Multi-region based radial GCN algorithm for real-time action recognition

Han-Byul Jang[1][0000-0003-4815-1513] and Chil-Woo Lee[1][0000-0002-3391-1631]

[1] Chonnam National University, Yongbong-ro 77, Buk-gu, Gwangju, Republic of Korea

Abstract. The MRGCN algorithm [1] uses simple optical flow and image gradient instead of skeleton data as inputs for deep learning, so that implementation of an action recognition system used in the real world can be possible. However, to be applied to various application systems, real-time processing is absolutely necessary. For example, in the case of an intelligent surveillance system that responds to crimes or accidents occurring on the street, real-time processing is essential. In this paper, we describe the parallel processing algorithm developed for the implementation of real-time behavior recognition system using MRGCN and the neural network structure that has variability according to the sampling time of input data. By analyzing the processing modules constituting the algorithm and executing the modules that can be processed simultaneously, it was possible to obtain a processing speed improved by nearly 50% compared to the existing sequential processing method.

Keywords: Human action recognition, Graph convolutional network, Real time system.

1 Introduction

If the action of a person appearing in a video can be recognized, it is able to be applied to various intelligent services. Since the purpose of such services is to provide immediate results without human intervention, real-time algorithms must be implemented. In particular, in an intelligent surveillance system, it is very important to implement a real-time behavior recognition algorithm with fast processing speed, as it is a key function to identify risky behaviors and quickly respond to them using behavioral recognition. For example, in order to find out a crime or accident situation in real time from a camera monitoring the street as shown in Figure 1, a recognition algorithm that can quickly process image data acquired through CCTV is essential.

Recent action recognition studies generally use deep learning approach. Deep learning makes it possible to effectively process complex data that is difficult to analyze through existing pattern recognition methods, so the recognition rate of algorithms and the performance of application services are further improved by using large amounts of data as inputs for deep learning. Recently, high-performance action recognition algorithms using GCN, the latest deep learning technique for learning data having a graph structure, have been announced, and they generally use skeleton joint coordinates as

learning data. However, the skeleton data has a problem in that it is difficult to obtain accurate values and so it is impossible to use in actual application services. In the previous study of this study [1], we proposed MRGCN algorithm using new input data that combines easily obtainable optical flow and image gradient instead of skeleton data. MRGCN achieved a higher recognition accuracy than previous research achievements through effective neural network learning by applying graphs tailored to the characteristics of new input data.

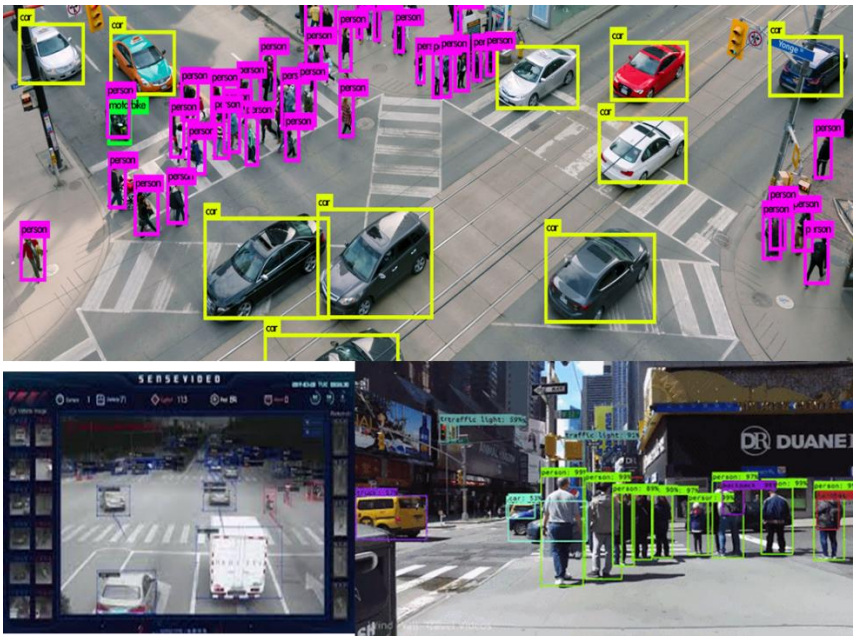

Fig. 1. Automatic surveillance system monitoring the street.

