# A Multi-Stream Graph Convolutional Networks-Hidden Conditional Random Field Model for Skeleton-Based Action Recognition

Kai Liu, *Student Member, IEEE*, Lei Gao, *Member, IEEE*, Naimul Mefraz Khan, *Member, IEEE*, Lin Qi, and Ling Guan, *Fellow, IEEE*

***Abstract*—Recently, Graph Convolutional Network(GCN) methods for skeleton-based action recognition have achieved great success due to their ability to preserve structural information of the skeleton. However, these methods abandon the structural information in the classification stage by employing traditional fully-connected layers and softmax classifier, leading to sub-optimal performance. In this work, a novel Graph Convolutional Networks-Hidden conditional Random Field (GCN-HCRF) model is proposed to solve this problem. The proposed method combines GCN with HCRF to retain the human skeleton structure information even during the classification stage. Our model is trained end-to-end by utilizing the message passing from the belief propagation algorithm on the human structure graph. To further capture spatial and temporal information, we propose a multi-stream framework which takes the relative coordinate of the joints and bone direction as two static feature streams, and the temporal displacements between two consecutive frames as the dynamic feature stream. Experimental results on three challenging benchmarks (NTU RGB+D, N-UCLA, SYSU) show the superior performance of the proposed model over state-of-the-art models.**

***Index Terms*—GCN, CRF, skeleton, hidden part state, action recognition.**

## I. Introduction

HUMAN action recognition is an active research area in multimedia computing with many applications such as human-computer interaction, video understanding and intelligent surveillance [49]. In essence, human action can be recognized from multimodal information sources, such as RGB, depth and skeletons, etc. Due to the succinctness of representation and robustness to variations of viewpoints and environment, skeleton-based human action recognition has attracted considerable attention in multimedia community recently [26], [31], [37], [47], [49]. In this paper, we focus on the problem of skeleton-based action recognition.

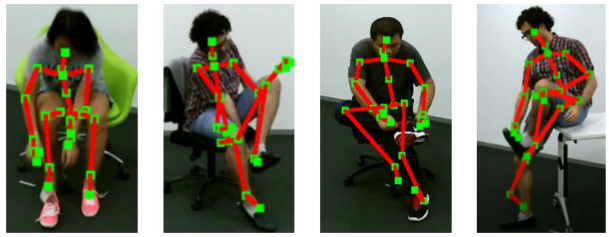


Fig. 1. Large intra-class variations exist as the same action can be performed in different ways. (samples taken from the action "wear a shoe" in NTU RGB+D dataset).

Convolutional deep neural network based methods like CNN or RNN usually represent the skeletal sequence as a vector of joint coordinates or a pseudo-image for applying the convolution operation, which can not capture the spatial structure of the human skeleton. Recently, graph convolutional networks, which generalize convolution from image to graph, have been successfully adopted in skeleton-based action recognition. GCN can preserve the connectivity among the skeleton joints, thus preserving the spatial structure. The spatial-temporal graph convolutional networks (ST-GCN) method in [35] is proposed to construct spatio-temporal skeleton graph, base on which multiple layers of spatio-temporal graph convolution operations are applied to extract the high-level feature maps for predicting the label by the softmax classifier. While the GCN-based method provided a significant increase in performance, we argue that a serious problem has been ignored. Although the graph structures and corresponding convolution operations have been designed well to preserve the natural structure of the human body, in the classification stage, the structural information is abandoned due to the utilization of fully connected layers and softmax classifier. This approach somewhat negates the effect of preserving structural information with GCN in the previous layers. Instead of using the traditional fully connected layer and softmax classifier, the proposed model attempts to retain the structure of the human body even in the classification stage. It is known that action recognition suffers from large intra-class variations as shown in Fig. 1. By keeping the structure of body

Manuscript received May 27, 2019; revised November 7, 2019; accepted February 3, 2020. Date of publication February 17, 2020; date of current version December 17, 2020. This work was supported by NSFC-Henan Joint Fund under Grant U1804152. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Elisa Ricci. *(Corresponding author: Lin Qi.)*

Kai Liu and Lin Qi are with the School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China (e-mail: 09liukai08@gmail.com; ielqi@zzu.edu.cn).

Lei Gao, Naimul Mefraz Khan, and Ling Guan are with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: iegaolei@gmail.com; n77khan@ee.ryerson.ca; lguan@ee.ryerson.ca).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2974323

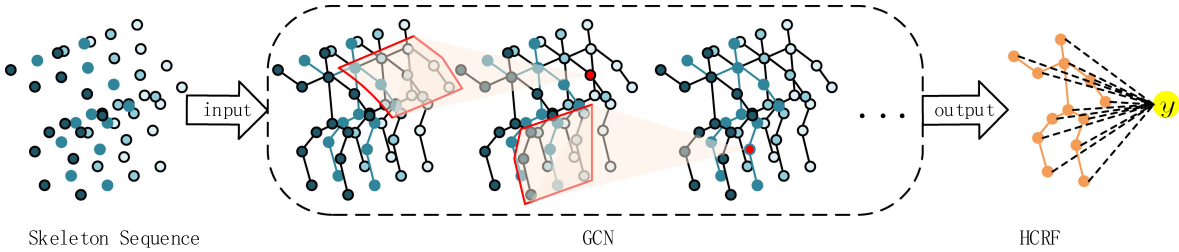



Fig. 2. The framework of the proposed GCN-HCRF model: We utilize a GCN network to extract feature of each body joint. Then the extracted feature is sent to HCRF for classification. The HCRF not only assigns a hidden part state (orange circle) for each joint's feature, but also models the compatibility among these hidden part states. The whole model is trained in an end-to-end manner with back propagation.

parts and modeling the compatibility among all joints' motion at classification stage, it can help recognize the same action performed in different ways. Furthermore, retaining the body structure in classification stage will in return guide the feature extraction of the GCN network, encouraging the network to extract more compatible joints' features for respective actions.

Inspired by hidden part models for action recognition in [43], we propose a novel GCN-HCRF model which combines GCN with Hidden Conditional Random Field to retain the human skeleton structure information during classification. An overview of the proposed method is shown in Fig. 2 on the next page. An HCRF model is combined with the GCN seamlessly for classification without abandoning the human structure. As can be seen in Fig. 2, the HCRF model's neuron connections can be structured in a way that preserves the spatial structure of the human skeleton. It is not possible to achieve such structure preservation with fully-connected layers, since every neuron in one layer is connected to every neuron in the next layer.

However, merely putting the GCN and HCRF model together is not enough to take advantage of the benefit of spatial structure preservation. The model needs to be trained end-to-end so that the HCRF model can guide the GCN to extract features that are semantically more meaningful. To achieve this task, we utilize a message passing strategy on the acyclic graph of human joints inspired by belief propagation [11].

Another notable problem in action recognition is that although static feature like joint's coordinate and bone direction within each frame has been proven to be effective in [23], [26], [37], the dynamic feature also contains rich information for skeletal action recognition since it provides explicit motion information. In this paper, motivated by classic two-stream CNN for action recognition in [22], [42], which takes appearance information as static feature stream and takes optical flow as dynamic feature stream. We take the relative coordinate of the joints and bone direction as two static feature streams and take the temporal displacements between the joints in two consecutive frames as the dynamic feature stream. All three streams are incorporated utilizing score level fusion, providing improved performance over utilizing a single stream.

To verify the superiority of the proposed model, we conduct experiments on three challenging datasets with different scale: NTU RGB+D (56,000 samples), N-UCLA (1,494 samples) and SYSU (480 samples). Since the size of N-UCLA and SYSU are too small to train the deep graph convolution network adequately, an intuitive option is to employ the pre-trained parameters from the large-scale datasets NTU RGB+D as the initial parameters. However, using the pre-trained parameters directly is not feasible as the number of body joints in the datasets varies, leading to different graph structures. In this work, we propose two adaptation strategies to solve this problem, the interpolation strategy and the trim strategy. Experiment results show that the proposed method outperforms state-of-the-art methods on all three datasets.

In this paper, our main contributions are summarized below:

1) We propose a novel GCN-HCRF model which retains the spatial structure of human joints from beginning to end. Unlike current GCN-based model, we discard the fully connected layers, since they destroy the spatial structure. Instead, we utilize an HCRF model to take full advantage of the spatial compatibility information among all human joints' motion for classification. An end-to-end training is carried out by using the message passing strategy on the acyclic graph of human joints, which enables the HCRF model to guide GCN to extract feature that is semantically more meaningful, improving the overall classification performance.
2) To further boost the performance, a three-stream framework is proposed, which takes the relative coordinate of joints and bone direction as two static feature streams and takes the temporal displacements between the joints in two consecutive frames as the dynamic feature stream.
3) To train the proposed model on two small-sized datasets N-UCLA and SYSU with the pre-trained parameters from NTU RGB+D, we proposed two adaptation strategies to solve the difference of graph structure, boosting the performance on N-UCLA and SYSU substantially.
4) Experimental results demonstrate the proposed model outperforms state-of-the-art approaches on three standard datasets.

