A Continuous Real-time Hand Gesture Recognition Method based on Skeleton

Tien-Thanh Nguyen, Nam-Cuong Nguyen, Duy-Khanh Ngo,
Viet-Lam Phan, Minh-Hung Pham, Duc-An Nguyen, Minh-Hiep Doan, Thi-Lan Le

School of Electrical and Electronic Engineering (SEEE),
Hanoi University of Science and Technology, Hanoi, Vietnam.

2022 11th International Conference on Control, Automation and Information Sciences (ICCAIS) | 978-1-6654-5248-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCAIS56082.2022.9990122

***Abstract*—While isolated hand gesture recognition methods aims to determine the type of gestures for a given sequence, continuous hand gesture recognition methods have to perform one more task: determining the starting point and ending point of the hand gesture. This task becomes challenging as the starting point and ending points of the gestures are not usually obvious even for human being. This paper presents a method for continuous hand gesture recognition based on skeleton information that consists of two phases: gesture detection and gesture recognition. In our method, to leverage the lightweight and the robustness of recognition models, TD-Net (Triple Feature Double Motion) model is employed in both gesture detection and recognition phases. Experimental results on IPN dataset have shown that the proposed method outperforms different state-of-the-art methods with 40.10% of Levenshtein accuracy and 0.1ms of inference time.**

***Index Terms*—hand gesture recognition, skeleton-based hand gesture recognition, continuous hand gesture recognition.**

I. Introduction

Hand gesture recognition (HGR) has obtained a great attention thanks to its widely applications in human-machine interaction, virtual reality, game and heath-care [1], [2]. Among various modality and sensor types used for capturing hand gesture, vision is the most widely used. In vision-based HGR methods, hand gestures can be recognized by exploiting directly RGB images or by extracting skeleton information from RGB sequences [3], [4] and then applying skeleton-based algorithms [5], [6]. Compared with RGB-based human gesture recognition, skeleton-based models have different advantages in term of computational time and memory required. However, although skeleton-based methods have obtained the state of the art results on different human activity datasets such as NTU RGB+D [5], few works have been applied for hand gesture recognition and the obtained performances are still poor due to the complex and highly dynamic configuration of the hand [6]. Moreover, hand gesture recognition methods can be further categorized into isolated hand gesture recognition and continuous ones. While isolated hand gesture recognition methods aims to determine the type of gestures for a given sequence, continuous hand gesture recognition methods have to perform one more task: determining the starting point and ending point of the hand gesture. This task becomes challenging as the starting point and ending points of the gestures are not usually obvious even for human being. This paper presents a method for continuous hand gesture recognition based on skeleton information. From RGB image sequence, skeleton information is automatically extracted by a hand pose estimation method. In our work, an off-the-shelf method that is deployed in Mediapipe is employed. The proposed method contains one detector and two classifiers. To build these detector and classifiers, the TD-Net architecture (Triple Feature Double Motion) is employed. Different experiments on IPN [3] - a dataset for continuous hand gesture recognition will confirm the robustness of the proposed method.

The remaining of the paper is organized as follows: Section II briefly surveys different approaches proposed for both isolated and continuous hand gesture recognition. Section III describes in detail the proposed method for hand gesture recognition. Finally, Section IV introduces the dataset and experimental results. Conclusions and future works are given in Section V.

II. Related work

In vision-based HGR methods, hand gestures can be recognized by exploiting directly RGB images or by extracting skeleton information from RGB sequences [3], [4] and then applying skeleton-based algorithms [5], [6]. Concerning RGB-based hand gesture recognition, different models have been proposed to take into account the appearance as well as the motion information [4], [7]. Besides RGB video-based methods, recently, many skeleton-based methods are also made viable thanks to low-cost depth cameras and progress on hand pose estimation [8]. Skeleton has its own advantage for gesture recognition as it is robust to lighting conditions and background cluttering.

