



ORIGINAL RESEARCH

Explore human parsing modality for action recognition

Jinfu Liu¹ | Runwei Ding^{2,3} | Yuhang Wen¹ | Nan Dai⁴ | Fanyang Meng² | Fang-Lue Zhang⁵ | Shen Zhao¹ | Mengyuan Liu³

¹School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China

²Peng Cheng Laboratory, Shenzhen, China

³State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, China

⁴Changchun University of Science and Technology, Changchun, China

⁵Victoria University of Wellington, Wellington, New Zealand

Correspondence

Shen Zhao and Mengyuan Liu.
Email: zhaosh35@mail.sysu.edu.cn and
nkliuyifang@gmail.com

Funding information

National Natural Science Foundation of China, Grant/Award Number: 62203476; Natural Science Foundation of Guangdong Province, Grant/Award Number: 2024A1515012089; Natural Science Foundation of Shenzhen, Grant/Award Number: JCYJ20230807120801002; Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars, Grant/Award Number: RCYX20231211090248064

Abstract

Multimodal-based action recognition methods have achieved high success using pose and RGB modality. However, skeletons sequences lack appearance depiction and RGB images suffer irrelevant noise due to modality limitations. To address this, the authors introduce human parsing feature map as a novel modality, since it can selectively retain effective semantic features of the body parts while filtering out most irrelevant noise. The authors propose a new dual-branch framework called ensemble human parsing and pose network (EPP-Net), which is the first to leverage both skeletons and human parsing modalities for action recognition. The first human pose branch feeds robust skeletons in the graph convolutional network to model pose features, while the second human parsing branch also leverages depictive parsing feature maps to model parsing features via convolutional backbones. The two high-level features will be effectively combined through a late fusion strategy for better action recognition. Extensive experiments on NTU RGB + D and NTU RGB + D 120 benchmarks consistently verify the effectiveness of our proposed EPP-Net, which outperforms the existing action recognition methods. Our code is available at <https://github.com/liujf69/EPP-Net-Action>.

KEYWORDS

computer vision, deep learning

1 | INTRODUCTION

Human action recognition is an important task in the field of computer vision, which also has great research value and broad application prospects in education [1], human–computer interaction [2] and content-based video retrieval [3]. It can also be integrated with fields such as motion prediction [4], pose estimation [5] and micro-expression generation [6] to accomplish more complex human-related tasks. Most related methods [7–10] use unimodal data for action recognition,

typically skeleton-based action recognition, which take human skeletons [11–14] as the input. In recent years, the methods [15, 16] of integrating multiple modalities have emerged to make effective use of multimodal features for better action recognition. Among them, the human skeleton and RGB modality are two widely adopted input modalities.

Actually the skeleton modality can be viewed as a natural topological graph, where the graph vertices and edges represent joints and bones of human body, respectively. The graph-structured skeleton can well represent body movements

Jinfu Liu and Runwei Ding means co-first authors.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). CAAI Transactions on Intelligence Technology published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

and is highly robust to environmental changes [17], thereby adopted in many studies using graph convolutional networks (GCNs) [4, 7, 8, 18–20]. Besides, the RGB modality has rich appearance information, therefore some prior studies in action recognition employ convolutional neural networks (CNNs) [15, 21] to model spatiotemporal features from RGB images and achieve the desired effect. In multimodal-based action recognition, the features and complementarity between modalities are vital. However, most existing methods [15, 16] based on skeleton and RGB modality have some limitations in representation due to the input modality. For instance, the skeleton modality lacks the ability to depict the appearance of human body parts, while the RGB modality is prone to being influenced by various sources of noise, including background interference and changes in illumination. So it is meaningful to explore another modality that incorporates body-part appearance depiction while remaining noiseless and robust.

Inspired by this motivation, we intend to introduce the human parsing into action recognition. Actually, the human parsing facilitates the recognition of distinct semantic parts [22] and effectively eliminates irrelevant details while retaining crucial extrinsic features of the human body. Meanwhile, we propose an ensemble human parsing and pose network (EPP-Net) for multimodal-based action recognition. Unlike previous methods using unimodal modality (Figure 1a) or using noisy RGB modality (Figure 1b) in multimodal learning, our EPP-Net (Figure 1c) is the first to combine human parsing and pose modality for action recognition.

In this work, we advocate integrating human parsing feature maps as a novel modality into multimodal-based action recognition and propose an EPP-Net, as shown in Figure 2. Specifically, our EPP-Net consists of two trunk branches called human pose branch and human parsing branch. In the human pose branch, skeleton data is transformed into various representations. These skeleton representations are then fed into a GCN to obtain high-level features, for example, classification scores. In the human parsing branch, the human parsing features from different frames are sequentially combined to construct a feature map, which is also subsequently inputted into a CNN to derive high-level features. These high-level features from both branches are effectively integrated via an ensemble layer for the final action recognition. By leveraging the proposed EPP-Net, we effectively integrate pose data and human parsing feature maps to achieve better human action recognition.

Our contributions are three-fold, summarised as follows:

1. We advocate leveraging human parsing feature maps as a novel modality for multimodal-based action recognition, which is depictive and can filter out most irrelevant noise. This introduced human parsing modality can be effectively combined with skeleton data for better action recognition.
2. We propose a new dual-branch framework called EPP-Net, which effectively combines human parsing feature map and pose modality for the first time. In detail, the skeleton data

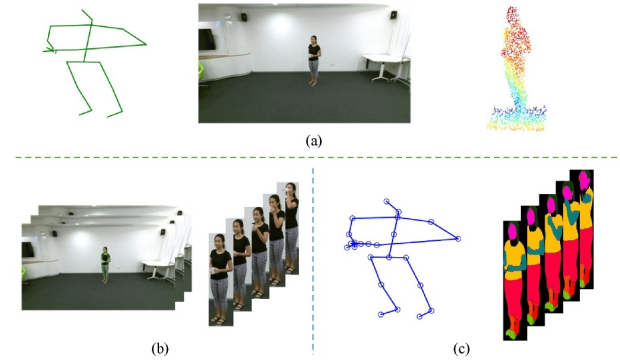


FIGURE 1 Action recognition using different modalities. (a) Previous methods using unimodal modality such as skeleton, RGB or point cloud data. (b) Previous multimodal methods using noisy RGB modality. (c) Our EPP-Net is the first to introduce novel human parsing modality and leverage both skeletons and human parsing feature maps for multimodal-based action recognition. EPP-Net, Ensemble Human Parsing and Pose Network.

and feature maps are inputted to GCN and CNN backbones for modelling high-level features, which will be effectively combined through a late fusion strategy for more robust action recognition.

3. Extensive experiments on benchmark NTU RGB + D and NTU RGB + D 120 datasets verify the effectiveness of our EPP-Net. In these two large-scale action recognition datasets, our model outperforms most existing unimodal and multimodal methods.