## II. Related Work

In this section, we briefly review the previous works on skeleton-based action recognition from two important components: neural networks based methods, CRF based methods.

### A. Neural Networks Based Methods

The neural network architectures relevant to this work include convolutional neural networks (CNNs), recurrent neural networks (RNNs) and graph convolution networks (GCNs). In [5], a CNN-based model transforms a skeleton sequence to a pseudo-image by treating the joint coordinate (x, y, z) as the R, G, B channels of a pixel. In [25], skeletons are visualized as a series of color images, in which visual and motion enhancement methods are applied to enhance their local patterns. In [9], a synchronous local and non-local approach is introduced to simultaneously capture the details and semantics in the spatio-temporal domain. In [37], to exploring the performance of different geometric features on action recognition, eight geometric features are presented for a three-layer LSTM. Moreover, a multi-stream LSTM architecture is proposed to fuse the classification results of different geometric feature streams. In [26], [27], [49], in order to achieve view invariant action recognition, the view adaptive neural networks is adopted to find the most suitable observation viewpoints. In [35], spatio-temporal GCN is proposed to apply graph convolution on the natural structure graph of the skeleton action sequences. In [23], non-local graph convolutional block is introduced to adaptively learn the graph structure for different layers and samples.

### B. CRF Based Methods

Sequence labeling models, like CRF [12], are suitable for modeling and analyzing human actions, because such Markov chain models are able to capture the structural dependencies among the outputs. Many extensions of CRF are applied to a wide range of fields. In [43], a hidden part-based CRF approach is proposed to combine both large-scale global features and local patch features for action recognition based on max-margin hidden conditional random fields (MMHCRF). In [39], by observing that each class has distinct sub-structures, a HCRF model is introduced to capture the sub-structures for gesture recognition. In [32], a mid-level video representation approach based on CRF is proposed for action recognition by viewing part-cluster assignments as hidden variables. In [33], a composite latent structure model is introduced to recognize skeleton sequences by representing each atomic action as a composite latent state. In [18], a nonparametric feature matching based CRF model is proposed for gesture recognition, in which the gesture classes are estimated for all frames to localize the outputs automatically in the unconstrained input video.

Joint model of CNNs and CRF is helpful in many applications. The fully connected CRF [19] performs well in semantic segmentation and human pose estimation, and it was trained jointly in [48] by unrolling iterations of the inference method. In [40], a spatio-temporal And-Or graph is proposed to reconfigure models during learning and inference. In [20], an extended 3D CNN model is proposed by incorporating structure alternatives to make it a structured deep architecture. In [38], a deep sequence model is introduced by extending the CRF models with CNN for higher level feature learning. In [13], a hybrid framework combines Convolutional Neural Network (CNN) and Latent Dynamic Conditional Random Field (LDCRF) to segment and recognize continuous actions simultaneously. In [8], a fully-connected temporal CRF model is proposed for reasoning over various aspects of activities that include objects, actions, and intentions, where the potentials are predicted by a deep network.

## III. GCN-HCRF Model

In this paper, the main aim is to retain the spatial structure of human joints in the classification. We view the spatial structure of human joints as a random field based in which a novel GCN-HCRF model is proposed. The proposed model can take full advantage of compatible information among all joints' motions. Therefore, it guides the GCN network to extract more semantically meaningful feature from all joints.

### A. Overview of The Proposed Framework

As shown in Fig. 2, the proposed model takes the sequences of body joins in the form of 3D coordinates as input. In the first part, a GCN network is applied to mine the feature and generate higher-level feature maps in the form of human structure graph. After that, a HCRF model is combined with the GCN seamlessly for classification without abandoning the human structure. The HCRF not only assigns a hidden part state for each joint's feature, but also models the compatibility among these hidden part states. The whole model is trained in an end-to-end manner with back propagation. The specific components of the GCN-HCRF model are explained in detail below.

### B. Feature Extraction With Graph Convolutional Network

Skeleton Graph Construction: Suppose a skeletal sequence includes $T$ frames and each frame has $N$ joints. Then we utilize the spatio-temporal graph convolutional network to extract feature of each joint. This GCN network is an undirected spatio-temporal graph $G = (V, E)$ which is constructed on the natural spatio-temporal structure of a skeletal sequence, where $V$ stands for the set of all the joints in a skeleton sequence, and $E$ is the edge set that includes the intra-body edges and the inter-frame edges. The intra-body edges are natural connections of human skeleton in each frame. The inter-frame edges are connections of the same joints between adjacent frames. The coordinate vectors of joints are used as this graph network's inputs. To exploit the feature with spatio-temporal information, the ST-GCN network [35] is constructed to perform convolution operation in two dimensions: spatial dimension and temporal dimension.

*Spatial Dimension:* The spatial graph convolution operation within one single frame is formulated as [35]:

$$f_{out}(v_i) = \sum_{v_j \in B(v_i)} \frac{1}{Z_i(v_j)} f_{in}(v_j) w(l_i(v_j)) \tag{1}$$

where $v_i$ is the $i$-th joint within a single skeleton frame, $f_{in}$ is the input feature map and $f_{out}$ is the output feature map. $B(v_i)$ represents the 1-distance neighbor set of the given joint $v_i$ within a single frame, and $v_j$ is a joint from the set $B(v_i)$. According

to the spatial configuration partitioning in [35], $B(v_i)$ is divided into 3 subsets: the root joint $v_i$ itself, the centripetal group and the centrifugal group. $l_i(v_j)$ represent a mapping function which is used for mapping each joint $v_j$ to the corresponding subset. $Z_i(v_j)$ is a normalizing term which equals the cardinality of the corresponding subset. $w$ is the weight function which provides a weight vector to compute the inner product.

*Temporal Dimension:* Since the number of neighbors for each joint on temporal dimension is fixed, it enables us to conduct the graph convolution operation with a simple strategy that is similar to the classic CNN convolution operation on regular images. Concretely, we conduct a $\tau \times 1$ convolution operation on the output feature map in Eq. (1), where $\tau$ is the temporal range called the temporal kernel size.

The spatial temporal graph convolution is implemented by representing Eq. (1) in the matrix form as follows [35]:

$$\mathbf{F}_{out} = \sum_j \mathbf{\Lambda}_j^{-\frac{1}{2}} (\mathbf{A}_j \otimes \mathbf{M}) \mathbf{\Lambda}_j^{-\frac{1}{2}} \mathbf{F}_{in} \mathbf{W}_j \tag{2}$$

where $\mathbf{F}_{in}$ represent the input feature map as a matrix of $(N, T, C_{in})$ dimensions, $N$ denotes the number of joints for each frame, $T$ denotes the number of frames for each action sequence and $C_{in}$ denotes the number of input channels. $\mathbf{F}_{out}$ represent the output feature map as a matrix of $(N, T, C_{out})$ dimensions, where $C_{out}$ denotes the number of output channels. For spatial configuration partitioning, $j \in \{0, 1, 2\}$. $\mathbf{W}_j$ is the $C_{in} \times C_{out} \times 1 \times 1$ weight matrix for 1-distance neighbors spatial temporal graph convolution operation. $\mathbf{A}_j$ denotes a $N \times N$ adjacency matrix. $\mathbf{A}_0$ denotes an identity matrix representing self-connections of joint itself. $\mathbf{A}_1$ denotes the connections of centripetal group and $\mathbf{A}_2$ denotes the connections of centrifugal group. $\mathbf{M}$ is a learnable weight matrix for learning the importance of all connections. The sign of $\otimes$ denotes element-wise product between two matrices. $\mathbf{\Lambda}_j$ is a diagonal matrix, whose element $\mathbf{\Lambda}_j^{ii} = \sum_k \mathbf{A}_j^{ik} + \alpha$, where $\mathbf{A}_j^{ik}$ is the element of matrix $\mathbf{A}_j$, and $\alpha$ is set to 0.001 to avoid empty rows in $\mathbf{A}_j$.

Now, we have a well-defined spatial and temporal convolution operation on the whole skeleton sequence, by which we extract feature with spatio-temporal information on every human joint. Different from [35] where the extracted feature is used directly for classification with fully connected layer, we send the extracted feature to HCRF for classification. Specifically, as shown in Fig. 3, the output feature map from the last layer of GCN is a matrix of $(N, T, C_{out})$ dimensions. After conducting the global average pooling on $T$ and on $[N, T]$ respectively, we obtain a local feature matrix $\mathbf{L}$ with $(N, C_{out})$ dimensions for joints' motion and a global feature vector $x_0$ with $C_{out}$ dimensions for the whole body's motion.