While a number of approaches have been dedicated to isolated hand gesture recognition, continuous hand gesture recognition is still immature. One of the most challenging issues in continuous gesture recognition is gesture-spotting, i.e. to find the starting and ending frame of an isolated gesture, which has always been a primary task of continuous gesture recognition. This task is also named temporal segmentation. In [9], Liu et al. proposed a method for continuous hand gesture recognition. In this method, hands are detected by a two-stream Faster R-CNN model. Then, temporal segmentation is performed based on the height of hand compared with a threshold. Then, a C3D feature extractor is employed to extract the spatio-temporal features for RGB an depth sequence of a gesture. The work in [10] presented a network that employs a recurrent three dimensional (3D)-CNN with connectionist temporal classification (CTC) for jointly segmentation and classification of dynamic hand gestures. In [4], the authors introduced an architecture consisting of two models: a detector and a classifiers. The aim of the detector is to detect whether the input sequence contain a gesture or not, then a classifier is built to identify the gesture from detected sequence.

As aforementioned, several attempts have been proposed for continuous hand gesture recognition, however, they mainly based on RGB sequences. Therefore, they required expensive computation resources. In this paper, a skeleton-based continuous hand gesture recognition will be proposed. Different experiments on IPN [3] - a dataset for continuous hand gesture recognition will confirm the robustness of the proposed method.

III. Proposed method

A. Overall

In this paper, we propose a framework for continuous hand gesture recognition based on skeleton information. Figure 1 illustrates the continuous hand gesture recognition this aims at detecting simultaneously the starting point, ending point and the class of hand gesture. It's worth noting that in the isolated hand gesture recognition, the starting and ending points are determined a prior. The main idea of the proposed method is to first perform gesture detection and then to identify the type of hand gestures. From RGB image, skeleton information is automatically extracted by a hand pose estimation method. In our work, an off-the-shelf method that is deployed in Mediapipe is employed. Figure 3 illustrates hand pose estimation using Mediapipe. Then, the detector and classifiers will be applied on the skeleton sequence. To build these detector and classifiers, the TD-Net architecture (Triple Feature Double Motion) is employed thanks to its outperformed results on different human action datasets [8]. In the following section, we will briefly describe TD-Net and the process proposed for continuous hand gesture recognition in IPN dataset.

B. TD-Net architecture

In [8], the author proposed TD-Net, a light-weight but effective deep model for skeleton based action recognition. TD-Net can perform well on both hand skeleton and full body skeleton data as the input of the network contains many advanced feature. The architecture of TD-Net is illustrated in Fig. 4. TD-Net contains four branches that are Joint Collection Distances (JCD), Normalized Coordinates of Joints, Slow Motion and Fast Motion. In the following section, we will briefly describe these branches.

Joint Collection Distances (JCD): For skeleton data, the Cartesian coordinate and geometric feature are commonly used. The Cartesian coordinate feature is location and view-point variant so it can by dramatically changed when the skeleton are shifted. On the other hand, the geometric is location-viewpoint invariant and thus is more popularly used for skeleton-based action recognition. For a skeleton with N joints, each joint has the corresponding Cartesian coordinates $p_i^t = (x, y, z)$ at frame t, JCD feature can be calculated as follows:

$$JCD^t = \begin{bmatrix} \left\|\overrightarrow{\mathbf{p_2^t p_1^t}}\right\| & & & \\ \vdots & \ddots & & \\ \vdots & \cdots & \ddots & \\ \left\|\overrightarrow{\mathbf{p_N^t p_1^t}}\right\| & \cdots & \cdots & \left\|\overrightarrow{\mathbf{p_N^t p_{N-1}^t}}\right\| \end{bmatrix} \quad (1)$$

where $\left\|\overrightarrow{\mathbf{p_i^t p_j^t}}\right\| (i \neq j)$ is the Euclidean distance between $\mathbf{p_i^t}$ and $\mathbf{p_j^t}$.

