Dual domain graph convolutional networks for skeleton based action recognition

B.K.N.Priyanka, Assistant Professor, Department of Data Science, SICET, Hyderabad
Suresh Ballala Assistant Professor, Department of Data Science, SICET, Hyderabad
M.Kavitha, Assistant Professor, Department of Data Science, SICET, Hyderabad

Abstract

Skeleton-based action recognition is attracting more and more attention owing to the gen-eral representation ability of skeleton data. The Graph Convolutional Networks (GCNs)methods extended from Convolutional Neural Networks (CNNs) are proposed to directlyextractspatial–temporalinformationfromthegraphs.PreviousGCNsusuallyaggregatethe skeleton information locally in the vertex domain. However, the focus on the localinformation brought about the limited representation ability in some actions containingoverall dynamics in both spatial and temporal, which pulled down the overall accuracy ofthe model. Therefore, this paper proposes a more comprehensive two-stream GCN archi-tecture containing the vertex-domain graph convolution and the spectral graph convolutionbased on Graph Fourier Transform (GFT). One stream utilizes an efficient vertex-domaingraph convolution to obtain effective spatial–temporal information via Graph Shift Blocks(GSB), while the other brings the global spectral information from our improved ResidualSpectral Blocks (RSB). According to the analysis of the experimental results, the actionmisalignmentforcertainactionsisreduced.Moreover,alongwithotherGCNmethodsthatonly focus on spatial–temporal information, our RSB strategies help improve their per-formance. DD-GCN is evaluated on three large skeleton-based datasets, NTU-RGBD 60,NTU-RGBD 120, and Kinetics-Skeleton. The experiment results demonstrate a compara-ble ability to the state-of-the-art.

KeywordsActionrecognition· Skeleton· Graphconvolutionalnetworks· Dual-domain· Spatial–temporal· Spectral

1 Introduction

Action recognition is a challenging task in the field of computer vision. And it is at theforefront of applications to understand the human social activity (Islam and Iqbal2020).ActionrecognitionbasedonRGBimages/videoshasbeenwidelyresearchedwithdeep learning methods, such as Convolution Neural Networks (CNNs). The motivation of mostaction recognition algorithms is to extract spatiotemporal feature representations fromRGBvideos.Andthen,aclassifieristrainedtodistinguishdifferentactions.SimonyanandZisserman (2014) proposed a two-stream method to extract spatial and temporal informa-tion separately. Also, to obtain temporal features, Ji et al. (2013) extended the traditional2D-CNNto 3D-CNN with a 3D convolutionkernel.Meanwhile, owing to the concise and compelling data source, skeleton-based actionrecognitionisattractingmoreandmoreattention.Concretely,skeleton-basedmethodscan effectively focus on the joint transformation of different actions by discarding redun-dant background information. A more robust and more efficient network based on skel-eton data can be designed to recognize human actions than the RGB-based methods. Andthe most important thing is that skeletal data can articulate joints connection status andtheirdynamicchanges.

Previous work construct the joint coordinates manually into a sequence of vectors(Vemulapallietal.2014;Jiangetal.2020).Thentherecurrentneuralnetwork(RNNs)is utilized to process the vectors (Liu et al. 2016; Song et al. 2017; Zhang et al. 2017;Zheng et al. 2019). Alternatively, the skeleton joints are composed into a 2D pseudo-image.ThenaCNN-basedmodelisabletogeneratethefinalprediction(Liuetal.2017;Lietal.2017a,b;Zhangetal.2019;Wangetal.2021).However,boththeRNN-basedandCNN-based methods do not explicitly take advantage of spatial relationships and temporaldynamics. Therefore, a series of graph convolutional networks (GCNs) are proposed forskeleton-based action recognition (Yan et al. 2018; Shi et al. 2019a, b; Tang et al. 2018;Chengetal.2020;Songetal.2021;Shietal.2020;Pengetal.2021;Liuetal.2021;Xieet al. 2021; Ahmad et al. 2021; Yoon et al. 2021). Inferred from CNNs, GCNs are able toprocess non-Euclidean data such as skeleton graphs through the regulation of the kernelsize and the promotion of the convolution operation. Subsequently, a graph convolutionmodule is widely used to construct the spatial–temporal GCN. Most of the GCN-basedmethods emphasize the improvement of a structure to obtain optimal spatial–temporalrepresentations.

Finally, the spectral features are combined with the spatial–temporal features extractedfrom the vertex stream to recognize the action. Compared with our previous SS-GCN, themaincontributions are summarized as follows:

- To extract spatial–temporal information more effectively, the shift operation on thegraphisemployedtoourvertexstreaminspiredbyShift-GCN(Chengetal.2020).Thisarticle further explores the effectiveness of the complementation of the vertex-domainand the spectral-domain features through a more efficient spatial–temporal stream,which proves the previous GCN is flawed in this task for some actions rely on globalinformation.
- AmorerobustspectralGCNsbackboneconsistingofRSBsisproposed,provingtobe more effective in extracting spectral features for action recognition. Though someexperimentsshowthatspectral-basedGCNperformsinferiortospatial-basedGCNin some computer vision tasks, our RSB shows particular improvement to the simplespectral-basedGCNsadoptedbyourpreviousSS-GCNowingtothedeeparchitecture.
- In previous work, the motivation of the combination of the spatial–temporal informa-tionandthespectralinformationisnotwellexpressedandsupported.Atthesametime,this paper proposes using spectral-domain information to make up for the weak rec-ognitionabilityofpreviousGCNsinsomeactions.Ananalysisoftheimprovement of each action category by the spectral-domain information is provided in the ablationstudy.
- More extensive experiments and more comprehensive analyses are performed. Owingto the improvement on both the spectral stream and the spatial–temporal stream, DD-GCNhasgreatlyimprovedourpreviousmodelSS-GCN.withanincreaseof5.3%/5.5%on the NTU-RGBD 60 dataset (Shahroudy et al. 2016). The top-1 and top-5 accuracyon the Kinetics-Skeleton dataset (Kay et al. 2017) are also improved by 0.9%/2.0%.Additional experiments on NTU-RGBD 120 (Liu et al. 2020) are performed and com-paredwith the SOTA.

