Gesture Recognition Techniques

Varun Sharma
Department Computer Science
Liverpool John Moores University
Liverpool, UK

Hoshang Kolivand
School of Computer Science and Maths
Liverpool John Moores University
Liverpool, UK
h.kolivand@ljmu.ac.uk

Shiva Asadianfam[*]
Faculty of Electrical & Computer Engineering, Qom University of Technology, Qom, Iran
Department of Computer Engineering
Islamic Azad University Qom Branch, Qom, Iran
Sh_asadianfam@yahoo.com

Dhiya Al-Jumeily
School of Computer Science and Maths
Liverpool John Moores University
Liverpool, UK

Manoj Jayabalan
School of Computer Science and Maths
Liverpool John Moores University
Liverpool, UK

***Abstract*—Gesture recognition is a topic in computer science and language technology with the goal of interpreting human gestures via mathematical algorithms. It is a subdiscipline of computer vision. In this paper, we describe some of Gesture recognition techniques such as Vision based gesture recognition and Graph based gesture recognition. Also, we explore these techniques with previous studies.**

Keywords—Gesture recognition, Vision based gesture recognition, Graph based gesture recognition, computer vision.

I. Introduction

Gesture recognition is a type of perceptual computing user interface that allows computers to capture and interpret human gestures as commands. The general definition of gesture recognition is the ability of a computer to understand gestures and execute commands based on those gestures. For example, imagine being able to check your home security camera as you drive home by simply making a hand gesture. Gestures could also be coupled with telematics systems, allowing the vehicle to provide information about nearby landmarks if it recognizes that an occupant is pointing at it[1].

Gesture recognition is the fast growing field in image processing and artificial technology. The gesture recognition is a process in which the gestures or postures of human body parts are identified and are used to control computers and other electronic appliances. In this paper, we describe some of Gesture recognition techniques such as Vision based gesture recognition and Graph based gesture recognition.

This paper is structured as follows. In Section 2, Recent Gesture Recognition techniques in this paper is discussed. In Section 3, the related studies on vision based gesture recognition are described in detail. In Section 4, the related studies on graph based recognition are described. In Section 5, the summary of two techniques are explain. Finally, the general conclusion of this paper is explained in section 6.

II. Recent Gesture Recognition techniques

With ever evolving technology, new methods are being developed and tested in the field of Gesture Recognition. The Gestures can be divided into three sections,

- Static 2D – simplest form and shape is basically used to identify gesture example fist or fingers. They are simpler poses complexities when tracking is required.
- Dynamic 2D – It’s enhancement to Static 2D recognition, where hand trajectories different features and their various combinations are used.
- 3D – Kinect sensor which allowed to capture depth in an image brough a revolution, depth map helps in identifying distance of all pixels from surface of the image, with many advantages of color images.

This section aims to detail down the latest Gesture Recognition techniques, pros, and cons of same.

III. Vision Based Gesture Recognition

Considering the advantages of 3D images, many studies were done on 3D images and methodologies were evolved to get better accuracy.

Zhu et al. [2], 3D shape context was used to represent 3D hand gestures. This basically was a technique to gather image information utilizing local shape context and global shape distribution of each 3D point. Once hand gestures are constructed using these 3D shape context, dynamic time warping algorithm was used to identify the gesture. Figure 2, shows pictorial representation of the method.

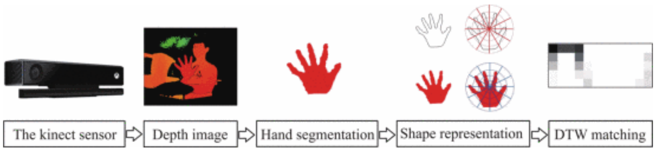


Figure 2. Representation of hand gesture using histogram and DTW algorithm [2]

Advantages of this algorithm includes robustness towards noise, articulated variations, and rigid transformation, speed (no need of GPU), application in real-time scenarios; Improvement to DTW algorithm by using Chi-Square coefficient[2].