### C. Hidden Conditional Random Field for Action Recognition

In this subsection, we describe how to model the HCRF based on the extracted features from above GCN. As described above, we obtained the local feature $\mathbf{L}$ that is a matrix with $(N, C_{out})$ dimensions and the global feature $x_0$ that is a vector with $C_{out}$ dimensions. Let $\mathbf{x}$ be the feature of the given skeletal sequence, which includes the local feature $\mathbf{L}$ and the global feature $x_0$, then the feature $\mathbf{x}$ is represented as $\mathbf{x} = (x_0, x_1, x_2, \ldots, x_i, \ldots, x_N)^{\mathrm{T}}$ where $x_0$ is the global feature vector of the whole body's motion, and $x_i (i = 1 \ldots N)$ denotes the local feature vector of the $i$-th joint's motion, which corresponds to the $i$-th row of the local feature matrix $\mathbf{L}$. Let $y$ be the corresponding class label of the given skeletal sequence, and $Y$ be the finite class label set, $y \in Y$, the total number of action classes is $|Y|$. For action recognition, our task is to predict the class label $y$ for the given feature $\mathbf{x}$. In order to model the compatibility among all joints' motion, we introduce a vector of hidden part states $\mathbf{h} = (h_1, h_2, \ldots, h_i, \ldots, h_N)^{\mathrm{T}}$, where $h_i$ is a hidden part state assigned to $x_i$, for $i = 1, 2, \ldots N$. Each $h_i$ takes value from a finite hidden part state set $H$ which cannot be observed in the training set but will be learned as the hidden variables of the model during training process, and the size of set $H$ is denoted as $|H|$. In addition to assigning a hidden part states $h_i$ to $x_i$, assume that there exist certain constraints between some pairs of $(h_j, h_k)$, where $j$ and $k$ are any two joints that are adjacent in human body. For each given action, the relative movement of the joints with respect to each other imposes these constraints. Taking the "wear shoe" action for example, as shown in Fig. 1, the hand joint and elbow joint might have the constraint that they both have to move down to the foot so that the hidden part states $h_j$ and $h_k$ at that two joints should be the same. Simultaneously, both the foot joint and the knee joint move up to the hand or both stay still by which they tend to have the same hidden part state.

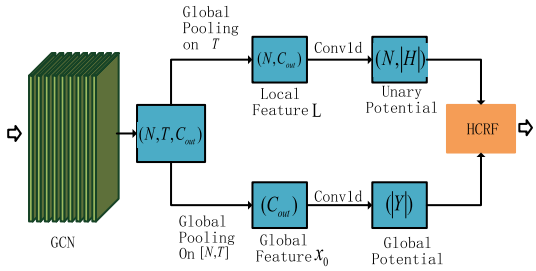



Fig. 3. Illustration of model implementing: Each blue rectangle represents a matrix or a vecor, and the orange rectangle represents the HCRF model. The output from the last layer of GCN is a matrix of $(N, T, C_{out})$ dimensions. After conducting the global average pooling on $T$ and on $[N, T]$, we obtain a local feature matrix $\mathbf{L}$ of $(N, C_{out})$ dimensions and a global feature vector $x_0$ of $(C_{out})$ dimensions. Then they are used as the unary potential $\phi(\cdot)$ and the global potential $\vartheta(\cdot)$ of HCRF after conducting 1D convolution operation with the kernel size of 1. More details about the potential of HCRF are described in Section III-C.

According to the theory of random fields in [1], given the feature $\mathbf{x}$, its corresponding hidden part states $\mathbf{h}$, and the class label $y$, a hidden conditional random field has the exponential form as follows:

$$P(y, \mathbf{h}|\mathbf{x}; \theta) = \frac{\exp(\Phi(y, \mathbf{h}, \mathbf{x}; \theta))}{\sum_{y' \in Y} \sum_{\mathbf{h} \in H^N} \exp(\Phi(y', \mathbf{h}, \mathbf{x}; \theta))} \tag{3}$$

where $\theta$ is the model parameter, and $\Phi(y, \mathbf{h}, \mathbf{x}; \theta)$ refers to potential function depending on the feature $\mathbf{x}$, the hidden part states $\mathbf{h}$, and the class label $y$. The denominator part is the partition function for normalization which sums over all possible hidden

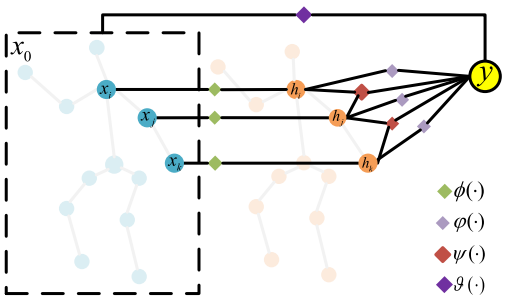



Fig. 4. HCRF structure for skeletal action: Unary potential $\phi(\cdot)$ measure the likelihood of $x_j$ is assigned as $h_j$. Unary potential $\varphi(\cdot)$ measures the compatibility between $h_j$ and $y$. Pairwise potential $\psi(\cdot)$ captures the compatibility between a pair of $(h_j, h_k)$ and $y$. Global potential $\vartheta(\cdot)$ measures the the likelihood of $x_0$ is assigned as $y$.

part states $\mathbf{h}$ and all possible class label $y'$. $H^N$ denotes the set of all possible hidden part states of $N$ hidden parts.

Then the probability of class label $y$ for the given feature $\mathbf{x}$ is the summation of Eq. (3) over all possible assignments of hidden part states $\mathbf{h}$:

$$P(y|\mathbf{x};\theta) = \sum_{\mathbf{h}\in H^N} P(y,\mathbf{h}|\mathbf{x};\theta) = \frac{\sum_{\mathbf{h}\in H^N}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))}{\sum_{y'\in Y}\sum_{\mathbf{h}\in H^N}\exp(\Phi(y',\mathbf{h},\mathbf{x};\theta))} \quad (4)$$

$\Phi(y,\mathbf{h},\mathbf{x};\theta)$ is defined as the summation of unary potential, pairwise potential and global potential in the following form:

$$\Phi(y,\mathbf{h},\mathbf{x};\theta) = \sum_{j\in\nu}\phi(x_j,h_j;\omega) + \sum_{j\in\nu}\varphi(y,h_j;\delta) + \sum_{(j,k)\in\epsilon}\psi(y,h_j,h_k;\xi) + \vartheta(y,x_0;\varpi) \quad (5)$$

where $\phi(\cdot)$ is unary potential with parameter $\omega$ that measures the likelihood of the local feature $x_j$ is assigned as the hidden part state $h_j$. $\varphi(\cdot)$ is unary potential with parameter $\delta$ that measures the compatibility between the hidden part state $h_j$ and the action label $y$. $\psi(\cdot)$ is pairwise potential with parameter $\xi$ that is meant for capture the compatibility between a pair of $(h_j, h_k)$ and the action label $y$. $\vartheta(\cdot)$ is global potential with parameter $\varpi$ that measures the likelihood of the global feature $x_0$ is assigned as the action label $y$. $\nu$ is the joint set of human body and $\epsilon$ is the edge set of human body. The HCRF structure for skeletal action is shown in Fig. 4. The details of these potential functions are described below.

Unary potential $\phi(x_j,h_j;\omega)$ measures the likelihood of the local feature $x_j$ is assigned as the hidden part state $h_j$, whose parameter $\omega$ is learned by above GCN network. In this work, the likelihood is obtained by applying common one-dimensional convolution operation with $1*1$ kernel size on the local feature matrix $\mathbf{L}$ with $(N, C_{out})$ dimensions, by which a probability matrix $\mathbf{M}$ with $(N,|H|)$ dimensions is drawn as shown in Fig. 3. Each element $\mathbf{M}_{jl}$ of the probability matrix $\mathbf{M}$ represents the probability that $x_j (j = 1\ldots N)$ is assigned as the $l$-th $(l = 1\ldots|H|)$ state of the hidden part state set $H$, since $h_j$ can take any value from the hidden part state set $H$. The parameter $\omega$ includes the parameter for above GCN network and the parameter for the one-dimensional convolution operation on the local feature matrix $\mathbf{L}$, which will be learned during training process.

Unary potential $\varphi(y,h_j;\delta)$ measures the compatibility between class label $y$ and hidden part state $h_j$. To compute this potential, we parametrize it as:

$$\varphi(y,h_j;\delta) = \sum_{a\in Y}\sum_{b\in H}\delta_{a,b}\cdot\mathbf{1}(y=a)\cdot\mathbf{1}(h_j=b) \quad (6)$$

where the parameter $\delta$ is a matrix of $(|Y|,|H|)$ dimensions, whose element $\delta_{a,b}$ represents how likely an action with class label $y=a$ contains a joint with hidden part state $h_j=b$, and it will be learned during training process. $\mathbf{1}(\cdot)$ is a indicator function which take the value 1 when its argument evaluates to true and 0 otherwise.

Pairwise potential $\psi(y,h_j,h_k;\xi)$ measures the compatibility between class label $y$ and a pair of hidden part states $(h_j,h_k)$, where $(j,k)$ corresponds to an edge in the human body graph as shown in Fig. 4. To compute this potential, we parametrize it as:

$$\psi(y,h_j,h_k;\xi) = \sum_{a\in Y}\sum_{b\in H}\sum_{c\in H}\xi_{a,b,c}\cdot\mathbf{1}(y=a)\cdot\mathbf{1}(h_j=b)\cdot\mathbf{1}(h_k=c) \quad (7)$$

where the parameter $\xi$ is a matrix of $(|Y|,|H|,|H|)$ dimensions, whose element $\xi_{a,b,c}$ means how likely an action with class label $y=a$ contains a pair of joints with hidden part states $h_j=b$ and $h_k=c$, and it will be learned during training process.

