Spatio-Temporal Sparse Graph Convolution Network for Hand Gesture Recognition

Omar Ikne, Rim Slama, Hichem Saoudi, Hazem Wannous

2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Spatio-Temporal Sparse Graph Convolution Network for Hand Gesture Recognition

Omar Ikne[1], Rim Slama[2], Hichem Saoudi[1] and Hazem Wannous[1]
[1] IMT Nord Europe, Institut Mines-Télécom, Univ. Lille, Centre for Digital Systems, F-59000 Lille, France
[2] LINEACT Laboratory, CESI Lyon, 69100 Villeurbanne, France

***Abstract*— Unlike whole-body action recognition, hand gestures involve spatially closely distributed joints, promoting stronger collaboration. This needs to be taken into account in order to capture complex spatial and temporal features. In response to these challenges, this paper presents a Spatio-Temporal Sparse Graph Convolution Network (ST-SGCN) for dynamic recognition of hand gestures. Based on decoupled spatio-temporal processing, the ST-SGCN incorporates Graph Convolutional Networks, attention mechanism and asymmetric convolutions to capture the nuanced movements of hand joints. The key novelty is the introduction of sparse spatio-temporal directed interactions, overcoming the limitations associated with dense, undirected methods. The sparse aspect models essential interactions between hand joints selectively, improving computational efficiency and interpretability. Directed interactions capture asymmetrical dependencies between hand joints, improving discernment of joint influences. Experimental evaluations on three benchmark datasets, including Briareo, SHREC'17 and IPN Hand, demonstrate ST-SGCN's state-of-the-art performance for dynamic hand gesture recognition. Codes are available at: `https://github.com/HichemSaoudi/ST-SGCN`.**

I. INTRODUCTION

In recent years, hand gesture recognition has become a key focus of research, due to its crucial role in improving video comprehension and supporting Natural User Interfaces (NUI)[12]. NUIs represent a transformative approach to human-computer interaction, shifting the reliance from traditional input devices, such as keyboards and mice, to the user's bodily expressions. In this context, accurate recognition of dynamic hand gestures, encompassing the fusion of poses and static hand movements, is of paramount importance, especially when based on 3D skeleton data.

Various methods have been explored for dynamic gesture recognition with skeletal data, including CNNs, RNNs, LSTMs [15], and recently, GCN-based approaches [21], [18], [5]. This shift towards GCN-based methods is largely attributed to their ability to incorporate spatial connectivity among hand joints, enhancing hand gesture recognition.

Spatio-temporal modeling has become increasingly important, particularly with Yan et al.'s Spatio-Temporal Convolutional Graph Networks (STGCNs) [21]. STCGN was mainly introduced to learn spatial and temporal dependencies between different body/hand joints within and between different frames in a sequence. This pioneering work has inspired subsequent advancements in the field of dynamic hand gesture recognition [24], [11], [19]. In the other hand, transformers, originally designed for NLP, have recently shown promise in this task, by their application in dynamic hand gesture recognition [24], [9]. Some approaches have focused on learning spatial and temporal graph adjacencies based on input graphs [2]. For example, in [1] authors introduced physical constraints to alter joint connections, and DG-STA [5] employed the attention mechanism to construct dynamic graphs.

Most recent attention-based methods [20] often use dense interaction to model interactions between hand joints, and assume that each joint interacts with all others. These methods generally represent interactions as undirected, treating interactions between joints as identical. However, we argue that these dense, undirected approaches introduce irrelevant interactions between different joints, ignoring the interlinked spatial and temporal trajectories of hand joints. It is obvious that conventional methods based on dense or undirected interactions will fail to capture these nuanced interactions effectively. To overcome these challenges, we propose a new approach that incorporates sparse spatio-temporal directed interactions, capturing both the directed sparse spatial and temporal interactions of the hand joints.

In this paper, we present a compact approach for dynamic hand gestures recognition using skeletal data. Our innovative approach called Spatio-Temporal Sparse Graph Convolution Network (ST-SGCN) puts a strong emphasis on the essential aspects of sparse and directed interactions between hand joints in a dynamic sequence. Our ST-SGCN initially employs a two-stream pipeline to process spatial and temporal graphs independently, underlining the decoupled aspect of our model. Subsequently, these streams merge into an integrated representation, forming a unified sparse and directed spatio-temporal graph that encompasses the dynamics of the gesture sequence. This end-to-end pipeline leverages attention mechanism, asymmetric convolutions and the strengths of CGNs to significantly improve the robustness of hand gesture recognition.