Table 1. 3D shape context - Performance on datasets

Dataset	Accuracy (%)
NTU Hand Digit Dataset	98.7
Kinect Leap Dataset	96.8
Senz3d Dataset	99.6
ASL-FS Dataset	87.1
ChaLearn LAP IsoGD Dataset	60.12

Despite performing well on many datasets, it was found that the accuracy was not good enough on many other datasets and it can be concluded that the approach though outperformed many other algorithms was not good enough to be generalized and will not be the best option in real-time applications.

A new experiment was done, utilizing NAO robot to evaluate the effectiveness of the model. The said technique used Leap Motion [3] to gather data and applied Kalman filter to the original data to remove unwanted noise. From the original coordinates there were three new features that were extracted and used namely, angle feature, angle velocity feature and length feature. Finally, these extracted features were fed to LSTM-RNN network to predict the gesture. Below Figure 3 is the pipeline demonstrating the overall architecture.

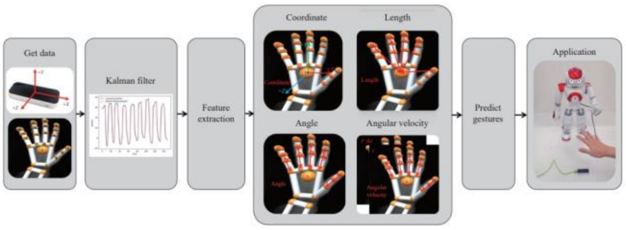


Figure 1. LSTM-RNN network pipeline [3]

Several experiments were done to find the accuracy of the model and it was found that using just 3D positions of finger joints resulted in 97.93% accuracy, only length resulted in 95.17%, Angle in 93.79% and Angular Velocity resulted in very less accuracy i.e. 79.31% due to speed the player used in movements. However, when all features were used together, highest accuracy of 99.31% was achieved.

Though the accuracy achieved in the research is very good but needs to be verified on other datasets and is lacking in the current research. Another aspect is noise introduced in Leap Motion and is something that can be studies further and instead of Kalman filter, other filters can be tried, and overall impact can be studied.

IV. Graph Based Gesture Recognition

In above approaches we saw how an image can be used to extract feature and deep learning techniques like LSTM etc can be applied. Feature creation is one of the most important facts and a good feature can help train a model and achieve a very good accuracy. Skeleton data now a days is widely used due to their robustness to accommodate complex and dynamic circumstances in action recognition. Along with feature creation and trying different datasets and data creation techniques, new techniques need to be evolved and one of such technique was use of Graph based techniques. In this section we will study different graph-based techniques that have been applied and pros and cons of same.

Conventional methods utilizing skeleton data relied heavily on hand crafted features to extract skeleton information, however with development of Deep Learning methods like CNN, RNN and GCN's, more advance mechanism was derived.

Earlier, joint information was used from skeleton data and temporal analysis was done for action recognition, however since they were lacking utilization of spatial information which is crucial in action determination, a new methodology ST-GCN [4] was developed which uses both spatial and temporal information to identify the action. Below Figure 4 shows graph capturing spatial-temporal information.

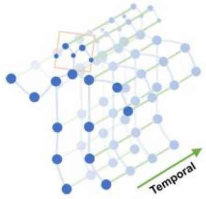


Figure 2. The spatial temporal graph of a skeleton sequence [4]

From the figure, we can infer that there are two types of edges, one form basis of spatial i.e.. Natural connectivity of joins and other that connects joins within different time frame and called as temporal edge. Below Figure 5 visually describes the method used in ST-GCN.

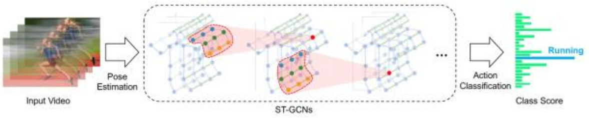


Figure 3. Construct spatial temporal graph on skeleton sequences [4]

ST-GCN was evaluated over two datasets namely Kinetics and NTU-RGB+D, ST-GCN reached accuracy of 52.8% and 88.3% respectively. The technique opened doors to experiments using GCN and further studies were done on how to create feature from skeleton data.