MRGCN is good to use in application services because it uses easily obtainable image information instead of skeleton data and has high performance, but it does not have a structure suitable for real-time processing. This paper describes a real-time algorithm that improves the structure of MRGCN to be suitable for application services. Action recognition application services are often not equipped with high-performance computers, but lack of processing capability can delay data acquisition timing and generate low frame rate data. Therefore, real-time action recognition algorithms should be tested so that they can be used even at low frame rates and improved to save processing time as much as possible. In this study, after re-training MRGCN according to input data of low frame rating, it was confirmed that proper recognition rate could be maintained up to 10 fps (frame per second) by experiment. In addition, in order to obtain faster processing speed, the modules of the algorithm were analyzed, and the structure was changed to parallelize the modules that could be processed simultaneously. As a result of applying the parallel processing structure, the processing speed could be improved by about 50%.

The order of this paper is as follows. First, this section introduces the outline of the paper. Section 2 examines the existing action recognition studies. Section 3 describes the structure and characteristics of MRGCN, the previous study of this paper. Section 4 proposes improvements of the MRGCN algorithm for real-time action recognition. Finally, Section 5 describes the conclusions and the factors that need improvement through future research.

2 Related Studies

Recently, most studies of human action recognition use deep learning approach. In the past, action recognition studies often used classic pattern recognition methods, but deep learning-based research has become mainstream because deep learning using large amounts of data has made it possible to analyze complex information inherent in action. Various neural network models such as CNN, LSTM, RNN, and GCN are used for action recognition. For example, [2] uses LSTM and CNN to classify feature data obtained from skeleton data with LSTM while performing CNN classification using data transformed into map images. The paper [3] defines the human body into five parts to generate input data of RNN. In each part, the skeleton data is processed to feature data, which is then input into the RNN to classify the actions. [4] uses a spatial-temporal LSTM algorithm using optimized skeleton data, and [5] uses TCN (Temporal Convolutional Network), a CNN method that processes feature information extracted from video images. In addition, [6] published by Deep Mind combined RGB image and optical flow image and used it as input data and achieved good recognition rate by using several deep learning techniques in combination.

Recent action recognition studies generally use GCN model. Since GCN can perform neural network learning using data that has a graph structure, it is very suitable for analyzing skeleton data expressed in graphs in the structure of human body joints. In [7] and [8], the authors introduced GCN into action recognition research for the first time and showed improved results than existing methods. The ST-GCN model presented in [7] achieved a high recognition rate by using a graph of skeleton data continuously connected over time. Since then, many studies have been published that improved ST-GCN. For example, [9] improves the recognition rate by dividing the skeleton data into four parts according to the composition of the human body. In [10], the authors presented AS-GCN (Actional structural GCN) that can better reflect the feature and the structure of actions. And [11] proposed a more mathematically optimized method GR-GCN (Graph regression based GCN) model.

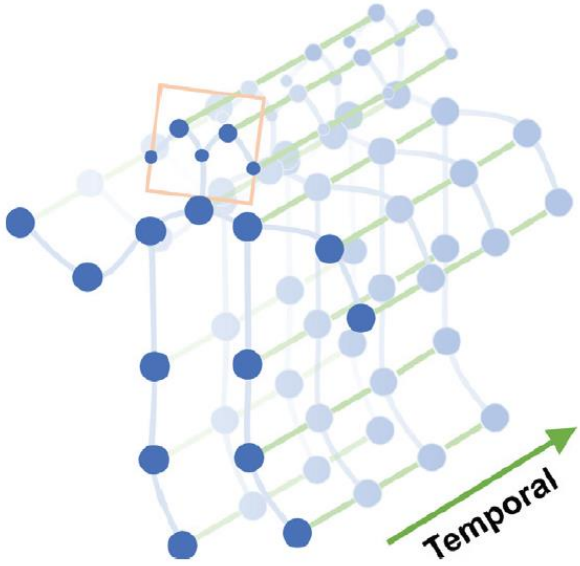

Fig. 2. A graph design same with the human joint connection form in ST-GCN [7].