2 | RELATED WORK

2.1 | Human action recognition

Human action recognition plays a pivotal role in video understanding, where the human actions convey essential information such as emotions and potential intentions, enabling a deeper understanding of individuals within the video content. Initially, researchers used hand-crafted features [23–25] for action recognition. With the rise of deep learning, more researchers turned their attention to neural networks, giving birth to many classical methods based on CNNs [26], RNNs [27] and GCNs [7–9, 28–30]. Meanwhile, action recognition has expanded from unimodal-based methods to multimodal-based methods, incorporating modalities such as skeletons [7, 30], RGB images [15, 31, 32], point clouds [33], depth images [34] and human parsing. Currently, achieving robust action recognition through multimodal fusion has gained significant attention. In this section, we will first introduce some unimodal-based and multimodal-based methods in action recognition. Then, we will emphasise introducing some knowledge about human parsing and incorporate the human parsing modality into action recognition. Finally, we will introduce some classical methods for multimodal fusion in human action recognition.

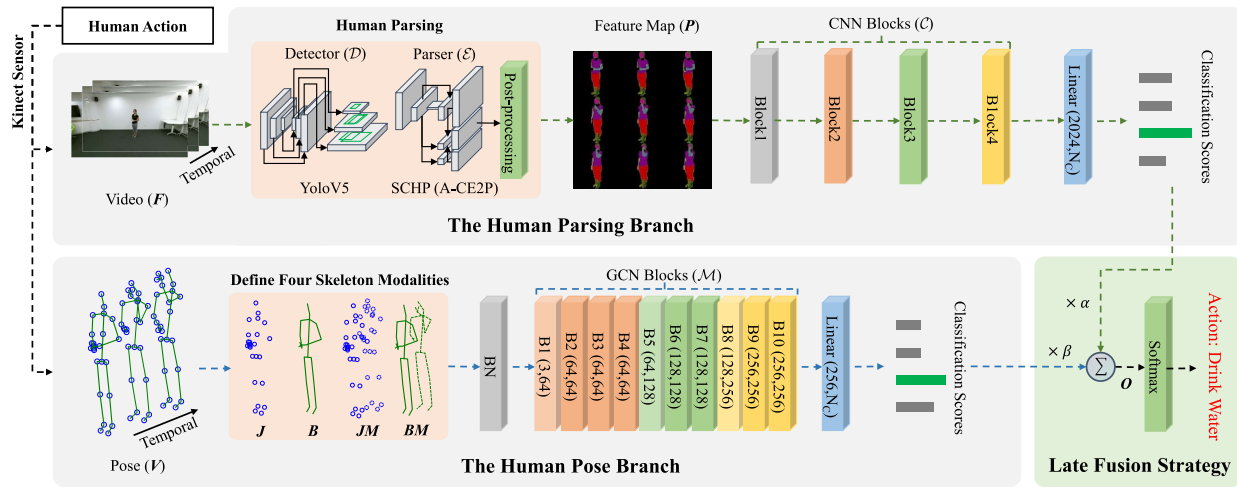


FIGURE 2 Framework of our proposed ensemble human parsing and pose network.

2.2 | Unimodal action recognition

Action recognition methods can be roughly divided into two ways: unimodal and multimodal. Usually the unimodal methods use a single modality such as pose [7, 19, 28, 29, 35, 36], RGB images [37, 38] or point clouds [33] as input. Among them, the skeleton is the most popular which is often used for skeleton-based action recognition. Because the graph-structured skeleton can well represent body movements and is highly robust to environmental changes, many studies [7–9, 28–30] have used it on GCNs and achieved high success. For example, Liu et al. [29] proposed a temporal decoupling graph convolutional network, which learn high-level features from skeleton data effectively via using the temporal-dependent adjacency matrix. Cheng et al. [8] proposed a shift-GCN, which used a simple shift operation to swap spatiotemporal features between skeletons. Chen et al. [7] proposed a CTR-GCN, which focuses on using channel features to model skeleton dependencies in unimodal action recognition. These unimodal methods usually focus on aggregating unimodal features from skeletons. However, unimodality has inherent limitations in representing human actions. For instance, the skeleton modality lacks the appearance of a human body and the RGB modality is easily influenced by various noises [39, 40], while the unordered point clouds lack fine-grained features. These flaws greatly limit the performance of unimodal-based methods and hinder the improvement of action recognition.

2.3 | Multimodal action recognition

Benefiting from the emergence of various multimodal datasets and improvements in computing resources, multimodal research involving representation, translation, alignment and fusion tasks [41] has become popular [42–44]. Commonly used modalities for multimodal-based action recognition include skeletons [7, 45], colour images [15, 16, 45, 46], depth maps

[34], text [47, 48] and point clouds [33]. For example, Shu et al. [16] propose a novel multimodal fusion network called ESE-FN to aggregate discriminative information of skeletons and colour images for better action recognition. To better use text modality to assist skeleton-based action recognition, Xu et al. [48] leverage large language models to generate text features to provide global prior knowledge for skeleton joints. Wu et al. [34] use 3D CNN to effectively fuse and complement the skeleton and depth information for robust multimodal action recognition. These multimodal-based methods take full advantage of the complementarity between modalities. However, they still suffer from the inherent imperfections of single modality.

Human parsing task is an important vision task that holds significant importance in video surveillance and human behaviour analysis. Human parsing involves the segmentation of a human image into fine-grained semantic parts, including the head, torso, arms, and legs. Several benchmarks have been proposed for the human parsing task, providing large-scale annotations of body parts [22, 49, 50]. A number of works concentrated on this problem and proposed novel models for better semantic parsing. The majority are based on ResNet architecture [22, 51–53], while some are based on HRNet architecture [54] and Transformer architecture [55]. Inspired by the human parsing task, we exploit the advantages of human parsing to get noiseless and concise representations, thus introducing the human parsing feature map into multimodal action recognition. Based on the novel human parsing modality, we propose an EPP-Net for multimodal-based action recognition. Our EPP-Net argues to utilise both pose data and feature map sequences of human parsing. Compared with the above approaches, the human parsing module filters out irrelevant information regarding illumination and backgrounds while selectively retaining spatiotemporal features of all body parts. This novel modality is proven effective in our ensemble framework.

Generally, the fusion strategies among modalities in multimodal action recognition can be summarised into early

fusion [15], intermediate fusion [45] and late fusion [34]. Early fusion integrates features immediately after they are extracted, while late fusion performs integration after each modality has made a decision. Intermediate fusion usually uses intermediate level features of each modality for multimodal interactions [45]. These fusion strategies are widely used in multimodal-based action recognition. For example, Joze et al. [45] introduced a multimodal transfer module (MMTM) into action recognition, which operates between CNNs and uses multimodal information to recalibrate features of different modality based on the intermediate fusion strategies. Wu et al. [34] proposed a multimodal two-stream 3D network framework, which fuses the classification scores of RGB and depth stream based on the late fusion strategy. Among these three common fusion strategies, late fusion tends to be efficient and concise, and there is no need to design complex fusion modules when using complementary modalities. In our proposed EPP-Net, we also use a late fusion strategy to integrate the classification scores of both human pose and human parsing modalities.