In addition to the local feature, It's well known that the global feature is also very important for recognizing actions. So we use the global potential $\vartheta(y,x_0;\varpi)$ to measure the compatibility of class label $y$ and the global feature vector $x_0$ of the whole action. As shown in Fig. 4, the global feature vector $x_0$ with $C_{out}$ dimensions is obtained by conducting global average pooling on the output of the GCN network on $N$ and $T$. Then we apply common one-dimensional convolution operation with $1*1$ kernel size on the global feature vector $x_0$. After that we draw a probability vector of $|Y|$ dimensions whose $i$-th item represent the probability that $x_0$ is assigned as the $i$-th class. The parameter $\varpi$ includes the parameter for above GCN network and the parameter for the one-dimensional convolution operation on the global feature vector $x_0$, which will be learned during training process.

Note that if we only consider the global potential $\vartheta(y,x_0;\varpi)$ and remove all the others, the proposed GCN-HCRF model will degenerate into the GCN with a fully connected layer for classification. If we don't consider the global potential $\vartheta(y,x_0;\varpi)$, the proposed model can still work well. But the performance of the proposed model will increase when taking the global potential $\vartheta(y,x_0;\varpi)$ into account, as shown in the results in Section IV-B1.

### D. End-to-End Training

One major goal of this work is the end-to-end training of the whole model so that the GCN can extract semantically meaningful features by learning from the HCRF. Assuming that the

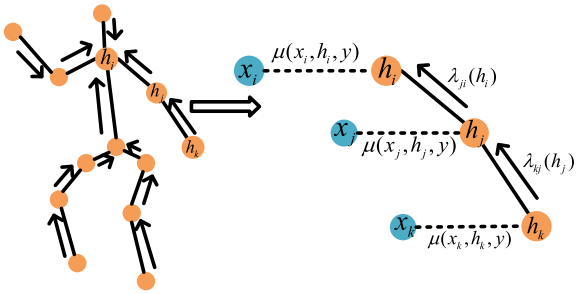



Fig. 5. Message passing route and passing rule.

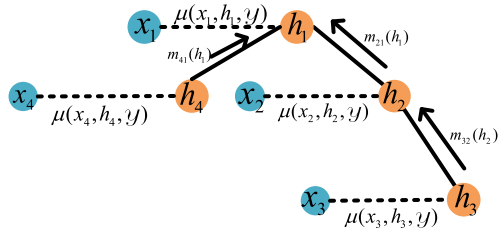



Fig. 6. An example of message passing on a graph with four joints.

training dataset includes S labeled action sequences. Following the minimum negative conditional log-likelihood rule, the loss function is shown in Eq. (8):

$$L(\theta) = -\sum_{s=1}^{\mathrm{S}} \log P(y^{(s)}|\mathbf{x}^{(s)};\theta) \tag{8}$$

where $y^{(s)}$ is the label of the $s$-th action sequence sample, $\mathbf{x}^{(s)}$ is the feature of the $s$-th action sequence sample, and S is the total number of samples. The definition of $P(\cdot)$ is shown in Eq. (4). To forward compute the loss, the key challenge is to calculate the summation over all possible assignments of hidden part state for body joints in the numerator and the denominator of Eq. (4), which is intractable by brute force as all the possible assignments of hidden part state are exponential. For example, if the human body includes $N$ joints, there are $|H|^N$ possible assignments. Fortunately, since the human joints structure is an acyclic graph, we can calculate the summation efficiently through belief propagation on this acyclic graph. According to the belief propagation algorithm for a tree structure [11], the summation can be computed efficiently by message passing to root joint from all the other joints. On the human structure graph, our message passing route and passing rule are shown in Fig. 5.

In Figure 5, $m_{ji}(h_i)$ is a message from joint $j$ to joint $i$ about how likely joint $i$ is in state $h_i$. $m_{ji}(h_i)$ takes all upstream messages to joint $i$, which is updated by the following rule:

$$m_{ji}(h_i) \leftarrow \sum_{h_j \in H} \mu(x_j,h_j,y)\eta(h_i,h_j,y) \prod_{b\in B(j)\setminus i} m_{bj}(h_j) \tag{9}$$

where $B(j) \setminus i$ means the neighboring joint set of joint $j$ except joint $i$. $m_{bj}(h_j)$ takes all neighbor joints' message going to joint $j$. $\mu(x_j,h_j,y)$ is the summation of both unary potential and the global potential at joint $j$, which is formulated in Eq. (10). Here the global potential $\vartheta(y,x_0;\varpi)$ is divided by $N$ so that the influence of global potential is evenly distributed among $N$ joints.

$$\mu(x_j,h_j,y) = \exp\Big(\phi(x_j,h_j;\omega) + \varphi(y,h_j;\delta) + \frac{1}{N}\vartheta(y,x_0;\varpi)\Big) \tag{10}$$

$\eta(h_i,h_j,y)$ is the pair potential between joint $i$ and joint $j$, which is formulated in Eq. (11):

$$\eta(h_i,h_j,y) = \exp(\psi(y,h_i,h_j;\xi)) \tag{11}$$

Using Eq. (10) and Eq. (11), we compute the $\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))$ part in Eq. (4) as follows:

$$\begin{aligned}
&\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta)) \\
&= \exp\Bigg(\sum_{j\in\nu}\phi(x_j,h_j;\omega) + \sum_{j\in\nu}\varphi(y,h_j;\delta) \\
&\quad + \sum_{(j,k)\in\epsilon}\psi(y,h_j,h_k;\xi) + \vartheta(y,x_0;\varpi)\Bigg) \\
&= \exp\Bigg(\sum_{j\in\nu}\phi(x_j,h_j;\omega) + \sum_{j\in\nu}\varphi(y,h_j;\delta) \\
&\quad + \sum_{j\in\nu}\frac{1}{N}\vartheta(y,x_0;\varpi) + \sum_{(j,k)\in\epsilon}\psi(y,h_j,h_k;\xi)\Bigg) \\
&= \exp\Bigg(\sum_{j\in\nu}\Big(\phi(x_j,h_j;\omega) + \varphi(y,h_j;\delta) \\
&\quad + \frac{1}{N}\vartheta(y,x_0;\varpi)\Big) + \sum_{(j,k)\in\epsilon}\psi(y,h_j,h_k;\xi)\Bigg) \\
&= \exp\Bigg(\sum_{j\in\nu}\Big(\phi(x_j,h_j;\omega) + \varphi(y,h_j;\delta) \\
&\quad + \frac{1}{N}\vartheta(y,x_0;\varpi)\Big)\Bigg) \cdot \exp\Bigg(\sum_{(j,k)\in\epsilon}\psi(y,h_j,h_k;\xi)\Bigg) \\
&= \prod_{j\in\nu}\mu(x_j,h_j,y) \prod_{(j,k)\in\epsilon}\eta(h_j,h_k,y)
\end{aligned} \tag{12}$$

Then we compute the $\sum_{\mathbf{h}\in H^N}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))$ part in Eq. (4) with Eq. (12) and Eq. (9). To make the computation easier to understand, we follow the way in [11] to explain the computation by giving an example. Fig. 6 shows a graph with four joints which is part of the human body in Fig. 5. For this example, the joint set $\nu = \{1,2,3,4\}$, the edge set $\epsilon = \{(1,2),(2,3),(1,4)\}$, and the corresponding vector of hidden part states $\mathbf{h} = \{h_1,h_2,h_3,h_4\}$.

The $\sum_{\mathbf{h}\in H^N}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))$ is computed as follows:

$$\begin{aligned}
&\sum_{\mathbf{h}\in H^N}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))\\
&=\sum_{h_1\in H}\sum_{\mathbf{h}\setminus h_1}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))\\
&=\sum_{h_1\in H}\sum_{h_2\in H}\sum_{\mathbf{h}\setminus h_1,h_2}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))\\
&=\sum_{h_1\in H}\sum_{h_2\in H}\sum_{h_3\in H}\sum_{\mathbf{h}\setminus h_1,h_2,h_3}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))\\
&=\sum_{h_1\in H}\sum_{h_2\in H}\sum_{h_3\in H}\sum_{h_4\in H}\prod_{j\in\nu}\mu(x_j,h_j,y)\prod_{(j,k)\in\epsilon}\eta(h_j,h_k,y)\\
&=\sum_{h_1\in H}\sum_{h_2\in H}\sum_{h_3\in H}\sum_{h_4\in H}\mu(x_1,h_1,y)\mu(x_2,h_2,y)\mu(x_3,h_3,y)\\
&\quad\times\mu(x_4,h_4,y)\eta(h_1,h_2,y)\eta(h_2,h_3,y)\eta(h_1,h_4,y) \quad (13)
\end{aligned}$$

where $\mathbf{h}\setminus h_1$ indicates all hidden part states except $h_1$, $\mathbf{h}\setminus h_1,h_2$ indicates all hidden part states except $h_1,h_2$, $\mathbf{h}\setminus h_1,h_2,h_3$ indicates all hidden part states except $h_1,h_2,h_3$. For this example, only $h_4$ is left when $h_1,h_2$ and $h_3$ are excepted.