The action recognition studies discussed above generally use skeleton data that records changes of coordinates of human joints as input data of the neural network. Since skeleton data contains well the context information of human body movement, it can be effectively used for action recognition and is particularly suitable when it uses the GCN algorithm with the graph shown in Figure 2. However, there is necessary conditions that special hardware such as Kinect [12] or posture recognition software such as OpenPose [13] must be used to obtain the accurate coordinate of the joints. Therefore, there is a problem in that it is difficult to acquire accurate skeleton data in a field where stable conditions are not be satisfied.

MRGCN [1], a previous study of this study, can learn and recognize using new input data that can be easily acquired from images instead of skeletons, while using the highly efficient GCN algorithm. In detail, MRGCN achieved higher recognition accuracy than that of the existing GCN-based action recognition by using a feature vector that combines optical flow and image gradient that was obtained from an input image through simple processing. This paper shows that this MRGCN algorithms can be effectively applied to action recognition application services by improving them to be suitable for real-time processing.

3 The Structure of MRGCN

3.1 The Overall Structure of MRGCN

Skeleton data, which is a stream of human 3D joint coordinates, is used as general input data for action recognition. To acquire the skeleton data, special hardware such as Kinect [12] or professional posture recognition software such as OpenPose [13] must be used. However, it is difficult to apply such special hardware or software when the data acquisition environment is poor because accurate results can be obtained only when the very restrictive environment is provided. Therefore, in [1], the previous study of this paper, we proposed MRGCN, a more effective action recognition algorithm using optical flow and image gradients, which can be simply calculated by processing 2D images instead of skeleton data.

The overall sequence of the MRGCN algorithm is shown in Figure 3. As shown in this figure, MRGCN first calculates the optical flow and image gradient in the person's area appearing in the input image at (a-b) stages, and then converts it into compressed data of HoFG (histogram of flow and gradient), code, mean, and standard deviation at (c) stage. The converted data is used for network training and recognition as an input to the neural network at (d) stage. In the recognition process, the recognition result is finally output through the classifier stage (e).

3.2 The Generation Method of Input Data

As described above, MRGCN uses optical flow and image gradient as input. The reason for combining the two data is that it is assumed that action can be described by combining human movement and shape information. Optical flow is good for expressing movement information, and image gradient is good for representing shape information.

After the two kinds of data are generated as 32-dimensional histograms based on the direction using the HoG (histogram of gradient) algorithm [14], they are reduced and converted into 3-dimensional respectively, total 6-dimensional vectors to enable fast neural network learning. The 3-dimensional vectors are the HoFG code obtained using Equations (1) and (2), and the mean and standard deviation of the histogram. In the Equation (1), the Mode value is sequentially obtained at the histogram index with the histogram value greater than the neighboring value. Next, the HoFG Code is calculated by applying the obtained Mode values to Equation (2). In this case, C is a constant using 20.