3 | METHOD

3.1 | Define four skeleton modalities

In our proposed EPP-Net, the human pose branch leverages human skeleton data collected by sensors. Conceptually, the skeleton sequence forms a natural topological graph in which joints are graph vertices, and bones are edges. The graph can be defined as $G = \{V, E\}$, where $E = \{e_1, e_2, \dots, e_N\}$ and $V = \{v_1, v_2, \dots, v_N\}$ represent N bones and N joints of the human body. The joint v_i is defined as $\{x_i, y_i, z_i\}$ in 3D pose data, where x_i, y_i and z_i locate v_i in a three-dimensional Euclidean space.

Here, we define skeleton data as four different modalities, namely joint (J), bone (B), joint motion (JM) and bone motion (BM). Given two joints data $v_i = \{x_i, y_i, z_i\}$ and $v_j = \{x_j, y_j, z_j\}$, a bone data of the skeleton is defined as a vector $e_{v_i, v_j} = (x_i - x_j, y_i - y_j, z_i - z_j)$. Given two joints data $v_{ti}, v_{(t+1)i}$ from two consecutive frames, the data of JM is defined as $m_{ti} = v_{(t+1)i} - v_{ti}$. Similarly, given two bones data $e_{v_{(t+1)i}, v_{(t+1)j}}, e_{v_{ti}, v_{tj}}$ from two consecutive frames, the data of BM is defined as $m_{v_{ti}, v_{tj}} = e_{v_{(t+1)i}, v_{(t+1)j}} - e_{v_{ti}, v_{tj}}$.

3.2 | Generate human parsing feature map

In our EPP-Net, the human parsing branch leverages the noiseless feature maps as the input modality. These feature maps are generated from noisy RGB videos via the human parsing models based on the following steps, which is shown in Figure 3. Given an RGB video stream $F = \{f_1, f_2, \dots, f_N\}$ with N frames, we first extract the features of each frame to filter most noise and irrelevant factors by

$$I_i = \mathcal{E}[\mathcal{D}(f_i)], i \in 1, 2, \dots, N, \quad (1)$$

where \mathcal{D} and \mathcal{E} denote an object detector and a feature extractor, respectively. Inspired by the top-down pose estimation, we use an object detector \mathcal{D} to filter out most ambient noise and focus on the human body. The feature extractor \mathcal{E} performs fine-grained human parsing from the detected human anchors to obtain the typical features of body parts. In the implementation, we use YoloV5 [56] and SCHP (A-CE2P) [53] as the object detector and human parser to extract useful human parsing features from RGB videos.

In order to save computing and storage resources, we resize the extracted frame features I_i and select t frames to construct the final feature map by Equation (2).

$$P = \mathcal{S}_t[\phi(\mathcal{R}(I_1), \mathcal{R}(I_2), \dots, \mathcal{R}(I_N))], 1 \leq t \leq N. \quad (2)$$

In Equation (2), the operation \mathcal{R} resizes each frame features to the specified size (H_s, W_s) and t frames, which will be arranged in chronological order to form the final feature map with a size of (H_m, W_m) via the function \mathcal{S}_t . The operator ϕ colourises the feature by assigning a unique RGB value in the colour domain to each parsing category. This colourisation step magnifies the inter-category difference in features between semantic parts of the human body. It also provides features close to natural RGB image for better vision backbone learning. Figure 4 visualises the feature maps of six action samples, namely *drink water*, *throwing*, *stand up*, *clapping*, *hand waving* and *salute*.

3.3 | Compute unimodal classification scores

Inspired by GCN's unique advantages in modelling graph-structured data, our EPP-Net uses GCN to obtain skeleton-based unimodal classification scores. Actually the GCNs consist of two parts: graph convolution module and temporal convolution module. The normal graph convolution can be formulated as follows:

$$H^{l+1} = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^l W^l), \quad (3)$$

where H^l is the joint features at layer l and σ is the activation function. $D \in R^{N \times N}$ is the degree matrix of N joints and W^l is the learnable parameter of the l th layer. A is the adjacency matrix representing joint connections. Generally, the A can be generated by using static and dynamic ways. The A is generated by using data-driven strategies in dynamic ways, while it is defined manually in static ways. In our EPP-Net, we use 10 dynamic GCN blocks. The four different skeleton modalities mentioned above will first go through a data normalisation layer, then input into the 10 GCN blocks for feature extraction. Finally, a linear layer is used to obtain the unimodal classification score.

CNNs have unique advantages in modelling Euclidean data (e.g. images), which can effectively learn local fine-grained features through the convolution kernel. These convolution layers with specific kernels sizes can further learn high-level features when applied to the body parts of human parsing. Therefore, we use four CNN blocks to process the human parsing feature maps and obtain high-level features for classification. Finally, a linear layer is also used to obtain the unimodal classification score based on the human parsing modality.

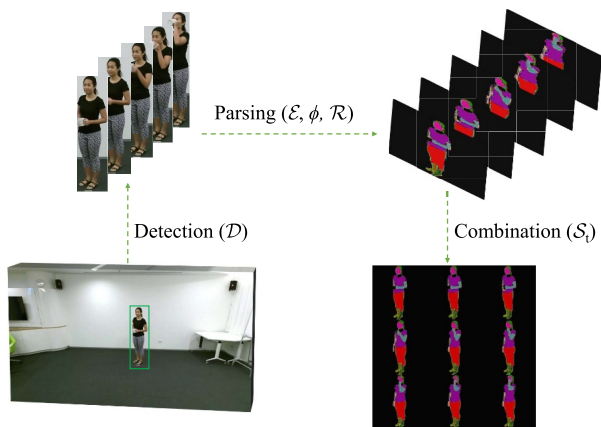


FIGURE 3 Generate noiseless human parsing feature map from noisy RGB videos.

In our proposed EPP-Net, we use cross-entropy loss as the classification loss:

$$\mathcal{L}_{cls} = - \sum_i^N \mathcal{Y}^i \log \mathcal{S}_{\mathcal{M}}^i, \quad (4)$$

where N is the number of samples in a batch and \mathcal{Y}^i is the one-hot presentation of the true label about the action sample i . The $\mathcal{S}_{\mathcal{M}}^i$ is the true unimodal classification score of action i output by the linear layer.