II. METHODOLOGY

In this section, we present our end-to-end ST-SGCN pipeline, as illustrated in figure 1. The process flow comprises three main steps. First, our ST-SGCN engages in decoupled learning of directed and sparse spatial and temporal interactions from a provided skeleton data sequence, using attention mechanism and asymmetric convolutions. The second step consists in acquiring cross-spatial-temporal graph representations for the given sequence, capturing both

979-8-3503-9494-8/24/$31.00

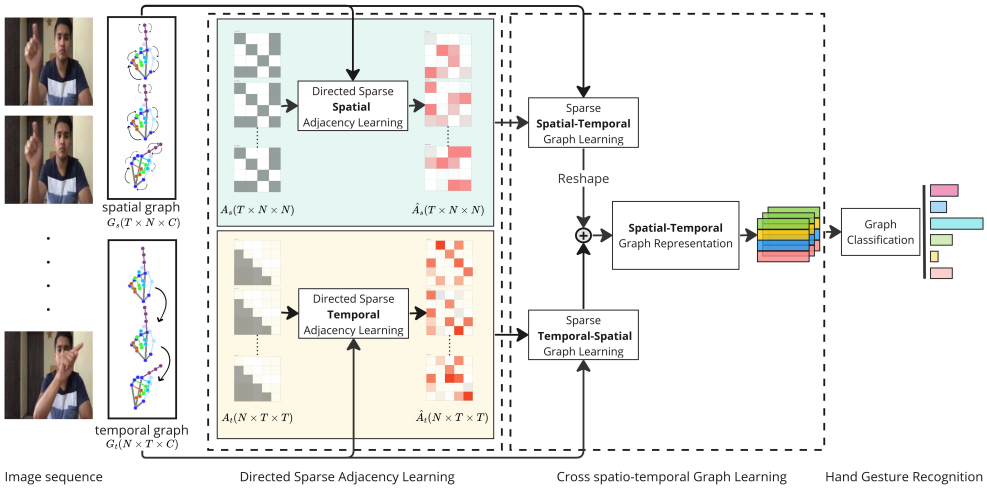

Fig. 1. **The proposed ST-SGCN approach**. Given a sequence of hand skeleton data extracted from RGB images, we initialize the spatial (A_s) and temporal (A_t) adjacency matrices, then refine them into learned matrices ($\hat{A}_s$ and $\hat{A}_t$) using directed sparse adjacency learning. Next, we introduce a cross-spatio-temporal module to capture the spatio-temporal graph representation. Finally, the resulting representation is processed by an MLP to produce the corresponding gesture.

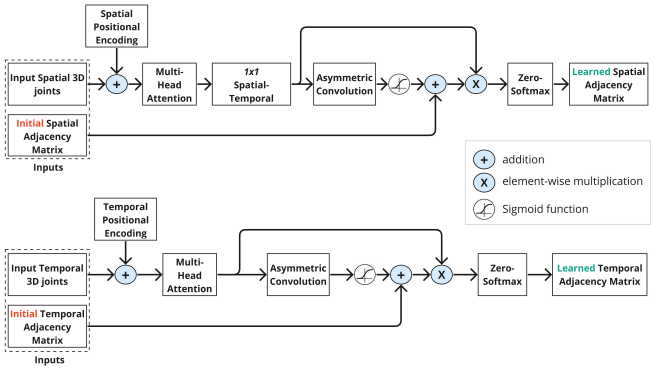

Fig. 2. Spatial (top) and temporal (bottom) directed sparse adjacency matrix learning process.

spatial-temporal and temporal-spatial interactions. Finally, the resulting graph representation is processed by a Multi-Layer Perceptron (MLP) to yield corresponding gesture. This structured approach ensures in-depth exploration of spatial and temporal dependencies, facilitating accurate and robust dynamic prediction of hand gestures.

A. Directed Sparse Adjacency Learning

The process of learning directed sparse adjacency aims to enhance the representation of interactions, encoded by the adjacency matrices, among distinct hand joints, as shown in Fig. 2. Both pipelines presented in Fig. 2 share a similar structure, differing only in the input data (spatial and temporal graphs) and the positional encoding applied to the spatial and temporal streams. The following is an in-depth explanation of the underlying process.

a) Spatial and Temporal Graphs: A 3D hand skeleton sequence X is denoted by 3D coordinates for each hand joint at each frame, represented as $X \in \mathbb{R}^{T \times N \times 3}$. Here, T is the number of frames, N indicates the number of joints in the skeleton data, and 3 represents the (x, y, z) coordinates.