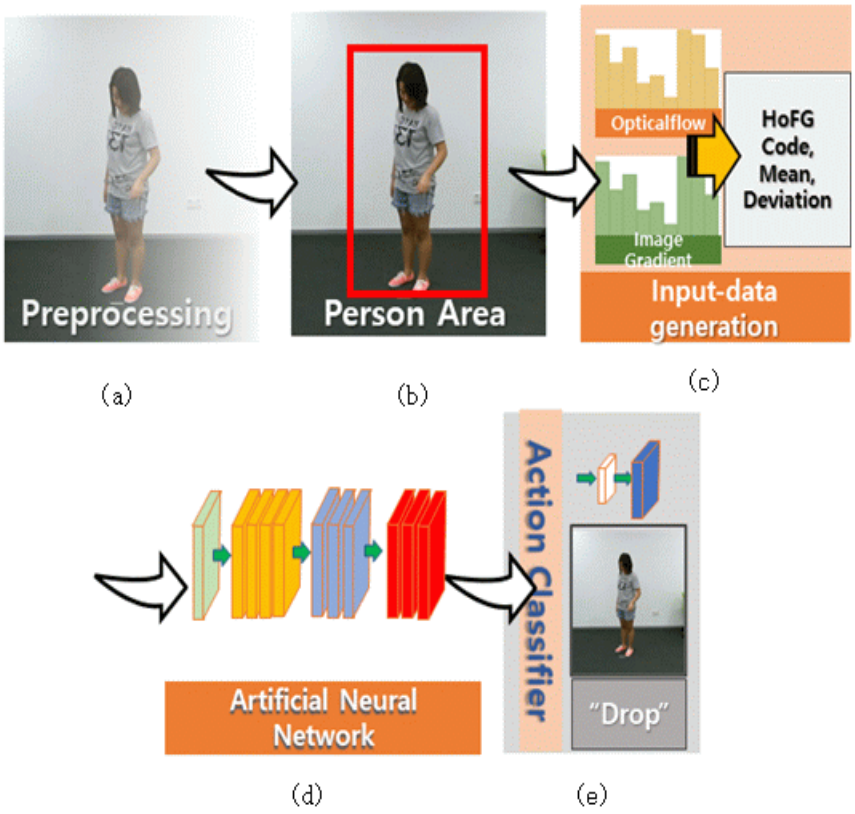

Fig. 3. Overall flowchart of MRGCN algorithm (The source of flowchart is [1] and the input image is from [15]). (a) Preprocessing of input image stage, (b) Finding "Person area" stage, (c) Generation input data stage, (d) Applying input data to the MRGCN stage and (e) Classifying the result action stage.

$$Mode(order) = \begin{cases} (bin\ index\ i) + 1 : if\ the\ bin\ is\ mode \\ 0 : otherwise \end{cases} \quad (1)$$

$$HoFG\ Code = \sum_{order=1}^{k} Mode(order) * C^{k-order} \quad (2)$$

3.3 Graph Design of MRGCN

Studies that learn GCN using skeleton data generally use the same graph as the joint structure of the human body. Since this graph can directly reflects the context of human body's movement, is very suitable for processing skeleton data. However, MRGCN

using new input data requires new graph has the design that matches the input data. Therefore, an artificially designed graph structure is used to efficiently apply the shape and the motion changes of the human body.

Figure 4 is the plan view of the graph designed for MRGCN. As shown in Figure 5, this graph is structured to process input data systematically by arranging local data acquisition regions defined as RoM (Region of motion) using three-layers. Each RoM can generate the data of one node in the graph, and when the entire graph structure connecting all RoM nodes is drawn on a surface, it looks like it is spreads out radially, so this algorithm is named multi-region based radial GCN algorithm(MRGCN). In addition, MRGCN was able to achieve Top1 recognition accuracy of 94.28%, higher than the 93.27% Top1 recognition accuracy of the existing ST-GCN algorithm, as a result of comparative recognition experiments for 10 actions.

