EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition

Yong Li[2*] [†], Zihang He[2†], Xiang Ye[2], Zuguo He[1] and Kangrong Han[3]

Abstract

Hand gesture recognition methods play an important role in human-computer interaction. Among these methods are skeleton-based recognition techniques that seem to be promising. In literature, several methods have been proposed to recognize hand gestures with skeletons. One problem with these methods is that they consider little the connectivity between the joints of a skeleton, constructing simple graphs for skeleton connectivity. Observing this, we built a new model of hand skeletons by adding three types of edges in the graph to finely describe the linkage action of joints. Then, an end-to-end deep neural network, hand gesture graph convolutional network, is presented in which the convolution is conducted only on linked skeleton joints. Since the training dataset is relatively small, this work proposes expanding the coordinate dimensionality so as to let models learn more semantic features. Furthermore, relative coordinates are employed to help hand gesture graph convolutional network learn the feature representation independent of the random starting positions of actions. The proposed method is validated on two challenging datasets, and the experimental results show that it outperforms the state-of-the-art methods. Furthermore, it is relatively lightweight in practice for hand skeleton-based gesture recognition.

Keywords: Graph convolutional network, Deep learning, Hand gesture recognition

1 Introduction

Hand gesture recognition plays a key role in human-computer interaction and is attracting increasing interests for its potential applications in various fields. The methods for recognizing hand gesture can be classified into two categories, static and dynamic ways. The first category identifies hand gestures from a single image and thus runs faster while the second identifies hand gestures from a sequence of images and thus yields high precision. Dynamic identification methods are getting rising attention with the fast development of hardware. Recently, the advancement of precise hand pose estimation [1–3] allows for the hand gesture skeleton sequences to be generated in real time. Consequently, skeleton sequences which have high semantic information and small data size start to replace RGB images and depth maps [4, 5] in dynamic hand gesture recognition.

Smedt et al. [6] proposed a new descriptor, shape of connected joints (SoCJ), for skeleton-based hand gesture recognition in 2016 and demonstrated better performance over the methods employing depth maps as input. In addition to SoCJ, many other hand-crafted features [7, 8] were also designed. However, the problem of hand-crafted features is the inadequate ability to describe high-level semantic information.

Recently, deep neural networks have been widely used in the field of hand gesture recognition. Nunez et al. [9] proposed a method to extract features of each frame using CNNs and aggregate the outputs of CNNs with a LSTM [10]. Chen et al. [11] proposed a new motion feature augmented recurrent neural network that firstly encodes the joints of each finger and then the joints of the whole hand. The method in [9] treated the skeleton joints as a pseudo-image and that in [10] treated skeleton joints as a vector

*Correspondence: yli@bupt.edu.cn
[†]Yong Li and Zihang He contributed equally to this work.
[2]School of Electronic Engineering, Beijing University of Posts and Telecommunications, No.10 Xitucheng Road, Haidian District, Beijing 100876 , People's Republic of China
Full list of author information is available at the end of the article

sequence, but the connectivity of joints was not explicitly considered.

Hou et al. [12] designed an end-to-end spatial-temporal attention residual temporal convolutional network (STA-Res-TCN) which modifies temporal convolutional networks [13] for skeleton-based dynamic hand gesture recognition. STA-Res-TCN models the connectivity between joints by multiplying the output of an additional branch with the original branch, which forms the soft attention mechanism. However, using 3-D coordinates as a 1-D sequence, i.e., embedded X, Y, and Z dimensional coordinates on the same channel, limited its performance. Avola et al. [14] aggregated the leap motion controller sensor (LMC) with deep LSTM for hand gesture recognition. Features that are highly discriminative for the recognition are extracted by LMC and input to deep LSTM. But LMC was an independent module and cannot be directly embedded in deep LSTM, which made the model a two-stage structure.

An end-to-end model, spatial temporal graph convolutional networks (ST-GCN) [15], was recently proposed for skeleton-based human activity recognition. This network used motion features and kept the connectivity of joints. Because these two advantages are both critical to hand gesture recognition, ST-GCN might also be suited for it. However, we notice the performance of ST-GCN might be limited in hand gesture recognition by the following facts: (1) Hand gesture is much finer than human gesture because the former typically has a higher intra-class variance and lower inter-class variance. (2) The datasets for hand gesture recognition are much smaller than human action recognition, which may more likely lead neural networks to overfitting.

