JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition

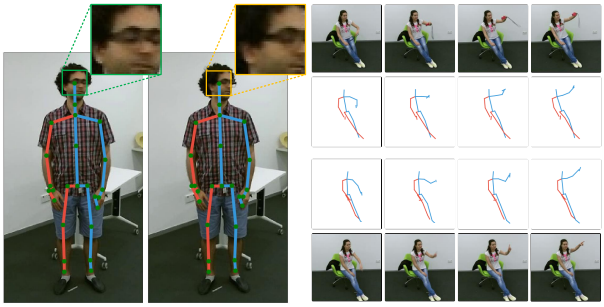
Jinmiao Cai[1*] Nianjuan Jiang[1] Xiaoguang Han[3] Kui Jia[2] Jiangbo Lu[1]
[1]SmartMore Corporation [2]South China University of Technology [3]The Chinese University of Hong Kong (Shenzhen)
{jinmiao.cai,nianjuan.jiang,jiangbo}@smartmore.com hanxiaoguang@cuhk.edu.cn kuijia@scut.edu.cn

Abstract

Skeleton-based action recognition has attracted research attentions in recent years. One common drawback in currently popular skeleton-based human action recognition methods is that the sparse skeleton information alone is not sufficient to fully characterize human motion. This limitation makes several existing methods incapable of correctly classifying action categories which exhibit only subtle motion differences. In this paper, we propose a novel framework for employing human pose skeleton and joint-centered light-weight information jointly in a two-stream graph convolutional network, namely, JOLO-GCN. Specifically, we use Joint-aligned optical Flow Patches (JFP) to capture the local subtle motion around each joint as the pivotal joint-centered visual information. Compared to the pure skeleton-based baseline, this hybrid scheme effectively boosts performance, while keeping the computational and memory overheads low. Experiments on the NTU RGB+D, NTU RGB+D 120, and the Kinetics-Skeleton dataset demonstrate clear accuracy improvements attained by the proposed method over the state-of-the-art skeleton-based methods.

1. Introduction

Human recognition [31, 28, 33, 8] is an active yet challenging task in the field of computer vision. Recently, with the advancement in depth sensors such as Microsoft Kinect and human pose estimation technology [4, 39], obtaining accurate human pose data is becoming much easier. Skeleton-based human action recognition has attracted a lot of attentions and made significant progress over the last decade. Human skeleton motion sequences retain useful high-level motion signal by eliminating the redundant information in the RGB video clips. Compared with the original RGB video clip, a skeleton sequence, with the human body joints in the form of 2D or 3D coordinates, is sparse. Thus, neural networks designed for skeleton-based action recognition can be lightweight and efficient. Recent methods [36, 13, 27, 12, 25, 26, 32, 10, 23, 24, 7, 15, 34] further exploit a variety of deep neural networks in the attempt to fully excavate the internal characteristics of dynamic human skeleton sequences.



(a) (b)

Figure 1. The motivations of the proposed method. (a) Sample frames in the "shake head" sequence from NTU RGB+D [22], where the skeletal joints of the head hardly capture the local subtle movements. (b) Different actions with similar skeleton sequences. Sample frames are taken from "taking a selfie" (top row) and "pointing to something" (bottom row).

As the input of a single-modal action recognition network, skeleton sequences can effectively describe the global human body motion. However, the local subtle motion cues are lost in the process of extracting human poses from video frames. Due to its extreme sparseness, a skeleton sequence can hardly capture subtle features in human motion. There are obvious disadvantages in recognizing human action relying solely on skeleton sequences.

Firstly, for action categories mainly characterized by local subtle movements, such as "shaking head" (Figure 1(a)), the difference between the skeletons extracted from two successive frames are so subtle that it is hardly useful for describing such actions. The lack of effective representation makes it more challenging for the network to predict related behavioral categories. Moreover, such local subtle move-

*This work was mainly done when Jinmiao, Nianjuan and Jiangbo were interning and working in Shenzhen Cloudream Technology Co., Ltd. Corresponding email: jiangbo@smartmore.com.

ments are easily obscured by the noisy pose estimation, when the body movement of the action is weak. Quantitatively, in regard to the actions such as "reading" and "writing", the performance of the skeleton-based single-modal methods seems to have encountered a bottleneck due to the aforementioned drawbacks.

Secondly, skeletons alone may not be as distinctive in describing certain action categories. For instance, as shown in Figure 1(b), "pointing to something" and "taking a selfie" have very similar skeleton sequence representations. Thus these categories are easily confused with each other by a skeleton-based single-modal method.

There exist methods that use multi-modal sources to augment skeleton information with some additional inputs, such as RGB features, depth maps, joint heat-maps, for action recognition [5, 20, 9, 2, 17, 18, 6, 1, 40, 35]. By introducing such visual information which contains rich motion information both in global and local domains, this kind of multi-modal methods can describe human body motion more completely, and help the neural network to recognize human actions better. However, as the complexity of the input increases, the number of network parameters and the computing resource consumption increase substantially compared to the single-modal counterparts.

To move beyond such limitations, we propose to augment the skeleton with the light-weight local visual information surrounding skeletal joints in the form of *Joint-aligned optical Flow Patches (JFP)*. Considering the joint coordinates as the anchors, the most relevant motion cues of the human body can be located in input images. We extract such most relevant motion cues around each human joint from the video as JFP, and keep it light-weight by removing the redundant background information.