Given an input sequence $X \in \mathbb{R}^{T \times N \times 3}$, we create spatial ($G_s$) and temporal ($G_t$) graphs by performing axis permutation on the input data, considering the first dimension as the batch dimension, and the subsequent dimensions corresponding to spatial/temporal and coordinates. Consequently, $G_s \in \mathbb{R}^{T \times N \times 3}$, and $G_t \in \mathbb{R}^{N \times T \times 3}$. The former emphasizes spatial interactions between hand joints (intra-frame interactions), independent of time, while the latter focuses on the temporal aspect, capturing the evolution of a specific hand joint independently over the time axis (inter-frame interactions).

An extra feature, incorporating joint connectivity information, particularly the *centrality encoding* [23], indicating the number of joints connected to a given joint, is added to the feature vector of each node. This inclusion aims to augment each node with its neighborhood information. A scaling step is performed to prevent disproportionality with regard to Euclidean coordinates (x, y, z).

b) Positional Encoding: Accurate positional information, both in spatial and temporal dimensions, is crucial when dealing with dynamic graphs encoded by skeletal data.

For spatial positional encoding, each joint is encoded with an integer indicating its position in the hand skeleton. While for temporal encoding, frames inherently lack attributes that indicate their positions. To address this, we introduce a positional encoding designed to assign a unique marker to each frame. Inspired by [17], we use sine and cosine functions with distinct frequencies as encoding functions, ensuring that joints in the same frame are identically encoded.

c) Directed Sparse Adjacency: In this stage, our objective is to introduce sparsity and directionality into the interactions between nodes in both the spatial and temporal graphs, where these interactions are represented by the adjacency matrices. The goal is to transform these matrices into directed and sparser forms.

We begin the process by binary initializing ($\{0, 1\}$) the spatial (A_s) and temporal (A_t) adjacency matrices, each following a distinct approach. The spatial matrix is initialized based on the natural topology of the hand, while the temporal matrix is initialized as an upper triangular matrix, signifying that each frame influences its subsequent frames.

Next, we refine the initialized adjacency matrices in two steps. Initially, a multi-head attention mechanism is employed to capture dense spatial ($I_s \in \mathbb{R}^{N\times N}$) and temporal ($I_t \in \mathbb{R}^{T\times T}$) interactions among the hand joints.

Driven by the intuition that the influence of a node i on a node j differs from that of j on i when modeling gestures dynamics, we choose asymmetric over symmetric convolution. Our ablation study empirically validates this choice. Consequently, a series of asymmetric convolution kernels is applied to both spatial (G_s) and temporal (G_t) graphs, along with I_s and I_t, yielding the corresponding directed spatial (DA_s) and temporal (DA_t) adjacency matrices.

To further refine these interactions, we generate a binary mask M by filtering out scores from DA_s and DA_t that fall below a user-defined hyperparameter, represented as the threshold $\epsilon \in [0, 1]$. In both the spatial and temporal domains, the masks, M_s and M_t, are defined as follows: $M_s = \{1 \mid DA_s > \epsilon\}$ and $M_t = \{1 \mid DA_t > \epsilon\}$, respectively.

In line with the recommendation in [16], we adopt the Zero-Softmax activation function. This choice is motivated by the propensity of Softmax to yield non-zero values for zero inputs, which could counteract the sparsity achieved in the adjacency matrices. The zero-softmax is defined as:

$$\text{Zero-Softmax}(x) = \frac{(\exp(x) - 1)^2}{\sum_i (\exp(x_i) - 1)^2 + \epsilon} \tag{1}$$

where x is an input value, ϵ is a small constant for stability.

Subsequently, we construct the sparse directed adjacency matrices, $\hat{A}_s$ and $\hat{A}_t$, as follows:

$$\hat{A}_s = \text{Zero-Softmax}\left((M_s + A_s) \odot \hat{DA_s}\right) \tag{2}$$

$$\hat{A}_t = \text{Zero-Softmax}\left((M_t + A_t) \odot \hat{DA_t}\right) \tag{3}$$

Here, A_s and A_t represent the initial spatial and temporal adjacency matrices. $\odot$ is the element-wise multiplication.