Normalized Coordinates of Joints (NCJ): To take into account the hand joint coordinate and enrich the spatial information, NCJ feature is proposed. Different from JCD, the NCJ feature only focus on changes in distance between the thumb joint to other joints. As the thumb is often has good estimation quality, this feature can well represent the different hand shapes. Denotes p_{thumb}^t is the coordinates of the thumb, coordinates of others joints are defined as follows:

$$NCJ_i^t = \mathbf{p_i^t} - \mathbf{p_{thumb}^t} \quad (2)$$

Slow and Fast Motion: Geometric features like NCJ and JCD only represent the spatial relations of the skeleton, whereas the temporal information, which is very important for continuous action recognition, is missed. Therefore, the global motion features are employed by differential the spatial position of hand skeleton joints between two consecutive frames. To cover both fast and slow execution speed of an action, the two-scale motions are also proposed with different frame-sampling rates. It can be calculated as follow:

$$M_{slow}^t = S^{t+1} - S^t, t \in 1, 2, 3, ..., T-1; \quad (3)$$

$$M_{fast}^t = S^{t+2} - S^t, t \in 1, 3, ..., T-2; \quad (4)$$

where M_{slow}^t and M_{fast}^t denote the Slow Motion the Fast Motion, S^{t+1} and S^{t+2} are one frame and two frames behind the S^t, respectively.

1D CNN for action recognition: After obtaining all feature from the original data, we transform all feature to vectors, each vector depict the feature at frame t. This 1D CNN operate fast and efficient, which is suitable for continuous action recognition. In the architecture of TD-Net as shown in Fig 4, "c1D(3,64)" denotes one 1D convolution layer with kernel size = (1x3), number of filters = 64. GAP denotes Global Average Pooling layer and FC is a Fully Connected layer.

C. Hand gesture detection and recognition

When working with IPN dataset, we observe that the hand gestures in this dataset can be categorized into two subsets: pointing and non-pointing subsets as pointing class has longer duration to control the position of the cursor while non-pointing class has shorter duration to manipulate the interface.