To adapt ST-GCN to the hand gesture recognition, this work proposed a new architecture named hand gesture graph convolutional networks (HG-GCN). The structure is shown in Fig. 1; the convolution operations are only executed between the joints which are linked. Also, the usually embedded 3-D coordinates are expanded 10-D to let models learn more semantic features. Moreover, relative coordinates are employed to help HG-GCN learn the feature representation independent of the random starting positions of actions. With these improvements, the proposed method achieves considerable performance on DHG-14/28 [16] and SHREC'2017 [17].

The rest of the paper is organized as follows: Section 2 presents the structure of ST-GCN and our modifications on it. Section 3 discusses the experimental results obtained. Finally, conclusions are outlined in Section 4.

2 Research methods

Convolutional networks perform well on processing images and skeleton joints sequences which can be seen as special images. However, arranging the joints as pixels in images leads to the destruction of the original human body topology. Since this defect has been noticed, Yan et al. [15] proposed a novel model named ST-GCN to limit the convolutional operations between the linked joints. Inspiring by their work, we proposed HG-GCN for the finger hand gesture recognition. To put the proposed model into context, a brief overview of this structure is firstly provided. Then, we describe our HG-GCN for hand gesture recognition. Also, we detail two strategies to process embedded data.

2.1 Overview of spatial temporal graph convNet

This section analyzes the structure of ST-GCN and its propagation. A normal convolutional network takes input as a four-dimensional matrix whose shape is $[N, H, W, C]$ where N denotes the batch size, C denotes the channel,

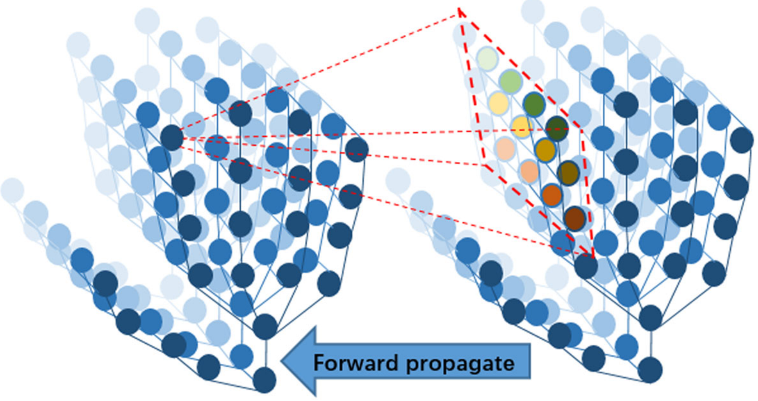


Fig. 1 An illustration of the spatial temporal network for dynamic hand gesture recognition proposed in this paper. Blue dots that present joints and links between them construct a whole human hand skeleton. The fading of color presents the forward of frames. Joints in the red box propagate their weights to the special joint in the next layer

and $H \times W$ denotes the area of the image. In order to use convolutional networks for skeleton-based action recognition, an embedded skeleton joints sequence is reshaped to $[N, T, V, C]$ where N denotes the batch size, T denotes the length of frames, V denotes the number of joints each frame, and C denotes the coordinate dimensions of joints. Although skeleton joints can be presented as an image in this way, it ignores the relationship between different parts of skeleton joints and hence propagate irrelevant information from one joint to another, which introduces noise between them.

To address this problem, Yan et al. [15] proposed ST-GCN to multiply an adjacent $[V, V]$ matrix A with feature maps after $t \times 1$ convolutional operations. The elements in this matrix are decided by the relationship of each two joints, e.g., column vectors denote joints themselves and row vectors denote the joints linked to them. The whole weights add to 1 for every joint and are the same for all linked joints, e.g., numbers of $A_{1,M}$ and $A_{N,M}$ are both 0.5 if joint V_M is only linked to joint V_N. An example of it is shown in Fig. 2.