By reorganizing the sums in Eq. (13), we rewrite Eq. (13) as follows:

$$\begin{aligned}
&\sum_{\mathbf{h}\in H^N}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))\\
&=\sum_{h_1\in H}\sum_{h_2\in H}\sum_{h_3\in H}\mu(x_1,h_1,y)\mu(x_2,h_2,y)\mu(x_3,h_3,y)\\
&\quad\times\eta(h_1,h_2,y)\eta(h_2,h_3,y)\sum_{h_4\in H}\mu(x_4,h_4,y)\eta(h_1,h_4,y)\\
&=\sum_{h_1\in H}\sum_{h_2\in H}\sum_{h_3\in H}\mu(x_1,h_1,y)\mu(x_2,h_2,y)\mu(x_3,h_3,y)\\
&\quad\times\eta(h_1,h_2,y)\eta(h_2,h_3,y)m_{41}(h_1)\\
&=\sum_{h_1\in H}\sum_{h_2\in H}\mu(x_1,h_1,y)\mu(x_2,h_2,y)\eta(h_1,h_2,y)m_{41}(h_1)\\
&\quad\times\sum_{h_3\in H}\mu(x_3,h_3,y)\eta(h_2,h_3,y)\\
&=\sum_{h_1\in H}\sum_{h_2\in H}\mu(x_1,h_1,y)\mu(x_2,h_2,y)\eta(h_1,h_2,y)\\
&\quad\times m_{41}(h_1)m_{32}(h_2)\\
&=\sum_{h_1\in H}\mu(x_1,h_1,y)m_{41}(h_1)\sum_{h_2\in H}\mu(x_2,h_2,y)\eta(h_1,h_2,y)\\
&\times m_{32}(h_2)\\
&=\sum_{h_1\in H}\mu(x_1,h_1,y)m_{41}(h_1)m_{21}(h_1)\\
&=\sum_{h_1\in H}\mu(x_1,h_1,y)\prod_{b\in B(1)}m_{b1}(h_1) \quad (14)
\end{aligned}$$

where $B(1)$ is the neighbouring joint set of joint 1. For this example, the neighbouring joints of joint 1 are joint 2 and joint 4.

Based on the theory of message passing in [11], it's easy to compute the $\sum_{\mathbf{h}\in H^N}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))$ for the whole human body in Fig. 5 as follows:

$$\begin{aligned}
&\sum_{\mathbf{h}\in H^N}\exp(\Phi(y,\mathbf{h},\mathbf{x};\theta))\\
&=\sum_{h_i\in H}\sum_{\mathbf{h}\setminus h_i}\prod_{j\in\nu}\mu(x_j,h_j,y)\prod_{(j,k)\in\epsilon}\eta(h_j,h_k,y)\\
&=\sum_{h_i\in H}\mu(x_i,h_i,y)\prod_{b\in B(i)}m_{bi}(h_i) \quad (15)
\end{aligned}$$

where $\mathbf{h}\setminus h_i$ indicates all hidden part states except $h_i$. $B(i)$ is the neighbouring joint set of joint $i$. $x_i$ is the local feature of joint $i$ which is included in the feature $\mathbf{x}$.

Using Eq. (15), we can rewrite Eq. (4) as:

$$P(y|\mathbf{x};\theta)=\frac{\sum_{h_i\in H}\mu(x_i,h_i,y)\prod_{b\in B(i)}m_{bi}(h_i)}{\sum_{y'\in Y}\sum_{h_i\in H}\mu(x_i,h_i,y')\prod_{b\in B(i)}m_{bi}(h_i)} \quad (16)$$

Once we computed $P(y|\mathbf{x};\theta)$ with Eq. (16), the loss in Eq. (8) will be obtained. Then the back propagation is applied to the loss function to perform end-to-end training.

### *E. Three Stream Fusion*

Motivated by classic two-stream CNN for action recognition in [22], where one stream takes static feature (appearance) from still images and the other stream takes dynamic feature (optical flow) from consecutive frames, we use static features and dynamic features for skeletal action recognition. According to [21], [34], [37], both relative coordinate and bone direction information of each frame are powerful static features. Temporal displacements of joints in [21] have been proven to be effective dynamic features for action recognition. In this paper, all three streams are combined to improve the skeleton action recognition performance namely, relative coordinate stream, bone stream and temporal displacements stream.

**Relative Coordinates Stream:** Relative coordinates provide translation invariant features as explained in [21], which performs much better than using absolute coordinates of joints. In this work, the relative coordinate of the joints are adapted with respect to hip (left hip and right hip) and shoulder (left shoulder and right shoulder), each joint is represented as a 12D vector.

**Bone Stream:** It is known that bone is the physical connections of body joints. Each bone is represented as a 3D vector that represents the direction information of itself. The bone direction vector is calculated by the difference between the coordinates of two adjacent body joints. Given an acyclic graph with $N$ body joints, there exist $N-1$ bones. Each body joint is assigned with a unique bone direction vector except for the hip joint, which will be filled with 0.

**Temporal Displacements Stream:** Temporal displacements provide information about the amount of motion happening between adjacent frames. For each joint, it is represented as a 3D vector which is calculated by the difference of each joint between

two consecutive frames. Displacement information provides explicit motion information to the model as a strong feature in the learning process.

All the three streams are trained with the same GCN-HCRF model separately. For the final prediction, we take the average fusion strategy to combine the score from the three streams.

## IV. Experiment Results and Analysis

To show the effectiveness of the proposed model, we conduct experiments on three benchmark datasets, the NTU RGB+D dataset, the Northwestern-UCLA dataset and the SYSU dataset.

In the following sections, we first introduce the datasets and experiment settings in Section IV-A. Then we perform the ablation studies in Section IV-B to analyze the contributions of the proposed model components to the recognition performance. Section IV-C presents the performance comparisons with state-of-the-art approaches on the three datasets respectively.

### *A. Datasets and Experimental Settings*

**NTU RGB+D Dataset:** Currently, the NTU RGB+D is the largest and most widely used multimodal dataset for human action recognition. The dataset provides 3D skeleton data with 3D coordinates of 25 joints detected by the Kinect V2 depth sensors. There are 56,000 clips totally in 60 classes which are performed by 40 subjects, and each action is captured by 3 cameras from different viewpoints. This dataset recommends two benchmarks: 1) Cross-Subject (CS): the dataset is divided into training set (40,320 clips) and validation set (16,560 clips), where the actors in two subsets are different. 2).Cross-View (CV): the training set (37,920 clips) are captured by camera 2 and 3 from different viewpoints, and the validation set (18,960 clips) are captured by camera 1. It is a challenging dataset for action recognition because of the large amount of samples, various subjects, and the difference in camera views.

**Northwestern-UCLA Dataset (N-UCLA):** This dataset is a small-sized multimodal dataset for human action recognition, which contains 1494 sequences covering 10 action classes. It provides 3D skeleton data with 3D coordinates of 20 joints detected by the Kinect V1 depth sensors. Each action is captured one to six times by ten subjects with three cameras from different viewpoints. In this work, we follow the standard protocol proposed by [45] to report accuracy on three settings, which chose every two of the views for training and the other for testing.

**SYSU 3D Human-Object Interaction Set (SYSU):** This dataset is a small-sized multimodal dataset for Human-Object Interaction action recognition, which contains 480 sequences covering 12 action classes. It provides 3D skeleton data with 3D coordinates of 20 joints detected by the Kinect V1 depth sensors. 40 participants were asked to perform 12 different activities freely. For each action, each subject manipulates one of the six different objects: phone, wallet, bag, chair, mop and besom. These action samples have different durations, ranging from 58 frames to 638 frames. This dataset is challenging for high similarity among actions. In this work, we follow the standard protocol proposed by [14] to report the averaged accuracy from 30-fold cross validation on both Cross Subject (CS) setting and Same Subject (SS) setting.

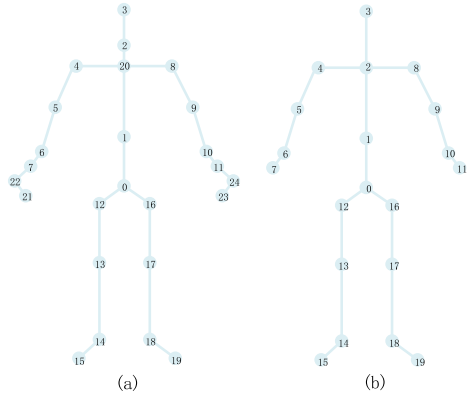



Fig. 7. (a) The skeleton of NTU RGB+D dataset with 25 joints; (b) The skeleton of N-UCLA dataset and SYSU dataset with 20 joints.