2 Relatedwork

Owing to the effectiveness data, there is more and more research focusing on skeleton-based action recognition. The skeleton data that indicates the coordinates dynamics showsrobustness against illumination change, background variation, and body diversity. Themethods are composed of the handcraft-feature methods and the deep learning methods.One typical handcraft feature is based on the theory of Lie Group (Vemulapalli et al. 2014;Jiang et al. 2020; Fernando et al. 2015). Vemulapalli et al. (2014) propose a Lie-groupskeletal representation that uses rotations and translations in 3D space to model the 3Dgeometric relationships between different body parts specifically. Inspired by this work,Jiang et al. (2020) create a new spatial–temporal skeleton transformation descriptor (ST-STD) to obtain a comprehensive view of the skeleton in both spatial and temporal domainfor each frame, followed by a denoising sparse long short term memory (DS-LSTM) net-work. Fernando et al. (2015) use the parameters from the ranking functions per video as anewvideo representation.

However, the deep learning features are more substantial than the handcraft-featuremethods due to various deep models such as RNNs and CNNs. RNNs-based methods canextract the dynamic information with the ability of modeling sequences (Du et al. 2015;Liu et al. 2016, 2018; Song et al. 2017; Zhang et al. 2017; Li et al. 2018; Zheng et al.2019). Du et al. (2015) propose an end-to-end hierarchical RNN for skeleton-based actionrecognition, based on the ability to model the long-term contextual knowledge of temporalsequences of the RNNs. Liu et al. (2016) further propose a tree-structure traversal methodbased on LSTM to deal with occlusion and noise in human skeleton data. To make betteruse of the multi-modal features extracted for each joint, then they (Liu et al. 2018) intro-duce a feature fusion method within the trust gate ST-LSTM unit. Song et al. (2017) com-bine the spatial attention subnetwork and the temporal attention subnetwork with the mainLSTM network to pay various levels of attention to different frames. Zhang et al. (2017)proposeatwo-streamViewAdaptivenetworkforskeletonactionrecognitiontoelimi-natetheinfluenceoftheviewpointsbycombiningRNNfeatureswithCNNfeatures.Lietal.(2018)introduceanindependentlyRNN(IndRNN)architecturetoovoidthegradient vanishing while learning long-term dependencies. Zheng et al. (2019) integrate the atten-tionmechanismintoLSTMtomodelspatialandtemporaldynamicssimultaneously.

Meanwhile,byformingtheskeletonintopseudo-images,CNN-basedmethodsarealso widely studied (Ke et al. 2017; Liu et al. 2017; Kim and Reiter 2017; Li et al. 2017a,b; Cao et al. 2019). Ke et al. (2017) introduce a manual clip generation method for theskeletonjointsofeachframewhichareplacedasachainbyconcatenatingthejoints.Liuet al. (2017) present an enhanced visualization method for skeleton data according to aview-invariant transform, an image colorization, and a CNN-based model. Kim and Reiter(2017) re-design the Temporal Convolutional Neural Networks (TCN) to learn the spa-tial–temporal representations of the human skeleton data. Li et al.

2.1 Graphconvolutionalneuralnetworks

Nevertheless,neitherCNNsnorRNNsprocessthenon-Euclideangraphsdirectly.Boththesequences in RNNs and the grids in CNNs have flaws while blending spatial and tempo-ral patterns. Therefore, several GCN-based models are proposed to capture spatiotemporalfeatures from graphs (Yan et al. 2018; Shi et al. 2019a, b; Peng et al. 2021; Liu et al. 2021;Xie et al. 2021; Ahmad et al. 2021). Inferred by CNNs, these GCNs avoid the handcraftedpart-assignment. Yan et al. (2018) propose to treat the skeleton sequences as spatiotempo-ral graphs and extend CNNs to the vertex domain of the graph by a spatiotemporal GCN(ST-GCN). The spatiotemporal information is shown vital for trajectory data in differentdomains (Knauf et al. 2016). Unlike CNNs, the convolution operation in the GCNs unitcontains the input data and learnable weights and the adjacency matrix of the graph dem-onstrating the spatiotemporal connection. By constructing a naturally connected skeletongraph, ST-GCN eliminates the need to specify the data structure manually. Si et al. (2019)combine vertex-domain graph convolution with LSTM to capture features in both spatialconfiguration and temporal dynamics. Based on ST-GCN, Shi et al. (2019a) raise a two-streamadaptiveGCN(2s-AGCN)toobtainthejointandthesecond-orderinformationof the skeleton data. They add learnable adaptive parameters to the adjacency matrix toimprove the limitations of natural connection in ST-GCN. Then 2s-AGCN is extended toMS-AAGCN (Shi et al. 2020) by a multi-stream architecture which combines the informa-tionfrombothjointsandbones,aswellastheirmotiontrends.AnotherworkfromShiet al. (2019b) propose a directed graph network (DGN) to model joints and bones in thenaturalhumanbody,whicharerepresentedasadirectedacyclicgraph(DAG).Chenget al. (2020) propose a novel shift operation for spatial GCNs based on the previous work,which greatly reduces the GFLOPs and increases the inflexibility of the receptive fields.Inspired by this work, our vertex-domain stream consists of spatiotemporal shift GCNblocks, which is more effective while extracting non-local relationships between spatialandtemporaldomains.

