2019 IEEE International Symposium on Multimedia (ISM)

Graph Convolutional Networks-Hidden Conditional Random Field Model for Skeleton-Based Action Recognition

Kai Liu[1], Lei Gao[2], Naimul Mefraz Khan[2], Lin Qi[1], Ling Guan[2]
[1]*School of Information Engineering, Zhengzhou University, Zhengzhou, China*
[2]*Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada*
Email: 09liukai08@gmail.com, iegaolei@gmail.com, n77khan@ee.ryerson.ca, ielqi@zzu.edu.cn, lguan@ee.ryerson.ca

***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 and HCRF to retain the human skeleton structure information during the classification stage. The proposed model is trained end-to-end by utilizing message passing from 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 coordinates of the joints and bone direction as two static feature streams, and the temporal displacements as the dynamic feature stream. Experimental results on two challenging benchmarks (NTU RGB+D, N-UCLA) show the superior performance of the proposed model over state-of-the-art models.**

***Keywords*-GCN; CRF; Skeleton; Hidden Part State; Action Recognition;**

I. Introduction

Human action recognition is an important topic in multimedia computing with many applications such as human-computer interaction, virtual reality and video understanding [16]. Recently, skeleton-based human action recognition has attracted considerable attention, since the skeleton data is succinct in representation and robust to variations of viewpoints and environment. In this paper, we focus on the problem of skeleton-based action recognition.

Convolutional deep neural network based methods like CNN or RNN have difficulty in fully expressing the dependency among joints in the human skeleton. To solve this issue, graph convolutional networks, which generalize convolution from image to graph, have been successfully adopted in skeleton-based action recognition recently [13, 19]. GCN can preserve the connectivity among the skeleton joints, thus preserving the spatial structure. While the GCN-based methods provided a significant increase in performance, 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 architecture 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.

Inspired by hidden part models for action recognition in [22], we propose a novel GCN-HCRF model to retain the human skeleton structure information during classification. An overview of the proposed method is shown in Fig. 1 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. 1, 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. This model is trained end-to-end so that the HCRF model can guide the GCN to extract features that are semantically more meaningful.

Another notable problem is that although static features within each frame like joint's coordinate and bone direction have been proven to be effective modalities for skeleton-based action recognition in [13, 19]. Dynamic feature from consecutive frames like temporal displacements also contain rich information for skeletal action recognition, since it provides explicit motion information. In this paper, we take the relative coordinate of the joints, bone direction and the temporal displacements as three streams. All three streams are fused with score level fusion, providing improved performance over utilizing a single stream.

To verify the superiority of the proposed model, we conduct experiments on two challenging datasets with different scale: NTU RGB+D (56,000 samples), N-UCLA (1,494 samples). Experimental results show that the proposed method outperforms state-of-the-art methods on both datasets.

In this paper, our main contributions are summarized below:

1) We propose a novel GCN-HCRF model which retains

978-1-7281-5606-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ISM46123.2019.00013

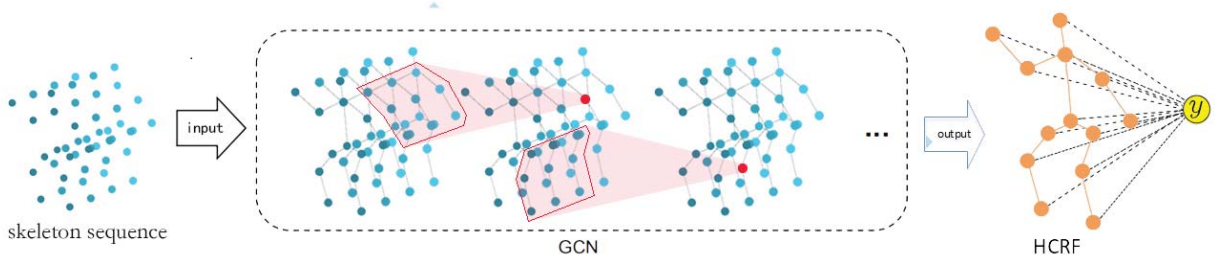

Figure 1. The framework of the proposed GCN-HCRF model: We utilize a GCN network to extract motion feature of each body joint. Then the extracted motion feature is sent to HCRF for classification. The whole model is trained in an end-to-end manner with back propagation.

the spatial structure of human joints from beginning to end. 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 motion feature that is semantically more meaningful, improving the overall classification performance.