**Experimental Settings:** For NTU RGB+D dataset, the architecture of ST-GCN in [35] is employed with 9 layers of spatial temporal graph convolution operators. For each layer, the spatial temporal graph convolution operator is followed by a Dropout with a rate of 0.5, a Batch Normalization and a ReLU activation function. The number of output channels for each layer is 64, 64, 64, 128, 128, 128, 256, 256 and 256 respectively, The strides of the 4-th and the 7-th temporal convolution layers are set to 2 as pooling layer. Instead of using the SoftMax classifier, the output of the last layer is sent to HCRF for classification as shown in Fig. 3, and the size of hidden states set is set to 100. Then we conduct end to end training with Stochastic Gradient Descent (SGD). The global parameters such as Nesterov momentum, weight decay, initial learning rate are set to 0.9, 0.0001 and 0.1 respectively. For the N-UCLA dataset and the SYSU dataset, the only different setting in GCN part is the number of output channels for each layer. For HCRF part, the size of hidden states set is set to 20 for both N-UCLA dataset and SYSU dataset (further explanation regarding the size of the hidden state set can be found in Section IV-B3). Given that these datasets have a much smaller sample size than the NTU RGB+D dataset, we cut the number of output channels for each layer in half which are 32, 32, 32, 32, 64, 64, 64, 128, 128 and 128 respectively. Even after doing so, the sample sizes of N-UCLA and SYSU are too small to adequately train the network. CNN or RNN based methods [25] that consider the skeleton sequence as images can take advantage of pre-trained model on large-scale image datasets such as ImageNet when training on small-sized datasets. Motivated by that, in our work we first conduct pre-training on the halved network with NTU RGB+D dataset. Then we use the learned parameters as the initial parameters for N-UCLA dataset and SYSU dataset, with initial learning rate 0.01. However, there are 25 joints captured in the NTU RGB+D dataset while only 20 joints captured in the N-UCLA and SYSU Dataset as shown in Fig. 7, which lead to different graph structure. So the pre-training parameters from NTU RGB+D dataset cannot be used for training on N-UCLA dataset and SYSU dataset directly. We propose two adaptation strategies to solve the problem in Section IV-B5.

TABLE I
COMPARE THE ACCURACY(%) OF THE PROPOSED MODEL WITH THE BASELINE ST-GCN

| Methods | CS | CV |
|---|---|---|
| ST-GCN[35](baseline) | 81.5 | 88.3 |
| GCN-HCRF(ours) without global potential | 83.7 | 91.2 |
| GCN-HCRF(ours) with global potential | **84.3** | **91.7** |

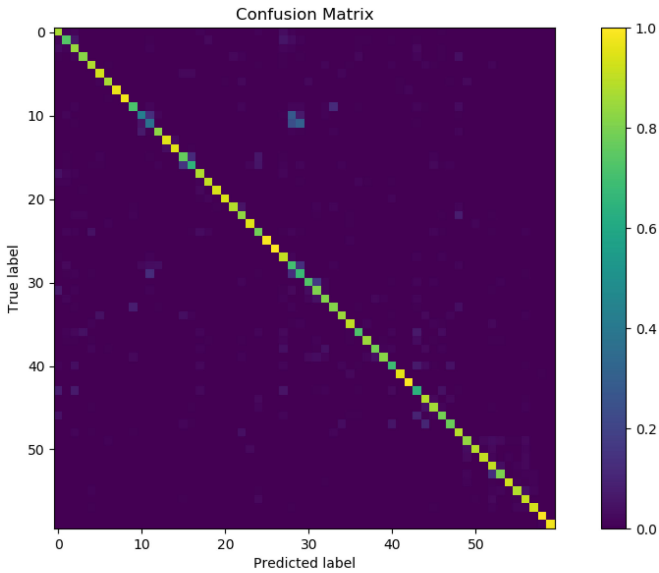



Fig. 8. The confusion matrix on the CS setting of NTU RGB+D dataset(overall accuracy is 84.3%).

## B. Ablation Study

*1) Effectiveness of The Proposed GCN-HCRF Model:* We first examine the effectiveness of the proposed GCN-HCRF model on the NTU RGB+D dataset. We compare the performance of the proposed GCN-HCRF model with ST-GCN model [35]. The results in Table I show that GCN-HCRF outperforms ST-GCN with same input (the absolute coordinate of joints) by 2.8% and 3.4% on the CS setting and the CV setting respectively. It also shows that the performance of the proposed model is improved by 0.6% and 0.5% on CS and CV settings respectively when taking the global potential into account. The confusion matrices of our results on the CS setting and the CV setting are shown in Fig. 8 and Fig. 9, respectively.

*2) Effectiveness of End-to-End Training:* To demonstrate the effectiveness of end-to-end training, we performed a separate training, whose performance is compared with that of end-to-end training in Table II. For the separate training manner, we first train the GCN separately, then use the output features of GCN to train the HCRF separately. As shown in Table II, compared with the separate training manner, the performance of end-to-end training is improved by 2.2% and 3.0% on CS and CV settings respectively. The comparison results demonstrated that the end-to-end training manner enables the HCRF part to guide GCN to extract more semantically meaningful feature from all joints, improving the overall classification performance.

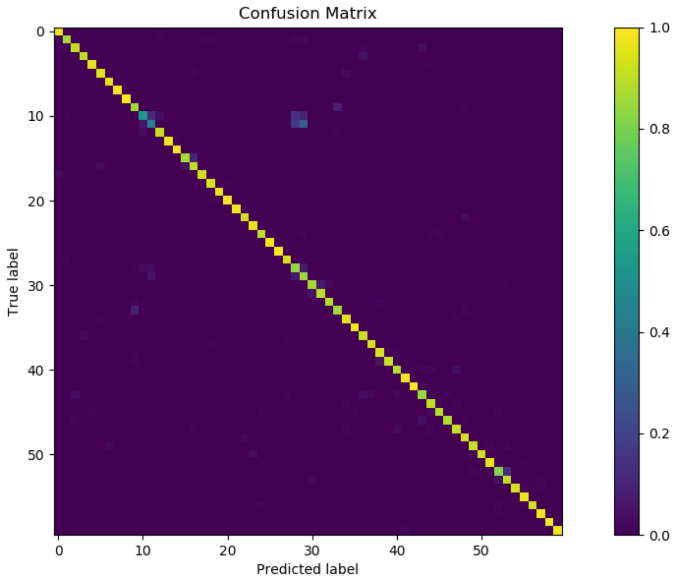



Fig. 9. The confusion matrix on the CV setting of NTU RGB+D dataset(overall accuracy is 91.7%).

TABLE II
ACCURACY(%) COMPARISON WITH DIFFERENT TRAINING MANNERS

| Training manner | CS | CV |
|---|---|---|
| GCN-HCRF separate training | 82.1 | 88.7 |
| GCN-HCRF end-to-end training | **84.3** | **91.7** |

TABLE III
COMPARE THE ACCURACY(%) WITH DIFFERENT SIZE OF HIDDEN STATES SET ON THE CS SETTING OF NTU RGB+D DATASET

| | $\|H\|$=90 | $\|H\|$=100 | $\|H\|$=110 | $\|H\|$=110 |
|---|---|---|---|---|
| GCN-HCRF | 82.9 | **84.3** | 83.1 | 82.7 |

*3) Influence of The Size of Hidden State Set:* In this subsection, we examine the influence of the size of hidden state set on the CS setting of NTU RGB+D dataset. According to [39], [43], the size of hidden states set $|H|$, is typically set to about twice the number of classes. For NTU RGB+D dataset, there are 60 classes. We learn the GCN-HCRF model four times on the CS setting of NTU RGB+D dataset with four different sizes of possible hidden states ($|H| = 90, 100, 110, 120$) given in Table III. In Table III, it shows that the proposed model achieves optimal accuracy when the sizes of possible hidden states is set to 100. So for the rest of the reported results, as described in experimental settings, we set the size of hidden states set to 100 for both CS and CV setting of NTU RGB+D dataset. As for N-UCLA and SYSU datasets, the size of hidden states set is set to 20 for both of them as they only have 10 and 12 action classes respectively.

TABLE IV
COMPARE THE ACCURACY(%) WITH DIFFERENT STREAM ON THE CS SETTING OF NTU RGB+D DATASET

| Streams | Accuracy(%) |
|---|---|
| GCN-HCRF(Rs) | 86.2 |
| GCN-HCRF(Bs) | 87.2 |
| GCN-HCRF(Ts) | 85.6 |
| GCN-HCRF(Rs+Bs) | 88.7 |
| GCN-HCRF(Rs+Ts) | 88.1 |
| GCN-HCRF(Bs+Ts) | 88.8 |
| GCN-HCRF(3s fusion) | 90.0 |

*4) Three-Stream Network:* Another important improvement is the introduction of three-stream framework. In this subsection, we examine the effectiveness of the proposed three-stream network on the CS setting of NTU RGB+D dataset. The performance of using the three streams, relative coordinate stream (Rs), bone stream (Bs), and temporal displacements stream (Ts) for the proposed model is shown in Table IV, confirming the importance of each stream in skeletal action recognition. Then we combine their scores pair-wisely and all three together, as described in Section III-E, to obtain fusion results, which are also presented in Table IV. The statistics clearly demonstrated the power of fusion, especially when all three streams are fused (3 s fusion), indicating the existence of complementary information among the three steams.