Once joint V_m is linked with other N joints, the forward propagation to one joint is presented:

$$V_{(l+1)m} = \sum_{t=1}^{T} V_{lmt} \frac{w_{lmt}}{1+N} + \sum_{t=1}^{T} \sum_{n=1}^{N} V_{lnt} \frac{w_{lnt}}{1+N}; \qquad (1)$$

where l denotes the layer of feature maps, N denotes the set of joints linked to v_m, w denotes the corresponding weights, and T denotes the temporal stride of the kernel. As for feature maps, the propagation is presented:

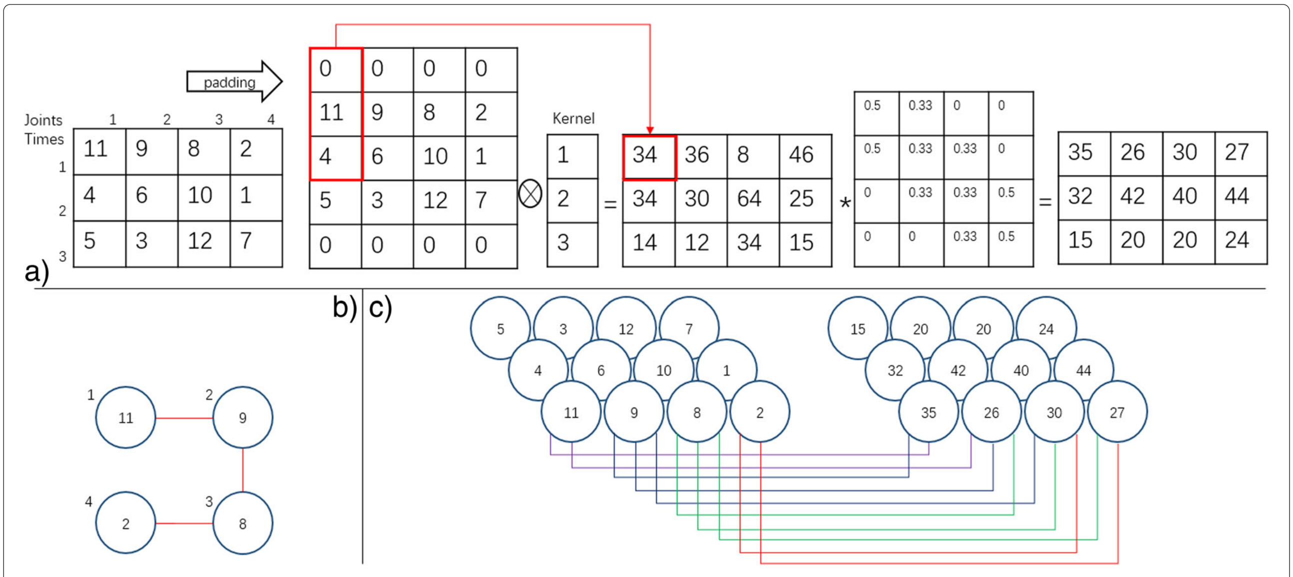
$$f_{out} = f_{in} WA; \qquad (2)$$

where f_{in} and f_{out} denote the input and output feature maps, respectively. A denotes the adjacent matrix and W denotes the weight matrix. Also, this model is composed of 9 layers of spatial temporal graph convolution operators and is lightweight enough to run in real time. Because hand gestures are much finer than human actions and the dataset of the former is much small. The suitable depth of HG-GCN is also explored, and the results are shown in Table 1. HG-GCN with 8 convolutional layers not only alleviates the overfitting, but also keeps enough parameters to precisely recognize hand gestures.

2.2 Hand gesture graph convNet

This section discusses embedding the ST-GCN model into hand gesture recognition. For DHG-14/28 [16] and SHREC'2017 [17], there are 22 skeleton joints for a hand as shown in Fig. 3a: four joints for each finger, one joint (1) for the palm, and one joint (0) for the wrist. The thumb contains joints $(5, 4, 3, 2)$, the index finger contains joints $(9, 8, 7, 6)$, the middle finger contains joints $(13, 12, 11, 10)$, the ring finger contains joints $(17, 16, 15, 14)$, and the pinkie contains joints $(21, 20, 19, 18)$.