3.4 | Late fusion strategy

Finally, a late fusion strategy is used to better fuse the classification scores of two modalities. Generally the late fusion is efficient and concise when using complementary modalities such as pose and human parsing feature maps. In detail, the classification scores from human pose branch and human parsing branch are fused via an ensemble layer, which is formulated as follows:

$$\mathcal{O} = \alpha \cdot \mathcal{C}(\mathcal{P}) + \beta \cdot \mathcal{M}\mathcal{V}, \quad (5)$$

where \mathcal{P} and \mathcal{V} denote human parsing feature maps and the human pose data, respectively. To obtain the classification scores of each modality, we use a CNN \mathcal{C} and GCN \mathcal{M} to process the feature maps and skeleton data, respectively.

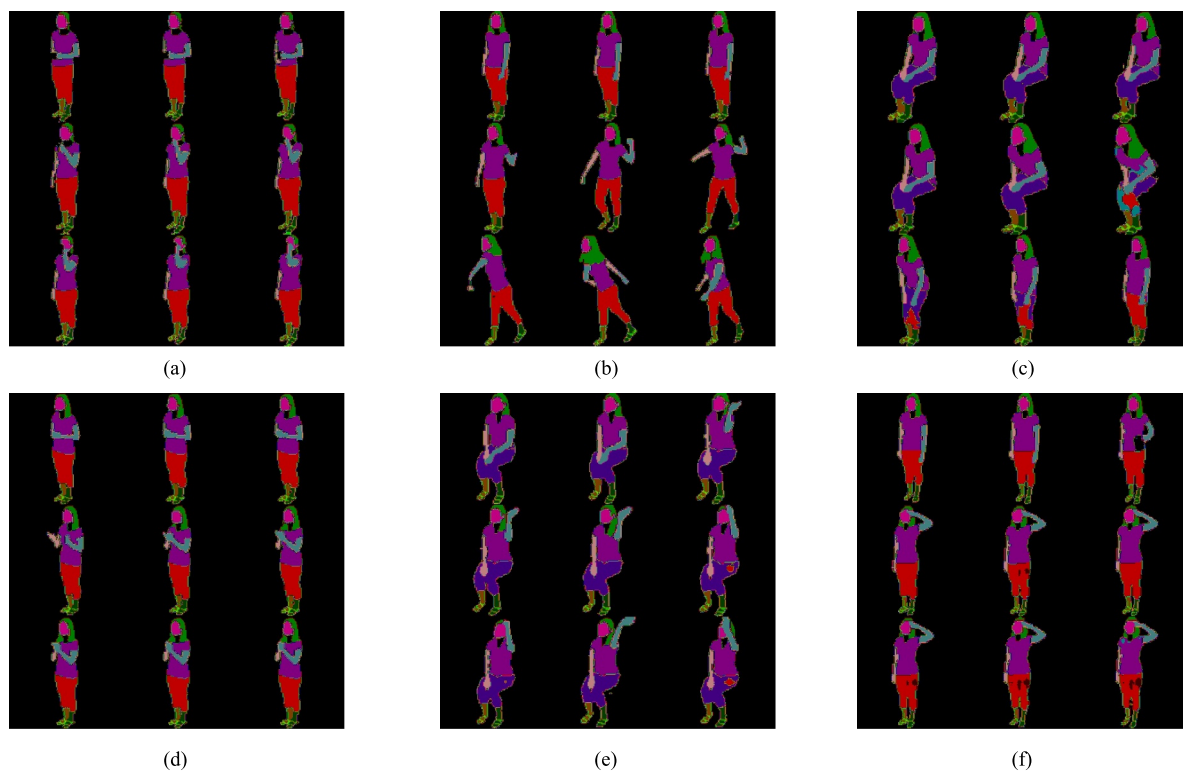


FIGURE 4 Visualisation of human parsing feature maps for different action samples. (a) Feature map of *drink water*. (b) Feature map of *throwing*. (c) Feature map of *stand up*. (d) Feature map of *clapping*. (e) Feature map of *hand waving*. (f) Feature map of *salute*.

Subsequently, the classification scores of two modalities are effectively fused by using two ensemble rates α and β . The final result O that integrates two modalities will be processed by a softmax layer for final action recognition.

4 | EXPERIMENTS

4.1 | Datasets

We use two widely used large-scale human action recognition datasets in our experiments: **NTU-RGB + D** (NTU 60) and **NTU-RGB + D 120** (NTU 120).

NTU-RGB + D [57] is a widely used 3D action recognition dataset containing 56,880 skeleton sequences and human action videos, which are categorised into 60 action classes and performed by 40 different performers. The original paper [57] suggests two benchmark scenarios for evaluation: (1) Cross-view (X-view), where the training data originates from cameras at 0° (view 2) and 45° (view 3), and the testing data is sourced from the camera at 45° (view 1). (2) Cross-subject (X-sub), where the training data comprises samples from 20 subjects, while the remaining 20 subjects are reserved for testing.

NTU-RGB + D 120 [58] is derived from the NTU-RGB + D dataset. A total of 114,480 video samples across 120 classes are performed by 106 volunteers and captured using three Kinect V2 cameras. The original work [58] also suggests two criteria: (1) Cross-subject (X-Sub), where the training data is sourced from 53 subjects, while the testing data originates from the other 53 subjects. (2) Cross-setup (X-Set), where the training data is composed of samples with even setup IDs, and the testing data comprises samples with odd setup IDs.

4.2 | Implementation details

All experiments are conducted on two NVIDIA GeForce RTX 3070 GPUs and four Tesla V100-PCIE-32GB GPUs. On the human pose branch, we adopt CTR-GCN [7] as the backbone to obtain pose-based classification scores. We use SGD for model training, conducting 65 epochs with a batch size of 64. The initial learning rate is set to 0.1 and is decayed by a factor of 0.1 at epochs 35 and 55. On the human parsing branch, we adopt InceptionV3 [59] as the backbone to obtain parsing-based classification scores. We also use SGD to train the model for 45 epochs with a batch size of 64. The initial learning rate is set to 0.03 and is decayed by a factor of 0.0001 every 10 epochs. During training, we select 9 frames randomly to construct the feature maps of human parsing and randomly change brightness, contrast, and saturation for data augmentation while nine frames at equal intervals in testing. When generating the human parsing feature map, the human parsing map of each frame will be resized to the specified size of [160, 160] and nine frames of which will be arranged in chronological order to form the final feature map with a size of

[480, 480]. Specifically, we fused the classification scores of joint, bone, JM, BM and human parsing in a ratio of 2:2:1:1:2.

4.3 | Comparison with related methods

In Table 1, we report the **Top-1** accuracy of our EPP-Net and compare with the existing methods on the NTU-RGB + D and NTU-RGB + D 120 datasets. In these two widely-used action recognition datasets, our EPP-Net outperforms most existing unimodal and multimodal methods under the recommended evaluation benchmarks. It is worth emphasising that our EPP-Net is the first to combine human parsing feature map and pose data for multimodal-based action recognition.