2) A three-stream framework is proposed to further improve the performance. it takes the relative coordinate of joints and bone direction from each frame as two static feature streams, and the temporal displacements between the joints in two consecutive frames as the dynamic feature stream. Then the score level fusion is adopted for final prediction.

3) Experimental results demonstrate the proposed model outperforms state-of-the-art approaches on two standard datasets.

II. Related Work

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your system, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

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

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 [3], 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 [15], the authors design two view adaptive neural networks, VA-RNN and VA-CNN, which reposition the observation viewpoints adaptively to facilitate better action recognition from skeleton data. In [19], spatio-temporal GCN is proposed to apply graph convolution on the graph of the skeleton action sequences. In [13], adaptive graph convolutional block is introduced to learn the graph structure for different layers and samples.

B. CRF Based Methods

Sequence labeling models, like CRF [8], 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 [22], a hidden part-based CRF approach is proposed to combine both large-scale global features and local patch features, and then actions are classified with max-margin hidden conditional random fields (MMHCRF). In [18], by observing that each class has distinct sub-structures, a HCRF model is introduced to capture the sub-structures for gesture recognition.

C. Combination of Neural Networks and CRF

Joint model of CNNs and CRF is helpful in many applications. The fully connected CRF [10] performs well in semantic segmentation and human pose estimation, and it was trained jointly in [24] by unrolling iterations of the inference method. In [20], a deep sequence model is introduced by extending the CRF models with CNN for higher level feature learning. In [9], a hybrid framework combines Convolutional Neural Network (CNN) and Latent Dynamic Conditional Random Field (LDCRF) to segment and recognize continuous actions simultaneously. In [6], 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 on which a novel GCN-HCRF model is proposed. The proposed model can take full advantage of compatible information among all joints' motions.

A. Overview of The Proposed Framework

As shown in Fig. 1, 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 extract 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 motion 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

Suppose we have a skeletal sequence including T frames and each frame has N joints. we employ the spatio-temporal graph convolutional network to extract motion feature of each joint. This GCN network is an undirected spatio-temporal graph $G = (V, E)$, 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. To exploit the motion feature with spatio-temporal information, the spatial temporal graph convolution is implemented with Eq. 1 as follows [19]:

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

where $\mathbf{F}_{in}$ represents 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. $\mathbf{F}_{out}$ represents 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 α is set to 0.001 to avoid empty rows in $\mathbf{A}_j$. For the temporal dimension, we conduct the graph convolution with the classical convolution operation on regular images, since the number of neighbors for each joint on temporal dimension is fixed.

Now, we have a well-defined spatial and temporal convolution operation on the whole skeleton sequence, by which we extract motion feature with spatio-temporal information on every human joint. Then we use it as the feature of the given action sequence. Specifically, as shown in Fig. 2, 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.

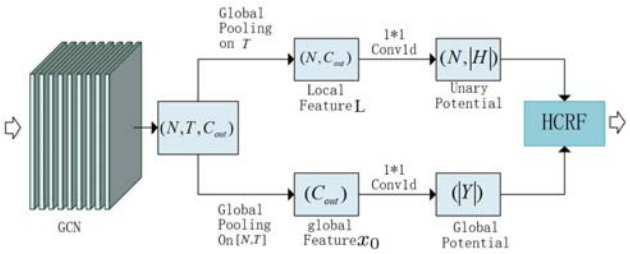

Figure 2. Illustration of model implementing: the output from the last layer of GCN is a matrix of (N, T, C_{out}) dimensions. After conducting the global average pooling on T and on $[N, T]$, we obtain a local feature matrix $\mathbf{L}$ and a global feature vector x_0. Then they are used as the unary potential $\phi(\cdot)$ and the global potential $\vartheta(\cdot)$ of HCRF after conducting convolution operation. More details are described in section 3.3.

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. Based on the aforementioned description, 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 motion feature of the given skeletal sequence, which includes the local feature $\mathbf{L}$ and the global feature x_0, then the motion feature $\mathbf{x}$ is represented as $\mathbf{x} = (x_0, x_1, x_2, ..., x_i, ..., x_N)^{\mathrm{T}}$ where x_0 is the global feature vector of the whole body's motion, and $x_i (i = 1...N)$ 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 motion feature $\mathbf{x}$. In order to model the compatibility among all joints' motion, we introduce a vector of hidden part states $\mathbf{h} = (h_1, h_2, ..., h_i, ..., h_N)^{\mathrm{T}}$, where h_i is a hidden part state assigned to x_i, for $i = 1, 2, ...N$. 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.

