subsequent dimensional transformation, the resulting features are standardized to have the same dimension.

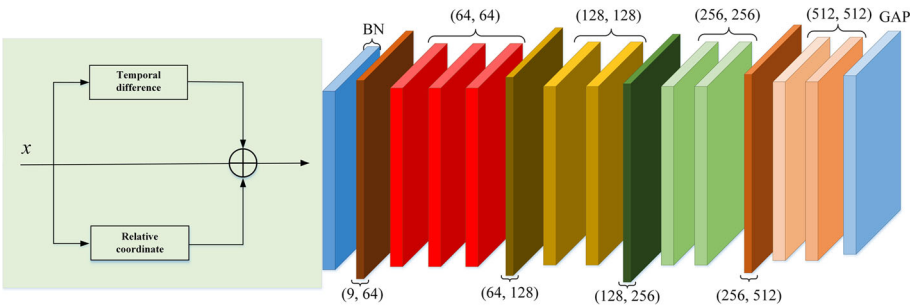


Fig. 2 An overall of skeleton preprocessing and the architecture of 2S-AGCN with temporal difference

3.3 Feature fusion module

RGB features and skeleton features are often fused in computer vision tasks, particularly in HAR and pose estimation, because they capture different aspects of visual information and complement each other. The fusion of these features is designed to enhance the overall understanding of the scene or task, rather than leading to information redundancy. Here are explanations of how redundancy is mitigated in the fusion of RGB and skeleton features:

1). Complementary Information: RGB features capture visual information about the appearance and color of objects and scenes in a video or image. On the other hand, skeleton features capture information about the spatial and temporal positions of key body joints or key points. These two types of features provide complementary information. RGB features can help recognize objects and their interactions, while skeleton features are valuable for understanding human pose and movement. 2). Robustness: Fusing RGB and skeleton features can enhance the robustness of a computer vision system. RGB features may be sensitive to changes in lighting conditions, occlusions, or cluttered backgrounds, which skeleton features are less affected by. By combining both, the system can better adapt to varying real-world scenarios. 3). Improved Accuracy: Integrating RGB and skeleton features often results in improved recognition or classification accuracy. The fusion allows the model to capture both the appearance and motion cues, which can lead to more discriminative and informative representations for tasks like action recognition, gesture recognition, or human pose estimation.

In this section, we introduce the incorporation of Transformer and Bilinear Pooling techniques [32, 33] to combine the features extracted by two preprocessing models. We begin by elucidating the Multi Head Pool Attention (MHPA), a self-attention mechanism that facilitates adaptable resolution modeling within the Transformer block. This enables the operation of multi-scale converters at gradually changing spatiotemporal resolutions. The VT-BPAN module we propose is visualized in Fig. 3. This module seamlessly integrates the familiar structure of the Transformer block, as depicted in Fig. 4, offering a lightweight enhancement that efficiently fuses multimodal information for RGB-D action recognition. Here, we denote F_{RGB} and F_{SKE} as the features extracted from the RGB and skeleton modules, respectively.

3.4 Multi head pooling attention

Consider the C-dimensional input tensor X with a sequence length of L and $X \in R^{L\times C}$. According to MHA [25], MHPA projects input X into the intermediate query tensor $\hat{Q} \in R^{L\times C}$, key tensor $\hat{K} \in R^{L\times C}$, and value tensor $\hat{V} \in R^{L\times C}$ through linear operations

$$\hat{Q} = XW^{Q} \quad \hat{K} = XW^{K} \quad \hat{V} = XW^{V} \tag{3}$$

Define the pooling operator $P(\cdot;\Theta)$, where $\Theta := (\mathrm{k}, \mathrm{s}, \mathrm{p})$, $P(X;\Theta) \in R^{\tilde{L}\times C}$, $\tilde{L} = \tilde{T} \times \tilde{H} \times \tilde{W}$. The operator uses the pool kernel k of dimension $k_T \times k_H \times k_W$, the stride size s of corresponding dimension $s_T \times s_H \times s_W$, and the padding p of corresponding dimension $p_T \times p_H \times p_W$ to reduce the input tensor of dimension $\mathrm{L} = T \times H \times W$ to $\tilde{\mathrm{L}}$,

$$\tilde{\mathrm{L}} = \left[\frac{\mathrm{L} + 2\mathrm{p} - \mathrm{k}}{\mathrm{s}}\right] + 1 \tag{4}$$