Yan et al. linked the 18 joints of the human body skeleton with 17 edges from head to foot in order. Due to the high intra-class variation of fine hand gestures, linking the 22 hand skeleton joints with 21 edges may be unable to sufficiently encode the connectivity information between joints and hence the learned feature from hand joint model may yield an inferior recognition performance. Observing this, we proposed a hand gesture



$$\begin{array}{c|cccc} \text{Joints / Times} & 1 & 2 & 3 & 4 \\ \hline 1 & 11 & 9 & 8 & 2 \\ 2 & 4 & 6 & 10 & 1 \\ 3 & 5 & 3 & 12 & 7 \end{array} \xrightarrow{\text{padding}} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 11 & 9 & 8 & 2 \\ 4 & 6 & 10 & 1 \\ 5 & 3 & 12 & 7 \\ 0 & 0 & 0 & 0 \end{bmatrix} \otimes \underset{\text{Kernel}}{\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}} = \begin{bmatrix} 34 & 36 & 8 & 46 \\ 34 & 30 & 64 & 25 \\ 14 & 12 & 34 & 15 \end{bmatrix} * \begin{bmatrix} 0.5 & 0.33 & 0 & 0 \\ 0.5 & 0.33 & 0.33 & 0 \\ 0 & 0.33 & 0.33 & 0.5 \\ 0 & 0 & 0.33 & 0.5 \end{bmatrix} = \begin{bmatrix} 35 & 26 & 30 & 27 \\ 32 & 42 & 40 & 44 \\ 15 & 20 & 20 & 24 \end{bmatrix}$$

Fig. 2 a A propagation example of an embedding matrix which contains four one-dimensional joints and three frames. This matrix first convolve with a 3 × 1 kernel and then multiply with the 4 × 4 link matrix. **b** How these four joints link. **c** The directly propagating way of the joints of the first frame

Table 1 Results of different numbers of convolutional layers in HG-GCN on SHREC'17 Dataset in 120 epochs

Layers	10	9	8
Accuracy	89.32	90.88	91.21
Layers	7	6	5
Accuracy	90.55	90.55	90.77

model to more accurately describe the "linkage" motion information of different joints and hence let CNNs learn more semantic features for better gesture recognition.

Firstly, a basic hand skeleton is built by linking joints in each finger, e.g., $(5,4)$, $(4,3)$, $(3,2)$ for thumb, the base of thumb with the wrist, e.g., $(2,0)$, and bases of other fingers with the palm, e.g., $(6,1)$. Then, the three types of edges are to be added: The first type is to link the tip of each finger except the pinkie with the base of the finger to its right, e.g., $(5,6)$; the tip of the pinkie is linked with the base of the ring finger. The second type is to link the third joint of each finger except the pinkie to the second joint of the little finger. The third type is to link the tip and third joint of the same finger. All added edges are shown in Fig. 4b.

The basic hand skeleton provides most of the topology and motion information. The first type of added edges allows the model to measure the distance of two adjacent fingers horizontally and vertically. This type of addition lets the information of one finger propagate to another and hence is able to encode the relationship between them, e.g., overlap or separate. The second type of added edges provides one more original point. It is connected with all fingers and hence better measures the open degree of the hand. The third type of added edges directly provides the information of some actions like grabbing in which fingers bend. The propagation of our model is as follows if these four types of edges are denoted as A_1, A_2, A_3, A_4 in order:

$$f_{\text{out}} = f_{\text{in}} W \cdot F_{rowavg}(A_1 + A_2 + A_3 + A_4); \tag{3}$$

where F_{rowavg} presents the function to make the sum of elements in every row to 1.

2.3 Data embedding strategy

This section discusses two data processing ways in order to address the two problems in recognizing hand gesture. The first presents overfitting by expanding the dimensions of coordinates. The second presents reducing location noises by translating absolute coordinates to relative coordinates.