According to the theory of random fields in [1], given the motion 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:

$$P(y, \mathbf{h}|\mathbf{x}; \theta) = \frac{\exp(\Phi(y, \mathbf{h}, \mathbf{x}; \theta))}{\sum_{y' \in Y} \sum_{\mathbf{h} \in H^N} \exp(\Phi(y', \mathbf{h}, \mathbf{x}; \theta))} \quad (2)$$

where θ is the model parameter, and $\Phi(y, \mathbf{h}, \mathbf{x}; \theta)$ refers to potential function depending on the motion 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 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 motion feature $\mathbf{x}$ is the summation of Eq. 2 over all possible assignments of hidden part states $\mathbf{h}$:

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

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

$$\begin{aligned} \Phi(y, \mathbf{h}, \mathbf{x}; \theta) = &\sum_{j \in \nu} \phi(x_j, h_j; \omega) + \sum_{j \in \nu} \varphi(y, h_j; \delta) \\ &+ \sum_{(j,k) \in \epsilon} \psi(y, h_j, h_k; \xi) + \vartheta(y, x_0; \varpi) \end{aligned} \quad (4)$$

The details of these potential functions in Eq. 4 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 ω 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. 2. Each element $\mathbf{M}_{jl}$ of the probability matrix $\mathbf{M}$ represents the probability that $x_j (j = 1...N)$ is assigned as the l-th $(l = 1...|H|)$ state of the hidden part state set H, since h_j can take any value from the hidden part state set H. The parameter ω 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:

$$\varphi(y, h_j; \delta) = \sum_{a \in Y} \sum_{b \in H} \delta_{a,b} \cdot \mathbf{1}(y = a) \cdot \mathbf{1}(h_j = b) \quad (5)$$

where the parameter δ 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. To compute this potential, we parametrize it as:

$$\begin{aligned} &\psi(y, h_j, h_k; \xi) \\ &= \sum_{a \in Y} \sum_{b \in H} \sum_{c \in H} \xi_{a,b,c} \cdot \mathbf{1}(y = a) \cdot \mathbf{1}(h_j = b) \cdot \mathbf{1}(h_k = c) \end{aligned} \quad (6)$$

where the parameter ξ 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.

Global potential $\vartheta(y, x_0; \varpi)$ measures the compatibility of class label y and the global feature vector x_0 of the whole action. As shown in Fig. 2, 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 represents the probability that x_0 is assigned as the i-th class. The parameter ϖ 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.

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 training dataset includes S labeled action sequences. Following the minimum negative conditional log-likelihood rule, the loss function is shown in Eq. 7:

$$L(\theta) = -\sum_{s=1}^{\mathrm{S}} \log P(y^{(s)}|\mathbf{x}^{(s)}; \theta) \quad (7)$$

where $y^{(s)}$ is the label of the s-th action sequence sample, $\mathbf{x}^{(s)}$ is the motion 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. 3. 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. 3, which is intractable by brute force as all the possible assignments of hidden part state are exponential. 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 [8], 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. 3.

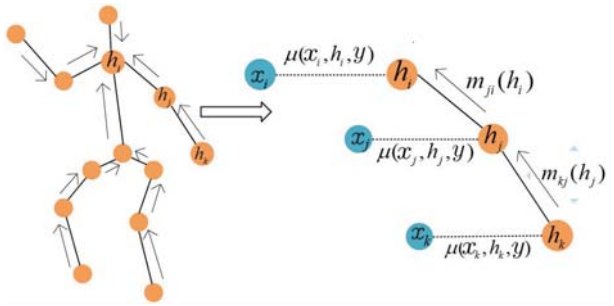

Figure 3. Message passing route and passing rule.