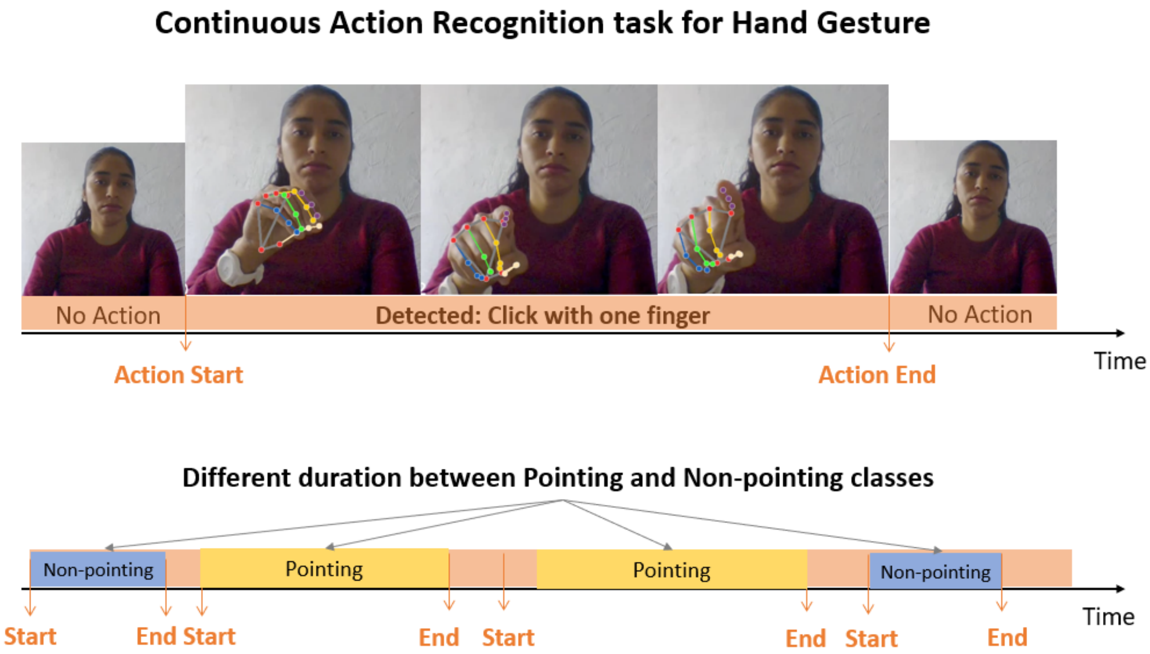

Fig. 1. Continuous hand gesture recognition for IPN-Hand dataset

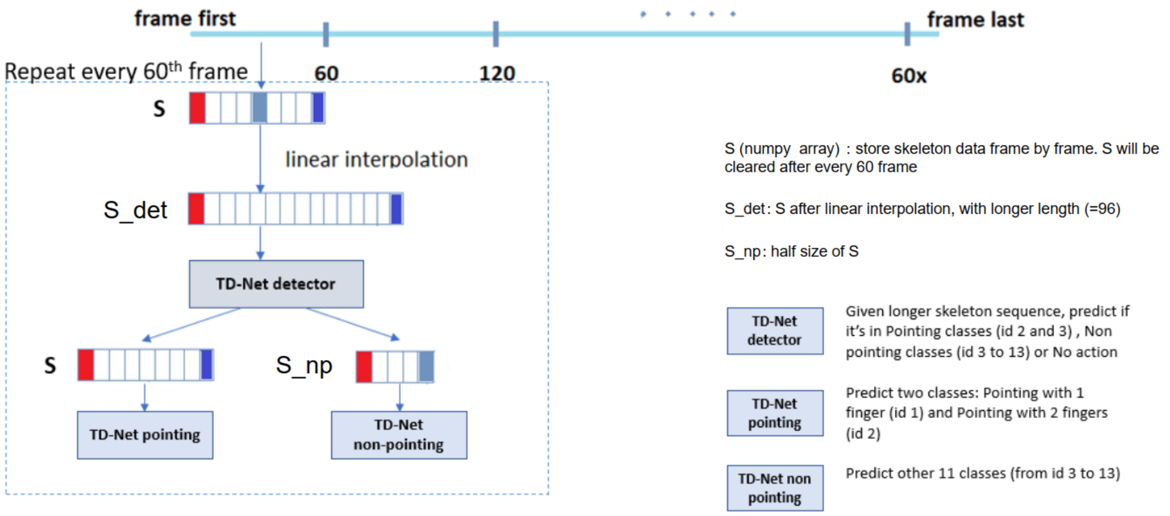

Fig. 2. The proposed process of recognizing continuous hand gesture using three TD-Net models.

Fig. 3. An example of skeleton sequence extracted from RGB images

Details about the duration of each class can be seen at Table I.

Therefore, we will train three different TD-Net models, each serves the different task: TD-Net detector, TD-Net pointing and TD-Net non-pointing. The process is illustrated in Fig. 2. In real-time, we collect sequence of skeleton data, frame by frame. The skeleton sequence is stored in an Numpy Multi-dimensional array **S**. As we select the IPN-Hand dataset for evaluation, we calculate the mean duration of all actions in the dataset. Based on that value, we set the maximum length for **S**. When **S** is not fulfilled, we perform the linear interpolation to up-sample the sequence of skeletons. At first, TD-Net detector will classify a sequence **S** into one of three classes: No gesture,