Estrach et al. (2014) exploit a global structure of the graph with the spectrum of itsgraph-Laplacian matrix to generalize the convolution operator from CNNs. A vanilla GCNin the spectral domain is proposed by constructing a graph spectral convolution layer, inwhich the spatial filter is replaced with a spectral filter. Henaff et al. (2015) develop animproved spectral GCN by smooth the spectral filters. By smoothing the spectral filters inthe spectral domain, a more localized filter in the space domain is obtained faster duringthedecay.Defferardetal.(2016)learnthefunctionsoftheLaplaciandirectlytoavoidtheeigendecompositionwhilecalculatingthespectralconvolution.InspiredbyChengetal.(2020), a dual-domain graph CNN is proposed to capture both spatiotemporal and spectralinformation with two kinds of graph convolution operators. Inferred by ResNet, a novelresidual-connectedspectralbackboneisproposedtoavoidgradientvanishing.

3 Graphconvolutionoperations

Thissectionintroducestwosortsofgraphconvolutionoperationsaccordingtographsignalprocessing(GSP) for skeleton actionrecognition.

3.1 Vertex-domaingraphconvolution

GCNs have been a widely used architecture since the work of Yan et al. (2018). By constructing the skeleton data into graph $G = (V, E)$ with N joints and T frames, a vertex-domaingraphconvolutionoperationisdefinedwiththethoughtoftemplatematching.
Because of the absence of node ordering and the structure diversity, the simplest way todesignatemplatetocalculatetheconvolutionistouseascalarforallneighbors.Givenan input vector h of l *th*layer in a GNN, the vertex-domain scalar convolution is shown asfollows:

$$h_i^{l+1} = \sigma\left(\sum_{j \in N_i} \langle w^l, h^l_{ij} \rangle\right), \tag{1}$$

where⟨, ⟩ is the product operation andis the activation function. $w^l \in R$ is the templatevector to obtain neighborhood information in layer l. And N_i denotes the set of all neighbornodesofnode$i$.Forageneralconvolutioningraphneuralnetworks,thefollowingformula
is obtained:

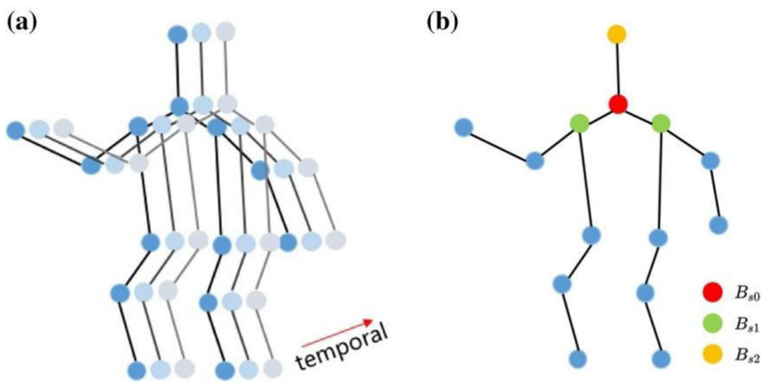

Fig.1 Illustrationoftheskeletongraphforvertex-domaingraphconvolution.Thebluedotsrepresentingthe body joints are connected in both spatial and temporal domain. For the vertex-domain convolution, theyare divided into three handcrafted subsets: root subset B_{s0}, centripetal subset B_{s1} and centrifugal subset B_{s2}(Colorfigure online)

3.2 Spectral-domaingraphconvolution

Inskeletonactionrecognition,thelatestmethodsalltreatjointsandbones,aswellastheirmotiontrajectories,asaspatiotemporalgraphtoperformvertex-domainconvo-lution operations. However, since the skeleton data is regarded as graphs, the ignoredspectral-domain information is also vital according to the Spectral Graph Theory. Theanalysis of the properties of a graph concerning the characteristic polynomial, eigenval-ues,andeigenvectorsoftheLaplacianmatrix,isthemainpartofspectralgraphtheoryinmathematics.

Thespectralconvolutionisperformedbythefollowingsteps:GraphLaplacian matrix,FourierfunctionsandFouriertransform,Convolutiontheorem.TheNthskel-etonsequenceintime$T$isconvertedtoaspatiotemporalgraph$G=(V,E)$.Accordingtospectralgraphtheory,TheAdjacencymatrixisrepresentedas$A$.Anotheressential operatoristhegraphisLaplacianmatrix$L$.AndthesimpleLaplacianmatrixisdefined as$L = D - A \in R^{n\times n}$.$D = diag(d(v_1), \ldots , d(v_N)) \in R^{n\times n}$isthediagonaldegreematrixand$d(\cdot)$isthedegreeofnodev_i.ThenthenormalizedLaplacianmatrixisdefinedas

$L=D^{-\frac{1}{2}}(D-A)D^{-\frac{1}{2}}=I-D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$.

Itisobviousthatthe Laplacianmatrix$L$isarealsymmetricmatrix.Givenavector relatedtovertexv_i,$\mathbf{h}$istheoutputvectorbycalculatingtheproductoftheLaplacianmatrix$L$and .Anditsphysicalimplicationcanbeclarifiedwiththefollowingformula:

$$\mathbf{h}=L=(D-A)=D-A,$$

$$\begin{aligned}\mathbf{h}[i]&=d(v_i)[i]-\sum_{v_j\in N(v_i)}A_{i,j}[i]\\&=\sum_{v_j\in N(v_i)}1\cdot[i]-\sum_{v_j\in N(v_i)}1\cdot[j]\\&=\sum_{v_j\in N(v_i)}([i]-[j]),\end{aligned}$$

wheretheoutputvector$\mathbf{h}$representsthedifferencebetween$v_i$anditsneighborvertexv_j.