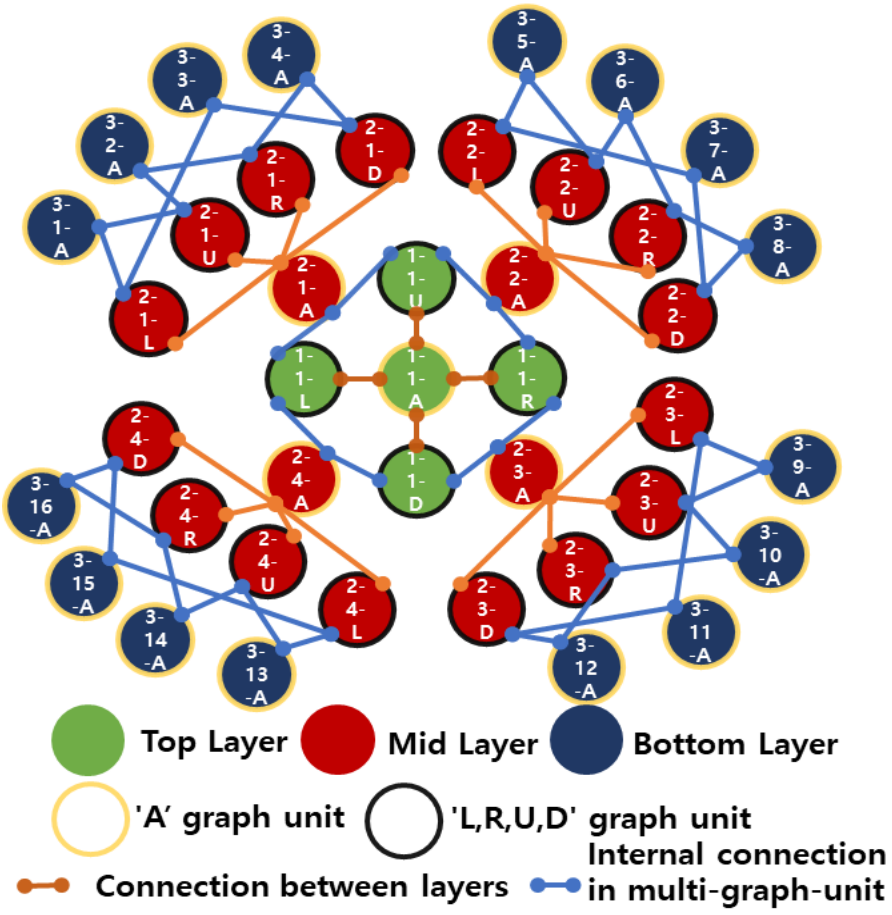

Fig. 4. Overall graph configuration of MRGCN in [1]. The graph has three-level layer structure (Top, Mid and Bottom Layer). Nodes in the graph have five types according to the characteristics of obtaining input data method (A: All, L: Left, R: Right, U: Up and D: Down).

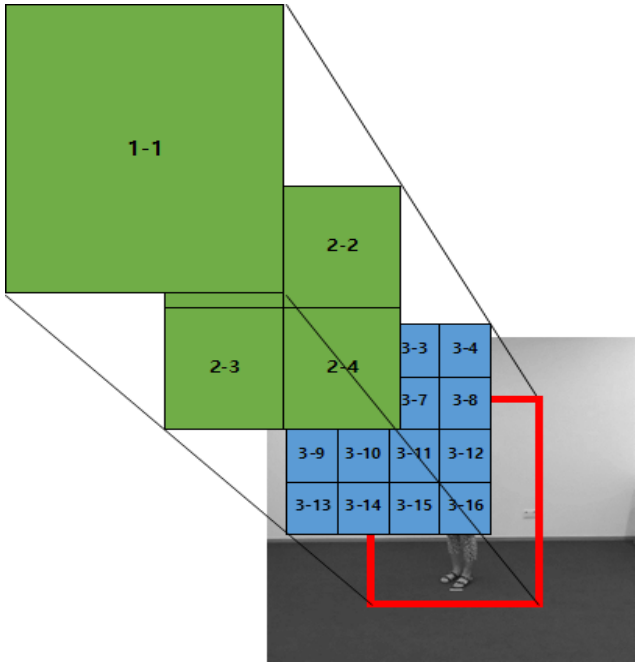

Fig. 5. Regions of data acquisition (RoM: region of motion) arranged using 3 layers in [1].