Though ST-GCN proved to be promising method, but it was found that node interaction does not necessarily provide the complementary required information, it also introduces possibility of noise. It was also found that use of GCN can become over-smoothing when multi-layer GCN is used.

To overcome the issues of GCN, a new methodology ST-GDN [5] was proposed. This method provides a better aggregation of messages by removing embedding redundancy, it addresses GCN's over-smoothing problem. High level working of this method is shown in Figure 6.

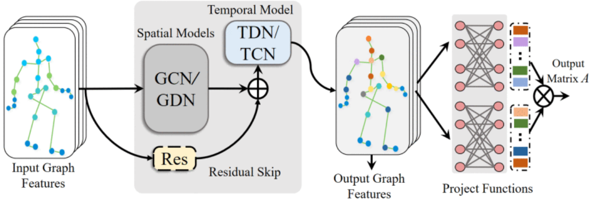


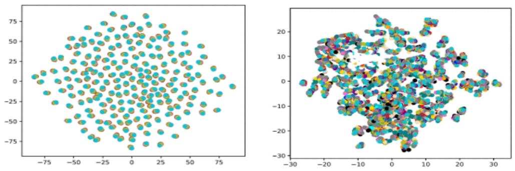
Figure 4. Illustration of the ST-GDNs block[5]

ST-GDNs basically has four building blocks, listed in table below with high level working

Table 2. ST-GDNs building blocks

Node-wise ST-GDN (ST-GDN2)	Represents the features in new feature-space (coordinates changed), the feature embeddings are standardized and correlation is removed, thereby removing the over-smoothing problem
frame-wise ST-GDN (ST-GDN-T)	Like ST-GDN2
element-wise ST-GDN (ST-GDN-E)	Like ST-GDN2
combination of GCN and GDN (ST-GDCN)	Graph representation learning is enhanced by use to of convolutional/deconvolutional feature embeddings

Figure 7 shows feature creation by both ST-GCN and ST-GDN methods [4, 5]



(a) Feature with ST-GCN

(b) Feature with ST-GDN^2

Figure 5. Feature creation

As we can see that in Figure 7 (a) we find that it is hard to distinguish nodes as they have similar representation, on the contrary when coordinates are changed, nodes can be easily distinguished hence confirming ST-GDN2, alleviate over-smoothing problem [4, 5].

Above method was evaluation over many challenging datasets and listed are accuracies against each.

- NTU RGB+D dataset – 95.9%
- NTU RGB+D 120 dataset – 82.3%
- Kinetics-skeleton dataset – 60.5%

Although, ST-GCN were successful in hand gesture recognition, it was found that they had certain limitation like usage of fixed graph be it spatial based on hand skeleton tree or temporal dimension which restricts hand gesture recognition. Actions like touching forehead is different from clapping or jumping, these examples strongly indicates that graph structure should be data dependent and this problem is not solved in ST-GCN. In order to overcome these issues two-stream graph attention convolutional network with spatial–temporal attention was proposed [6]which was based on [7]

Beginning with Adaptive Graph Convolutional Network [7] , it basically constructs two types of graphs, one can be called as global graph which represents common data patterns and other is unique to local(each) data points. These two graphs are then optimized individually which helps in better fitting of the model's hierarchical structure. ST-GCN focusses on first order information which is feature vector of vertex, but it does not consider second order information which is basically feature of bones between joints. This information is crucial as bone length and direction plays an important role in action recognition, AGCN on other hand utilizes this information. AGCN formulates length of bones and their direction as vector, this vector is then fed to AGCN to predict the class. Basic idea of AGCN is to utilize both networks and increase the efficiency.