Laplacianmatrixisalsoapositivesemidefinitematrixandcanbeprovedwith Eq.5 bythefollowingformula,thequadraticformof$L$:

$$\begin{aligned}f^{\top}Lf&=\sum_{v_j\in V}[i]\sum_{v_j\in N(v_i)}([i]-(j))\\&=\sum_{v_i\in V}\sum_{v_i\in N(v_i)}([i]\cdot[i]-[i]\cdot[j])\\&=\sum_{v_i\in V}\sum_{v_i\in N(v_i)}\frac{1}{2}[i]\cdot[i]-[i]\cdot[j]+\frac{1}{2}[j]\cdot[j]\\&=\frac{1}{2}\sum_{v_i\in V}\sum_{v_i\in N(v_i)}([i]-[j])^2.\end{aligned}$$

As shown in Eq. 7, the quadratic form of the Laplacian matrix L is the sum of the squaresofthedifferencebetweeneachvertexanditsneighborhoodsinagraph.Frombothperspec-tives in Eqs.5 and 7, the physical implication of the Laplacian matrix is that it is a measureofthedifferencebetweeneachnodeanditsneighbornodesinthegraph.Thisisquite differentfromtheAdjacencymatrixappliedinvertex-domaingraphconvolutionoperation,whichprovides thestrength ofthe connectionof theedge betweennodes.

ThevitalLaplacianmatrix$L$ispreciselythebasiccontentofgraphspectralconvolu-tionoperation.Theconvolutioninthevertexdomaincannotbeexpressedasameaning-fuloperatorroughly.However,theconvolutionoperator $*_G$ iseasilydefinedinthespectraldomainaccording to graph convolution theorem:

$$w *_G h = U \; U^T w \odot U^T, h \; ,$$

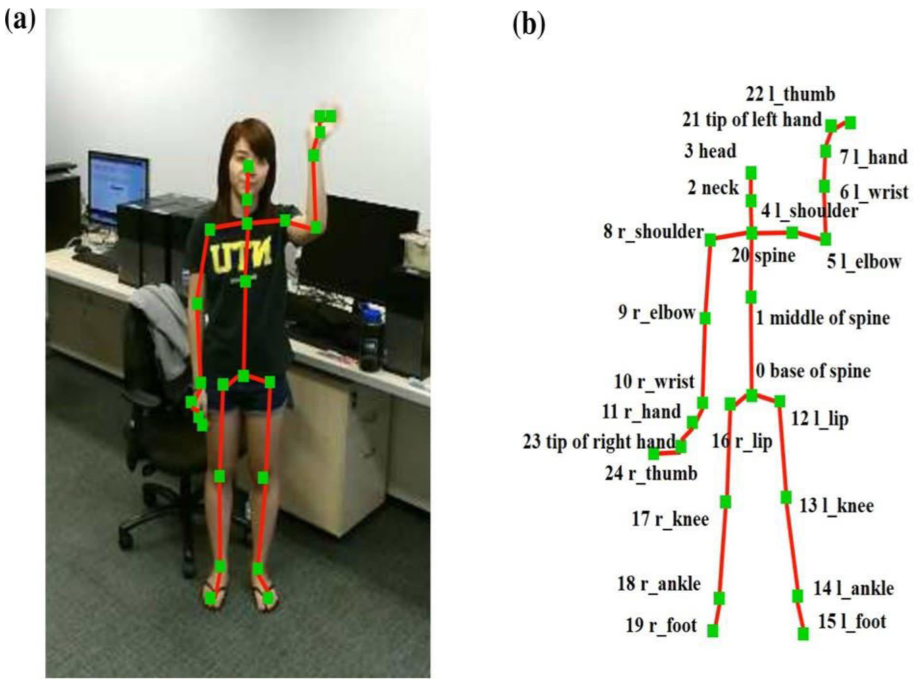

Fig.6 Examplesforclass“handwaving”.Theredlineandgreendotsrepresenttheskeletons(Colorfigureonline)

w

$*$

G

h

$=$

U

3.2.1 NTU-RGBD120dataset

NTU RGBD 120 dataset is an extended version of the NTU-RGBD 60 dataset by addinganother60classesandanother57,600video/skeletonsamples.Itconsistsof114,480action samples divided into 120 action classes. The number of persons of different ages increasesto106.ThesamplesarecapturedinthreeangleswhichisthesameasNTU-RGBD60.Theskeleton data employed in this work consists of 25 human joints, as shown in Fig. 6. Thetwo benchmarks are also defined as CS and CV. The action can be categories into DailyActions(82),MedicalConditions(12),andMutualActions/TwoPersonInteractions(26).

3.2.2 Kinetics-Skeletondataset

Kinetics is an activity recognition dataset for RGB-based action recognition, which consists of 300,000 videos clips in 400 classes (Kay et al. 2017). Yan et al. (2018) construct askeleton data based on it by extracting 18 body joints for each frame with an open-sourcetoolbox OpenPose. Then the large-scale skeleton-based dataset called Kinetics-Skeleton isobtained. The training data is set to 240,000 skeleton clips, and the test data consists of20,000 clips. This dataset is challenging, so both the top-1 and top-5 accuracies are presentasother methods do.

3.3 ImplementdetailsofDD-GCN

The DD-GCN is implemented with Pytorch deep learning framework. Some hyperparameters are needed for both the vertex-domain stream and the spectral-domain stream. ForNTU-RGBD 60 dataset and NTU-RGBD 120 dataset, the optimizer is SGD (stochasticgradientdescent)method.Andthelossfunctioniscross-entropyloss.SimilartoChenget al. (2020), the weight decay and initial learning rate of the vertex-domain stream are setto 0.0001 and 0.1. The learning rate decays by 10 at epoch of 60th, 80th, 100th. For spec-tral-domain stream on NTU-RGBD 60 dataset and NTU-RGBD 120 dataset, the weightdecay and initial learning rate of the vertex-domain stream are set to 0.003 and 0.1. Thelearningrate decays by 10 at epochof 30th, 40th.