4 Improvement of MRGCN for Real-time Recognition

4.1 Improvement of Processing Speed using Multi-processing

An intelligent surveillance system, which is a representative application of action recognition, can be very important in modern society because it can automatically recognize dangerous situations such as crimes and accidents to prevent greater damage. However, in order to usefully use the action recognition algorithm in the application system, real-time processing must be possible.

As shown in Figure 3, the MRGCN algorithm has a structure in which each processing module is sequentially followed, so that the next stage can be executed only after passing through the previous stage. Therefore, faster processing speed can be obtained by processing modules that can be simultaneously processed in parallel. For example, if the “image acquisition module” is independently parallel processed, the waiting time to acquire the input image can be saved. In addition, “optical flow calculation”, “image gradient generation”, and “person area position estimation” can be simultaneously processed in the previous stage of input data generation, so a lot of processing

time can be saved. The flow chart of the improved algorithm by applying parallel processing is shown in Figure 6. As a result of improving the structure of the algorithm as shown in the flowchart, the processing speed could be increased by about 50% compared to the previous one.

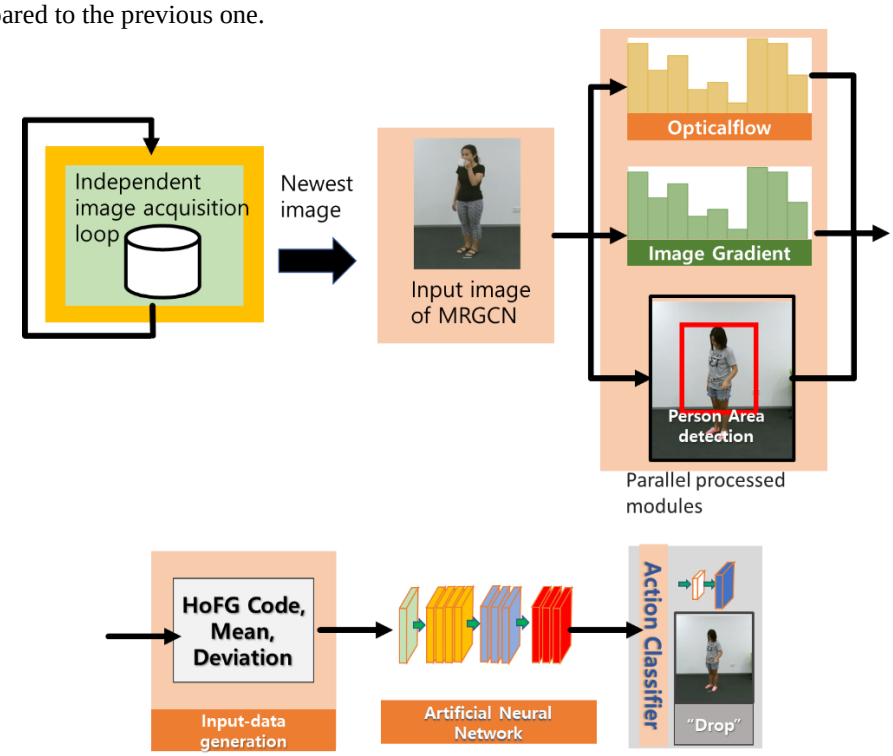

Fig. 6. Flowchart of MRGCN algorithm applying parallel processing.

4.2 Training a Neural Network with Frame Rate Change

Another issue to be considered for algorithm improvement is the quality of input data. The acquisition quality of input data may be degraded by many factors, such as the processing power of the application system, whether the using place is indoors or outdoors, or whether the lighting is bright or dark. In particular, the lack of processing power can cause lower frame rate of data acquisition because it affects the data acquisition timing. Therefore, the algorithm needs to be improved to generate accurate data even at low frame rate. Also, it is necessary to experiment whether recognition performance is maintained at low frame rate.