We default to using overlapping kernels k with conformal padding p in the pooling attention operator, so that the sequence length of $\tilde{\mathrm{L}}$, the output tensor $P(\cdot;\Theta)$, undergoes an overall

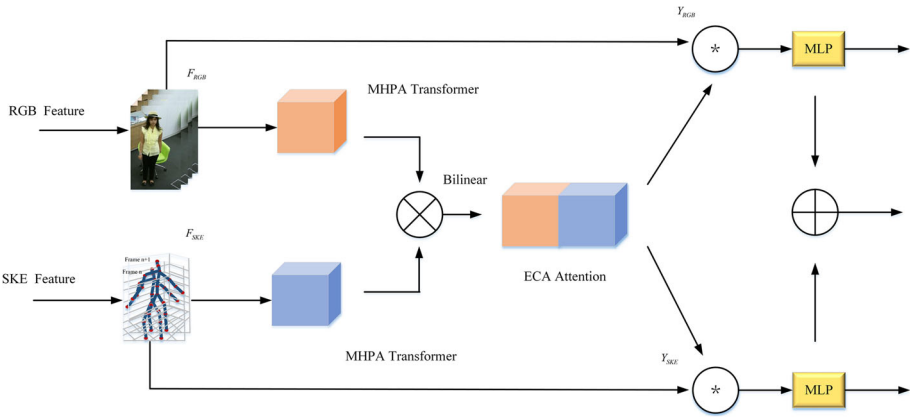


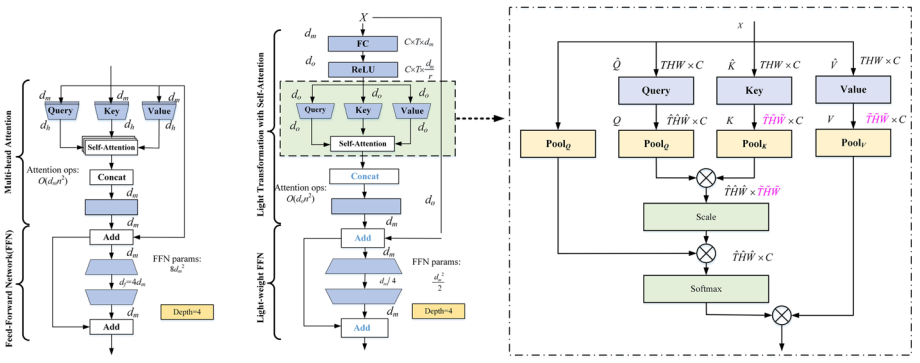
Fig. 3 The VT-BPAN model features a comprehensive architecture with two distinct streams: a skeleton sequence and an RGB frame. The 2S-AGCN network is dedicated to extracting skeleton features, whereas the R(2+1)D network specializes in extracting RGB features. These extracted features are categorized into three components: RGB features, skeleton features, and the ultimate fusion features. The network is meticulously designed to facilitate effective feature fusion

reduction of $s_T s_H s_W$ factors.

$$Q = P\left(\hat{Q}; \Theta_Q\right) \quad K = P\left(\hat{K}; \Theta_K\right) \quad V = P\left(\hat{V}; \Theta_V\right) \tag{5}$$

A standard Transformer block typically comprises several components: multi-head attention layers, feed-forward networks (FFN), layer normalization, and shortcut connections. In practice, a group of query attention functions must be calculated concurrently and organized into matrices denoted as Q, with K and V representing the key and value matrices. The attention output and pooling attention can then be computed as follows:

$$\text{Attention}\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{6}$$



(a) Standard Transformer Block (b) Light-weight Vision Transformer Block with MHPA