In the NTU-RGB + D dataset, the classification accuracy is 94.7% and 97.7% on the benchmark of X-sub and X-view, respectively, which outperforms InfoGCN [19] by 1.7% and 0.6%. On the tougher benchmark namely X-Sub, our EPP-Net outperforms LST [47] and LA-GCN [48] by 1.8% and 1.2% although the LST and LA-GCN model additionally introduces a new modality of text data. In the NTU-RGB + D 120 dataset, the classification accuracy is 91.1% and 92.8% on the benchmark of X-sub and X-view, respectively, which outperforms CTR-GCN by 2.2% and 2.2%. Like most multimodal methods, VPN [15] and TSMF [31] use RGB and skeleton modalities for better action recognition. On two benchmarks of the NTU-RGB + D 120 dataset, our EPP-Net outperforms the VPN by 4.8% and 5.0% and outperforms the TSMF by 4.1% and 3.7%.

We present the confusion matrices of our EPP-Net in Figure 5 and select the most challenging benchmark (i.e. NTU120 X-Sub) for analysis. In the confusion matrix of the NTU120 X-Sub benchmark, a total of 62 samples achieved an accuracy exceeding 95%, accounting for approximately 51.67% of the total samples. Meanwhile, there are 98 action samples with an accuracy exceeding 85%, accounting for approximately 81.67%. Among 120 action samples, the ‘staggering’, ‘jump up’, ‘arm circles’, ‘take off jacket’, ‘walking towards’ and ‘cheers and drink’ action samples have the highest recognition accuracy, reaching 100%, while the ‘staple book’ action sample has the lowest recognition accuracy, just reaching 44.66%.

4.4 | Ablation study

In Table 2, we present the recognition accuracy of four skeleton modalities and human parsing feature map in NTU60 and NTU120 datasets, respectively. We also explore the ensemble accuracy of different skeleton modalities and feature maps. For example, when the joint, bone and human parsing feature maps are combined (J + B + P), the ensemble recognition accuracy of two benchmarks in the NTU60 dataset is 94.6% and 97.6%, which outperforms the setting that combining four skeleton modalities (J + B + JM + BM) by 2.2% and 1.1% respectively. Similarly, the setting (J + B + P) outperforms the setting (J + B + JM + BM) by 1.9% and 2.2% on the two

TABLE 1 Accuracy comparison with state-of-the-art methods on the NTU-RGB + D and NTU-RGB + D 120 dataset.

Modality	Method	Source	NTU 60 (%)		NTU 120 (%)	
			X-Sub	X-View	X-Sub	X-Set
Skeleton	Shift-GCN [8]	CVPR'20	90.7	96.5	85.9	87.6
Skeleton	Dynamic GCN [9]	MM'20	91.5	96.0	87.3	88.6
Skeleton	DSTA-Net [35]	ACCV'20	91.5	96.4	86.6	89.0
Skeleton	MS-G3D [60]	CVPR'20	91.5	96.2	86.9	88.4
Skeleton	MST-GCN [61]	AAAI'21	91.5	96.6	87.5	88.8
Skeleton	CTR-GCN [7]	ICCV'21	92.4	96.8	88.9	90.6
Skeleton	GS-GCN [28]	CICAI'22	90.2	95.2	84.9	87.1
Skeleton	PSUMNet [36]	ECCV'22	92.9	96.7	89.4	90.6
Skeleton	InfoGCN [19]	CVPR'22	93.0	97.1	89.8	91.2
Skeleton	STSA-Net [62]	Neuro.'23	92.7	96.7	88.5	90.7
Skeleton + RGB	VPN [15]	ECCV'20	93.5	96.2	86.3	87.8
Skeleton + RGB	TSMF [31]	AAAI'21	92.5	97.4	87.0	89.1
Depth + RGB	DRDIS [34]	TCSVT'22	91.1	94.3	81.3	83.4
Skeleton + text	LST [47]	ICCV'23	92.9	97.0	89.9	91.1
Skeleton + RGB	STAR-Transformer [32]	WACV'23	92.0	96.5	90.3	92.7
Skeleton + text	LA-GCN [48]	arXiv'23	93.5	97.2	90.7	91.8
Skeleton + parsing	Ours (J + B + P)		94.6	97.6	90.9	92.7
Skeleton + parsing	Ours		94.7	97.7	91.1	92.8

Note: Our best results are in bold.

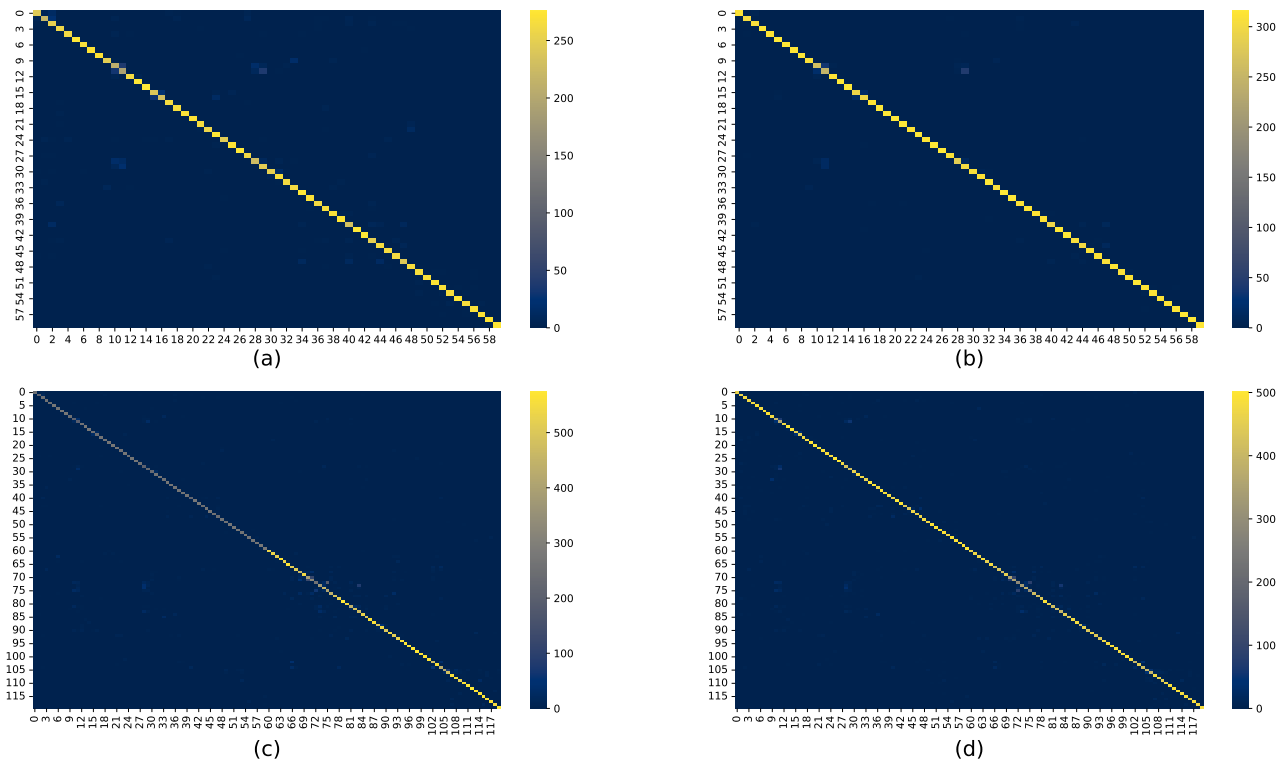


FIGURE 5 The confusion matrix of the NTU60 dataset and NTU120 dataset. The more yellow the squares on the diagonal, the more accurate the recognition. (a) NTU60 dataset on the benchmark of X-sub. (b) NTU60 dataset on the benchmark of X-view. (c) NTU120 dataset on the benchmark of X-sub. (d) NTU120 dataset on the benchmark of X-set.