B. Cross Spatio-Temporal Graph Representation Learning

Having obtained the sparse spatial and temporal adjacency matrices through learning, our objective is to construct a comprehensive spatio-temporal graph representation for the given skeleton sequence. This involves the utilization of two separate yet parallel networks: the spatio-temporal graph representation (STGR), which is dedicated to capturing interaction-pattern features denoted as L_{IP}, and the temporal-spatial graph representation (TSGR), designed to capture pattern-interaction features represented as L_{PI}. The implementations of STGR and TSGR are detailed as follows:

$$L_{IP}^{(l)} = \delta\left(\hat{A}_t \cdot \delta\left(\hat{A}_s L_{PI}^{(l-1)} W_{s_1}\right) W_{t_1}\right) \tag{4}$$

$$L_{PI}^{(l)} = \delta\left(\hat{A}_s \cdot \delta\left(\hat{A}_t L_{IP}^{(l-1)} W_{t_2}\right) W_{s_2}\right) \tag{5}$$

Here, $\hat{A}s$ and $\hat{A}t$ represent the learned spatial and temporal adjacency matrices, while $Ws_1, Ws_1, W_{t_1}, W_{t_2}$ correspond to the weights of the GCN. The variable l signifies the l^{th} layer of the GCN, and $\delta(\cdot)$ denotes a non-linear activation function. The learned STGR and TSGR representations are merged to reconstruct the global spatio-temporal representation. The merging process consists of reshaping the spatial-temporal graph representation and subsequently adding it to the temporal-spatial graph representation.

C. Graph Classification and Loss Function

The resulting spatio-temporal graph representation captures crucial features from the input sequence. In the learning process, this representation undergoes mean pooling along the attention dimension followed by a classification head represented by an MLP layer, enabling the model to categorize the sequence into distinct hand gestures. Training is performed using the Cross-Entropy loss, and the model's performance is evaluated based on recognition accuracy.

III. Experimental results

In this section, we cover evaluation datasets, ablation studies and comparison with state-of-the-art methods.

A. Datasets

a) SHREC'17 TRACK [6]: It contains sequences of 14 hand gestures performed in two ways: using one finger and the whole hand. Each gesture is performed between 1 and 10 times by 28 participants resulting in 2800 sequences.

b) Briareo [13]: It includes 12 gestures performed by 40 subjects using their right hand, with each gesture repeated thrice. The dataset comprises a total of 1440 sequences. Subjects 1 to 26 are allocated for training, 27 to 32 for validation, and 33 to 40 for testing.

c) IPN Hand [3]: It contains over 4,000 gesture instances from 50 subjects. Each subject continuously performed 21 gestures with three random breaks in a single video. 13 gestures are defined to control the pointer and actions focused on the interaction with touchless screens.

B. Implementation Details

In our experimental setup, we conducted our experiments using two NVIDIA GeForce RTX 3090 GPUs. We employed the AdamW optimizer. The hyperparameters include the initial learning rate set at 0.001, which is adjusted gradually during training. The training process lasted for 200 epochs on all datasets. For the model architecture, we incorporated 6 asymmetric convolutions, a multi-head attention with 4 heads, set the dimensionality d_{model} to 64, and applied a mask threshold of 0.5. To prevent overfitting, we introduced dropout with a rate of 0.5 as a regularization measure. The choice of the sequences length is dataset-specific, with 60 frames used for SHREC'17 and Briareo and 80 frames for the IPN Hand dataset. To enhance generalization, we implemented data augmentation through random moving [21]. This technique involves applying random affine transformations to the sequence, simulating changes in viewpoint angles.

TABLE I

ACCURACY (%) COMPARISON WITH STATE-OF-THE-ART METHODS ON EVALUATION DATASETS.

BRIAREO DATASET.

Method	Modality	Accuracy
C3D-HG [13]	ir	87.5%
TBN-HGR [7]	ir + normals	97.2%
LSTM-HG [13]	Skeleton	94.4%
3D-Jointsformer [25]	Skeleton	95.4%
Ours	Skeleton	**98.0%**

IPN HAND DATASET.

Method	Modality	Accuracy
ResNet-50 [3]	RGB-Seg	75.1%
C3D [10]	RGB	77.7%
ResNeXt-101 [3]	RGB-Flow	86.3%
Dist-Time [8]	Skeleton	87.5%
Ours	Skeleton	**89,0%**

SHREC'17 DATASET (ONLY SKELETON).