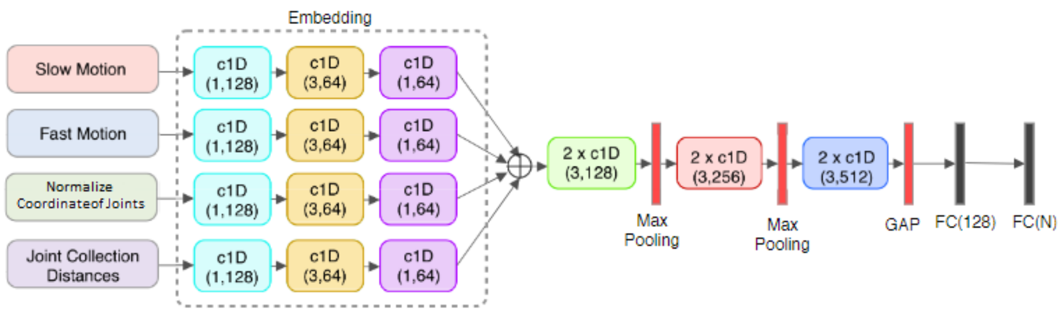

Fig. 4. Architecture of TD-Net for skeleton-based gesture recognition [8].

Pointing and Non-pointing.

If the detector classifies the action to Pointing class, the TD-Net trained for Pointing will be called to detect whether the action is Pointing with one finger or Pointing with two fingers. TD-Net pointing requires longer temporal length in **S** as the duration of the Pointing class is approximately four times as other actions in Non-pointing class. If the detector classify into the Non-pointing class, the TD-Net non-pointing will classify the action into one of 11 classes (From id 3 to 13 in Table I. TD-Net non-pointing only requires half size of **S** to recognize as actions in this group happens very fast.

After successfully classify the action, **S** will be cleared and continuously refilled. This operation only happens when **S** reaches the maximum length to avoid calling the recognition models many times. By doing this, the frame per second of our framework can remain high to smoothly handle real-time video stream.

TD-Net detector, TD-Net pointing and TD-Net non-pointing are trained with the skeleton data of corresponding actions. For the detector, we select the same amount of sample for each class: No gesture, Pointing (id 1 and id 2) and Non-Pointing (id from 3 to 13). TD-Net detector also has small size as we reduce number of filters in each 1D Convolutional layer.

IV. Experimental results

A. Dataset

To evaluate the proposed method, IPN dataset introduced in [3] is used. The main aim of IPN dataset is to design and collect a dataset for hand gesture recognition. Two main functional classes with 13 hand gestures are designed. The first class contains hand gestures that are used to control the position of the cursor on the screen (pointer) while the the second class contains gestures that used to manipulate the interface. Videos are collected from 50 subjects in 28 scenarios. There are totally 4000 gesture samples with more than 800,000 frames.

TABLE I
STATICS PER GESTURE OF IPN HAND DATASET.

ID	Gestures	Name	Instances	Duration
0	No gesture	No-gest	1431	147
1	Pointing with one finger	Point-1f	1010	219
2	Pointing with two fingers	Point-2f	1007	224
3	Click with one finger	Click-1f	200	56
4	Click with two fingers	Click-2f	200	60
5	Throw up	Th-up	200	62
6	Throw down	Th-down	201	65
7	Throw left	Th-left	200	66
8	Throw right	Th-right	200	64
9	Open twice	Open-2	200	76
10	Double click with one finger	Double-1f	200	68
11	Double click with two fingers	Double-2f	200	70
12	Zoom in	Zoom-in	200	65
13	Zoom out	Zoom-o	200	64

B. Isolated hand gesture recognition result

We first evaluate the performance of TD-Net for isolated hand gesture recognition. The Mediapipe hand solution is used to extract the skeleton data of hand in each frame. The output after the skeleton estimation consists of 4180 data files - each data file is the skeleton information of each frame, these frames belong to the time of representing a particular class of hand gestures have gone through the stages of bone extraction, segmentation of the gesture classes, normalization in the pre-processing step. The above 4180 skeleton information files are divided into two parts: a training set consisting of 3087 files and a test set of 1093 files. The learning task is to predict class labels for each gesture sample. We use classification accuracy, which is the percent of correctly labeled examples, and the confusion matrix of the predictions, as evaluation metrics for this test. Then, we conducted training data set of 3087 files containing hand skeleton information for the TD-Net and three other GCN models, namely ST-GCN [11], 2s-AGCN [12], for comparison.