After a simple format conversion procedure, a JFP sequence can be represented as sparsely as the skeleton sequence. This locally dense but globally sparse representation makes it possible to capture the local subtle motion of human body movement, while keeping the neural network light-weight. The proposed JFP sequence encodes local visual cues with a kinetically meaningful structure inherited from the human pose skeleton. Therefore, same as the skeleton sequence, JFPs can be aggregated by very sparse graph connections via a GCN formulation. In this paper, we propose a two-stream GCN architecture, *JOLO-GCN*, to fuse local motion information from a JFP sequence with global motion information from a skeleton sequence.

To demonstrate the effectiveness of our proposed method, extensive experiments are conducted on three popular large-scale human action recognition datasets, namely, NTU-RGB+D [22], NTU RGB+D 120 [14] and Kinetics-Skeleton [5, 36]. Our method outperforms state-of-the-art skeleton-based methods on these datasets. The main contributions in this paper are summarized as follows:

- We propose a novel scheme to represent the visual information surrounding each skeletal joint as JFP, which contains rich local subtle body motion cues in a relatively sparse format.
- We demonstrate that our JOLO-GCN which jointly uses the local motion information from a JFP sequence with the global motion information from a skeleton sequence, gains significant accuracy improvements compared with the single-modal baselines.
- Extensive experiments are conducted on three large-scale human action recognition datasets. Our method obtains the state-of-the-art results on these datasets.

2. Related Work

In this section, we briefly review the approaches that have utilized human skeletons as a key information source for action recognition. These approaches can be divided roughly into single-modal skeleton-based methods and multi-modal skeleton-based methods.

2.1. Single-Modal Skeleton-Based Methods

Taking only the skeleton data as the input for the human action recognition task, several single-modal methods [36, 13, 27, 12, 25, 26, 32, 10, 23, 24, 7, 15] have been proposed. These methods are usually light-weight and computationally efficient.

Graph convolutional networks (GCN) use a general and effective framework for learning representation of graph structured data. Various GCN variants [24, 36, 25, 12, 26] have achieved the state-of-the-art results on skeleton-based action recognition. Yan *et al*. proposed ST-GCN [36], a generic graph-based formulation for modeling dynamic skeletons, which can capture the patterns embedded in the spatial configuration of the joints as well as their temporal dynamics. The topology of the graph employed in ST-GCN is fixed and defined according to the human skeletal structure. Thus, it is not guaranteed to be optimal flexibly for recognizing specific actions. In the attempt to address this drawback, adaptive graph CNNs [25, 12, 24], which can adaptively learn the graph topology, are used for automatically inferring spatial dependency among pose joints. In AS-GCN [12], the actional links (A-link) is inferred in a data-driven manner to capture the latent dependencies between any pose joints. In 2S-AGCN [25], three parts – a fixed physical structure graph, a data-dependent attention graph and a global learned graph – are designed in the adjacency matrix of an adaptive graph, which increase the model flexibility and also the generality. In addition, the second-order information of joint coordinates (namely "Bones") is aggregated together with the "Joints" as a parallel source input for providing more motion cues.

The RNN and LSTM structures are effective for analyzing time series of streaming data. Different methods

based on a RNN or LSTM framework are explored in recent works [7, 13, 26, 27, 32, 15]. In IndRNN [13], the neurons in the same layer are independent of each other, which has great stability against gradient vanishing and exploding. Si *et al*. [27] proposed SR-TSL, which utilizes a LSTM module in series with GCN for spatial reasoning and temporal stack learning. AGC-LSTM [26] applies a graph convolution operator to replace the fully connected operator within LSTM, so as to explore the co-occurrence relationship between spatial and temporal domains.

Since the skeleton input is sparse, these single-modal skeleton-based methods are advantageous in terms of computational complexity and the network scale. However, subtle motion characteristics that cannot be well captured by skeleton dynamics remain challenging for single-modal skeleton-based networks to learn.

2.2. Multi-Modal Skeleton-Based Methods

Multi-modal approaches are widely used in the field of action recognition [5, 20, 9, 2, 17, 18, 6, 1, 40, 35]. Augmenting data sources, such as RGB images, optical flow, depth maps, joint heat-maps, and pose skeletons, provide richer semantic cues for neural networks to infer the human action in the scene. Since a comprehensive literature review is beyond the scope of this paper, here we only focus on skeleton-based multi-modal methods.

The chained multi-stream network [40] computes and integrates several important visual cues in a Markov chain model which adds complementary cues successively. The C3D architecture [30] is used as the base architecture of the multi-stream network. Hu *et al*. [9] proposed a novel deep bilinear framework for learning multi-modal temporal features. Multi-modal inputs include RGB images, depth maps, and skeleton sequences. Note that the depth and RGB input utilized in this work [9] are image patches aligned with the skeletal joints. Luvizon *et al*. [20] proposed a multi-task deep architecture to perform 2D and 3D pose estimation jointly with action recognition. The idea is to aggregate image visual features, joint probability maps and the estimated pose sequence to infer the human action in the recognition module. Joint heat-maps are utilized in [17, 18, 6] to augment the sparse skeleton information. High-dimensional heat-map sequences are compressed in different representations. Additional motion cues might be inferred from such probabilistic representations for recognizing action.