Fig. 4 (a, b) Block-wise comparison between the standard Transformer block and the MHPA light-weight Transformer

$$\mathrm{PA}(\cdot) = softmax\left(\frac{P\left(Q;\Theta_Q\right)P(K;\Theta_K)^T}{\sqrt{d}}\right)P\left(V;\Theta_V\right) \tag{7}$$

where d_k represents the dimension of the Key vector, and for simplicity, we set $d = d_k$. Subsequently, each distinct head receives its unique set of query matrices, key matrices, and value matrices. This arrangement enables the input vectors to be projected into distinct representation subspaces. Furthermore, the MHA mechanism permits the model to emphasize information from various subspaces at different spatial locations. This process is illustrated as follows.

$$\mathrm{MHA}(Q, K, V) = \mathrm{Concat}\left(\mathrm{head}_1, \ldots, \mathrm{head}_h\right)W^O \tag{8}$$

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 $W_i^O \in R^{hd_v\times d_{\mathrm{model}}}$.

Every encoder layer consists of both a fully connected FFN and an attention layer. We can represent each encoder layer as the following function:

$$\mathrm{FFN}(x) = max\left(0, xW_1 + \beta_1\right)W_2 + \beta_2 \tag{9}$$

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

The design of the spatial lightweight Transformer encoder layer draws inspiration from ST-TR [24] and DelighT [34]. It incorporates an enhanced network architecture and adjusted weight dimensions, as depicted in Fig. 4(b).

We consider the input tensor X, which has a shape of $T \times d_m \times C$, where $d_m = H \times W$. The initial focus is on compressing the output feature information.

$$\mathbf{S}_k = \sigma\left(\tilde{W}_1 ReLU\left(X\tilde{W}_2\right) + b\right) \tag{10}$$

where $\tilde{W}_1 \in R^{\frac{d_m}{r}\times d_m\times C}$ and $\tilde{W}_2 \in R^{C\times 1\times d_m}$ represent weight vectors, $d_o = \frac{d_m}{r}$. σ represents the Sigmoid activation function. b is the bias vector.

Subsequently, we obtain 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 $X_V \in R^{T\times C\times 1\times d_v}$ by rearranging the input data. This transformation is applied independently to each frame within the T dimension. The matrices Q, K, and V are derived by multiplying $W_Q \in R^{d_o\times 1\times C}$, $W_K \in R^{d_o\times 1\times C}$, and $W_V \in R^{d_o\times 1\times C}$ with their respective inputs and subsequently undergoing dimensional transformation.

In contrast to the conventional spatial self-attention module, our approach begins by compressing the embedded dimension within each spatial lightweight Transformer encoder layer. Here is the formula used to compute the self-attention matrix:

$$\mathrm{Attention}(Q, K, V) = softmax\left(\frac{\left(W_Q X_Q\right)\left(W_K X_K\right)^T}{\sqrt{d_k}} + \mathbf{S}_k\right)\left(W_V X_V\right) \tag{11}$$

$$\mathrm{Attention}(Q, K, V) = softmax\left(\frac{P\left(W_Q X_Q;\Theta_Q\right)\left(W_K X_K;\Theta_K\right)^T}{\sqrt{d_k}} + \mathbf{S}_k\right)P\left(W_V X_V;\Theta_V\right) \tag{12}$$

Because the dimension of T varies within the batch, it is possible to efficiently share parameters along the time dimension, applying transformations independently to each frame. Following the transformations described above, the self-attention matrix yields an output

shape of $T \times d_o \times d_v$. Consequently, we can obtain the output of our model through a straightforward reshaping process.

Figure 4(b) illustrates the integration of a spatial lightweight transformation into the Transformer block. When comparing this configuration with the standard Transformer block and the lightweight Transformer block, the computational cost associated with calculating attention is denoted as $O(d_m n^2)$ and $O(d_o n^2)$, respectively, where it's important to note that $d_o < d_m$. Consequently, the lightweight Transformer block succeeds in reducing the attention calculation cost by a factor of $\frac{d_m}{d_o}$. In our experiments, we specifically set $d_o = \frac{d_m}{16}$, which resulted in a 16-fold reduction in multiplicative addition operations when compared to the original Transformer structure. It's worth highlighting that the advantage of the dot product lies in its speed and spatial efficiency during operation, mainly because it allows for the utilization of highly optimized matrix multiplication routines in its implementation [7].