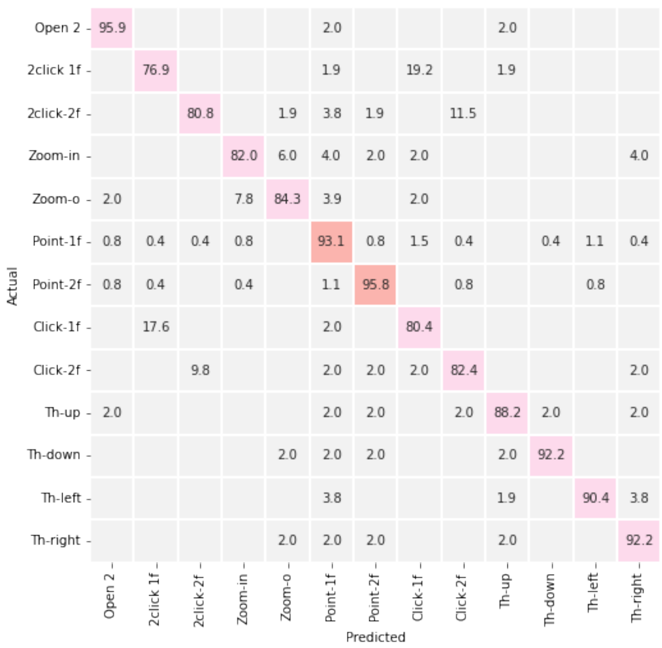

Fig. 5. Confusion matrix of TD-Net of IPN-Hand Isolated HGR

TABLE II
ACCURACY AND F1-SCORE OF ISOLATED HGR TASK USING IPN DATASET WHEN USING SKELETON INFORMATION.

Model	Acc (%)	F1-score (%)
ST-GCN [11]	46.94	-
2s-AGCN [12] (Joint)	82.16	74.20
2s-AGCN (Bone)	81.52	73.80
2s-AGCN (Ensemble)	83.81	-
Proposed method	**84.98**	**79.01**

From the results of TABLE II, the first original model ST-GCN brought about low accuracy when it only reached 46.94%. However, 2s-AGCN model, which is the improve of ST-GCN spatial and temporal graph convolutional networks, the 2s-AGCN model after merging two streams of bones and joints has increased accuracy to 36.87%, compared to ST-GCN. However, it can be seen that TD-Net outperformed the two GCNs model in both Accuracy and F1-score metrics. TD-Net achieved 84.98 % for Accuracy and 79.01 % for F1-score. This result prove that TD-Net not only has fast computation speed, but also excellent recognition accuracy. The confusion matrix of TD-Net is shown on Figure 5. The main diagonal of the confusion matrix has shown us that this model recognizes quite well with an accurate prediction rate of 60-96%, especially hand movements with a long duration or movements are less obscured with a correct ratio of 90-96%, e.g. Pointing-1f. Accuracy slightly decreased when recognizing gestures that have a short execution time, moving perpendicular to the screen.

Table III compares the results of the proposed method with others methods that based on other modalities such as RGB images, optical flow images. Obviously, models using a skeleton information-based recognition method achieved higher accuracy results. The RGB image-based recognition method has limitations due to camera movement, lighting and background conditions. Hand gesture recognition method based on skeleton information can overcome the above disadvantages because skeleton information is not affected by color, light factors and environment.

TABLE III
COMPARISON RGB MODALITY AND SKELETON MODALITY USING IPN HAND DATASET FOR ISOLATED HGR TASK

Model	Input Sequence	Modality	Acc(%)
C3D [3]	32-frames	RGB	77.75
ResNeXt-101 [3]	32-frames	RGB	83.59
2s-AGCN (Ensemble) [13]	-	Skeleton	83.81
Proposed method	-	**Skeleton**	**84.98**

C. *Continuous hand gesture recognition result*

The authors of IPN-Hand dataset provide an additional train - test split for evaluate Continuous HGR. Different from Isolated HGR, the dataset is split into 147 videos for training and 51 videos for testing. We also use Mediapipe to extract the skeleton data from each training video to train three different models: TD-Net detector, TD-Net pointing and TD-Net non-pointing. Each model use samples from specific classes as described in the previous section.