*5) Effectiveness of Different Adaptation Strategies on N-UCLA:* As shown in Fig. 7, the N-UCLA dataset and the SYSU dataset were captured by Kinect V1, which returns 20 body joints. However, there are 25 skeletal joints captured in the NTU RGB+D dataset by Kinect V2. So it is not straightforward to use the pre-trained parameters from NTU RGB+D dataset as the initial parameters for N-UCLA and SYSU dataset. To solve this issue, we conduct two different adaptation strategies. Since the N-UCLA dataset and the SYSU dataset both report 20 joints, we only use the N-UCLA dataset to explain the process. For the interpolation strategy, we interpolate 5 extra joint 20, 21, 22, 23, 24 to the skeleton of the N-UCLA dataset. The joint 20 is set to the middle of joint 4 and joint 8, joint 21 and 22 is set to the same position as joint 7, and joint 23, 24 is set to the same position as joint 11. After the interpolation, it has the same number of joints as NTU RGB+D dataset during training. Then the proposed model is trained on the N-UCLA dataset with the pre-trained parameters from NTU RGB+D dataset. For the trim strategy, we remove joint 20, 21, 22, 23, 24 from all samples in NTU RGB+D dataset. Then the proposed model is trained with the trimmed NTU RGB+D dataset, and the learned parameters are employed as the initial parameter for N-UCLA dataset. According to Table IV, since the B-stream is the strongest stream, performance of the two strategies on the B-stream is tabulated in Table V. In Table V, it shows the trim strategy is better than the interpolation strategy. For the rest of the reported results, we utilized the trim strategy for both N-UCLA and SYSU dataset.

TABLE V
COMPARE THE ACCURACY(%) OF TWO STRATEGIES WITH B-STREAM ON THE N-UCLA DATASET

| Test View | V3 | V2 | V1 | Average |
|---|---|---|---|---|
| Interpolation | 91.8 | 86.3 | 82.4 | 86.8 |
| Trim | **94.2** | **88.9** | **86.0** | **89.7** |

TABLE VI
COMPARISON WITH STATE-OF-THE-ART METHODS ON NTU RGB+D DATASET

| Methods | CS | CV | Year |
|---|---|---|---|
| Lie Group[33] | 50.1 | 82.8 | 2014 |
| CNN based methods: | | | |
| Temporal Conv.[38] | 74.3 | 83.1 | 2017 |
| ESV[25] | 80.0 | 87.2 | 2017 |
| Motion+Trans+CNN[3] | 83.2 | 89.3 | 2017 |
| HCN[5] | 86.5 | 91.1 | 2018 |
| SLnL-rFA [9] | 89.1 | 94.9 | 2018 |
| SGN[30] | 86.6 | 93.4 | 2019 |
| RNN/LSTM based methods: | | | |
| HBRNN[44] | 59.1 | 64.0 | 2015 |
| VA-LSTM[27] | 79.2 | 87.7 | 2017 |
| LSTM+FA+VF[49] | 73.8 | 85.9 | 2018 |
| FGF-LSTM[37] | 76.4 | 87.7 | 2018 |
| SR-TSL[6] | 84.8 | 92.4 | 2018 |
| AGC-LSTM[7] | 89.2 | 95.0 | 2019 |
| VA-fusion(aug.)[28] | 89.4 | 95.0 | 2019 |
| GCN based methods: | | | |
| ST-GCN[35] | 81.5 | 88.3 | 2018 |
| Generalized GCN[41] | 87.5 | 94.3 | 2018 |
| Part-based GCN[21] | 87.5 | 93.2 | 2018 |
| 2s-NLGCN[23] | 88.5 | 95.1 | 2018 |
| Ours: | | | |
| GCN-HCRF(Rs) | 86.2 | 92.1 | |
| GCN-HCRF(Bs) | 87.2 | 91.9 | |
| GCN-HCRF(Ts) | 85.6 | 92.7 | |
| GCN-HCRF(fusion) | **90.0** | **95.5** | |

### C. Comparisons to Other State-of-the-Art Approaches

In this section, we compare the proposed three streams GCN-HCRF model with state-of-the-art approaches on the NTU RGB+D, N-UCLA and SYSU, separately.

*1) NTU RGB+D Dataset:* We follow the standard CS and CV protocols introduced in [2] to evaluate the proposed method. Specifically, we compare the proposed model with other models for skeletal action recognition, including one hand-crafted feature based method [33], several CNN based methods [3], [5], [9], [25], [30], [38], several RNN/LSTM based methods [6], [7], [27], [28], [37], [44], [49], and several GCN based methods [21], [23], [35], [41]. The results in Table VI show that the proposed model achieves the best performance, outperforming the best GCN based method in [23] by 1.5% and 0.4% on CS and CV settings, respectively. In addition, the proposed model outperforms state-of-the-art approach [28] without using data-augmentation.

*2) N-UCLA Dataset:* There are three views in this dataset. According to the standard protocol proposed in [45], there are

TABLE VII
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE N-UCLA DATASET

| Setting(test view) | V3 | V2 | V1 | Avg | Year |
|---|---|---|---|---|---|
| HOJ3D[24] | 54.5 | - | - | - | 2012 |
| AE[15] | 76.0 | - | - | - | 2014 |
| Lie Group[33] | 74.2 | - | - | - | 2014 |
| VA-LSTM[27] | 70.7 | - | - | - | 2017 |
| HBRNN-L[45] | 78.5 | 83.5 | 79.3 | 80.5 | 2016 |
| ESV[25] | 92.6 | - | - | - | 2017 |
| E-TS-LSTM[10] | 89.2 | - | - | - | 2017 |
| E-GRU(aug.)[29] | 90.7 | - | - | - | 2018 |
| SGN[30] | 92.5 | - | - | - | 2019 |
| AGC-LSTM[7] | 93.3 | - | - | - | 2019 |
| VA-Fusion(aug.)[28] | 95.3 | 88.7 | 80.2 | 88.1 | 2019 |
| Ours: | | | | | |
| GCN-HCRF(Rs) | 92.9 | 86.9 | 81.6 | 87.1 | |
| GCN-HCRF(Bs) | 94.2 | 88.9 | 86.0 | 89.7 | |
| GCN-HCRF(Ts) | 93.5 | 85.5 | 79.6 | 86.2 | |
| GCN-HCRF(fusion) | **96.3** | **90.2** | **88.2** | **91.5** | |

TABLE VIII
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SYSU DATASET

| Setting | CS | SS | Year |
|---|---|---|---|
| LAFF[16] | 54.2 | - | 2016 |
| Dynamic Skeletons[30] | 75.5 | 76.9 | 2015 |
| ST-LSTM+ Trust Gate[17] | 76.5 | - | 2016 |
| VA-LSTM[27] | 76.9 | 77.5 | 2017 |
| DPRL+GCNN[46] | 76.9 | 83.5 | 2018 |
| Generalized GCN[41] | 77.9 | - | 2018 |
| SR-TSL[6] | 80.7 | 81.9 | 2018 |
| E-GRU(aug.)[29] | 85.7 | 85.7 | 2018 |
| VA-Fusion(aug.)[28] | 86.7 | 86.2 | 2019 |
| SGN[30] | 86.9 | 86.5 | 2019 |
| Ours: | | | |
| GCN-HCRF(Rs) | 90.2 | 89.3 | |
| GCN-HCRF(Bs) | 90.8 | 90.2 | |
| GCN-HCRF(Ts) | 91.1 | 90.7 | |
| GCN-HCRF(fusion) | **93.4** | **92.7** | |

three settings, each using two of the views for training and the other for testing. In [28], [45], experiments were conducted on all the three settings, while, in [7], [10], [15], [24], [25], [27], [29], [30], [33], only the first two views were used for training and the other for testing. We compare our results with all those works shown in Table VII, where V3 means that training on the first two views and testing on the third, V2 means that training on the first and the third views and testing on the second, and V1 means that training on the last two views and testing on the first. The result shows that the proposed model achieves the best performance. Notably, without data-augmentation, the proposed model significantly outperforms state-of-the-art methods in [28] by 8.0% and 3.2% on setting V1 and average, respectively.

*3) SYSU Dataset:* We follow the standard protocols proposed by [14] to evaluate the performance. For the Cross Subject (CS) setting, half of the subjects are used for training and the others for testing. For the Same Subject (SS) setting, half of the videos of each subject are used for training and the others for testing. According to the standard protocols proposed by [14], for each setting, the averaged results from 30-fold cross validation are required to report, which shown in Table VIII. The proposed model outperforms the other methods, particularly, the performance is 6.5% and 6.2% higher than the best reported in [30] for CS setting and SS setting, respectively.