Consider a lightweight FFN architecture. In this structure, the first layer reduces the input dimension from d_m to $\frac{d_m}{r}$, and the second layer subsequently expands it back to d_m, where r represents the reduction factor. This lightweight FFN design significantly decreases the number of parameters and computational operations in the FFN, achieving a reduction factor of $r \cdot d_f = d_m$, where d_f is the original FFN dimension. In a standard Transformer model, the FFN typically expands the dimension by a factor of 4. For our experiments, we have chosen to set $r = 4$. Consequently, this lightweight FFN results in a remarkable 16-fold reduction in the number of FFN parameters.

3.5 Compact bilinear pooling

In this study, we employ the compact bilinear pooling (CBP) method [33] to handle the fusion feature. The CBP method is explained as follows

$$\begin{aligned}\langle F_{RGB}(\mathcal{X}), F_{SKE}(\mathcal{Y})\rangle = \sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}}\langle x_s, y_u\rangle^2 &\approx \sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}}\langle \phi(x), \phi(y)\rangle \\ &\equiv \langle C(\mathcal{X})C(\mathcal{Y})\rangle \end{aligned} \tag{13}$$

Define $C(\mathcal{X})$ as the summation of $\phi(x_s)$ over all elements s in set $\mathcal{S}$, and $C(\mathcal{Y})$ as the summation of $\phi(y_u)$ over all elements u in set $\mathcal{U}$. Here, C represents the feature channel. We introduce F_{fusion} as the fusion feature. Subsequently, it can be subjected to L2 regularization for normalization.

$$F_{fusion} = \frac{F_{fusion}}{\left\| F_{fusion} \right\|_2} \tag{14}$$

Next, we take the obtained F_{fusion} and pass it through both a fully connected layer and a ReLU layer. Consequently, we describe the resulting new fusion feature as $Z_{fusion} \in \mathrm{R}^{C\times 1}$.

$$Z_{fusion} = ReLU\left(\tilde{W} \cdot F_{fusion}\right) \tag{15}$$

where $\tilde{W} \in \mathrm{R}^{C\times C^2}$ represents the weight matrix, the network is structured to enhance feature expressiveness through attention mechanisms, following a design similar to ECANet. To compute attention weights, Conv1d is employed.

Then, we can compute Y_{RGB} and Y_{SKE} as

$$Y_{RGB} = F_{RGB} \odot \sigma\left(Conv1D\left(Z_{fusion}\right)\right) \tag{16}$$

$$Y_{SKE} = F_{SKE} \odot \sigma\left(Conv1D\left(Z_{fusion}\right)\right) \tag{17}$$

where σ represents the Sigmoid activation function, and $\odot$ denotes the Hadamard product.

In this study, we incorporate a multilayer perceptron (MLP) as our classification module. In this module, a batch normalization layer and the ReLU activation function are sequentially connected. Subsequently, a softmax function is applied to normalize the predictions into probabilistic distributions. Finally, the two features are combined through summation to yield the ultimate fusion feature denoted as 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.

$$L_{total} = \lambda_1 L_{RGB} + \lambda_2 L_{SKE} \tag{18}$$

where λ_1 and λ_2 are weighting factors, while L_{RGB} signifies the loss from the RGB stream, and L_{SKE} represents the loss from the skeleton stream.

4 Experiments

We perform a series of experiments on three publicly available datasets to assess the efficacy of our proposed VT-BPAN for HAR. Additionally, we conduct a comprehensive ablation study to examine the performance of each individual module.