Despite the benefits of bringing more cues to facilitate the action recognition task, these multi-modal approaches can only handle small temporal windows due to the significant amount of network parameters and computing resources required.

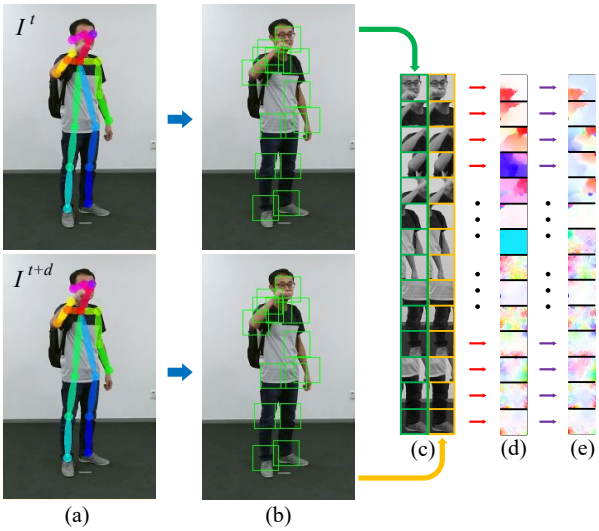


Figure 2. The Joint-aligned optical Flow Patches (JFP) estimation from two successive frames with corresponding 2D pose joints. (b) Patches centered around body joints obtained for (a) a successive frame pair are first cropped to form (c) JAP pairs. (d) Residual optical flow is estimated from each JAP pair, followed by zero-mean normalization to obtain the corresponding (e) JFP. Noted that the zero-mean normalization is applied for eliminating the influence of joint localization imprecision on the joint alignment.

3. Method

Given a video sequence, we denote the image frames as $\mathcal{I} = \{I^t \in \mathbb{R}^{H\times W\times 3} \mid t = 1, ..., T\}$, where H is the height of the video frame, and W denotes the frame width. The corresponding skeleton sequence containing the 2D/3D coordinates of K joints in all T frames, is denoted as $\mathcal{J} = \{J_k^t \mid k = 1, ..., K; t = 1, ..., T\}$. We first use JFP to augment the 2D/3D pose skeleton sequence, and then use GCN for feature embedding and learning.

3.1. Joint-Aligned Optical Flow Patches (JFP)

As analyzed in the introduction, the sparse skeleton information alone is not sufficient to fully characterize human motion. In the attempt to address this limitation, we treat the human joint coordinates as anchor points in the image, and explicitly describe subtle motion cues corresponding to these locations in the form of optical flow patches.

Specifically, a local square-shaped patch W_k^t is obtained for each joint J_k^t in frame I^t by a *joint-centered* (w.r.t. the 2D coordinates of the joint) cropping operation:

$$W_k^t = \boldsymbol{Crop}(I^t, J_k^t, l) \,, \tag{1}$$

where $\boldsymbol{Crop}(\cdot)$ denotes the joint-centered cropping operation, and l is a custom parameter that denotes the side length of the cropped patch W_k^t. We term W_k^t as Joint-centered Appearance Patch (JAP), and the corresponding JAP sequence for the skeleton sequence $\mathcal{J}$ is denoted by $\mathcal{W} = \{W_k^t \mid k = 1, ..., K; t = 1, ..., T\}$.

It is clear that the appearance of body parts can be explicitly captured in JAPs. In addition, a JAP sequence also presents the local subtle motion of each body part. Intuitively, optical flow explicitly reflects the dense motion field between successive video frames, and is a more effective representation for subtle motion cues. Motivated by this, we propose to estimate the optical flow between successive JAP pairs of each body joint to obtain the Joint-aligned optical Flow Patch (JFP) (see Figure 2). The JFP sequence corresponding to $\mathcal{J}$ is denoted as $\mathcal{F} = \{F_k^t \mid k = 1, ..., K; t = 1, ..., T\}$.

Unlike the conventional dense optical flow, JFP only focuses on capturing the local subtle motion. For a clear description of the local motion field captured in JFP, we decompose the general motion field between consecutive frames as follows. As illustrated in Figure 3(a), an object (or a body part concerned in this work) moves from bottom-left to top-right between two successive video frames I^t and I^{t+d}, where d denotes a temporal interval. In this process, the object may undergo translation, rotation and deformation. The motion vector $U_f(p)$ of the matching pixels p and p' can be obtained by estimating the conventional optical flow between these two full frames. In fact, the global displacement of the object can be represented simply as the movement of the object center, i.e., a motion vector $V_j(c)$ from the object center location c to its counterpart c'.

After center-aligning the JAP pairs W^t and W^{t+d} in Figure 3(b), the global displacement is actually eliminated. When computing optical flow from the JAP pair, the resulting residual motion field U_r in the JFP F^t as shown in Figure 3(c) only contains subtle local motion such as rotation and deformation. Based on the local motion field U_r and the global displacement $V_j(c)$ by differencing the corresponding joints' coordinates, the full motion $U_f(p)$ at pixel p can be approximated as follows:

$$U_f(p) \approx V_j(c) + U_r(p) \ . \tag{2}$$

Around an object (or a body part concerned in this work), Eqn. (2) gives an approximated expression on the relationship of the residual local motion field U_r captured in JFP with the full motion field U_f in the conventional optical flow and the global displacement vector $V_j(c)$. Please note that JFP is estimated from the JAP pair directly, but not computed as the approximated expression in Eqn. (2). The residual local motion field U_r captured in JFP is the key information used in our method for action recognition.