## V. CONCLUSION

In this work, we propose a GCN-HCRF model for skeleton-based action recognition, which can retain the spatial structure of human joints from beginning to end. It takes full advantage of the spatial compatibility information among all joints' motion. Then we carried out end-to-end training on the whole model so that the GCN part of the proposed model extracts semantically meaningful features by the guidance of the HCRF part. Furthermore, we proposed a three-stream framework to further boost the performance, which employs the relative coordinate of the joints and bone direction as two static feature streams and the temporal displacements between the joints in two consecutive frames are adopted as the dynamic feature stream. To train the proposed model on two small-sized datasets N-UCLA and SYSU with the pre-trained parameters from NTU RGB+D, we proposed two adaptation strategies to solve the difference of graph structure, the interpolation strategy and the trim strategy. The proposed model is evaluated on three challenging standard action recognition dataset, achieving state-of-the-art results for all of them.

## REFERENCES

[1] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, vol. 29, no. 10, pp. 1848–1852, Oct. 2007.
[2] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1010–1019.
[3] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, 2017, pp. 597–600.
[4] C. Liu and J. Liu, "Convolutional neural random fields for action recognition," *Pattern Recognit.*, vol. 59, pp. 213–224, 2016.
[5] C. Li and Q. Zhong, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1–8.
[6] C. Si and Y. Jing, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Eur. Conf. Comput. Vision*, 2018, pp. 103–118.
[7] C. Si and W. Chen, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1–10.
[8] G. A. Sigurdsson and S. Divvala, "Asynchronous temporal fields for action recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1–20.
[9] G. Hu and B. Cui, "Skeleton-based action recognition with synchronous local and non-local spatio-temporal learning and frequency attention," in *Proc. Stanford Inst. Comput. Math. Eng.*, 2019, pp. 1–6.
[10] I. Lee and D. Kim, " Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 1012–1020.

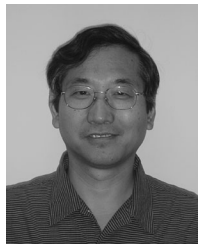
[11] J. S. Yedidia, W. T. Freeman, and Y. Weiss, *Exploring Artificial Intelligence in the new Millennium*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 239–269.
[12] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 282–289.
[13] J. Lei and G. Li, "Continuous action recognition based on hybrid CNN-LDCRF model," in *Proc. 4th IEEE Int. Conf. Image, Vision Comput.*, 2016, pp. 63–69.
[14] J.-F. Hu and W.-S. Zheng, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, pp. 2186–2200.
[15] J. Wang, Z. Liu, and Y. Wu, "Learning actionlet ensemble for 3D human action recognition," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, pp. 914–927.
[16] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, and J. Lai, "Real-time RGB-D activity prediction by soft regression," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 280–296.
[17] J. Liu, A. Shahroudy, D. Xu, and G. Wang, " Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 816–833.
[18] J. Chang, "Nonparametric feature matching based conditional random fields for gesture recognition from multi-modal video," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, pp. 1612–1625.
[19] K. Philipp and K. Vladlen, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Conf. Neural Inf. Process. Syst.*, 2012, pp. 109–117.
[20] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo, "3D human activity recognition with reconfigurable convolutional neural networks," in *Proc. ACM Multimedia*, 2014, pp. 97–106.
[21] K. Thakkar and P. J. Narayanan, "Part-based graph convolutional network for action recognition," in *Proc. Brit. Mach. Vision Conf.*, 2018, pp. 1–19.
[22] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[23] L. Shi and Y. Zhang, "Non-local graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vision Patten Recognit.*, 2019, pp. 12027–12035.
[24] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2012, pp. 20–27.
[25] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," in *Proc. Pattern Recognit.*, 2017, pp. 346–362.
[26] M. Li and H. Leung, "Multiview skeletal interaction recognition using active joint interaction graph," in *Proc. IEEE Trans. Multimedia*, 2016, pp. 2293–2302.
[27] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2117–2126.
[28] P.-F. Zhang and C. Lan, "View adaptive neural networks for high performance skeleton-based human action recognition," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, 2019, pp. 1–15.
[29] P.-F. Zhang, J. Xue, and C. Lan, "Adding attentiveness to the neurons in recurrent neural networks," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 135–151.
[30] P.-F. Zhang *et al.*, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE Conf. Comput. Vision Patten Recognit.*, 2020, pp. 1–10.
[31] P. Wei, H. Sun, and N. Zheng, "Learning composite latent structures for 3D human action representation and recognition," in *Proc. IEEE Trans. Multimedia*, 2019, pp. 1–1.
[32] R. Michails, I. Kokkinos, and S. Soatto, "Discovering discriminative action parts from mid-level video representation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 1242–1249.
[33] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 588–595.
[34] S. Wang and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection," in *Proc. IEEE Trans. Image Process.*, 2018, pp. 4382–4394.
[35] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. Assoc. Advancement Artif. Intell.*, 2018, pp. 7444–7452.
[36] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. Assoc. Advancement Artif. Intell.*, 2017, pp. 4263–4270.
[37] S. Zhang, Y. Yang, and J. Xiao, "Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks," in *Proc. IEEE Trans. Multimedia*, 2018, pp. 2330–2343.
[38] T. S. Kim and A. Reiter, "Interpretable 3 d human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2017, pp. 1623–1631.
[39] W. Sybor, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2006, pp. 1521–1527.
[40] X. Liang, L. Lin, and L. Cao, "Learning latent spatio-temporal compositional model for human action recognition," in *Proc. ACM Multimedia*, 2013, pp. 263–272.
[41] X. Gao, W. Hu, and J. Tang, "Generalized graph convolutional networks for skeleton-based action recognition," in *Proc. ACM Int. Conf. Mutimedia*, 2019, pp. 1–9.
[42] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-D convnet fusion for action recognition in videos with arbitrary size and length," in *Proc. IEEE Trans. Multimedia*, 2018, pp. 634–644.
[43] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic vs. max-margin," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, pp. 1310–1323.
[44] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1110–1118.
[45] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," in *Proc. EEE Trans. Image Process.*, 2016, pp. 3010–3022.
[46] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 5323–5332.
[47] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, X. Gao, "Discriminative multi-instance multi-task learning for 3D action recognition," in *Proc. IEEE Trans. Multimedia*, 2017, pp. 519–529.
[48] Z. Shuai *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. Int. Conf. Comput. Vision*, 2015, pp. 1529–1537.
[49] Z. Fan, X. Zhao, T. Lin, and H. Su, "Attention based multiview re-observation fusion network for skeletal action recognition," in *Proc. IEEE Trans. Multimedia*, 2018, pp. 363–374.

**Kai Liu** (Student Member, IEEE) received the M.S. degree in technology of computer application from Zhengzhou University, Zhengzhou, China, in 2012, where he is currently working toward the Ph.D. degree in communication and information systems. His research interests include computer vision and machine learning.

**Lei Gao** (Member, IEEE) received the M.S degree in communication and information systems from Zhengzhou University, Zhengzhou, China, in 2011, and the Ph.D. degree in electrical and computer engineering from Ryerson University, Toronto, ON, Canada, in 2017. He is currently with the Department of Electrical and Computer Engineering, Ryerson University. His research interests include multimedia signal processing, pattern recognition, machine learning, image processing, and information fusion. He is the recipient of a Visiting Fellowship from Microsoft Research Asia in 2016, the Top 10% Papers Award of 2015 IEEE International Conference on Image Processing, The National Scholarship of China, Ryerson International Student Scholarship, and The Excellent Graduate Research Award of Ryerson University.

**Naimul Mefraz Khan** (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Ryerson University, Toronto, ON, Canada. He is currently an Assistant Professor of Electrical and Computer Engineering with Ryerson University, where he co-directs the Ryerson Multimedia Research Laboratory. His research focuses on creating user-centric intelligent systems through the combination of novel machine learning and human-computer interaction mechanisms. He is the recipient of the best paper award at the IEEE International Symposium on Multimedia, the OCE TalentEdge Postdoctoral Fellowship, and the Ontario Graduate Scholarship. He is a member of the ACM.

**Lin Qi** received the B.Sc. degree in radio engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, the M.A.Sc. degree in computer science from the Zhengzhou University, Zhengzhou, China, and the Ph.D. degree in information and communication engineering from Beijing Institute of Technology, Beijing, China. He is currently a Professor with the School of Information Engineering, Zhengzhou University. His current research interests include image or video analysis and processing, pattern recognition, and signal detection and estimation.



**Ling Guan** (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of British Columbia, Vancouver, BC, Canada, in 1989. He is currently a Professor with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada. In 2001, he was appointed as a Tier I Canada Research Chair in Multimedia and Computer Technology. He also held visiting positions with British Telecom (1994), Tokyo Institute of Technology (1999), Princeton University (2000), National ICT Australia (2007), Hong Kong Polytechnic University (2008–2009), and Microsoft Research Asia (2002, 2009, and 2017). He has authored extensively in multimedia processing and communications, human-centered computing, machine learning, adaptive image and signal processing, and more recently, multimedia computing in the immersive environment. He is an IEEE Circuits and System Society Distinguished Lecturer, an Elected Member of the Canadian Academy of Engineering, and the recipient of 2014 IEEE Canada C.C. Gotlieb Computer Medal and the 2005 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS Best Paper Award.