4.1 Datasets

1) NTU-RGB+D dataset [36] :The dataset is considered one of the most widely utilized datasets in the field of HAR. It comprises a total of 56,880 video samples, encompassing 60 distinct action classes. Furthermore, it offers two evaluation scenarios: cross-subject (CS) and cross-view (CV). The experiments were conducted with respect to both of these evaluation scenarios, and we report the highest achieved recognition accuracy in each case.
2) MSR daily activity dataset [37] : The MSR daily activity dataset serves as a benchmark dataset focused on 3D human-object interaction actions. It encompasses 320 RGB-D videos featuring 16 unique actions performed twice by each of the 10 participants. These actions were executed once in a sitting position and once in a standing position.
3) NTU-RGB+D 120 dataset [38] : 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.

4.2 Implementation details

We conducted all experiments using two NVIDIA TITAN RTX GPUs within the PyTorch framework. Specifically, for the RGB stream, we resized RGB frames to 112×112, set the video sequence length to 20, and utilized an RGB model pre-trained on the Microsoft Kinect V2 [31]. Regarding the skeleton stream, we kept the skeleton sequence length at 50, and the remaining parameters followed the choices detailed in [21].

Our model's weight decay and learning rate were set to 0.0001 and 0.01, respectively, and we employed stochastic gradient descent for optimization with cross entropy serving as the loss function for backpropagation. The learning rate was set at 0.01.

4.3 Ablation study

In this section, to showcase the model's performance, we assess training and testing at the conclusion of each epoch, and we meticulously record the model's accuracy for each epoch.

1) Impact of fusion schemes: We evaluated several feature fusion methods, including averaging, multiplication, summation, concatenation, and maximum fusion. The outcomes are presented in Table 1. Notably, as Table 1 illustrates, the fusion technique led to a substantial enhancement in recognition accuracy, reaching an impressive 95%. Remarkably, the VT-BPAN fusion module exhibited the highest accuracy compared to the aforementioned models.
2) Impact of feature extraction mdels: In this paper, our video HAR approach can be succinctly summarized in two main paradigms. On one hand, we commonly employ a 2D+1D framework, wherein a 2D CNN is used to process each frame individually, followed by a 1D module that consolidates features from each frame. On the other hand, we also explore an alternative approach utilizing a 3D CNN with stacked 3D convolutions to jointly model temporal and spatial semantics.

   Drawing from the insights presented in [29], we conduct a comparative analysis of recognition accuracy across three convolutional neural network architectures: the 3D ResNet, MC3 ResNet [30], and the R(2+1)D model built upon the foundation of ResNet-18. The results unambiguously demonstrate that the R(2+1)D model consistently achieves the highest accuracy among the tested architectures.
3) Visual analysis: To enhance the illustration of VT-BPAN's self-attention impact, we conduct a comparative evaluation between the ECANet attention and the fusion self-attention generated by VT-BPAN. The outcomes of this analysis are presented in Fig. 5. Notably, the saliency maps derived from these attention mechanisms provide insights into the significance of individual pixels within each input image. It is evident from the results that VT-BPAN excels in capturing more meaningful pixels, underscoring its effectiveness in highlighting important image features.

4.4 Comparisons with the state-of-the-art

To ensure a fair and meaningful comparison, we assess our method against algorithms that also combine RGB and skeleton features, sharing similarities with our approach. Adhering to the evaluation criteria established in the original framework, we provide accuracy results for the NTU-RGB+D, NTU-RGB+D 120, and MSR daily activity datasets, as shown in

Table 1 Impact of fusion scheme on performance

Methods	Accuracy
Average	93.07
Multiplication	92.20
Sum	93.09
Concatenation	93.16
BPAN (Resnet 18)	94.85
HAMLET (MAT-CONCAT) [48]	95.12
VT-BPAN [ours]	**95.55**