We observe that the proposed JFP representation has two nice properties. Firstly, JFP provides only local subtle motion cues. As discussed above, the motion of human body parts can be represented as the global joint displacement plus the local subtle motion. The information captured by JFP is orthogonal to that of skeleton coordinates, which reflect the global joint displacement. Secondly, compared

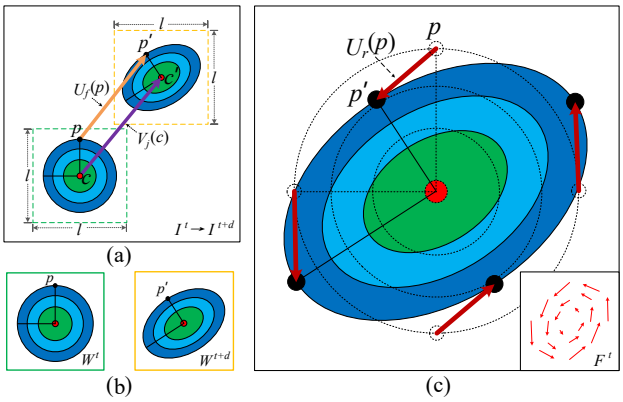


Figure 3. Schematic diagram of Joint-aligned optical Flow Patch (JFP) for two successive frames. (a) The object motion $\boldsymbol{U_f}$ from frame I^t to frame I^{t+d}, including translation, rotation and deformation. (b) The corresponding JAPs: W^t and W^{t+d}. (c) The alignment of two JAPs and the local subtle motion field U_r, which is computed to generate JFP F^t.

with general dense visual data, a JFP sequence encodes visual cues with a kinetically meaningful structure inherited from the human pose skeleton, namely, JFPs have a one-to-one correspondence with multiple skeletal joints. Although there are alternative visual features with a similar property, such as human pose heat-maps. Their data resolution is relatively large or they need to be learned by a neural network with substantially more parameters.

3.2. Graph Convolutional Networks (GCN)

In order to better capture the kinetically structured property encoded in the JFP sequences, we adopt GCN as our backbone network architecture. Given a K joints skeletal graph with a node collection of $\mathcal{V} = \{v_k \mid v_k = J_k, k = 1, ..., K\}$, the neighborhood of a node v_i is defined as $\mathcal{N}(v_i) = \{v_j \mid d(v_i, v_j) \leq D\}$, where $d(v_i, v_j)$ is the shortest path length from v_j to v_i. In a ST-GCN network setup [36], the spatial graph convolution unit is the key component, which is constructed to capture the spatial feature among joints. More specifically, given the graph adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{K \times K}$, $\boldsymbol{A}(j, i) = w$ if $v_j \in \mathcal{N}(v_i)$ and $\boldsymbol{A}(j, i) = 0$ otherwise. The adjacency matrix is normalized using a degree matrix $\boldsymbol{\Lambda}$ as:

$$\boldsymbol{\Lambda}(i, i) = \sum_{j=1}^{K} \boldsymbol{A}(i, j); \quad \boldsymbol{A^{norm}} = \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Lambda}^{-\frac{1}{2}} \ . \tag{3}$$

For the spatial graph convolution can be written in terms of the adjacency matrix as follows:

$$\boldsymbol{Y}_{out} = \sigma(\boldsymbol{W}(\boldsymbol{X}_{in}(\boldsymbol{A}^{norm} \odot \boldsymbol{M}))) \ , \tag{4}$$

where $\boldsymbol{Y}_{out} \in \mathbb{R}^{d_{out} \times K}$, $\boldsymbol{X}_{in} \in \mathbb{R}^{d_{in} \times K}$ and d denotes the feature dimension. $\boldsymbol{M}$ denotes a trainable mask for the adaptive re-weighting on the $\boldsymbol{A}^{norm}$, and $\odot$ denotes the

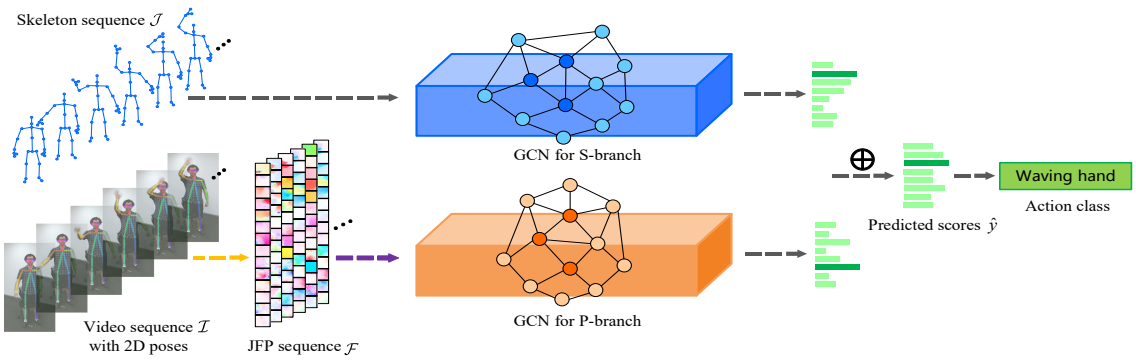


Figure 4. Framework of the proposed JOLO-GCN. A two-stream GCN-based network architecture is used to process the skeleton sequence (S-branch) and the JFP sequence (P-branch) independently. The predicted scores from these two branches are merged into the final predicted scores for each action class by linear blending.