We use the Levenshtein accuracy introduced in [4] as evaluation metric. This metric bases on the Levenshtein distance that calculates the number of modifications to convert from a predicted gesture classes sequence to the groundtruth one.

TABLE IV
RESULTS OF LEVENSHTEIN ACCURACY WITH DIFFERENT MODALITIES.

Model	Modality	Results	Inference speed (ms)
Proposed method	**Skeleton**	**40.10%**	**0.1**
ResNeXt-101 [3]	RGB-Flow	42.47%	53.7
ResNet-50 [3]	RGB-Flow	39.47%	43.1
ResNext-101 [3]	RGB-Seg	39.01%	39.9
ResNet-50 [3]	RGB-Seg	33.27%	29.2
ResNext-101 [3]	RGB	25.34%	30.1
ResNet-50 [3]	RGB	19.78%	20.4

The confusion matrices of our TD-Net for Pointing and Non-pointing are shown in Figure 7 and Figure 6 respectively. As seen from the matrices, our two recognition models obtains high recognition accuracy: 97.7 % for TD-Net Pointing and 86.2 % for TD-Net Non-pointing.

The Levenshtein accuracy of our framework when integrate all models is shown on Table IV. We also compare our method with the baseline results from IPN-Hand dataset in both accuracy and inference speed (inference time was measured in a single GTX 1080ti GPU). The Levenshtein accuracy of ours is 40.1 %, which is very close to other methods using RGB modality. This can be explained that the modality we used is the hand skeleton data. Although skeleton data can well represent the actions, they can not distinguish between no action and action, especially when the subject moves the hand out of frame. Therefore, the recognition accuracy of our

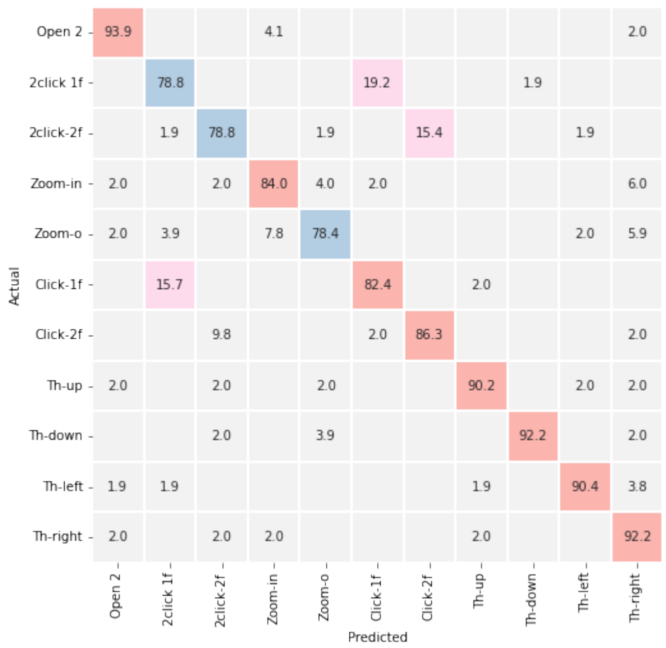

Fig. 6. Confusion matrix of TD-Net with non pointing classes

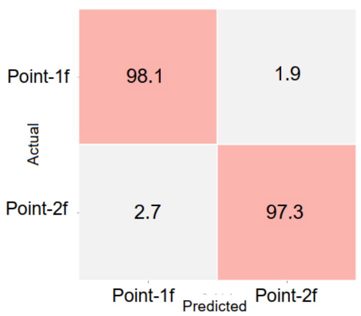

Fig. 7. Confusion matrix of TD-Net with Pointing classes

TD-Net detector suffers that issue. On the other hand, our framework outperformed other baseline methods in terms of inference speed. With the fast computation speed of the TD-Net, it takes only 0.1 ms for our framework to recognize an action. While those of other baseline methods are ranging from 20ms to 50ms.