The spatiotemporal graph convolution is based on predefined graph and as described above it has limitations, AGCN solves this problem by formulating a way to optimize all other parameters together and forms an end-to-end learning. It basically works on connection between two vertexes and their connectivity strength. Below Figure 8 shows overall architecture of 2 Stream AGCN, where B-Stream stands for network of bones and J-Stream stands for network of joints[7]. From both J and B stream we get scores which are finally added and fed to SoftMax layer for prediction.

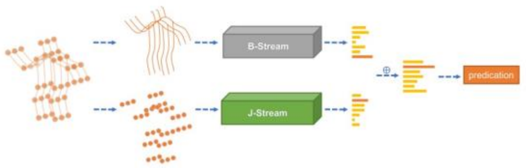


Figure 6. 2S-AGCN architecture [7]

The two stream – AGCN method was evaluated over two datasets. It was found that the model showed 95.1% accuracy over NTU-RGBD and 58.7% on Kinetics-Skeleton dataset.

Two stream- AGCN influenced two-stream graph attention convolutional network with spatial–temporal attention (STA-GCN) (Zhang et al., 2020), basic idea was to use motion over bone stream and better results were achieved.

STA-GCN, in addition to utilizing concepts of two stream – AGCN, it can be summarized in two steps as below:

- Temporal graph attention module was used, so that hand gesture encoding can be done with multi-scale temporal features.
- Two-stream hand gesture network was used as briefed below:
    - Pose stream – it uses joints from each frame as input
    - Motion stream – it uses joint offsets between neighbouring frames

Below Figure 9 details out network architecture for single stream[8].

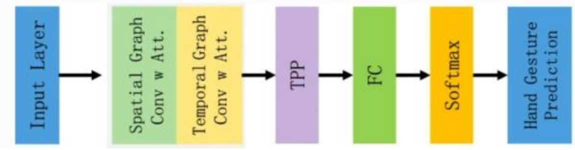


Figure 7. Single Stream architecture [8]

As shown in above figure, is the network architecture used by both Pose and Motion streams. First, we initialize the skeleton graph, the input then is fed to Spatial Graph convolution with spatial graph attention mechanism and temporal graph attention to extract the spatial temporal features.

Figure 10 shows spatial temporal features, where black line denotes spatial connections and blue denotes temporal connections[8].

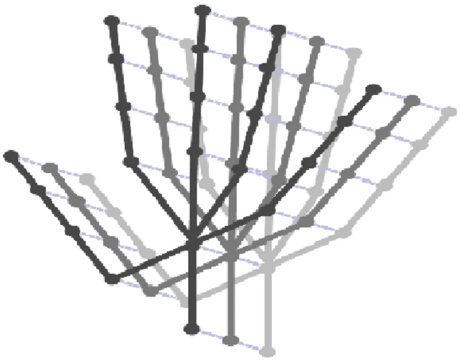


Figure 8. The spatial–temporal joint connections of the initial graph [8]

The output feature of GCN with spatial graph attention will be fed to the GCN with temporal graph attention. Temporal pyramid pooling layer (TPP) [9], is then used to extract multi-scale features, the output of TPP is then fed to fully connected (FC) layer and SoftMax is used for classification.

Below Table 3 shows accuracy of STA-GCN when used of SHREC'17 [8].

Table 3. STA-GCN accuracy matrix

Stream	14 gestures
Pose stream	93.2
Motion stream	94.4
Two streams	94.5

On DHG14/28 dataset, on 14 gestures accuracy of 91.5% was achieved.

STA-GCN when compared with 2s-AGCN on SHREC'17, it was found to be better by 2.1% on 14 gestures recognition and 0.7% on 28 gestures recognition, while when compared on DHG14/28, STA-GCN outperformed 2s-AGCN by 1.6% on 14 gestures recognition and 1.2% on 28 gestures recognition.