Hadamard product. The operation of $(\boldsymbol{X}_{in}(\boldsymbol{A}^{norm} \odot \boldsymbol{M}))$ guarantees the features corresponding to different skeletal joints to interact by following the skeleton topology given by $\boldsymbol{A}$. $\boldsymbol{W}$ denotes a 1×1 convolution layer to expand the feature dimension and $\sigma(\cdot)$ is a non-linear activation layer (e.g., ReLU). We refer interested readers to [36, 21] for a detailed discussion on GCNs for skeleton-based action recognition and GCNs in the context of the spatial graph theory.

3.3. JOLO-GCN

The framework of JOLO-GCN is illustrated in Figure 4. We use a two-stream network architecture for processing the skeleton sequence and the JFP sequence independently. The skeleton sequence $\mathcal{J}$ is fed into the first GCN branch (called *S-branch*) and the prediction scores are generated subsequently. In parallel, the precomputed JFP sequence $\mathcal{F}$ is fed into the second GCN branch (called *P-branch*). We train S-branch and P-branch independently. The cross-entropy loss $\mathcal{L}$ is used for the training:

$$\mathcal{L} = -y^T log(\hat{y}) \tag{5}$$

where y denotes a one-hot label vector of the ground-truth action class and $\hat{y}$ denotes the predicted scores. The final action class prediction scores are obtained by linear-blending the predicted scores from the two GCN branches.

3.4. Data Format Conversion of JFP

A RGB video sequence $\mathcal{I}$ takes on a four-dimensional format of $T \times H \times W \times 3$. Compared with the original video sequence, the skeleton sequence $\mathcal{J}$ is in a smaller four-dimensional format of $T \times K \times 3 \times N$, with the dimension N representing the maximum number of persons that may appear in the scene. The magnitude of $K \times N$ is much smaller than $H \times W$. As a consequence, though the scale of the neural network for a video sequence would be large usually, the skeleton sequence can be processed with some lightweight networks, such as RNNs or GCNs. Based on the earlier discussion, the proposed JFP sequence is considered as an orthogonal cue to the skeleton sequence. In terms of the data format, the JFP sequence is designed to hold the sparsity property as the skeleton sequence. This assures that it can be processed in a lightweight network accompanied with the skeleton sequence. Therefore, we downsample each JFP using a bilinear interpolation function from the resolution of $l \times l$ to a smaller one of $\mu \times \mu$. Then, we convert the data format of a JFP sequence similar to that of the skeleton sequence. As the format of the skeleton sequence is $T \times K \times 3 \times N$, the JFP sequence ($T \times K \times \mu \times \mu \times 2 \times N$) is converted to a four-dimensional format of $2T \times K \times \mu^2 \times N$. This design choice keeps a good trade-off between the strength of the JFP representation and the computational complexity.

4. Datasets and Implementation Details

4.1. Datasets

We evaluate the performance of our method on three benchmark skeleton-based action recognition datasets.

NTU RGB+D: "NTU RGB+D" [22] is a widely-used benchmark dataset in the field of skeleton-based human action recognition. In this dataset, 56,880 video samples corresponding to 60 action classes are provided. All samples are performed by 40 distinct performers and recorded in 17 different indoor scene setups by three cameras from different views. The provided data of each sample include an RGB video, a 3D human skeleton sequence, a depth map sequence and a IR video. Two official evaluation protocols are adopted in our experiments, i.e., Cross-subject (X-sub) and Cross-view (X-view).

NTU RGB+D 120: "NTU RGB+D 120" is the extended version of the NTU RGB+D dataset by adding another 60 more challenging classes with another 57,600 video samples. All video samples are performed by 106 distinct performers in a wide range of age distribution, and recorded in 32 different indoor scene setups by three cameras from different views. Two evaluation protocols are recommended: Cross-subject (X-sub) and Cross-setup (X-setup).