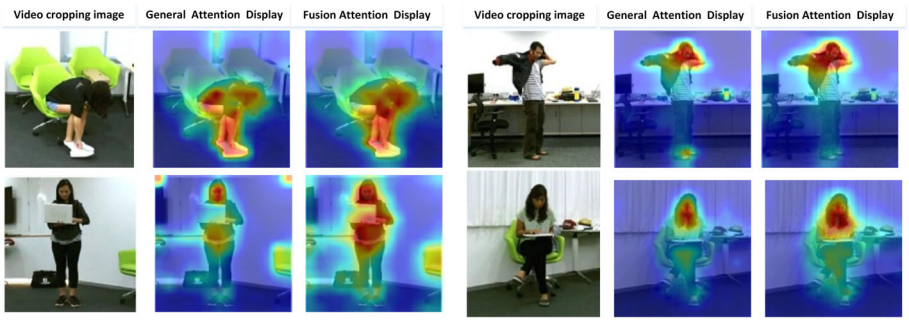


Fig. 5 Visualization results of ViT fused MPHA module

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

Methods	year	Multi-model	CS	CV
BI-LSTM [39]	2019	Yes	85.4	91.6
FUSION [40]	2020	Yes	91.8	94.9
MSAF [41]	2020	Yes	92.24	-
MMTM [42]	2020	Yes	91.9	95.3
VPN(I3D) [43]	2020	Yes	93.5	96.2
BPAN (Resnet 18) [29]	2021	Yes	94.85	97.4
VT-BPAN [ours(a)]	2022	Yes	**95.12**	**97.37**

Table 3 Comparison of MSR daily activity dataset model with the state-of-the-art

Methods	year	Multi-model	Top-1 Accuracy
SBR [44]	2019	✓	91.10
MCRL [45]	2019	✓	94.38
VT-BPAN [ours]	2022	✓	94.53

Table 4 Comparison of attention module and Transformer module in the MSR daily activity dataset

Methods	Multi-model	Top-1 Accuracy
VT-BPAN [ECA attention]	✓	94.50
VT-BPAN [ECA attention+MPHA Transformer]	✓	95.55

Table 5 Comparison of NTU-RGB+D 120 dataset model with the state-of-the-art

Methods	year	Multi-model	C-Sub	C-Set
separable STA [46]	2019	Yes	83.8	82.5
Verma et al. [47]	2020	Yes	76.7	77.9
VPN(I3D) [43]	2020	Yes	86.3	87.8
BPAN (Resnet 18) [29]	2021	Yes	86.6	88.1
VT-BPAN [ours(a)]	2022	Yes	**86.3**	**88.2**

Tables 2, 3, 4 and 5. We conduct an extensive comparative analysis between our method and other techniques, including simpler fusion-based methods (e.g., BI-LSTM [39]) and attention-based approaches (e.g., MMTM [42] and BPAN [29]).

The comparison between the ECA attention module and the ViT module on the MSR daily activity dataset is presented in Table 4. The experimental findings indicate that the combination of the Transformer module and ECA attention enhances accuracy by approximately 1.05%. Based on the above experimental results, the methodology employed in this paper combines Transformer-based bilinear pooling with an attention-driven approach, utilizing Resnet 18 as the backbone network. This amalgamation consistently outperforms the current state-of-the-art results, highlighting the advancement in action recognition achieved by our approach.

5 Conclusion

This paper introduces a novel multimodal HAR model, integrating both temporal and spatial feature extraction techniques by utilizing R(2+1)D and 2S-AGCN. Additionally, we propose the VT-BPAN module, designed to enhance feature expressiveness through feature fusion and the incorporation of a vision Transformer attention mechanism. Furthermore, we introduce a streamlined Transformer improvement. To validate the efficacy of the VT-BPAN module, we conduct a comprehensive comparison of various fusion strategies. Ultimately, we employ fully connected perceptrons to derive the final fusion features, enabling an end-to-end training process.

We evaluate our model on three benchmark datasets: NTU-RGB+D, NTU-RGB+D 120, and MSR daily activity dataset. Our experimental results demonstrate superior performance compared to existing methods. Considering the existing network limitations, our future research will focus on refining fusion methods for multimodal data and addressing challenges posed by heterogeneous networks.

Acknowledgements The author thanks for the National Nature Science Foundation of China (No.12175194).