The public datasets of skeleton-based dynamic hand gesture recognition are usually small and thus are not sufficient to well train a model, which always leads overfitting. Researchers adopt data augmentation strategy like scaling, shifting, time interpolation, and adding noises to alleviate it. In addition to these, we proposed a coordinate conversion way to augment the embedding data. Training sets of DHG-14/28 [16] and SHREC'2017 [17] both provide skeleton joints with three and two-dimensional coordinates. We convert them to spherical and polar coordinates, respectively. Formally,

$$r = \sqrt{x^2 + y^2 + z^2} \quad \theta = arccos\frac{z}{r} \quad \varphi = arctan\frac{y}{x} \tag{4}$$

$$\rho = \sqrt{\tilde{x}^2 + \tilde{y}^2} \quad \tilde{\theta} = arctan\frac{\tilde{y}}{\tilde{x}}; \tag{5}$$

where (r, θ, φ) and $(\rho, \tilde{\theta})$ are vectors of spherical and polar coordinate system, respectively. Due to the insufficiency of training data, the deep neural networks may only learn the feature representation based on the cartesian coordinate, while the semantic features from the spherical or polar coordinates are not well learned when the networks overfit. Incorporating the joint coordinates under the different system will effectively alleviate the overfitting problem, making the embedded encode more

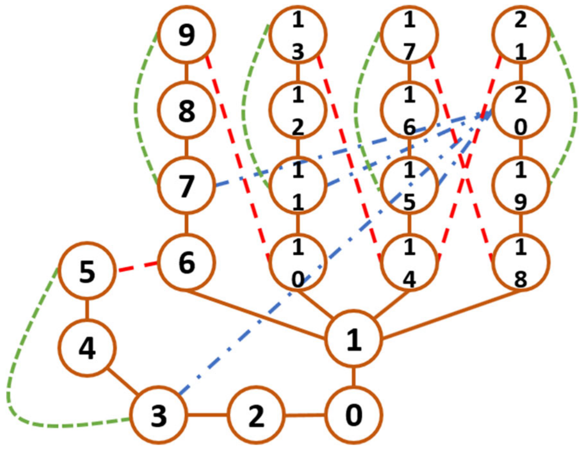


Fig. 3 An illustration of the codes of hand joints. Three types of additional edges colored in red, blue, and green

rotation and distance (length) information. To further analyze the effect of complex coordinate combinations in HG-GCN, an ablation experiment without relative CO is carried out as shown in Table 2. When single kind of coordinates is implemented, word, spherical, image, and polar coordinate yield 84.98%, 72.64%, 86.43%, and 81.20% classification accuracy, respectively. This reveals that Cartesian and 2-D coordinates better describe hand gestures. When working with non-Cartesian coordinates, 2-D and 3-D coordinates improve 2.90% and 3.78% accuracy, respectively. When four kinds work together, the classification accuracy further improve, verifying that each kind of coordinates can be a good supplementary to another one. The results also show that implementing handcraft coordinate transformations helps neural networks to understand the hand gestures. Consequently, the channel number C of our matrix expands from 3 to 10, in which exist 3-D, 2-D, spherical, and polar coordinates, e.g., $\{x, y, z, r, \theta, \varphi, \tilde{x}, \tilde{y}, \rho, \tilde{\theta}\}$.

In addition to the insufficient training data, the actions of hands start from varying places in skeleton-based dynamic hand gesture recognition datasets. The random starting positions have an impact on the recognition performance as part attention of CNNs will be paid to the variation of positions. To address this problem, we propose substituting relative CO for absolute CO. The coordinates of the wrist joint at the first frame are subtracted from the coordinates of all the joints. Formally:

$$\widetilde{X_{vt}} = X_{vt} - X_{00} \quad v \in [0, 21], t \in [0, T-1]; \tag{6}$$

where X denotes the absolute CO vector $\{x, y, z, ...\}$ and $\widetilde{X}$ denotes the relative CO vector. X_{vt} denotes the coordinate vector of v_{th} joint at t frame. By this means, the coordinate vector of the wrist joint at the first frame of every action will be of the form $\{0, 0, 0, ...\}$. Thus, the starting positions of hand gestures in different videos are the same as each other, which relieves the burden on CNNs to exclude the noises by the variation of locations and hence helps CNNs extract features of a better generalization ability.