Method		Pose	Visual	NTU X-sub (%)	NTU X-view (%)	NTU 120 X-sub (%)	NTU 120 X-setup (%)	KS Top-1 (%)	KS Top-5 (%)
GCA-LSTM[16]		✓	-	76.1	84.0	61.2	63.3	-	-
SkeleMotion[3]		✓	-	76.5	84.7	67.7	66.9	-	-
Chained Net[40]		✓	✓	80.8	-	-	-	-	-
ST-GCN[36]		✓	-	81.5	88.3	-	-	30.7	52.8
Ind-RNN[13]		✓	-	81.8	88.0	-	-	-	-
SR-TSL[27]		✓	-	84.8	92.4	-	-	-	-
Deep Bilinear[9]		✓	✓	85.4	90.7	-	-	-	-
2D/3D pose[20]		✓	✓	85.5	-	-	-	-	-
AS-GCN[12]		✓	-	86.8	94.2	-	-	34.8	56.5
2S-AGCN[25]		✓	-	88.5	95.1	83.7*	85.8*	36.1	58.7
SGN[38]		✓	-	89.0	94.5	79.2	81.5	-	-
AGC-LSTM[26]		✓	-	89.2	95.0	-	-	-	-
MS-G3D[19]		✓	-	91.5	96.2	86.9	88.4	38.0	60.9
Posemaps[18]		✓	✓	91.7	95.3	64.6	66.9	-	-
	Backbone	Pose	Visual	X-sub(%)	X-view(%)	X-sub(%)	X-setup(%)	Top-1(%)	Top-5(%)
Ours	ST-GCN	✓	✓	90.4	95.6	-	-	33.7	57.6
	2S-AGCN	✓	✓	**93.8**	**98.1**	**87.6**	**89.7**	**38.3**	**62.3**

Table 1. Quantitative comparisons of the validation accuracy on the NTU RGB+D, NTU RGB+D 120 and Kinetics-Skeleton datasets (referred to as "NTU", "NTU 120", and "KS", respectively). The proposed JOLO-GCN is built and evaluated over two different GCN-based backbones (ST-GCN [36] and 2S-AGCN [25]), and we report both end results accordingly. The second and third columns ("Pose" and "Visual") indicate whether a method under evaluation utilizes skeleton and/or visual data (e.g. images, depth maps, and optical flow), respectively. * denotes the results obtained using the released codes.

Kinetics-Skeleton: Kinetics [5] is a large-scale dataset for action recognition. It contains around 300,000 video clips. The actions cover 400 classes ranging from daily activities, sports scenes, to complex actions with interactions. A related dataset named Kinetics-Skeleton [36] is generated for skeleton-based action recognition. Adopting the public OpenPose toolbox [4], the authors estimated the 2D pose of 18 joints for every frame of the video clips, and also attained each joint's estimation confidence. Following the evaluation method in [36, 25, 24, 12], the dataset is divided into a training set (240,000 samples) and a validation set (20,000 samples). Top-1 and Top-5 accuracies are reported.

4.2. Implementation Details

Implementing JFP: In order to improve experimental efficiency, the JFP sequences are estimated from video sequences and 2D pose sequences in advance. Following the setting of [17, 18], we use the OpenPose toolbox [4] to extract the corresponding 2D pose sequence (i.e., 18 joints) from videos. We remove the joints of eyes and ears to eliminate the redundancy of the patch overlap between adjacent joints. As a result, a total of 14 joints are chosen for processing the corresponding JFPs. In this work, we use the classic TV-L1 algorithm [37] for the optical flow estimation due to its simplicity and effectiveness.

For the NTU RGB+D and NTU RGB+D 120 dataset, the scale of human bodies in the scene is relatively stable, so the patch size $l \times l$ of JFP is empirically set to a fixed value i.e., $l = 32$. For the Kinetics-Skeleton dataset, because the scale of the human body has a large variation, we use the average bone length of each sample to define the patch size. Finally, all of these JFPs are downsampled to a smaller resolution of $\mu \times \mu$, with $\mu = 8$ in our experiment.

Two-Stream GCN-Based Network: In order to fully verify the effectiveness and general compatibility of the JOLO-GCN, we opt for two different GCN-based backbones in our experiments: 1) ST-GCN [36], and 2) 2S-AGCN [25]. In our algorithm design, the joint number of the JFP sequences is 14, while the skeleton sequences use a different 25-joint skeleton structure. Therefore, different graph structures following the respective skeleton structures are constructed for the S-branch and the P-branch.

In the S-branch, an input skeleton sequence is required to contain 300 frames in the original settings of ST-GCN and 2S-AGCN. In the P-branch, the input JFP sequences are downsampled by a temporal downsampling factor of 2, and the sequence length of the corresponding JFP is 64. The JFP sequence ($T \times K \times \mu \times \mu \times 2 \times N = 64 \times 14 \times 8 \times 8 \times 2 \times 2$) is converted to $2T \times K \times \mu^2 \times N = 128 \times 14 \times 64 \times 2$.

The predicted scores from the two branches are added with blending weights to obtain the final prediction. The scores from the S-branch and the P-branch are merged by linear-blending with the weights of 0.5 and 0.5.

Backbone	Data used	X-sub (%)
ST-GCN	Joints (S-branch)	81.5
	JFP (P-branch)	86.6
	Joints + JFP	**90.4**
2S-AGCN	Joints	86.6
	Bones	86.2
	Joints + bones (S-branch)	88.5
	JFP (P-branch)	88.1
	Joints + bones + JFP	**93.8**

Table 2. Comparisons of the accuracy obtained by the S-branch, the P-branch and their combination on the NTU RGB+D dataset.