TABLE 2 Accuracy of different modalities on the NTU-RGB + D and NTU-RGB + D 120 dataset.

Modality					NTU 60 (%)		NTU 120 (%)	
J	B	JM	BM	P	X-Sub	X-View	X-Sub	X-Set
✓					90.2	95.0	85.0	86.7
	✓				90.5	94.7	86.2	87.5
		✓			88.1	93.2	81.2	83.0
			✓		87.3	92.0	81.7	82.9
				✓	84.5	86.4	69.8	74.6
✓	✓				92.2	96.2	88.7	90.2
✓	✓	✓	✓		92.4	96.5	89.0	90.5
✓				✓	93.8	97.0	87.9	90.6
	✓			✓	93.6	96.9	89.1	91.2
✓	✓			✓	94.6	97.6	90.9	92.7
✓	✓	✓	✓	✓	94.7	97.7	91.1	92.8

Note: Our best results are in bold.

Abbreviations: B, bone; BM, bone motion; J, joint; JM, joint motion.

benchmarks of the NTU120 dataset, respectively. We also observed that when the skeleton modality is combined with the feature map (such as J + P and B + P), it achieves better accuracy than the unimodal method, which indicates the feature map is complementary to the skeleton modality and can contribute to the action recognition task.

In Table 3, we explore the impact of the coloured and uncoloured parsing feature maps on classification accuracy. When using the coloured feature maps, the accuracy of unimodality is 69.8%, which outperforms the uncoloured setting by 0.3%. Our EPP-Net combines the coloured feature maps and skeleton data for multimodal-based action recognition, which is better than using uncoloured parsing feature maps. We argue that the colourisation magnifies the inter-category difference in features between semantic parts of the human body, which is helpful for action recognition.

Table 4 shows the impact of using different CNN backbones to process human parsing feature maps on recognition accuracy. Among them, the setting using InceptionV3 [59] achieved the best recognition accuracy, which outperforms the VGG13 [63] and ResNet34 [64] by 4.8% and 4.7% in unimodal accuracy. Table 5 shows the impact of using different human parser to generate human parsing feature maps on recognition accuracy. In Table 6, we also explore the impact of using different frames to construct feature maps on recognition accuracy. To avoid lack of information or redundancy, our proposed EPP-Net selects 9 frames feature to construct feature maps, which achieves 91.1% recognition accuracy in the X-sub benchmark of the NTU120 dataset. Additionally, we showcase the parameter and computation cost required for the proposed EPP-Net in Table 7.

Our observation is three-fold based on the experimental results in Tables 2 and 4. First, the feature maps show desirable

TABLE 3 Performance comparison between the coloured and uncoloured parsing feature maps on the NTU-RGB + D 120 X-sub benchmark. We conducted multiple experiments and reported the average results.

Colourisation	Modality	
	Parsing (%)	Ensemble (%)
w/	69.8	91.1
w/o	69.5	90.8

Note: Our best results are in bold.

TABLE 4 Performance of various vision backbones on the NTU-RGB + D 120 X-sub benchmark. The parsing (%) represents the accuracy of the human parsing branch when using different vision backbones.

Backbone	Input size	Modality	
		Parsing (%)	Ensemble (%)
VGG11 [63]	224 × 224	64.6	90.7
VGG13 [63]	224 × 224	65.0	90.8
ResNet18 [64]	224 × 224	64.6	90.6
ResNet34 [64]	224 × 224	65.1	90.7
ConvnextV2 Tiny [65]	224 × 224	66.7	90.8
ConvnextV2 Tiny [65]	384 × 384	68.9	91.0
EVA-02 Small [66]	224 × 224	62.2	90.5
InceptionV3 [59]	299 × 299	69.8	91.1

Note: Our best results are in bold.

TABLE 5 Performance of various human parsers on NTU-RGB + D 120 datasets.

Modality	Human parser	NTU 120 (%)	
		X-Sub	X-Set
Skeleton	-	89.4	91.2
Parsing	PSP-Net [52]	53.6	65.8
Skeleton + parsing	PSP-Net [52]	90.0	91.7
Parsing	SCHP (A-CE2P) [53]	69.8	74.6
Skeleton + parsing	SCHP (A-CE2P) [53]	91.1	92.8

Note: Our best results are in bold.

TABLE 6 Comparison of different numbers of frames in feature map construction on the NTU-RGB + D 120 X-sub benchmark.

#Frame	Modality	
	Parsing (%)	Ensemble (%)
4	62.0	90.9
9	69.8	91.1
16	72.6	91.0
25	71.4	91.0

Note: Our best results are in bold.

TABLE 7 Comparison of parameter and computation costs when training and inferring on a single action sample.

Modality	Method	Param.	Flops
Depth + RGB	DRDIS [34]	171.98M	27.92G
	(Res3D-101 + Res3D-101)		
Skeleton + RGB	STAR-Transformer [32]	58.49M	18.67G
	(ResNet-MC18 + Pure-VIT)		
Skeleton + parsing	Ours	25.27M	7.84G
	(CTR-GCN + Inception-V3)		

Note: Our best results are in bold.

performance across different CNN backbones, which demonstrates that the human parsing modality is robust. Second, better recognition performance is obtained when the feature maps are effectively combined with different skeleton modalities (e.g. joint and bone), which also shows that the human parsing feature map is effective. Third, when the human parsing modality is combined with a single skeleton modality, better recognition accuracy will be obtained than the combination of multiple different skeleton modalities. This phenomenon confirms that human parsing modality can better supplement the skeleton information compared with the homogeneous modality.