Method	Accuracy
STA-RES-TCN [9]	93.6%
ST-GCN [21]	92.7%
DG-STA [5]	94.4%
DD-NET [22]	94.6%
FPPR-PCD [4]	96.1%
DSTA-Net [17]	**97.0%**
Ours	92,9%

C. Ablation study

In our ablation study, conducted on the SHREC'17 dataset, we aim to investigate the impact of different components of our model, specifically focusing on the directed interactions (DI) learned through asymmetric convolutions, and the sparse interactions (SI). The results are provided in table II.

TABLE II

ABLATION STUDY ON DIRECTED INTERACTIONS (DI) AND SPARSE INTERACTIONS (SI) ON SHREC'17 TRACK DATASET.

Method	Accuracy
ST-SGCN w/o DI, w/o SI	91.3%
ST-SGCN w/o DI	92.3%
ST-SGCN	**92.9%**

As shown in table II, removing both directed and sparse interactions (ST-SGCN w/o DI, w/o SI), i.e. using non-directed dense interactions, leads to an accuracy of 91.3%. When sparse interactions are incorporated without directed interactions (ST-SGCN w/o DI), denoting the use of undirected sparse interactions, the recognition accuracy rises to 92.3%. Finally, our full model incorporating both directed and sparse interactions (ST-SGCN) achieves the highest accuracy, at 92.9% underlining the valuable contribution of these components to improving model's performance.

D. Comparison with state-of-the-art and discussion

We conducted an extensive performance comparison of our method against recent state-of-the-art methods based on recognition accuracy and an in-depth performance analysis.

a) Recognition accuracy comparison: The results, presented in Table I, underline the potential of our method on evaluation datasets. Our approach yields competitive results on the SHREC'17 dataset and achieves an accuracy of 98.0% on the Briareo dataset and 89.0% on the IPN Hand dataset, outperforming state-of-the-art methods.

In the Briareo and IPN Hand datasets, our approach demonstrates superior performance, relying exclusively on 3D coordinates of the hand joints (skeleton), compared to methods using either 3D joints coordinates, or additional modalities such as RGB, Flow or Depth.

The lower accuracy observed on the SHREC'17 dataset can be attributed to the dataset's mixed nature, featuring both coarse and fine gestures. Our model faces challenges, especially for gestures like Tap, Expand, and Pinch. Coarse gestures like Tap may lack distinctive features for precise differentiation within our spatio-temporal graph approach. Additionally, fine gestures like Expand and Pinch involve intricate hand movements, requiring precision in modeling sparse directed interactions. To enhance performance on SHREC'17, further refinement in capturing fine-grained details and improving the model's ability to discern subtle variations in hand dynamics is crucial.

b) Performance Analysis: We carried out a performance analysis of our method in comparison with existing state-of-the-art methods on the Briareo dataset. This analysis covers key parameters such as the number of model parameters, inference time and video RAM (VRAM) demand on the graphics card. The results are presented in Table III.

TABLE III

PERFORMANCE ANALYSIS ON BARIAREO DATASET.

Method	Parameters(M)	Inference(ms)	VRAM(GB)
R3D-CNN [14]	38,0	30,0	1.3
C3D-HG [13]	26.7	55,0	1,0
TBN-HGR [7]	24.3	26.7	1.8
3D-Jointsformer [25]	**8.8**	16.4	-
Ours	10.4	**12.5**	**0.1**

Our model stands out in terms of efficiency, with the lowest number of parameters (10.4 million), guaranteeing high recognition accuracy in a lightweight design. It achieves a fast inference time of 12.5 milliseconds, making it suitable for real-time applications, and requires minimal VRAM usage (0.1 gigabytes). This efficiency extends its deployment compatibility to various hardware configurations, even those with limited VRAM capacity.

IV. CONCLUSION AND FUTURE WORK

In this paper, we introduced an innovative approach for dynamic hand gesture recognition. Our approach incorporated sparse spatio-temporal directed interactions, capturing both the directed sparse spatial and temporal interactions of the hand joints, effectively addressing the limitations of dense undirected interaction methods. Our experiments demonstrated the state-of-the-art performance of our approach on benchmark datasets, including Briareo and IPN Hand.

In future work, we aimed to develop a real-time implementation of our model to support untrimmed sequences and interactive applications in RA/RV context.