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

Data Availability Statement Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

1. Diskin Y, Nair B, Braun A, Duning S, Asari V K (2013) Vision-based navigation system for obstacle avoidance in complex environments. In 2013 IEEE applied imagery pattern recognition workshop (AIPR) (pp. 1-8). IEEE
2. Lin W, Sun MT, Poovandran R, Zhang Z (2008) Human activity recognition for video surveillance. Proc IEEE Int Symp Circuits Syst 2737–2740
3. Othman NA, Aydin I (2021) Challenges and Limitations in Human Action Recognition on Unmanned Aerial Vehicles: A Comprehensive Survey. Trait Signal 38(5)
4. Adewopo V, Elsayed N, ElSayed Z, Ozer M, Abdelgawad A, Bayoumi M (2022) Review on action recognition for accident detection in smart city transportation systems. arXiv:2208.09588
5. Zhao R, Ali H, Vander Smagt P (2017) Two-stream RNN/CNN for action recognition in 3D videos Proc IEEE/RSJ Int Conf Intell Robots Syst (IROS): 4260–4267
6. Song S, Lan C, Xing J, Zeng W, Liu J (2018) Skeleton-indexed deep multi-modal feature learning for high performance human action recognition. Proc IEEE Int Conf Multimedia Expo (ICME): 1–6

7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Proc Adv Neural Inf Process Syst: 5998–6008
8. Chen J, Ho CM (2022) MM-ViT: Multi-modal video transformer for compressed video action recognition. Proc IEEE/CVF Winter Conf Appl Comput Vis (WACV): 786–797
9. Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, Ma S, Xu C, Gao W (2021). Pre-trained image processing transformer. Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR): 12299–12310
10. Parmar N et al (2018) Image transformer. [Online]. Available: arXiv:1802.05751
11. Zhou L et al (2018) End-to-end dense video captioning with masked transformer. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR): 8739–8748
12. Zeng Y et al (2020) Learning joint spatial-temporal transformations for video inpainting. Proc Eur Conf Comput Vis (ECCV): 528–543
13. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Houlsby N (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929
14. Beal J, Kim E, Tzeng E, Park D H, Zhai A, Kislyuk D (2020) Toward transformer-based object detection. arXiv:2012.09958
15. Yu Q, Wang H, Kim D, Qiao S, Collins M, Zhu Y, Chen LC (2022). Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2560–2570)
16. Lin K, Wang L, Liu Z (2021) End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1954–1963)
17. Chen J, Lu, Y, Yu Q, Luo X, Adeli E, Wang Y, Zhou Y (2021). Transunet: transformers make strong encoders for medical image segmentation. arXiv:2102.04306
18. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. Proc Adv Neural Inf Process Syst: 568-576
19. Donahue J, Hen LA, Saenko K (2015) Long-term recurrent convolutional networks for visual recognition and description. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR): 2625-2634
20. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. Proc 32nd AAAI Conf Artif Intell: 1-9
21. Shi L, Zhang Y, Cheng J, Lu H (2019) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR): 12026-12035
22. Chi H-g, Ha M H, Chi S, Lee SW, Huang Q, Ramani K (2022) Infogcn: Representation learning for human skeleton-based action recognition." Proc IEEE Conf Comput Vis Pattern Recognit (CVPR): 20154–20164
23. Qiu H, Hou B, Ren B, Zhang X (2022) Spatio-temporal tuples transformer for skeleton-based action recognition. [Online]. Available: arXiv:2201.02849
24. Plizzari C, Cannici M, Matteucci M (2021) Skeleton-based action recognition via spatial and temporal transformer networks. Comput, Vis. Image Understand. Art. no, p 103219
25. Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, Feichtenhofer C (2021) Multiscale vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6824–6835)
26. Zolfaghari M, G. Oliveira L, Sedaghat N, Brox T (2017) Chained multistream networks exploiting pose, motion, and appearance for action classification and detection. Proc IEEE Int Conf Comput Vis (ICCV):2904–2913
27. Li J, Xie X, Pan Q, Cao Y, Zhao Z, Shi G (2020) SGM-Net: Skeletonguided multimodal network for action recognition. Pattern Recognit, Art. no, p 107356
28. Das S , Dai R, Koperski M, Minciullo L, Garattoni L, Bremond F, Francesca G (2019) Toyota smarthome: Real-world activities of daily living. Proc IEEE Int Conf Comput Vis (ICCV):833–842
29. Xu W, Wu M, Zhao M, Xia T (2021) Fusion of skeleton and RGB features for RGB-D human action recognition. IEEE Sens J 21(17):19157–19164
30. Tran D , Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. Proc IEEE/CVF Conf Comput Vis Pattern Recognit: 6450–6459
31. Zhang Z (2012) Microsoft Kinect sensor and its effect. IEEE MultiMedia Mag 19(2):4–10
32. Hu JF, Zheng WS, Pan J, Lai J, Zhang J (2018) Deep bilinear learning for RGB-D action recognition. Proc Eur Conf Comput Vis (ECCV):335–351
33. Gao Y, Beijbom O, Zhang N, Darrell T (2016) Compact bilinear pooling. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR): 317–326
34. Mehta S et al (2021) DeLighT: Deep and Light-weight Transformer. [Online]. Available: arXiv:2008.00623
35. Wang Q , Wu B, Zhu P, Li P, Zuo W, Hu Q (2020) ECA-Net: Efficient channel attention for deep convolutional neural networks. Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR):11531–11539