FortheKinetics-Skeletondataset,theSGDisadoptedastheoptimizer.Thesettingsofweightdecayandtheinitiallearningratearethesamewith NTU-RGBDdatasetsinthe vertex-domain stream. For spectral stream, the weight decay, Nesterov momentum forSGD, the base learning rate is set to 0.001, 0.9, 0.001. The learning rate decays by 10 atepochof 45th, 55th.

3.3.1 RSBstrategies

The effectiveness of the spectral-domain backbone, which adopts the residual-connectedspectral block, is evaluated in Table 1. Compared with the stream adopting simple spectralgraph convolution, the residual spectral stream demonstrates a better performance with anincrease of 15.1% and 12.9% on NTU-RGBD 60 CS and CV. Some recent experimentsshow that spectral-based GCN performs inferior to spatial-based GCN in some computervision tasks. However, our experiments based on the RSB backbone show a certain devel-opmentpotentialofthespectralconvolution.Thecriticalproblemofthepreviousspec-tral convolution network lies in relatively shallow architecture. At the same time, residual-connected architecture for the spectral-domain stream of DD-GCN is capable of capturingdeep spectral information. While combined with the vertex-domain stream, which focuseson the spatiotemporal information, the residual DD-GCN has a superior performance withanincreaseof1.1%and0.7%onCSandCV.Incontrast,thesimpleDD-GCNseemsstren-uoustoobtain adequatespectral informationforthe vertex-domainstream.

Table1 TheablationstudyonNTU-RGBDdatasetdenotingtheeffectiveness of the Res-SpectralUnit

Methods	CS(%)	CV(%)
SimpleSpectralStream	55.2	65.3
ResidualSpectralStream	70.3	78.2
Vertex-domainStream(Shift)	87.8	95.1
DD-GCN(Simple)	88.6	95.3
DD-GCN(Residual)	88.9	95.8

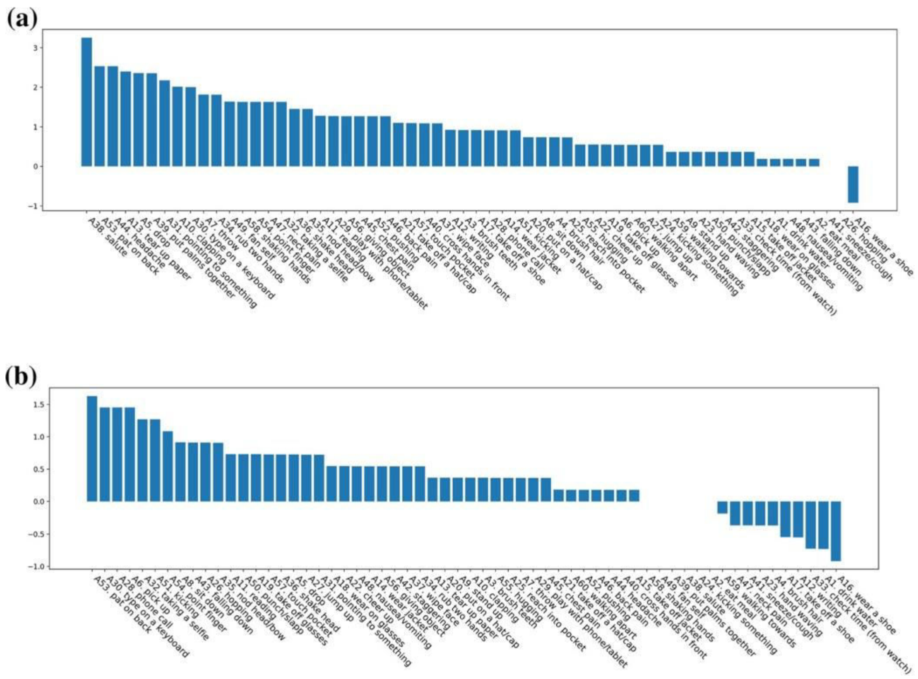

Fig.7 Illustrationoftheperformancegain(%)ofthespectral-domainstreamwithrespecttothevertex-domain stream on the NTU-RGBD 60 dataset for the CS setting. The vertical axis is calculated by subtract-ing the DD-GCN accuracy of each action from the vertex-domain stream. The horizontal axis denotes theclassof action as provided in Shahroudy et al. (2016)

3.3.2 ExperimentsonNTU-RGBD120dataset

On NTU-RGBD 120 Dataset, two standard evaluation protocols are applied in Liu et al.(2020). The comparison results are shown in Table 5. The experiment accuracy of DD-GCN is 84.9% for the CS set, 86.0% for the CV set. Compared with 3s RA-GCN, ourtwo-stream model has a 3.8%/3.3% increase on CS and CV set. The performance oftwo-streamST-TR-AGCN(Plizzarietal.2021)concatenatingspatial–temporalmod-ule with self-attention mechanism is 2.2%/1.3% lower than DD-GCN. The DD-GCNachieves 0.7% higher accuracy on CS set and 0.5% higher on CV set than the work inWangetal.(2021).ThisdemonstratesthesuperiorityofourGCNmodelthatutilizestheresidualspectralstreambasedonthespectral-domaingraphconvolution.

The results of DD-GCN on the NTU-RGBD 120 dataset are 0.4% lower than 2s Shift-GCN, which superimposes the same backbone repeatedly with additional preprocesseddata,thebonegraphs(thedifferentialofspatialcoordinates).ComparedwiththeSOTA4s Shift-GCN, our results are slightly inferior but with much lesser parameters. Neverthe-less,ourworkhasbenefitedbyfusingtwodistinguishinggraphconvolutionoperators.The experiment results show that our two-stream network is reasonable and practical toobtainthelocaldiversitiesandtheglobaldynamicsevenwithoutadditionaldata.