Table 2 Results of different combinations of the coordinate on SHREC'17 Dataset

Word	Spherical	Image	Polar	Accuracy
✓	×	×	×	84.98
×	✓	×	×	72.64
×	×	✓	×	86.43
×	×	×	✓	81.20
✓	✓	×	×	87.88
×	×	✓	✓	90.21
✓	✓	✓	✓	91.32

The ✓ denotes the implemented operation and × denotes not

3 Results and discussion

We evaluated HG-GCN on two public datasets, DHG-14/28 [16] and SHREC'2017 [17]. The experimental results show the considerable performance of HG-GCN. The two datasets and experimental results are outlined in the rest of this section. Also, analysis of results and limitations of the proposed methods are discussed.

3.1 Datasets

DHG-14/28 Dataset and SHREC'17 Track Dataset are both public dynamic hand gesture datasets. Each of them contains 2800 sequences of 14 hand gestures performed in 2 finger configurations (hence can also be seen as 28 classes). For DHG-14/28, each configuration of one gesture is performed 5 times by 20 participants. For SHREC'17, each configuration of one gesture is performed between 1 and 10 times by 28 participants. Both of them provide coordinates of 22 hand joints in the 3D word space and 2D image space per frame.

3.2 Experimental results

The experiment on DHG-14/28 follows a leave-one-subject-out cross-validation strategy. The final result is the average of the outputs of 20 experiments.

The performance comparisons of HG-GCN with other advanced methods is shown in Table 3. The proposed method achieves 89.2% and 85.3% classification accuracy on 14 gestures and 28 gestures setting, respectively. DSTM [14] is excluded from this experiment for not giving the reference data and experiment result. HG-GCN improves 0.3% accuracy over STA-Res-TCN [12] for the complicated 28 gestures setting, which shows the advantage of the proposed method on recognizing fine gestures. As for the comparison with other methods, HG-GCN shows its advantages obviously.

The experiment on SHREC'17 Track Dataset follows the division of the training set and the testing set that have 1960 training and 840 testing sequences, respectively. The performance comparisons are shown in Table 4. Except for DLSTM which uses hand-crafted angular features

Table 3 Results on DHG-14/28 Dataset

Method	14 gestures	28 gestures
SoCJ+HoHD+HoWR [16]	83.1	80.0
De Smedt et al. [6]	82.5	68.1
CNN+LSTM [9]	85.6	81.1
Chen et al. [11]	84.6	80.3
DPTC [18]	85.8	80.2
STA-Res-TCN [12]	89.2	85.0
HG-GCN	89.2	85.3

The second column describes the accuracy rates of 14 gestures setting. The last column describes the accuracy rates of 28 gestures setting

Table 4 Results on SHREC'17 Dataset

Method	14 gestures	28 gestures
HIF3D [19]	90.4	80.4
De Smedt et al. [6]	88.2	81.9
Devineau et al. [20]	91.2	84.3
STA-Res-TCN [12]	93.6	90.7
DLSTM [14]	97.6	91.4
HG-GCN	92.8	88.3

The second column describes the accuracy rates of 14 gestures setting. The last column describes the accuracy rates of 28 gestures setting

from hand joints as the input, other results are similar to the results on DHG-14/28. HG-GCN yields 92.8% and 86.3% accuracy on 14 gestures and 28 gestures setting, respectively. The worse performance compared with STA-Res-TCN may be caused by the reduced ratio of the training set to the testing set (from 19 : 1 to 7 : 3) [21–23]. In this case, the inverse kinematics used by STA-Res-TCN works more effectively than the polar and spherical coordinates which also need to be learned.

The results shown in Table 5 exhibits the surprising speed of the proposed method. DLSTM [14] is excluded from this comparison for not belonging to an end-to-end structure. The Skeletons/s denotes the number of frames we can calculate per second and the Gestures/s denotes the number of gesture examples we can calculate per second. The proposed method achieves 31605 Skeletons/s and 525.0 Gestures/s, exceeding the speed of CNN + LSTM and STA-Res-TCN by a large part, also far exceeding the standard of real-time analysis (i.e., 30 skeletons per frame) in the video.