V. Conclusions

This paper introduced a method for continuous hand gesture recognition based on skeleton information. Different experiments on IPN [3] - a dataset for continuous hand gesture recognition has confirmed the robustness of the proposed method. For isolated hand gesture recognition, the proposed method obtained 84.98% and 79.01% in term of accuracy and F1-score. Compared with others methods, the proposed method outperformed not only skeleton-based methods and but also other methods based on RGB and optical flow. Concerning continuous hand gesture recognition, the proposed method produces the best results with 40.10% of Levenshtein accuracy and 0.1ms of inference time. However, the current results are still poor and have room for improvement. In the future, end-to-end network will be investigated for both hand gesture spotting and recognition.

Acknowledgment

This work was supported in part by the National Program: Support for research, development, and technology application of industry 4.0 (KC-4.0/19-25), under the grant for the project: Research, design and manufacture of Cobot applied in industry and some other fields with human-robot interaction (code: KC-4.0-35/19-35).

References

[1] M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: a review of techniques," *journal of Imaging*, vol. 6, no. 8, p. 73, 2020.

[2] Q. Lei, H. Zhang, Z. Xia, Y. Yang, Y. He, and S. Liu, "Applications of hand gestures recognition in industrial robots: a review ," in *Eleventh International Conference on Machine Vision (ICMV 2018)*, A. Verikas, D. P. Nikolaev, P. Radeva, and J. Zhou, Eds., vol. 11041, International Society for Optics and Photonics. SPIE, 2019, pp. 455 – 465. [Online]. Available: https://doi.org/10.1117/12.2522962

[3] G. Benitez-Garcia, J. Olivares-Mercado, G. Sánchez-Pérez, and K. Yanai, "Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition," *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4340–4347, 2021.

[4] 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)*, 2019, pp. 1–8.

[5] D.-T. Pham, Q.-T. Pham, T.-L. Le, and H. Vu, "An efficient feature fusion of graph convolutional networks and its application for real-time traffic control gestures recognition," *IEEE Access*, vol. 9, pp. 121 930–121 943, 2021.

[6] V.-D. Le, V.-N. Hoang, T.-T. Nguyen, V.-H. Le, T.-H. Tran, H. Vu, and T.-L. Le, "A unified deep framework for hand pose estimation and dynamic hand action recognition from first-person rgb videos," in *2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2021, pp. 1–6.

[7] H.-G. Doan, T.-H. Tran, H. Vu, T.-L. Le, V.-T. Nguyen, S. V. Dinh, T.-O. Nguyen, T.-T. Nguyen, and D.-C. Nguyen, "Multi-view discriminant analysis for dynamic hand gesture recognition," in *Pattern Recognition*, M. Cree, F. Huang, J. Yuan, and W. Q. Yan, Eds. Singapore: Springer Singapore, 2020, pp. 196–210.

[8] H. V. T.-L. L. Tien-Thanh Nguyen, Dinh-Tan Pham, "A robust and efficient method for skeleton-based human action recognition and its application for cross-dataset evaluation," *IET Computer Vision*, 2022.

[9] Z. Liu, X. Chai, Z. Liu, and X. Chen, "Continuous gesture recognition with hand-oriented spatiotemporal feature," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3056–3064.

[10] 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 networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4207–4215.

[11] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *CoRR*, vol. abs/1801.07455, 2018. [Online]. Available: http://arxiv.org/abs/1801.07455

[12] H. Liu, J. Tu, and M. Liu, "Two-stream 3d convolutional neural network for skeleton-based action recognition," *arXiv preprint arXiv:1705.08106*, 2017.

[13] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.