Table5 ThecomparisonsofexperimentresultsonNTU-RGBD120dataset

Methods	CS(%)	CV(%)	Year
ST-LSTM(Liuetal.2016)	55.7	57.9	2016
SkeleMotion(Caetanoetal.2019)	67.7	66.9	2019
TSRJI(Caetanoetal.2019)	67.9	62.8	2019
Part-AwareLSTM(Liuetal.2020)	55.7	57.9	2020
2sShift-GCN(+bones)(Chengetal.2020)	85.3	86.6	2020
4sShift-GCN(+bonesandmotions)(Chengetal.2020)	**85.9**	**87.6**	2020
FuzzyCNN (Banerjee et al. 2021)	74.8	76.9	2021
AMV-GCN(Liuetal.2021)	76.7	79.0	2021
3sRA-GCN(Songet al.2021)	81.1	82.7	2021
ST-TR-AGCN(Plizzarietal.2021)	82.7	85.0	2021
SEMN(Wangetal.2021)	84.2	85.5	2021
DD-GCN(ours)	**84.9**	**86.0**	2021

Experimental resultsandthe state-of-the-artarehighlightedin bold

4 Conclusion

In this paper, a dual-domain GCN (DD-GCN) for skeleton-based action recognition isproposed. We integrate spectral-domain information with spatial–temporal informationthrough an end-to-end two-stream architecture. A spectral-GCN backbone is proposedbased on the spectral-domain graph convolution. Compared with the previous GCN, whichonly focuses on the spatial–temporal information of the skeleton graphs, we explore thecomplementary spectral-GCN architecture and the necessity. With a deep residual-con-nected RSB backbone, the accuracy of most actions has been improved, primarily theactions with broader dynamic changes in global. The experiment results on three large-scale datasets demonstrate the effectiveness of our DD-GCN. The ablation studies explorethe reasons for the superiority of DD-GCN for the task of skeleton-based action recog-nition. The extensive experiments on three large-scale datasets, NTU-RGBD 60, NTU-RGBD 120, and Kinetics-Skeleton, show competitive or state-of-the-art performance. Inthe future, we will optimize the spectral-domain backbone for skeleton-based action rec-ognitionandhopetoinspiremoreworktofocusonthedual-domaingraphconvolutions.

Acknowledgements ThisworkisfundedbytheNatureNaturalScienceFoundationofChina(62002220).XinghaoJiang is the corresponding author.

Authorcontributions SC:Conceptualization,Methodology,Writing-originaldraft,Software.KX:Supervi-sion,Validation.ZM:DataCuration.XJ:Investigation,Visualization.TS:Writing-reviewandediting.

Funding ThisworkisfundedbytheNatureNaturalScienceFoundationofChina(62002220).

Data availability Thedatasetssupportingtheresultsofthisarticleareincludedwithinthearticleanditsadditionalfiles.

References

Ahmad,T.,Jin,L.,Lin,L.,&Tang,G.(2021).Skeleton-basedactionrecognitionusingsparsespatio-tem-poralGCN withedge effectiveresistance.*Neurocomputing,423,* 389–398.

Banerjee, A., Singh, P. K., &Sarkar, R. (2021). Fuzzy integral-based CNN classifier fusion for 3D skel-eton action recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 31*(6),2206–2216.

Caetano, C., Brémond, F., & Schwartz, W. R. (2019).Skeleton image representation for 3D action recognition based on tree structure and reference joints. In *2019 32nd SIBGRAPI conference on graphics, patternsand images (SIBGRAPI)* (pp. 16–23). IEEE.

Caetano, C., de Souza, J. S., Brémond, F., dos Santos, J. A., & Schwartz, W. R. (2019).SkeleMotion: Anew representation of skeleton joint sequences based on motion information for 3D action recognition.In *16th IEEE international conference on advanced video and signal based surveillance, AVSS 2019*,Taipei,Taiwan, September 18–21, 2019(pp. 1–8). IEEE.

Cao, C., Lan, C., Zhang, Y., Zeng, W., Lu, H., & Zhang, Y. (2019). Skeleton-based action recognition withgated convolutional neural networks.*IEEE Transactions on Circuits and Systems for Video Technol-ogy,29*(11), 3247–3257.

Chen, S., Xu, K., Xinghao, J., &Tanfeng, S. (2021). Spatiotemporal-spectral graph convolutional networksfor skeleton-based action recognition. In *2021 IEEE international conference on multimedia and expoworkshops,ICME workshops, virtual*,July 5–9, 2021 (pp.1–6).

Cheng,K.,Zhang,Y.,He,X.,Chen,W.,Cheng,J.,&Lu,H.(2020).Skeleton-basedactionrecognitionwithshift graph convolutional network. In *2020 IEEE/CVF conference on computer vision and pattern recognition,CVPR 2020*,Seattle, WA,USA, June13–19, 2020(pp. 180–189).

Cho, S., Maqbool, M. H., Liu, F., &Foroosh, H. (2020).Self-attention network for skeleton-based humanactionrecognition.In*IEEEwinterconferenceonapplicationsofcomputervision,WACV2020*,Snow-massVillage, CO, USA, March1–5, 2020 (pp. 624–633).

Chung,F.R.,&Graham,F.C.(1997).*Spectralgraphtheory*.No.92.AmericanMathematicalSociety.

Defferrard,M.,Bresson,X.,&Vandergheynst,P.(2016).Convolutionalneuralnetworksongraphswithfastlocalized spectral filtering. In *Advances in neural information processing systems 29: Annual confer-ence on neural information processing systems 2016*, December 5–10, 2016, Barcelona, Spain (pp.3837–3845).