In Fig. 3, $m_{ji}(h_i)$ 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:

$$m_{ji}(h_i) \leftarrow \sum_{h_j \in H} \mu(x_j, h_j, y)\eta(h_i, h_j, y) \prod_{b \in B(j) \setminus i} m_{bj}(h_j) \tag{8}$$

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. 9. 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 .

$$\mu(x_j, h_j, y) = \exp(\phi(x_j, h_j; \omega) + \varphi(y, h_j; \delta) + \frac{1}{N}\vartheta(y, x_0; \varpi)) \tag{9}$$

$\eta(h_i, h_j, y)$ is the pair potential between joint i and joint j, which is formulated in Eq. 10:

$$\eta(h_i, h_j, y) = \exp(\psi(y, h_i, h_j; \xi)) \tag{10}$$

Using Eq. 4, Eq. 9 and Eq. 10, we compute the $\exp(\Phi(y, \mathbf{h}, \mathbf{x}; \theta))$ part in Eq. 3 as follows:

$$\exp(\Phi(y, \mathbf{h}, \mathbf{x}; \theta)) = \prod_{j \in \nu} \mu(x_j, h_j, y) \prod_{(j,k) \in \epsilon} \eta(h_j, h_k, y) \tag{11}$$

Based on the theory of message passing in [8], it is 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. 3 as follows:

$$\sum_{\mathbf{h} \in H^N} \exp(\Phi(y, \mathbf{h}, \mathbf{x}; \theta)) = \sum_{h_i \in H} \mu(x_i, h_i, y) \prod_{b \in B(i)} m_{bi}(h_i) \tag{12}$$

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 motion feature $\mathbf{x}$.

Using Eq. 12, we can rewrite Eq. 3 as:

$$P(y|\mathbf{x}; \theta) = \frac{\sum_{h_i \in H} \mu(x_i, h_i, y) \prod_{b \in B(i)} m_{bi}(h_i)}{\sum_{y' \in Y} \sum_{h_i \in H} \mu(x_i, h_i, y') \prod_{b \in B(i)} m_{bi}(h_i)} \tag{13}$$

Once we computed $P(y|\mathbf{x}; \theta)$ with Eq. 13, the loss in Eq. 7 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 [12], 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. Both relative coordinate [11] and bone direction information [13] of each frame are powerful static features. Temporal displacements of joints in [11] have been proven to be effective dynamic features for action recognition. In this paper, all three streams, namely, relative coordinate stream(Rs), bone stream(Bs) and temporal displacements stream(Ts), are fused with score level fusion.

IV. Experiment and Performance Evaluation

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

A. Experimental Settings

For NTU RGB+D dataset, the architecture of ST-GCN in [19] is employed with 9 layers of spatial temporal graph convolution operators. Instead of using the SoftMax classifier, the output of the last layer is sent to HCRF for classification as shown in Fig. 2, 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, the only different setting in GCN part is the number of output channels for each layer. Since this dataset has a much smaller sample size than the NTU RGB+D dataset, we cut the number of output channels for each layer in half which is 32, 32, 32, 32, 64, 64, 64, 128, 128 and 128 respectively. For HCRF part, the size of hidden states set is set to 20. In addition, it is worth noting that there are 25 joints captured in the NTU RGB+D dataset while only 20 joints captured in the N-UCLA, which lead to different graph structure. To solve the problem, we trim SpineShoulder, HandTipRight, ThumbRight, HandTipLeft and ThumbLeft joints from all samples in NTU RGB+D dataset. Then we conduct pre-training on the trimmed NTU RGB+D dataset,

and use the pre-training parameters as initial parameters to train on N-UCLA dataset, with initial learning rate 0.01.

B. NTU RGB+D Dataset and Results

The NTU RGB+D is a large scale 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. We follow the standard CS and CV protocols introduced in [2] to evaluate the proposed method. 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 [19]. The results in Table 1 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.

Methods	CS	CV
ST-GCN[19](baseline)	81.5	88.3
GCN-HCRF(ours)	**84.3**	**91.7**

Table 1: Compare the accuracy(%) of the proposed model with the baseline ST-GCN.

Then we compare the proposed model with other models for skeletal action recognition. The results in Table 2 on the next page show that the proposed three-stream model (3s fusion) achieves the best performance, outperforming state-of-the-art approach [16] without using data-augmentation.

C. N-UCLA Dataset and Results

N-UCLA 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 [23] to report accuracy on three settings, which choose every two of the views for training and the other for testing. We compare our results with other works shown in Table 3. The result shows that the proposed three-stream model (3s fusion) achieves the best performance. Notably, without data-augmentation, the proposed model significantly outperforms state-of-the-art methods in [16] by 8.0% and 3.2% on setting V1 and average, 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 an 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 improve 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. The proposed model is evaluated on two challenging standard action recognition datasets, achieving state-of-the-art results.