4.5 | Visualisation

To better illustrate the fusion effect of the human pose branch and human parsing branch, we visualised the overall accuracy of using only pose and the fusion accuracy between two branches on the NTU 120 X-sub benchmark. We also showcased the top-8 action categories with the highest improvement after adding the human parsing branch as shown in Figure 6. Among all actions, the action ‘move heavy objects’ has the highest improvement, with an accuracy of 87.15% when using only the human pose branch. Upon fusing the human pose branch and the human parsing branch, the accuracy of the action ‘move heavy objects’ increases to 95.77%, resulting in an improvement of 8.62%. To further demonstrate the impact of the Human parsing branch, we selected four samples from four action categories in Figure 6 for analysis, which are shown in Table 8. When two complementary modalities (pose and parsing) are fused, it can improve the classification score of the sample in the ground truth category, which substantiate the positive influence of the human parsing branch.

In addition, in order to better demonstrate the modelling ability and effectiveness of GCN on skeleton data, we performed the t-distributed Stochastic Neighbor Embedding visualisation of the features about the 8 action categories in Figure 6 to analyse the relationship between different action categories in the feature space, which is shown in Figure 7. We observed that the skeleton features of the same category are close and the skeleton features of different categories are far

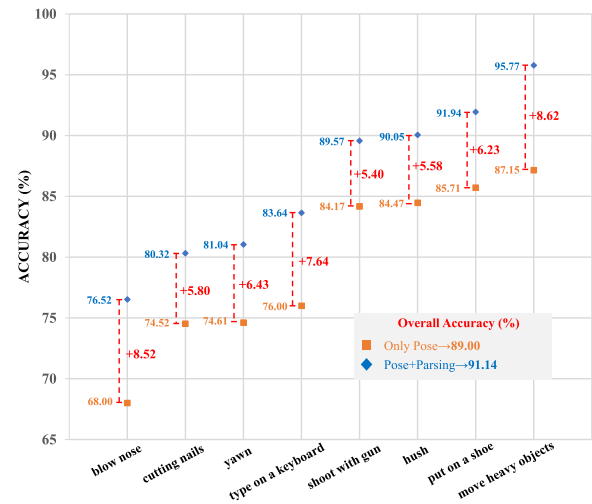


FIGURE 6 Visualisation of the top-8 actions with the highest improvement when the human pose branch is integrated with the human parsing branch.

TABLE 8 Comparison of classification scores when using different modalities. The table shows the classification score of the ground truth category with values ranging from 0.0 to 1.0.

Action samples	Classification scores		
	Pose	Parsing	Pose + parsing
Type on a keyboard (S001C001P007R002A030)	0.9892	0.0060	0.9980
Cutting nails (S027C001P006R002A075)	0.0005	0.8720	0.9990
Yawn (S027C001P006R001A103)	0.0456	0.9709	0.9999
Blow nose (S027C003P006R001A105)	0.8674	0.0931	0.9999

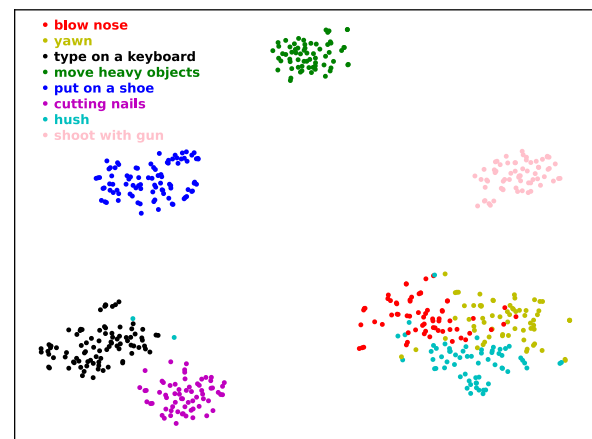


FIGURE 7 t-SNE visualisation of skeleton features modelled by GCN for different action categories. GCN, graph convolutional network; t-SNE, t-distributed Stochastic Neighbor Embedding.

apart in the feature space. Meanwhile, action samples of different categories but with similarities are close in the feature space. For example, the 'blow nose', 'yawn' and 'hush' actions in Figure 7 are very close in the feature space. We argue that this phenomenon is due to the fact that the above three actions mainly occur on the head and face of the human body, leading to their features being close in the feature space.

5 | CONCLUSIONS

We propose a new dual-branch framework called EPP-Net for multimodal-based action recognition, which introduces the depictive and noiseless human parsing feature maps as a novel modality to represent human actions. Different from previous methods using noisy modality, our EPP-Net is the first to leverage both skeletons and human parsing feature maps with the aim of robust action recognition. The effectiveness of EPP-Net is verified on the NTU-RGB + D and NTU-RGB + D 120 datasets, where our EPP-Net outperforms most existing methods.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data openly available in a public repository that does not issue DOIs.