Dhillon, I. S., Guan, Y., &Kulis, B. (2007). Weighted graph cuts without eigenvectors A multilevelapproach.*IEEETransactionsonPatternAnalysisandMachineIntelligence,29*(11),1944–1957.

Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE conference on computer vision and pattern recognition, CVPR 2015*, Boston, MA,USA,June 7–12, 2015 (pp. 1110–1118).

Estrach, J. B., Zaremba, W., Szlam, A., &LeCun, Y. (2014). Spectral networks and deep locally connectednetworksongraphs.In*2ndInternationalconferenceonlearningrepresentations,ICLR*(Vol.2014).

Fernando, B., Gavves, E., Jose Oramas, M., Ghodrati, A., &Tuytelaars, T. (2015).Modeling video evolution for action recognition. In *IEEE conference on computer vision and pattern recognition, CVPR2015*,Boston,MA,USA, June7–12,2015 (pp.5378–5387).IEEEComputer Society.

Hammond,D.K.,Vandergheynst,P.,&Gribonval,R.(2009).Waveletsongraphsviaspectralgraphtheory. *CoRR*, abs/0912.3848.

He, K., Zhang, X., Ren, S., & Sun, J. (2016).Deep residual learning for image recognition. In *2016 IEEEconference on computer vision and pattern recognition, CVPR 2016*, Las Vegas, NV, USA, June27–30,2016 (pp. 770–778). IEEE ComputerSociety.

Henaff, M., Bruna, J., &LeCun, Y. (2015).Deep convolutional networks on graph-structured data.*CoRR*,abs/1506.05163.

Islam, M. M., &Iqbal, T. (2020). HAMLET: A hierarchical multimodal attention-based human activityrecognition algorithm. In *IEEE/RSJ international conference on intelligent robots and systems, IROS2020*,LasVegas,NV,USA, October24–January24,2021 (pp.10285–10292).

Ji, S., Xu, W., Yang, M., &Yu, K. (2013).3D convolutional neural networksfor human action recognition. *IEEETransactionsonPatternAnalysisandMachineIntelligence,35*(1),221–231.

Jiang, X., Xu, K., & Sun, T. (2020).Action recognition scheme based on skeleton representation with DS-LSTMnetwork.*IEEETransactionsonCircuitsandSystemsforVideoTechnology,30*(7),2129–2140.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green,T., Back,T.,Natsev,P.,Suleyman,M.,&Zisserman,A.(2017).Thekineticshumanactionvideodataset.*CoRR*,abs/1705.06950.

Ke, Q., Bennamoun, M., An, S., Sohel, F. A., &Boussaïd, F. (2017). A new representation of skeletonsequences for 3D action recognition. In *2017 IEEE conference on computer vision and pattern recog-nition,CVPR 2017*, Honolulu, HI,USA, July 21–26, 2017(pp. 4570–4579).

Kim, T. S., & Reiter, A. (2017). Interpretable 3D human action analysis with temporal convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition workshops, CVPR workshops2017*, Honolulu, HI, USA, July21–26, 2017 (pp. 1623–1631).

Knauf, K., Memmert, D., &Brefeld, U. (2016).Spatio-temporal convolution kernels.*Machine Learning,102*(2),247–273.

Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., & He, M. (2017a). Skeleton based action recognition usingtranslation-scale invariant image mapping and multi-scale deep CNN. In *2017 IEEE internationalconference on multimedia and expo workshops, ICME workshops*, Hong Kong, China, July 10–14,2017(pp.601–604).

Li, S., Li, W., Cook, C., Zhu, C., &Gao, Y. (2018). Independently recurrent neural network (IndRNN):Building a longer and deeper RNN. In *2018 IEEE conference on computer vision and pattern rec-ognition,CVPR2018*,SaltLakeCity,UT,USA,June18–22,2018(pp.5457–5466).

Li,C.,Zhong,Q.,Xie,D.,&Pu,S.(2017b).Skeleton-basedactionrecognitionwithconvolutionalneu-ral networks.In *2017 IEEE international conference on multimedia and expo workshops, ICMEworkshops*,HongKong,China,July10–14,2017(pp.597–600).

Liu, X., Li, Y., & Xia, R. (2021).Adaptive multi-view graph convolutional networks for skeleton-basedactionrecognition.*Neurocomputing,444,*288–300.

Liu, M., Liu, H., & Chen, C. (2017).Enhanced skeleton visualization for view invariant human actionrecognition.*PatternRecognition,68,*346–362.

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L., &Kot, A. C. (2020). NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis andMachineIntelligence,42*(10),2684–2701.

Liu, J., Shahroudy, A., Xu, D., Kot, A. C., & Wang, G. (2018).Skeleton-based action recognition usingspatio-temporalLSTMnetworkwithtrustgates.*IEEETransactionsonPatternAnalysisandMachineIntelligence,40* (12),3007–3021.

Liu,J.,Shahroudy,A.,Xu,D.,&Wang,G.(2016).Spatio-temporalLSTMwithtrustgatesfor3Dhuman action recognition. In B. Leibe, J. Matas, N. Sebe& M. Welling (Eds.), *Computer Vision—ECCV 2016—14th European conference, proceedings, Part III: Lecture notes in computer science*,Amsterdam,TheNetherlands,October11–14,2016(Vol.9907,pp.816–833).

Peng,W.,Shi,J.,Varanka,T.,&Zhao,G.(2021).RethinkingtheST-GCNsfor3Dskeleton-basedhumanactionrecognition.*Neurocomputing,454,*45–53.

Plizzari,C.,Cannici,M.,&Matteucci,M.(2021).Skeleton-basedactionrecognitionviaspatialandtemporaltransformernetworks.*ComputerVisionandImageUnderstanding,208–209,*103219.

Rahmani,H.,&Bennamoun,M.(2017).Learningactionrecognitionmodelfromdepthandskeletonvideos. In *IEEE international conference on computer vision, ICCV 2017*, Venice, Italy, October22–29,2017(pp.5833–5842).

Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D humanactivityanalysis.In*2016IEEEconferenceoncomputervisionandpatternrecognition,CVPR2016*,LasVegas,NV,USA,June27–30,2016(pp.1010–1019).

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019a). Two-stream adaptive graph convolutional networks forskeleton-based action recognition. In *IEEE conference on computer vision and pattern recognition,CVPR2019*,LongBeach,CA,USA,June16–20,2019(pp.12026–12035).

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019b). Skeleton-based action recognition with directed graphneuralnetworks.In*IEEEconferenceoncomputervisionandpatternrecognition,CVPR2019*,LongBeach,CA,USA,June16–20,2019(pp.7912–7921).

Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2020).Skeleton-based action recognition with multi-streamadaptivegraphconvolutionalnetworks.*IEEETransactionsonImageProcessing,29,*9532–9545.

Si,C.,Chen,W.,Wang,W.,Wang,L.,&Tan,T.(2019).AnattentionenhancedgraphconvolutionalLSTM network for skeleton-based action recognition. In *IEEE conference on computer vision andpattern recognition, CVPR 2019*, Long Beach, CA, USA, June 16–20, 2019 (pp. 1227–1236). Com-puterVisionFoundation/IEEE.

Simonyan,K.,&Zisserman,A.(2014).Two-streamconvolutionalnetworksforactionrecognitioninvideos.In*Advancesinneuralinformationprocessingsystems27:Annualconferenceonneuralinformationprocessingsystems2014*,December8–13,2014,Montreal,QC,Canada(pp.568–576).Song,S.,Lan,C.,Xing,J.,Zeng,W.,&Liu,J.(2017).Anend-to-endspatio-temporalattentionmodelforhumanactionrecognitionfromskeletondata.InS.P.Singh&S.Markovitch(Eds.),*Proceed-ingsofthethirty-firstAAAIconferenceonartificialintelligence*,February4–9,2017,SanFran-cisco,CA,USA(pp.4263–4270).

Song,Y.,Zhang,Z.,Shan,C.,&Wang,L.(2021).Richlyactivatedgraphconvolutionalnetworkforrobustskeleton-basedactionrecognition.*IEEETransactionsonCircuitsandSystemsforVideoTechnology,31*(5),1915–1925.

Tang, Y., Tian, Y., Lu, J., Li, P., & Zhou, J. (2018). Deep progressive reinforcement learning for skeleton-based action recognition. In *2018 IEEE conference on computer vision and pattern recognition, CVPR2018*,SaltLakeCity,UT,USA,June18–22,2018(pp.5323–5332).IEEEComputerSociety.

Vemulapalli,R.,Arrate,F.,&Chellappa,R.(2014).Humanactionrecognitionbyrepresenting3Dskeletonsas points in a Lie Group. In *2014 IEEE conference on computer vision and pattern recognition, CVPR2014*,Columbus, OH, USA, June 23–28, 2014 (pp. 588–595).

Wang,H.,Yu,B.,Xia,K.,Li,J.,&Zuo,X.(2021).Skeletonedgemotionnetworksforhumanactionrecog-nition.*Neurocomputing, 423,* 1–12.

Wu, B., Wan, A., Yue, X., Jin, P. H., Zhao, S., Golmant, N., Gholaminejad, A., Gonzalez, J., &Keutzer, K.(2018). Shift: A zero flop, zero parameter alternative to spatial convolutions. In *2018 IEEE conferenceon computer vision and pattern recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018(pp.9127–9135). IEEE Computer Society.

Xie,J.,Miao,Q.,Liu,R.,Xin,W.,Tang,L.,Zhong,S.,&Gao,X.(2021).Attentionadjacencymatrixbasedgraphconvolutionalnetworksforskeleton-basedactionrecognition.*Neurocomputing,440,*230–239.

Yan, S., Xiong, Y., & Lin, D. (2018).Spatial temporal graph convolutional networks for skeleton-basedactionrecognition.In*Proceedingsofthethirty-secondAAAIconferenceonartificialintelligence*,NewOrleans,Louisiana, USA, February 2–7, 2018 (pp. 7444–7452).

Yang, Y., & Li, D. (2020). NENN: Incorporate node and edge features in graph neural networks. In S. J.Pan & M. Sugiyama, (Eds.), *Proceedings of the 12th Asian conference on machine learning: Proceed-ings of machine learning research, PMLR*, Bangkok, Thailand, November 18–20, 2020 (Vol. 129, pp.593–608).

Yoon, Y., Yu, J., &Jeon, M. (2021). Predictively encoded graph convolutional network for noise-robustskeleton-basedaction recognition. *Applied Intelligence, 52,*1–15.

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., &Zheng, N. (2017). View adaptive recurrent neural net-works for high performance human action recognition from skeleton data. In *IEEE international con-ferenceoncomputervision,ICCV2017*,Venice,Italy,October22–29,2017(pp.2136–2145).

Zhang,P.,Lan,C.,Xing,J.,Zeng,W.,Xue,J.,&Zheng,N.(2019).Viewadaptiveneuralnetworksforhighperformance skeleton-based human action recognition. *IEEE Transactions on Pattern Analysis andMachineIntelligence, 41*(8), 1963–1978.

Zhang, X., Xu, C., & Tao, D. (2020).Context aware graph convolution for skeleton-based action recogni-tion. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020*, Seattle,WA,USA, June 13–19,2020 (pp. 14321–14330).

Zheng, W., Li, L., Zhang, Z., Huang, Y., & Wang, L. (2019).Relational network for skeleton-based actionrecognition.In *IEEE international conference on multimedia and expo, ICME 2019*, Shanghai, China,July8–12, 2019 (pp. 826–831).