Since the previous MRGCN study was experimented with the data set, it used a fixed and accurate frame rate of 30 fps (frame per second). However, if the processing speed is reduced due to the large amount of calculation, the input is delayed, and recognition fails as a result. Among the input data of the MRGCN, since the optical flow is calculated between two frames before and after, data becomes unstable when the time interval between frames is too large. Therefore, even if the frame rate is low, the image

acquisition for calculating the optical flow must be adjusted to the 30 fps speed. To solve this problem, as shown in Figure 7, it is necessary to separate the image acquisition modules independently and process module in parallel. This problem can be easily solved by using parallel processing to acquire input images at 30 fps speed regularly regardless of the main processing loop.

In addition, the degradation of data quality due to the influence of the data acquisition environment has a great influence on the recognition result. In other words, a higher recognition rate can be obtained for an image with high contrast and clear human motion. In this paper, among the previously used training dataset, actions with large motion and high contrast were separately classified and used for training. As a result, the types of actions in the learning process were configured to be more clearly distinguishable, so that more distinct experimental results could be obtained.

The newly constructed experimental data includes 10 actions from NTU RGB+D dataset [15], and the types of selected actions are shown in Table 1 below. And Table 2 shows the results of experiment with frame rates of 5 fps, 10 fps, 15 fps, and 30 fps for the selected action. As shown in the table 2, the performance drop of MRGCN is larger than that of ST-GCN using skeleton data, but it maintains the same level of performance up to 10 fps or more. However, at a lower frame rate of 5 fps, there is a explicit performance degradation, and judging from this, it can be seen that the input data of MRGCN has a characteristic that the context of the action is relatively easy to miss at low frame rate.

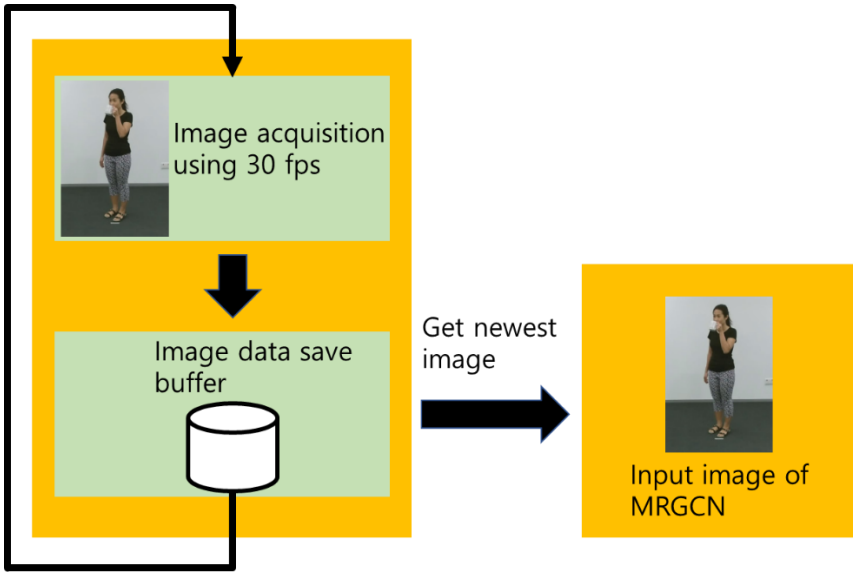

Fig. 7. Independent execution of image acquisition module.

Table 1. Selected 10 actions from NTU RGB+D dataset.

No.	Action	No	Action
1	Brush hair	6	Put on jacket
2	Throw	7	Take off jacket
3	Sit down	8	Kicking something
4	Stand up	9	Jump up
5	Reading	10	Staggering

Table 2. Comparison of recognition rates between ST-GCN and MRGCN according to frame rate

Frame rate	Top1 Accuracy of ST-GCN (Skeleton data)	Top1 Accuracy of MRGCN (6 channel data of optical flow and image gradient)
30 fps	90.31 %	93.99 %
15 fps	95.85 %	92.57 %
10 fps	93.66 %	92.86 %
5 fps	93.85 %	90.74 %

5 Conclusion

This paper describes an improved real-time MRGCN algorithm suitable for use in action recognition application systems. MRGCN is a high-performance GCN-based action recognition algorithm using new input data that can be easily acquired to solve the problem of performance degradation due to inaccurate data acquisition of existing skeleton data-based action recognition. However, to use it in an application system, it is necessary to improve the structure of the algorithm to have real-time processing capability. This is because it is possible to provide appropriate services only when the processing results can be obtained in real time in the application field of action recognition, such as an intelligent surveillance system.