ORCID

Jinfu Liu  <https://orcid.org/0009-0001-0994-9146>

Yuhang Wen  <https://orcid.org/0000-0001-6644-4075>

REFERENCES

- Li, M., et al.: Rhythm-aware sequence-to-sequence learning for labanotation generation with gesture-sensitive graph convolutional encoding. *IEEE Trans. Multimed.* 24, 1488–1502 (2022). <https://doi.org/10.1109/tmm.2021.3066115>
- Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(1), 14–29 (2016). <https://doi.org/10.1109/tpami.2015.2430335>
- Li, C., et al.: Hierarchical latent concept discovery for video event detection. *IEEE Trans. Image Process.* 26(5), 2149–2162 (2017). <https://doi.org/10.1109/tip.2017.2670782>
- Wang, X., et al.: Dynamic dense graph convolutional network for skeleton-based human motion prediction. *IEEE Trans. Image Process.* 33, 1–15 (2024). <https://doi.org/10.1109/tip.2023.3334954>
- Wang, Y., et al.: Global and local spatio-temporal encoder for 3D human pose estimation. *IEEE Trans. Multimed.* 26, 4039–4049 (2023). <https://doi.org/10.1109/tmm.2023.3321438>
- Zhang, Y., et al.: Facial prior guided micro-expression generation. *IEEE Trans. Image Process.* 33, 525–540 (2024). <https://doi.org/10.1109/tip.2023.3345177>
- Chen, Y., et al.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2021)
- Cheng, K., et al.: Skeleton-based action recognition with shift graph convolutional network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)
- Ye, F., et al.: Dynamic GCN: context-enriched topology learning for skeleton-based action recognition. In: *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)* (2020)
- Zhang, J., et al.: Zoom transformer for skeleton-based group activity recognition. *IEEE Trans. Circ. Syst. Video Technol.* 32(12), 8646–8659 (2022). <https://doi.org/10.1109/tcsvt.2022.3193574>
- Liu, M., et al.: Novel motion patterns matter for practical skeleton-based action recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2023)
- Liu, M., Meng, F., Liang, Y.: Generalized pose decoupled network for unsupervised 3D skeleton sequence-based action representation learning. *Cyborg Bionic Syst.* 2022, 0002 (2022). <https://doi.org/10.34133/cbsystems.0002>
- Wen, Y., et al.: Interactive spatiotemporal token attention network for skeleton-based general interactive action recognition. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)* (2023)
- Qiu, H., et al.: Spatio-temporal tuples transformer for skeleton-based action recognition. *arXiv:2201.02849* (2022)
- Das, S., et al.: VPN: learning video-pose embedding for activities of daily living. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2020)
- Shu, X., et al.: Expansion-squeeze-excitation fusion network for elderly activity recognition. *IEEE Trans. Circ. Syst. Video Technol.* 32(8), 5281–5292 (2022). <https://doi.org/10.1109/tcsvt.2022.3142771>
- Xing, Y., Zhu, J.: Deep learning-based action recognition with 3D skeleton: a survey. *CAAI Trans. Intell. Technol.* 6(1), 80–92 (2021). <https://doi.org/10.1049/cit2.12014>
- Tu, Z., et al.: Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition. *IEEE Trans. Multimed.* 25, 1819–1831 (2022). <https://doi.org/10.1109/tmm.2022.3168137>
- Chi, H.G., et al.: InfoGCN: representation learning for human skeleton-based action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022)
- Zhang, J., et al.: A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *CAAI Trans. Intell. Technol.* 7(1), 46–55 (2022). <https://doi.org/10.1049/cit2.12012>
- Tu, Z., et al.: Action-stage emphasized spatiotemporal VLAD for video action recognition. *IEEE Trans. Image Process.* 28(6), 2799–2812 (2019). <https://doi.org/10.1109/tip.2018.2890749>
- Liang, X., et al.: Look into person: joint body parsing & pose estimation network and a new benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(4), 871–885 (2019). <https://doi.org/10.1109/tpami.2018.2820063>
- Veeriah, V., Zhuang, N., Qi, G.J.: Differential recurrent neural networks for action recognition. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015)
- Wang, J., et al.: Mining actionlet ensemble for action recognition with depth cameras. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2012)
- Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2012)
- Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recogn.* 68, 346–362 (2017). <https://doi.org/10.1016/j.patcog.2017.02.030>
- Wang, H., Wang, L.: Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)
- Zhu, S., Zhan, Y., Zhao, G.: Multi-model lightweight action recognition with group-shuffle graph convolutional network. In: *Proceedings of the CAAI International Conference on Artificial Intelligence (CICAI)* (2022)
- Liu, J., et al.: Temporal decoupling graph convolutional network for skeleton-based gesture recognition. *IEEE Trans. Multimed.* 26, 1–13 (2023)

30. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2018)
31. Yu, B.X., Liu, Y., Chan, K.C.: Multimodal fusion via teacher-student network for indoor action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2021)
32. Ahn, D., et al.: STAR-transformer: a spatio-temporal cross attention transformer for human action recognition. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) (2023)
33. Fan, H., et al.: PSTNet: point spatio-temporal convolution on point cloud sequences. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021)
34. Wu, H., Ma, X., Li, Y.: Spatiotemporal multimodal learning with 3D CNNs for video action recognition. *IEEE Trans. Circ. Syst. Video Technol.* 32(3), 1250–1261 (2022). <https://doi.org/10.1109/tcsvt.2021.3077512>
35. Shi, L., et al.: Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (2020)
36. Trivedi, N., Sarvadevabhatla, R.K.: PSUMNet: unified modality part streams are all you need for efficient pose-based action recognition. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2023)
37. Li, K., et al.: Uniformer: unified transformer for efficient spatiotemporal representation learning. *arXiv:2201.04676* (2022)
38. Liu, Z., et al.: Video swin transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
39. Wang, K., et al.: Survey on person re-identification based on deep learning. *CAAI Trans. Intell. Technol.* 3(4), 219–227 (2018). <https://doi.org/10.1049/trit.2018.1001>
40. Liu, M., et al.: Salient pairwise spatio-temporal interest points for real-time activity recognition. *CAAI Trans. Intell. Technol.* 1(1), 14–29 (2016). <https://doi.org/10.1016/j.trit.2016.03.001>
41. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41(2), 423–443 (2019). <https://doi.org/10.1109/tpami.2018.2798607>
42. Yan, Z., et al.: Multimodal feature fusion based on object relation for video captioning. *CAAI Trans. Intell. Technol.* 8(1), 247–259 (2023). <https://doi.org/10.1049/cit2.12071>
43. Si, J., et al.: A multi-modal clustering method for traditional Chinese medicine clinical data via media convergence. *CAAI Trans. Intell. Technol.* 8(2), 390–400 (2023). <https://doi.org/10.1049/cit2.12230>
44. Fang, Y., et al.: ST-SIGMA: spatio-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting. *CAAI Trans. Intell. Technol.* 7(4), 744–757 (2022). <https://doi.org/10.1049/cit2.12145>
45. Joze, H.R.V., et al.: MMTM: multimodal transfer module for CNN fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
46. Tu, Z., et al.: DTCM: joint optimization of dark enhancement and action recognition in videos. *IEEE Trans. Image Process.* 32, 3507–3520 (2023). <https://doi.org/10.1109/tip.2023.3286254>
47. Xiang, W., et al.: Language supervised training for skeleton-based action recognition. *arXiv:2208.05318* (2022)
48. Xu, H., et al.: Language knowledge-assisted representation learning for skeleton-based action recognition. *arXiv:2305.12398* (2023)
49. Chen, X., et al.: Detect what you can: detecting and representing objects using holistic models and body parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
50. Liang, X., et al.: Deep human parsing with active template regression. *IEEE Trans. Pattern Anal. Mach. Intell.* 37(12), 2402–2414 (2015). <https://doi.org/10.1109/tpami.2015.2408360>
51. Ruan, T., et al.: Devil in the details: towards accurate single and multiple human parsing. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2019)
52. Zhao, H., et al.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
53. Li, P., et al.: Self-correction for human parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* 44(6), 3260–3271 (2020). <https://doi.org/10.1109/tpami.2020.3048039>
54. Sun, K., et al.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
55. Chen, W., et al.: Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
56. Ultralytics: ultralytics/yolov5: v7.0 - YOLOv5 SOTA realtime instance segmentation (2022)
57. Shahroudy, A., et al.: NTU RGB+D: a large scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
58. Liu, J., et al.: NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* 42(10), 2684–2701 (2020)
59. Szegedy, C., et al.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
60. Liu, Z., et al.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
61. Chen, Z., et al.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2021)
62. Qiu, H., et al.: Spatio-temporal segments attention for skeleton-based action recognition. *Neurocomputing* (2023)
63. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* (2015)
64. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
65. Woo, S., et al.: Convnext V2: co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
66. Fang, Y., et al.: EVA-02: a visual representation for neon genesis. *arXiv:2303.11331* (2023)

How to cite this article: Liu, J., et al.: Explore human parsing modality for action recognition. *CAAI Trans. Intell. Technol.* 1–11 (2024). <https://doi.org/10.1049/cit2.12366>