The effects of our improvements on 14 gesture setting of SHREC'17 are shown in Table 6. For comparison, ST-GCN without any modification yields nearly 89.0% accuracy. The model with additional edges improves 0.33% accuracy over the baseline ST-GCN. The addition of relative CO and coordinate conversion improve 0.23% and nearly 1.00% accuracy over the model with additional edges, respectively. The aggregation of them improves 3.78% over the baseline ST-GCN. The results not only show that every strategy we adopt is useful, but also reveals that they work well with each other.

Table 5 Speed of the methods

Method	Skeletons/s	Gestures/s
CNN+LSTM [9]	7615	126.5
STA-Res-TCN [12]	9691	161.0
HG-GCN	31605	525.0

The second column describes the max number of skeletons that can be processed by the network per second. The last column describes the max number of gestures that can be processed by the network per second

Table 6 Results of 14 gesture setting on SHREC'17 Dataset

Additional edges	Relative CO	CO conversion	Accuracy
×	×	×	88.99
✓	×	×	90.32
✓	✓	×	90.55
✓	×	✓	91.32
✓	✓	✓	92.77

The ✓ denotes the implemented operation and × denotes not

3.3 Discussion

This section discusses the implications of the findings in the context of existing research and highlights a limitation of the study.

The results in Tables 3 and 4 indicate the strong classification ability of the proposed method for hand gestures and results in Table 5 indicate the fast classification speed of the proposed method. Supporting by these two features, HG-GCN can be implemented into devices for video surveillance, human-computer interaction, robot vision, autonomous driving, and so on and provide precision classification results in real-time.

Although HG-GCN shows intriguing effectiveness, a drawback still limits its better performance which is to be improved in the future. The adjacent matrix is heuristically pre-designed and the same for all layers. Considering that different layers contain different-level semantic information, the adjacent matrix should be adaptively changed for different layers.

4 Conclusions

This paper proposed hand gesture graph convolutional network which is modified from spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition. A special adjacent matrix is designed to be multiplied with feature maps to amend the propagating directions of joint weights. Also, the dimensions of joint coordinates are expanded to better use domain knowledge for CNNs. Moreover, the coordinates are normalized to make every gesture start from the same position. Hand gesture graph convolutional network achieves high accuracy on both two challenge datasets with very fast speed. This shows the effectiveness of the proposed method. As for future development, more complex link ways of joints can be designed for a larger dataset. Other domain knowledge can also be introduced.

Abbreviations

CNNs: Convolutional neural networks; DPTC: Deformable pose traversal convolution; HG-GCN: Hand gesture graph convolutional network; LMC: Leap motion controller sensor; LSTM: Long time short memory; SOCJ: Shape of connected joints; ST-GCN: Spatial temporal graph convolutional networks; STA-Res-TCN: Spatial-temporal attention residual temporal convolutional networks

Acknowledgements

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Authors' contributions

ZHH came up with this idea and did the experiments. YL and XY were the major contributors in writing the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No. NSFC-61471067, 81671651), the National Great Science Specific Project (No. 2015ZX03002008), the Beijing Municipal Natural Science Foundation (No. 4172024), Beijing University of Posts and Telecommunications (No. 2013XD-04, 2015XD-02), and the Beijing Key Laboratory of Work Safety and Intelligent Monitoring Foundation.

Availability of data and materials

Please contact the author for data requests.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Science, Beijing University of Posts and Telecommunications, No.10 Xitucheng Road, Haidian District, Beijing 100876, People's Republic of China. [2]School of Electronic Engineering, Beijing University of Posts and Telecommunications, No.10 Xitucheng Road, Haidian District, Beijing 100876 , People's Republic of China. [3]Ye Peida School of Innovation and Enterprencurship, Beijing University of Posts and Telecommunications, No.10 Xitucheng Road, Haidian District, Beijing 100876, People's Republic of China.