36. Shahroudy A, Liu J, Ng T. T, Wang, G (2016) NTU RGB+D: A large scale dataset for 3D human activity analysis. Proc Comput Vis Pattern Recognit (CVPR):1010–1019
37. Wang J, Liu Z, Wu Y, Yuan J (2014) Learning actionlet ensemble for 3D human action recognition. IEEE Trans Pattern Anal Mach Intell 36(5):914–927
38. Liu J, Shahroudy A, Perez ML, Wang G, Duan L-Y, Chichung AK (2019) NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Trans Pattern Anal Mach Intell 42(10):2684–2701
39. Liu G, Qian J, Wen F, Zhu X , Ying R, Liu P,(2019) Action recognition based on 3D skeleton and RGB frame fusion. Proc IEEE/RSJ Int Conf Intell Robots Syst (IROS): 258–264
40. De Boissiere A M, Noumeir R (2020) Infrared and 3D skeleton feature fusion for RGB-D action recognition. IEEE Access: 168297–168308
41. Su L, Hu C, Li G, Cao D (2020) MSAF: Multimodal split attention fusion. [Online]. Available: arXiv:2012.07175
42. Joze HRV, Shaban A ,Iuzzolino ML, Koishida K (2020) MMTM: Multimodal transfer module for CNN fusion. Proc. IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR): 13289–13299
43. Das S, Sharma S, Dai R, Bremond F, Thonnat M (2020) VPN: Learning video-pose embedding for activities of daily living. Proc Eur Conf Comput Vis: 72-90
44. Shahroudy A, Ng T, Gong Y, Wang G (2018) Deep multimodal feature analysis for action recognition in RGB+D videos. IEEE Trans Pattern Anal Mach Intell 40(5):1045–1058
45. Liu T, Kong J, Jiang M (2019) RGB-D action recognition using multimodal correlative representation learning model. IEEE Sensors J 19(5):1862–1872
46. Das S , Dai R, Koperski M, Minciullo L, Garattoni L, Bremond F, Francesca G (2019) Toyota smarthome: Real-world activities of daily living. Proc IEEE Int Conf Comput Vis (ICCV): 833–842
47. Verma P, Sah A, Srivastava R (2020) Deep learning-based multimodal approach using RGB and skeleton sequences for human activity recognition. Multimed Syst 26(6):671–685
48. Islam MM, Iqbal T(2020) HAMLET: A hierarchical multimodal attention-based human activity recognition algorithm. Proc IEEE/RSJ Int Conf Intell Robots Syst (IROS): 1-8. 406–413