Methods	CS	CV	Year
ESV [14]	80.0	87.2	2017
VA-LSTM [15]	79.2	87.7	2017
HCN [3]	86.5	91.1	2018
SR-TSL [4]	84.8	92.4	2018
ST-GCN [19]	81.5	88.3	2018
Genarallized GCN [21]	87.5	94.3	2018
Part-based GCN [11]	87.5	93.2	2018
SGN [18]	86.6	93.4	2019
2s-AGCN [13]	88.5	95.1	2019
AGC-LSTM [5]	89.2	95.0	2019
VA-fusion(aug.) [16]	89.4	95.0	2019
ours:			
GCN-HCRF(Rs)	86.2	92.1	
GCN-HCRF(Bs)	87.2	91.9	
GCN-HCRF(Ts)	85.6	92.7	
GCN-HCRF(3s fusion)	**90.0**	**95.5**	

Table 2: Comparison with state-of-the-art methods on NTU RGB+D.

Setting(test view)	V3	V2	V1	Avg	Year
VA-LSTM[15]	70.7	-	-	-	2017
HBRNN-L[23]	78.5	83.5	79.3	80.5	2016
ESV[14]	92.6	-	-	-	2017
E-TS-LSTM[7]	89.2	-	-	-	2017
E-GRU(aug.)[17]	90.7	-	-	-	2018
SGN[18]	92.5	-	-	-	2019
AGC-LSTM[5]	93.3	-	-	-	2019
VA-Fusion(aug.)[16]	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(3s fusion)	**96.3**	**90.2**	**88.2**	**91.5**	

Table 3: Comparison with state-of-the-art methods on the N-UCLA dataset.

REFERENCES

[1] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," IEEE PAMI, pages 1848–1852, 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 CVPR, pages 1010-1019, 2016.

[3] C. Li, Q. Zhong, "Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation," In IJCAI, pages 1-8, 2018

[4] C. Si, Y. Jing, "Skeleton-Based Action Recognition with Spatial Reasoning and Temporal Stack Learning," In ECCV, pages 103-118, 2018.

[5] C. Si, W. Chen, "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition," In CVPR, pages 1-10, 2019.

[6] G. A. Sigurdsson, S. Divvala, "Asynchronous Temporal Fields for Action Recognition," In CVPR, pages 1-20, 2017.

[7] I. Lee, D. Kim, "Ensemble Deep Learning for Skeleton-based Action Recognition using Temporal Sliding LSTM networks," In ICCV, pages 1012–1020, 2017.

[8] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Exploring artificial intelligence in the new millennium," pages 239-269, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[9] J. Lei, G. Li, "Continuous Action Recognition Based on Hybrid CNN-LDCRF Model," In ICIVC, pages 63-69, 2016.

[10] K. Philipp, and K. Vladlen, "Efficient inference in fully connected crfs with gaussian edge potentials," In NIPS, pages 109-117, 2012.

[11] K. Thakkar, P J Narayanan, "Part-based Graph Convolutional Network for Action Recognition," In BMVC, pages 1-19, 2018.

[12] K. Simonyan, A. Zisserman, "Two-Stream convolutional Networks for Action Recognition in Videos," In NIPS. Pages 568-576, 2014.

[13] L. Shi, Y. Zhang, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," In CVPR, 2019.

[14] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," In PR, pages 346-362, 2017.

[15] 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 CVPR, pages 2117-2126, 2017.

[16] P.-F. Zhang, C. Lan, "View Adaptive Neural Networks for High Performance Skeleton-based Human Action Recognition," In IEEE PAMI, pages 1-15, 2019.

[17] P.-F. Zhang,. J. Xue, C. Lan, "Adding Attentiveness to the Neurons in Recurrent Neural Networks," In ECCV, pages 135-151, 2018.

[18] P.-F. Zhang, C. Lan, "Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition," CoRR, abs/1904.01189, 2019.

[19] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for skeleton-Based Action Recognition," In AAAI, pages 7444-7452, 2018.

[20] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," In CVPRW, pages 1623-1631, 2017.

[21] X. Gao, W. Hu, J. Tang, "Generalized Graph Convolutional Networks for Skeleton-based Action Recognition," CoRR, abs/1811.12013, 2018

[22] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic vs. max-margin," In IEEE PAMI, pages 1310-1323, 2010.

[23] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," In TIP, pages 3010-3022, 2016.

[24] Z. Shuai, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P.H. S. Torr, "Conditional random fields as recurrent neural networks," In ICCV, pages 1529-1537, 2015.