In 2019 [10], another improvement to graph-based methods was done, earlier methods used bones and joints separately and hence making skeleton as undirected graph, this posed limitations. To overcome these limitations, skeleton was representing as DAG, joints were used as vertexes and bones represents edges. Using this information, a DGCN was designed, which could propagate the information in adjacent joints and bones and hence better represent the current state, thereby giving opportunity to better identify the action. This study also solves the problem where graph cannot be directly used to populate the coordinates, for example clapping or hugging. Since there is strong dependency between two body parts be it two hands or hugging someone, such feature cannot be constructed using graph. This problem was solved using adaptive graph, in which topology of graph is parameterized and optimization takes place during training. Below Figure 11 is graph representation of human body, where blue circle indicates the root vertex [10].

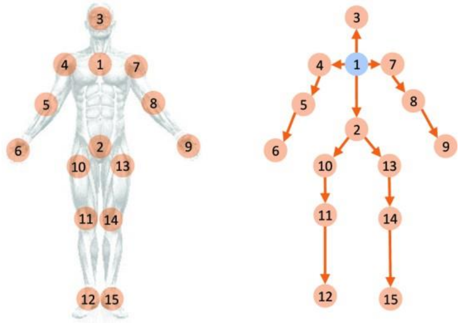


Figure 9. Illustration of the graph construction for skeleton data [10]

Following are the high-level steps involved:

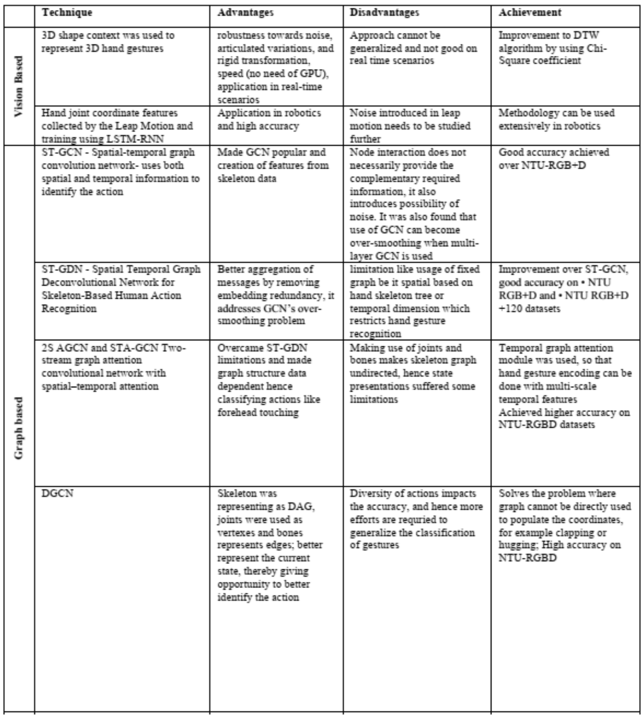
- Prepare Skeleton data frames containing joint coordinates
- Extract bone information
- Represent Spatial information of joint and bones as vertexes and edges within a DAG
- DGCN – Extract the action recognition features

DGCN method was applied on two famous datasets NTU-RGBD and Skeleton-Kinetics and accuracy of 96.1% and 59.6% was achieve respectively. Main thing to note here is due to diversity of actions in Skeleton-Kinetics dataset we find that accuracy is not very good, and this indicates that more study in the field is required and methods which can be generalized with higher accuracy. This study paves a path where generalization is key, use of skeleton + RGB data together can be studied and possibly a better accuracy can be achieved.

V. SUMMARY

Hand gestures recognition has gained popularity across many applications and many methods are being tried to try to get to more accurate real time predictions and at the same time trying to keep computational cost as low as possible. With devices like mobile phone, smart watches it has become utmost priority to develop software such that with limited hardware it can reach the goal. Various efforts that are being put are mentioned below. We have tried to capture the latest trends with advantages and disadvantages of various methods used and a summary is shown in Table 4.