Modality	2S-AGCN (%)	ST-GCN (%)
JAP	83.8	81.1
JDP	85.9	82.9
JFP	88.1	**86.6**
Skeleton	**88.5**	81.5
Skeleton + JAP	92.0	87.7
Skeleton + JDP	92.1	88.2
Skeleton + JFP	**93.8**	**90.4**
Skeleton + JFP + JAP	93.9	91.2
Skeleton + JFP + JDP	**94.1**	**91.8**

Table 3. Comparisons of the accuracy obtained by different modalities and their combinations evaluated on the NTU RGB+D dataset using the X-sub protocol. When different modalities are combined, their respective weights are set equal.

5. Experiments and Analysis

5.1. Experimental Results

We evaluate the proposed JOLO-GCN on the NTU RGB+D, NTU RGB+D 120, and Kinetics-Skeleton datasets and compare our method with the state-of-the-art skeleton-based action recognition methods. As shown in Table 1, our best model based on the 2S-AGCN backbone obtains action classification accuracy of 93.8% and 98.1% for the X-sub and X-view protocols on NTU RGB+D, 87.6% and 89.7% of classification accuracy for X-sub and X-setup protocols on NTU RGB+D 120 dataset, and 38.3% and 62.3% of Top-1 and Top-5 accuracy on Kinetics-Skeleton. These results validate the new state-of-the-art accuracy achieved by the proposed JOLO-GCN on the three benchmark datasets. In addition, when compared with the original GCN-based baselines, i.e., ST-GCN [36] and 2S-AGCN [25], our proposed method integrated with the JFP stream significantly improves their accuracy by 8.9% and 5.3% on the NTU RGB+D X-sub protocol, respectively.

It is clear to see that our method outperforms all the single-modal methods [10, 29, 36, 13, 27, 11, 12, 25, 26, 24] owing to the introduced JFP stream, which provides useful local motion information. Meanwhile, the proposed JOLO-GCN also outperforms all the multi-modal methods, which utilize skeleton and/or visual data, such as human pose heatmap [17, 18], depth maps [9], and optical flow [40].

5.2. Ablation Study

In this subsection, we examine the proposed JFP stream in more depth, and in particular the performance improvement brought by JFP and its advantage over other possible modalities for action recognition.

Comparison between S-branch and P-branch: As shown in Table 2, when ST-GCN is used as the GCN backbone, the result of the P-branch using the JFP input is better than that of the S-branch using the skeleton sequence. After combining two branches, the final recognition accuracy is improved by 8.9% over the S-branch. When examined on a stronger GCN backbone, 2S-AGCN, the proposed JFP stream again shows its great value in improving the recognition accuracy. Specifically, our proposed P-branch achieves an X-sub accuracy of 88.1%, which is comparable to 88.5% obtained by the S-branch (using joints and bones). Combining the two complementary branches together is beneficial and leads to a much better recognition result (93.8%) than that of each single branch alone. The accuracy gain brought by the JFP stream is around 5.3% over the original S-branch. To put this in context, merging the "Joints" stream and the "Bones" stream yields only an accuracy gain of less than 2% in 2S-AGCN.

Evaluation of the JFP Gain over Each Action Class: To appreciate the accuracy gain by including the JFP stream in more depth, we compare the performances of different branches on every action category. As shown in Figure 5, the performances of the single S-branch on 2S-AGCN ("joint+bone") backbone encounter an accuracy bottleneck in the actions characterized primarily by local subtle movements, due to the limitations of the sparse skeleton sequence discussed in Section 1. After merging S-branch and P-branch, these actions mainly characterized by local subtle movements, such as "clapping"(#10), "reading"(#11), "writing"(#12), "play with phone/tablet"(#29), and "type on a keyboard"(#30), have gained significant accuracy improvements, i.e., 11.3%, 16.1%, 19.8%, 6.9%, and 19.2% for the experiments using the 2S-AGCN backbone, respectively. This per-action class recognition accuracy comparison shows clearly that the proposed JFP sequences effectively capture the local subtle motion information and help the network to make more accurate recognition.

Comparison of Using Different Patch Modalities: In fact, it remains intriguing to find out whether other patch modalities can possibly be better options over JFP, such as JAP or JDP (Joint-centered Depth Patches). To this end, we conduct a series of experiments to evaluate different patch modalities, and their combinations with the input skeleton stream. The results are reported in Table 3.

A few key observations can be found from Table 3. First, when a single modality is used for GCN-based action recog-

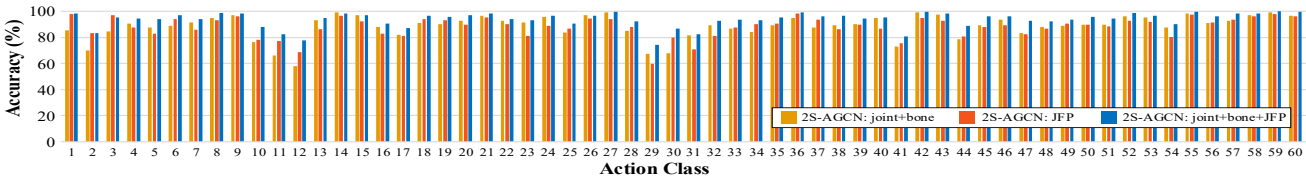


Figure 5. Class-by-class action recognition accuracy comparisons of the S-branch, P-branch and their combinations evaluated with the X-sub protocol on the NTU RGB+D dataset.

