Analyzing Hand Action Recognition using Spatial Temporal Graph Convolutional Networks (ST-GCN)

Khin Sabai Htwe [*1] Takaaki ISHIKAWA [*2] Hiroshi WATANABE [*1, 2]
[*] Waseda University

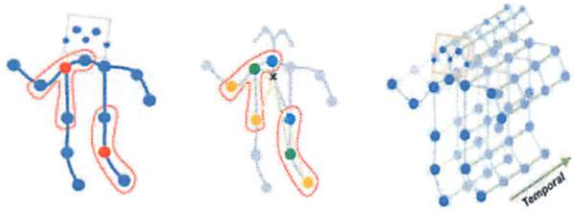
Abstract In this paper, we created hand keypoints dataset for hand action videos using pose estimation and analyzed hand action using Spatial-Temporal Graph Convolutional Network (ST-GCN).

1. Introduction

Nowadays, skeleton information is widely used in action recognition since there are robust pose estimation methods to detect skeleton data. Moreover, these skeletons are represented as a graph based on their movements and trained it in convolutional network to classify action recognition.

2. Overview of ST-GCN

As mentioned above, to represent the graph using skeletons, it is necessary two steps such as sampling and partition strategies shown in Fig 1a and Fig 1b for spatial configuration.



(a) (b) (c)

Fig. 1. Spatial-Temporal Information,
(a) Sampling (b) Partition Strategy and, (c) Temporal Configuration [1]

To sample on 2D image, the sampling area of the convolutional filter can be represented by delimiting the neighborhood around a central point in the rectangle grid. On graph, it considers the neighborhood of the center point as points that are directly connected by a vertex. In Partition strategy, it is based on the location of the joints and the characteristics of the movement of these joints. The points in the sampling region are partitioned into three subsets:

- The root node (center point) which are marked with green dots,
- The centripetal group (blue dots) which are the neighborhood nodes that are closest to the center of gravity of the skeleton (black cross), and
- The centrifugal group (yellow dots), which are the nodes farther from the center of the gravity.

The center of gravity is taken to be the mean coordinate of all joints of the skeleton in one frame. For temporal shown in Fig 1c, blue dots denote the body joints. First, the joints within one frame are connected with edges according to the connectivity of human body structure. Then, each joint will be connected to the same joint in the consecutive frame. After obtaining these spatial-temporal dimensions on this way, 10 ST-GCN layers apply convolutions on them.

3. Creating Hand Keypoints Dataset

Before training hand action using ST-GCN, we need to create hand keypoints dataset by using pose estimation method which is our previous work for hand joints detection [2] using baseline OpenPose. In this experiment, we download egocentric hand action videos [3] and the 144 videos are used to estimate hand poses. There are three steps to preprocess the data before training in model.

3.1. Segmenting the videos

To generate a video sample for each action, the original video is preprocessed almost 9 seconds (almost 240 frames, 640x480 pixels) per video since there are different lengths of frames.

3.2. Estimating the skeleton joints

In this step, the hand joints detection [2] is used to obtain the coordinates of the individuals' joints for all frames since it can detect 21 hand joints shown in Fig 2a for egocentric videos without even body or elbow.

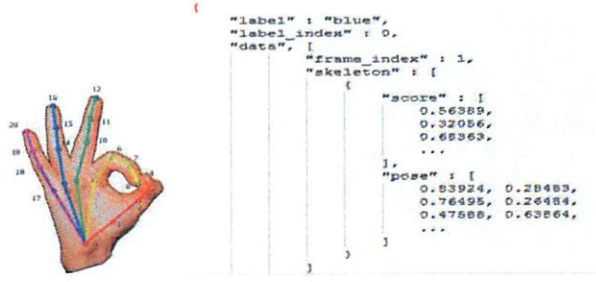


Fig. 2. (a) Hand Joints [2], and (b) Example x,y coordinates and its confidences

The provided skeleton form shown in Fig 2b is "hand pose": x, y coordinates and "score": confidence of that coordinates.

3.3. Filtering the skeleton joints

To train the skeleton sequences, the x,y coordinates are normalized into [0, 1] form by dividing width and height value of frame.