Table4. Summary of different methods used in Gesture recognition



	Technique	Advantages	Disadvantages	Achievement
Vision Based	3D shape context was used to represent 3D hand gestures	robustness towards noise, articulated variations, and rigid transformation, speed (no need of GPU), application in real-time scenarios	Approach cannot be generalized and not good on real time scenarios	Improvement to DTW algorithm by using Chi-Square coefficient
	Hand joint coordinate features collected by the Leap Motion and training using LSTM-RNN	Application in robotics and high accuracy	Noise introduced in leap motion needs to be studied further	Methodology can be used extensively in robotics
Graph based	ST-GCN - Spatial-temporal graph convolution network- uses both spatial and temporal information to identify the action	Made GCN popular and creation of features from skeleton data	Node interaction does not necessarily provide the complementary required information, it also introduces possibility of noise. It was also found that use of GCN can become over-smoothing when multi-layer GCN is used	Good accuracy achieved over NTU-RGB+D
	ST-GDN - Spatial Temporal Graph Deconvolutional Network for Skeleton-Based Human Action Recognition	Better aggregation of messages by removing embedding redundancy, it addresses GCN's over-smoothing problem	limitation like usage of fixed graph be it spatial based on hand skeleton tree or temporal dimension which restricts hand gesture recognition	Improvement over ST-GCN, good accuracy on • NTU RGB+D and • NTU RGB+D +120 datasets
	2S AGCN and STA-GCN Two-stream graph attention convolutional network with spatial–temporal attention	Overcame ST-GDN limitations and made graph structure data dependent hence classifying actions like forehead touching	Making use of joints and bones makes skeleton graph undirected, hence state presentations suffered some limitations	Temporal graph attention module was used, so that hand gesture encoding can be done with multi-scale temporal features Achieved higher accuracy on NTU-RGBD datasets
	DGCN	Skeleton was representing as DAG, joints were used as vertexes and bones represents edges; better represent the current state, thereby giving opportunity to better identify the action	Diversity of actions impacts the accuracy, and hence more efforts are requried to generalize the classification of gestures	Solves the problem where graph cannot be directly used to populate the coordinates, for example clapping or hugging; High accuracy on NTU-RGBD

VI. CONCLUSION

Based on comparative study among many existing methods to classify the gesture, vision based and Graph based are the one which have a good potential to accuracy classify the gesture and keeping the computational cost low.

REFERENCES

1. Mitra, S. and T. Acharya, *Gesture recognition: A survey.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2007. **37**(3): p. 311-324.
2. Zhu, C., et al., *Vision based hand gesture recognition using 3D shape context.* IEEE/CAA Journal of Automatica Sinica, 2019. **8**(9): p. 1600-1613.
3. Wu, B., J. Zhong, and C. Yang, *A visual-based gesture prediction framework applied in social robots.* IEEE/CAA Journal of Automatica Sinica, 2021. **9**(3): p. 510-519.
4. Yan, S., Y. Xiong, and D. Lin. *Spatial temporal graph convolutional networks for skeleton-based action recognition*. in *Thirty-second AAAI conference on artificial intelligence*. 2018.
5. Peng, W., J. Shi, and G. Zhao, *Spatial temporal graph deconvolutional network for skeleton-based human action recognition.* IEEE Signal Processing Letters, 2021. **28**: p. 244-248.
6. Zhang, Y., et al. *Polarnet: An improved grid representation for online lidar point clouds semantic segmentation*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
7. Fan, Y., et al. *Multi-Scale Adaptive Graph Convolutional Network for Skeleton-Based Action Recognition*. in *2020 15th International Conference on Computer Science & Education (ICCSE)*. 2020. IEEE.
8. Zhang, W., et al., *STA-GCN: two-stream graph convolutional network with spatial–temporal attention for hand gesture recognition.* The Visual Computer, 2020. **36**(10): p. 2433-2444.
9. Wang, P., et al., *Temporal pyramid pooling-based convolutional neural network for action recognition.* IEEE Transactions on Circuits and Systems for Video Technology, 2016. **27**(12): p. 2613-2622.
10. Shi, L., et al. *Skeleton-based action recognition with directed graph neural networks*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.