Received: 24 January 2019 Accepted: 19 July 2019
Published online: 13 September 2019

References

1. D. Tang, J. Taylor, P. Kohli, C. Keskin, T. K. Kim, J. Shotton, Opening the black box: hierarchical sampling optimization for estimating human hand pose. Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 3325–3333 (2015)
2. Q. Ye, S. Yuan, T. K. Kim, Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. Eur. Conference on Computer Vision (ECCV), 346–261 (2016)
3. G. Wang, X. Chen, H. Guo, C. Zhang, Region ensemble network: towards good practices for deep 3d hand pose estimation. J. Vis. Commun. Image Represent. **55**, 404–414 (2018)
4. P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, J. Kautz, Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. Comput. Vis. Pattern Recog. (CVPR), 4207–4215 (2016)
5. N. Neverova, C. Wolf, G. Taylor, F. Nebout, Moddrop: Adaptive multi-modal gesture recognition. IEEE Trans. Pattern. Anal. Mach. Intell. **38**, 1692–1706 (2014)
6. Q. D. Smedt, H. Wannous, J. P. Vandeborre, 3d hand gesture recognition by analysing set-of-joints trajectories. Eurographics Work. 3D Object Retr., 86–97 (2017)
7. X. Yang, Y. Tian, Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops, 14–19 (2012)
8. H. Chen, G. Wang, J. Xue, A novel hierarchical framework for human action recognition. Pattern Recognit. **55**, 148–159 (2016)
9. J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemyaor, J. F. Velez, Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognit. **76**, 80–96 (2018)
10. A. Graves, Long Short-Term Memory. Supervised Sequence Labelling Recurrent Neural Netw., 37-45 (2012). Springer Berlin Heidelberg
11. X. Chen, H. Guo, G. Wang, L. Zhang. Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition, (2017), pp. 2881–2885
12. J. Hou, G. Wang, X. Chen, J. Xue, R. Zhu, H. Yang, Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition. Proc. Fourth Int. Work. Observing Underst. Hands Action, 273–286 (2018)
13. C. Lea, M. D. Flynn, R. Vidal, A. Reiter, G. D. Hager, Temporal convolutional networks for action segmentation and detection. Comput. Vis. Pattern Recognit. (CVPR), 1003–1012 (2017)
14. D. Avola, M. Bernardi, L. Cinque, G. L. Foresti, C. Massaroni, Exploiting recurrent neural networks and leap motion controller for sign language and semaphoric gesture recognition. Computer Vision Pattern Recognition, 234–245 (2018)
15. S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition. Association for the Advance of Artificial Intelligence (AAAI), 7444–7452 (2018)
16. Q. D. Smedt, H. Wannous, J. P. Vandeborre, Skeleton-based dynamic hand gesture recognition. Comput. Vis. Pattern Recog. Workshops, 1206–1214 (2016)
17. Q. D. Smedt, H. Wannous, J. P. Vandeborre, J. Guerry, B. L. Saux, D. Filliat, Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset. Eurographics, 86–97 (2018)
18. J. Weng, M. Liu, X. Jiang, J. Yuan, Deformable pose traversal convolution for 3d action and gesture recognition. Eur. Conf. Comput. Vis. (ECCV), 142–157 (2018)
19. S. Y. Boulahia, E. Anquetil, F. Multon, R. Kulpa, in *International conference on image processing*. Dynamic hand gesture recognition based on 3D pattern assembled trajectories, (2017), pp. 1–6
20. G. Devineau, F. Moutarde, W. Xi, J. Yang, in *IEEE International Conference on Automatic Face Gesture Recognition*. Deep Learning for Hand Gesture Recognition on Skeletal Data, (2018), pp. 106–113
21. C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, Q. Dai, A fast uyghur text detector for complex background images. IEEE Trans. Multimed. **20**, 1–1 (2018)
22. C. Yan, L. Li, C. Zhang, B. Liu, Q. Dai, Cross-modality bridging and knowledge transferring for image understanding. IEEE Trans. Multimed, 1–1 (2019)
23. C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Q. Dai, Stat: Spatial-temporal attention mechanism for video captioning. IEEE Trans. Multimed, 1–1 (2019)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.