4. Experiment

In this experiment, the hand joints of 120 videos are used for training and 24 videos are used for testing and classify 8 actions such as blue, clicking, double-clicking, green, j-sign, scissors, swiping, and yellow.

4.1. Modification the model architecture

Although the graph representation approach by ST-GCN is very flexible, the model is trained by adding some edge links and adjacent neighborhood link to know whether or not influencing edge information. Therefore, we analyzed by training ST-GCN with new edge information. The edge information is inspired from [4] and defined as follows:

- If one of two joints is end site, one joint is two steps away from other joint. For example, if J0 is end site, J2 is two steps away from J0. Therefore, the edge links are [0,2],[0,6],[0,10],[0,14],[0,18],[2,4],[6,8], [10,12], and [14,16].
- If both joints are end site, these joints are connected. Therefore, the edge links are [0, 4], [0, 8], [0, 12], [0, 16], [0, 20], [4, 8], [4, 12], [4, 16], [4, 20], [8, 12], [8,16], [8,20], [12, 16], [12, 20], and [16, 20].

In the baseline edge information of ST-GCN for hand movements, self-links and adjacent links are 21 and 20 respectively. In this way, the total edge links are 41 links for baseline edge information and 66 links for new edge information.

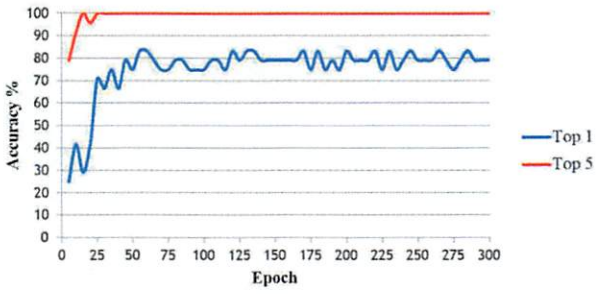


Fig. 3. Accuracy on Hand Action Recognition using baseline edge information

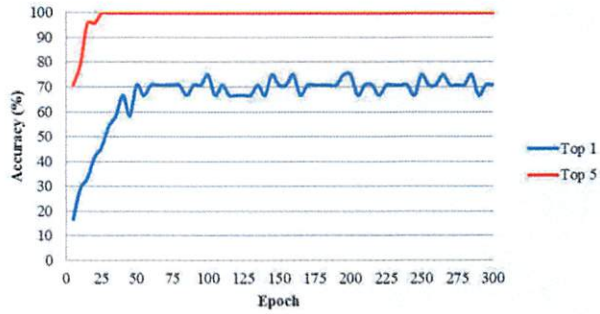


Fig. 4. Accuracy on Hand Action Recognition using new edge information

In Fig 3 and 4, the blue and red lines are represented top-1 and top-5 accuracy, respectively for hand action recognition. We can observe in final epoch that the top-1 accuracy is 79.17 % and 70.83% with baseline and new edge information, respectively. Based on analysis results, the model is influenced by changing edge information on the graph.

5. Conclusions

We analyzed skeleton-based hand action recognition using ST-GCN in this paper. We introduced new edge information by adding to baseline edge information for constructing graph. Although new edge information causes the accuracy to be decrease, we will consider geometric features using these edge links for next investigation on hand action recognition.

REFERENCES

[1] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition", AAAI2018.

[2] K. S. Htwe, T. Ishikawa, and H. Watanabe, "Improving Detection of Hand Joints in RGB Images using Maximum Confidence Value", The ITE Winter Annual Convention 2018.

[3] C. Xu, L. N. Govindarajan, and L. Cheng, "Hand Action Detection from Ego-centric Depth Sequences", Pattern Recognition, Volume 72, pp 494-503, December 2017.

[4] S. Zhang, X. Liu and J. Xiao, "On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks", The IEEE Winter Conference on Applications of Computer Vision (WACV), 24-31 March 2017.

*1Waseda University, Graduate School of Fundamental Science and Engineering, Department of Communications and Computer Engineering

*2Waseda University, Global Information and Telecommunication Institute

〒169-0072 Tokyo, Shinjuku, Okubo, 3-14-9, Shillman Hall-401

Phone: 03-5286-2509

E-mail: khinsabaihtwe@toki.waseda.jp