ACKNOWLEDGEMENT

This work is co-funded by the AI@IMT program of the Agence Nationale de la Recherche (ANR) and the region Hauts-de-France in France.

REFERENCES

[1] T. Ahmad, L. Jin, L. Lin, and G. Tang. Skeleton-based action recognition using sparse spatio-temporal gcn with edge effective resistance. *Neurocomputing*, 423:389–398, 2021.

[2] F. Al Farid, N. Hashim, J. Abdullah, M. R. Bhuiyan, W. N. Shahida Mohd Isa, J. Uddin, M. A. Haque, and M. N. Husen. A structured and methodological review on vision-based hand gesture recognition system. *Journal of Imaging*, 8(6):153, 2022.

[3] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, and K. Yanai. Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *25th International Conference on Pattern Recognition, ICPR 2020, Milan, Italy, Jan 10–15, 2021*, pages 4340–4347. IEEE, 2021.

[4] A. Bigalke and M. P. Heinrich. Fusing posture and position representations for point cloud-based hand gesture recognition. In *2021 International Conference on 3D Vision (3DV)*, pages 617–626. IEEE, 2021.

[5] Y. Chen, L. Zhao, X. Peng, J. Yuan, and D. N. Metaxas. Construct dynamic graphs for hand gesture recognition via spatial-temporal attention. *arXiv preprint arXiv:1907.08871*, 2019.

[6] Q. De Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, B. L. Saux, and D. Filliat. 3d hand gesture recognition using a depth and skeletal dataset: Shrec'17 track. In *Proceedings of the Workshop on 3D Object Retrieval*, 3Dor '17, page 33–38, Goslar, DEU, 2017. Eurographics Association.

[7] A. D'Eusanio, A. Simoni, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara. A transformer-based network for dynamic hand gesture recognition. In *2020 International Conference on 3D Vision (3DV)*, pages 623–632. IEEE, 2020.

[8] G. Fronteddu, S. Porcu, A. Floris, and L. Atzori. A dynamic hand gesture recognition dataset for human-computer interfaces. *Computer Networks*, 205:108781, 2022.

[9] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang. Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[10] O. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[11] Y. Li, Z. He, X. Ye, Z. He, and K. Han. Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition. *EURASIP Journal on Image and Video Processing*, 2019(1):1–7, 2019.

[12] W. Liu. Natural user interface- next mainstream product user interface. In *2010 IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design 1*, volume 1, pages 203–205, 2010.

[13] F. Manganaro, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara. Hand gestures for the human-car interaction: The briareo dataset. In E. Ricci, S. Rota Bulò, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, editors, *Image Analysis and Processing – ICIAP 2019*, pages 560–571, Cham, 2019. Springer International Publishing.

[14] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016.

[15] J. P. Sahoo, A. J. Prakash, P. Pławiak, and S. Samantray. Real-time hand gesture recognition using fine-tuned convolutional neural network. *Sensors*, 22(3):706, 2022.

[16] L. Shi, L. Wang, C. Long, S. Zhou, M. Zhou, Z. Niu, and G. Hua. SGCN: sparse graph convolution network for pedestrian trajectory prediction. *CoRR*, abs/2104.01528, 2021.

[17] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[18] R. Slama, W. Rabah, and H. Wannous. Str-gcn: Dual spatial graph convolutional network and transformer graph encoder for 3d hand gesture recognition. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2023.

[19] J.-H. Song, K. Kong, and S.-J. Kang. Dynamic hand gesture recognition using improved spatio-temporal graph convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6227–6239, 2022.

[20] A. Ulhaq, N. Akhtar, G. Pogrebna, and A. Mian. Vision transformers for action recognition: A survey. *arXiv preprint arXiv:2209.05700*, 2022.

[21] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[22] F. Yang, Y. Wu, S. Sakti, and S. Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the ACM multimedia asia*, pages 1–6. 2019.

[23] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu. Do transformers really perform bad for graph representation?, 2021.

[24] W. Zhang, Z. Lin, J. Cheng, C. Ma, X. Deng, and H. Wang. Sta-gcn: two-stream graph convolutional network with spatial–temporal attention for hand gesture recognition. *The Visual Computer*, 36:2433–2444, 2020.

[25] E. Zhong, C. R. del Blanco, D. Berjón, F. Jaureguizar, and N. García. Real-time monocular skeleton-based hand gesture recognition using 3d-jointsformer. *Sensors*, 23(16), 2023.