To implement a real-time system, first, parallel processing was applied to the processing module of the MRGCN algorithm to achieve a speed improvement of about 50%. In addition, the input image acquisition module is independently executed in parallel so that accurate data can be calculated regardless of the variable data acquisition frame rate. To obtain clearer experimental results, the action of the experimental data set was configured to be clearly distinguishable, and as a result of verification experiments after learning at frame rates of 5 fps, 10 fps, 15 fps, and 30 fps, the recognition rate remained the same until about 10 frames.

To complete a more effective real-time action recognition algorithm in the future, more elements need to be improved. First, it is necessary to improve the training data of the neural network to better represent the action that occurred in the real environment. Currently used learning data is not optimized in a form suitable for the application system, so it is necessary to enhance the configuration of data to be more suitable for the purpose. In addition, robustness must be proven for various variables such as the

acquisition location of the input image, the intensity of lighting, and the type of background. In the input data, it is also necessary to introduce a method of adding a new type of data that can supplement information or augmenting and improving already generated data. As described above, we plan to improve MRGCN into a real-time algorithm that can be used in the real world by combining elements that can adapt to various environmental changes.

ACKNOWLEDGMENTS

This research is supported by Ministry of Culture, Sports, and Tourism (MCST) and Korea Creative Agency (KOCCA) in the Culture Technology (CT) Research & Development Program (R2020060002) 2022.

References

1. Jang, H. B., & Lee, C. W.: Multi-region Based Radial GCN Algorithm for Human Action Recognition. In *International Workshop on Frontiers of Computer Vision*, Springer, Cham, pp. 325-342. (2022).
2. Li, Chuankun, et al.: Skeleton-based action recognition using LSTM and CNN. *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, pp. 585-590. (2017).
3. Du, Yong, Wei Wang, and Liang Wang: Hierarchical recurrent neural network for skeleton based action recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1110-1118. (2015).
4. Liu, Jun, et al.: Spatio-temporal lstm with trust gates for 3d human action recognition. *European conference on computer vision*, Springer, Cham. (2016).
5. Kim, Tae Soo, and Austin Reiter: Interpretable 3d human action analysis with temporal convolutional networks. *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, IEEE, pp. 1623-1631. (2017).
6. Carreira, Joao, and Andrew Zisserman: Quo vadis, action recognition? a new model and the kinetics dataset. *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724-4733. (2017).
7. Yan, Sijie, Yuanjun Xiong, and Dahua Lin: Spatial temporal graph convolutional networks for skeleton-based action recognition. *Thirty-second AAAI conference on artificial intelligence*, pp. 7444-7452. (2018).
8. Li, Bin, et al.: Spatio-temporal graph routing for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pp. 8561-8568. (2019).
9. Thakkar, Kalpit, and P. J. Narayanan: Part-based graph convolutional network for action recognition. *arXiv preprint arXiv:1809.04983*. (2018).
10. Li, Maosen, et al.: Actional-structural graph convolutional networks for skeleton-based action recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3590-3598. (2019).
11. Gao, Xiang, et al.: Optimized skeleton-based action recognition via sparsified graph regression. *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 601-610. 2019.

12. Wikipedia, “https://en.wikipedia.org/wiki/Kinect,” Nov. 2021.
13. Cao, Zhe, et al.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence 43.1*, pp. 172-186. (2019).
14. Dalal, Navneet, and Bill Triggs: Histograms of oriented gradients for human detection. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, IEEE, Vol. 1. (2005).
15. Shahroudy, Amir, et al.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010-1019. (2016).