Method	Parameters	GPU Memory Consumption	Runtime	FLOPS
ST-GCN	3.10M	341MB	93ms	16.3G
2S-AGCN	6.94M	704MB	138ms	37.3G
JOLO-GCN (ST-GCN+P-branch)	3.10+3.10M	341+353MB	93+22ms	16.3+3.9G
JOLO-GCN (2S-AGCN+P-branch)	6.94+3.48M	704+468MB	138+29ms	37.3+4.5G

Table 4. Comparisons of the proposed JOLO-GCN and baselines in parameter, runtime, and resource consumption. The above statistics report the average values for one forward inference on a Tesla K80 GPU. FLOPS denotes floating-point operations per second.

nition, the input skeletons or the proposed JFP stream tend to give better accuracies than JAP or JDP, as they capture motion dynamics more important and direct. Second, if one additional patch modality is added to complement the original skeleton stream, JFP is again the best choice over JAP and JDP, leading to the recognition accuracy of 93.83% on the 2S-AGCN backbone. Such a combination actually corresponds to our proposed method evaluated earlier in Table 1. Third, it can be found that including one more extra patch modality besides JFP at the cost of increased network complexity is not necessary, because the accuracy gain is quite marginal, e.g., 94.1% (Skeleton + JFP + JDP) versus 93.8% (Skeleton + JFP). Also, depth maps are not always available as the input to extract the JDP stream from.

Runtime Analysis: The following discussions are conducted using the NTU RGB+D dataset. In our proposed method, the JFP pre-processing includes loading full-HD video, temporal downsampling, joint-centered cropping of image patches (JAP), and TV-L1 optical flow estimation for computing JFP. The original videos of NTU RGB+D have 80 frames on average, and we downsampled them by a temporal factor of 2. Our JFP pre-processing takes about 1.5 seconds for a video averagely on a laptop with a Intel i5-7300HQ CPU without using GPUs. Our method using the 2S-AGCN backbone (three streams: "Joints"+"Bones"+"JFP") takes about 167 ms for a full prediction using a Tesla K80 GPU.

Model Complexity Analysis: Table 4 shows the comparison of JOLO-GCN and baseline methods (ST-GCN and 2S-AGCN) in model complexity. Compared with the baseline methods, the computational complexity introduced by the P-branch is relatively low. The difference of the joints numbers and frames numbers between skeleton sequences (25 joints, 300 frames) and JFP sequence (14 joints, 64 frames) makes P-branch fast computationally compared with both baselines.

6. Conclusion

In this paper, we proposed a novel approach of representing the visual information surrounding each skeletal joint as Joint-aligned optical Flow Patches (JFP), effectively capturing the useful local subtle body motion cues for skeleton-based action recognition. The derived JFP sequence has the advantage of a compact representation and inherits a kinetically meaningful structure from the human pose skeleton. Based on the proposed JOLO-GCN framework, we jointly exploit local subtle motion cues from the JFP sequence and global motion cues from the skeleton sequence for action recognition. The proposed method obtains state-of-the-art results on the three large-scale action recognition datasets. Our experiments further show that when applied on two different GCN-based backbones (ST-GCN [36] and 2S-AGCN [25]), the proposed method improves both of them by large performance margins. This validates the generalization ability of applying our scheme on different light-weight single-modal skeleton-based networks.

Acknowledgement

The work was supported in part by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the National Key R&D Program of China with grant No. 2018YFB1800800, by Natural Science Foundation of China with grant NSFC-61902334 and NSFC-61629101, by Guangdong Zhujiang Project No. 2017ZT07X152, and by Shenzhen Key Lab Fund No. ZDSYS201707251409055, by the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No.: 2017ZT07X183), the National Natural Science Foundation of China (Grant No.: 61771201), and the Guangdong R&D key project of China (Grant No. 2019B010155001).

References

[1] Fabien Baradel, Christian Wolf, and Julien Mille. Human action recognition: Pose-based attention draws focus to hands. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 604–613, 2017.

[2] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 469–478, 2018.

[3] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A Dos Santos, and William Robson Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[6] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018.

[7] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[9] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018.

[10] Zhiwu Huang, Chengde Wan, Thomas Probst, and Luc Van Gool. Deep learning on lie groups for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6099–6108, 2017.

[11] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 786–792. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[12] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5457–5466, 2018.

[14] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[15] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.

[16] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2017.

[17] Mengyuan Liu, Fanyang Meng, Chen Chen, and Songtao Wu. Joint dynamic pose image and space time reversal for human action recognition from videos. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[18] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[20] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018.

[21] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023, 2016.

[22] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[23] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Non-local graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1805.07694*, 2018.

[24] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019.

[25] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019.

[26] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[27] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.

[28] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[29] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28, 2017.

[30] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[31] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR 2011*, pages 3169–3176. IEEE, 2011.

[32] Hongsong Wang and Liang Wang. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Transactions on Image Processing*, 27(9):4382–4394, 2018.

[33] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[34] Zhikai Wang, Chongyang Zhang, Wu Luo, and Weiyao Lin. Key joints selection and spatiotemporal mining for skeleton-based action recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3458–3462. IEEE, 2018.

[35] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7922–7931, 2019.

[36] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[37] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007.

[38] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[39] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2344–2353, 2019.

[40] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2904–2913, 2017.