Enhancing Continuous Skeleton-Based Human Gesture Recognition by Incorporating Text Descriptions

Thi-Lan le[1], Viet-Duc Le[1], and Thuy-Binh Nguyen[2]

[1] SigM lab, School of Electrical and Electronics Engineering Hanoi University of Science and Technology, Hanoi, Vietnam
[2] University of Transport and Communications, Hanoi, Vietnam
lan.lethi1@hust.edu.vn

Abstract. Continuous gesture recognition is a crucial task in human-computer interaction. Unlike isolated gesture recognition, where individual gestures are analyzed independently, continuous recognition involves detecting and classifying multiple gestures seamlessly from continuous video streams. In this paper, we propose a method for continuous gesture recognition. Our proposed model operates in two stages: isolated gesture recognition and a sliding window-based approach for continuous gesture recognition. For isolated gesture recognition, we propose a dual encoder method named TDDNet, stand for Text-Enhanced DDNet, that integrates a skeleton encoder based on the DDNet model [7] with a text encoder based on CLIP. We evaluate our model on a self-collected dataset comprising 19 gestures relevant to human-COBOT interaction, collected from 50 subjects. Experimental results demonstrate that our model improves isolated gesture recognition accuracy from 84.2% to 85.5%, while for continuous gesture recognition, the model achieves a performance of 66.60%, compared to 66.00% of the baseline model. The source code is publicly available at https://github.com/duclvQ/improved_DDNet.

Keywords: continuous action recognition · text description · skeleton-based action recognition

1 Introduction

Continuous gesture recognition is a crucial task in human-computer interaction (HCI) [20]. Unlike isolated gesture recognition, which analyzes individual gestures independently, continuous recognition involves detecting and classifying multiple gestures seamlessly from continuous video streams [9]. Although significant progress has been made, most existing approaches rely on deep models trained on RGB sequences. These models, while accurate, often contain a large number of parameters, making them computationally heavy and challenging to deploy in real-world applications.

Recently, the increasing availability of low-cost 3D cameras and advancements in human pose estimation have led to the rise of 3D skeleton-based action

recognition as an active area of research. This approach offers a lightweight and informative representation for continuous gesture recognition [5, 17]. However, despite considerable progress, skeleton-based action recognition remains challenging, particularly when dealing with highly similar gesture classes. Additionally, they mainly focus on modeling the relation of human joints in a unimodal training scheme [28]. Therefore, they do not fully exploit the semantic relationships between actions. In some cases, the primary distinction between different gestures lies solely in the shape or movement of the hands. This information may be ignored by visual models due to the low resolution of the hands. However, in such cases, action descriptions could reveal these details. Therefore, recent researches have attempted to incorporate action descriptions in isolated human action recognition from trimmed videos [25].

Recently, Contrastive Language-Image Pre-training (CLIP) has emerged as an efficient method for image representation learning using natural language supervision. CLIP jointly trains an image encoder and a text encoder to predict correct pairings between an image and its corresponding text. In this paper, we explore the use of CLIP for continuous skeleton-based gesture recognition. Our proposed model operates in two stages: isolated gesture recognition and a sliding window-based approach for continuous gesture recognition. For isolated gesture recognition, we propose a dual encoder method named TDDNet that integrates a skeleton encoder based on the DDNet model [7] with a text encoder based on CLIP. We evaluate our model on a self-collected dataset comprising 19 gestures relevant to human-COBOT interaction, collected from 50 subjects. Experimental results demonstrate that our model improves isolated gesture recognition accuracy from 84.2% to 85.5%, while for continuous gesture recognition, the model achieves a performance of 66.60%, compared to 66.00% of the baseline model.

The rest of this paper is structured as follows. In Sec. II, some notable work related to human gesture recognition (HGR) are briefly presented. The proposed model is described in Sec. 3. Experiments and results on our self-built dataset are provided in Sec. 4.

2 Related works

In the literature, existing Hand Gesture Recognition (HGR) methods are generally categorized into two main approaches: isolated and continuous hand gesture recognition. While isolated recognition approach focuses on identifying gestures from segmented videos, each containing a single hand gesture. Continuous recognition approach involves recognizing multiple hand gestures from a continuous video stream, requiring the system to detect and differentiate gestures as they occur in real-time. Compared to isolated recognition, continuous recognition task is considered more challenging due to the added complexity of temporal action localization (TAL), identify when each gesture starts and ends within the video stream.

For isolated recognition, early studies primarily employed CNN-based models to extract spatial features from each frame of a segmented video and RNN-based

architectures to capture the temporal consistency between consecutive frames. This combination aimed to model both spatial and temporal aspects of hand gestures [10, 21, 29]. More recent works have paid attention on building more effective spatio-temporal descriptors by using 3D-CNN networks to simultaneously capture spatial information. As a result, 3D-CNN networks have significantly enhanced the performance of isolated hand gesture recognition systems, providing more robust and accurate representations of hand gestures in a variety of contexts [1, 14].

Although 3D-CNN networks have achieved important milestones by leveraging both spatial and temporal information, Yu et al. [29] declared that these networks are often too complex, leading to lower efficiency. Consequently, some other works focused on dealing with skeleton data to reduce computation complexity while ensuring the recognition rate [11, 16, 27, 30]. In these work, skeleton data can be captured by using either a depth camera Microsoft Kinect or a deep-learning model that can estimate skeleton information from RGB images, such as OpenPose [4], MediaPipe [12], AlphaPose [8], etc. To build a hand gesture representation, spatial features can be extracted using Convolutional Neural Networks (CNNs), while the temporal correlation of skeleton sequences can be captured by deep learning models designed to handle sequential data. Among these models, Graph Convolutional Network (GCN) bassed approach including the Spatial Temporal Graph Convolutional Network (STGCN) [26] and its variants are regarded as one of the most effective approaches. The main objective of GCN-based approach is to learn simultaneously both spatial and temporal information from skeleton data by directly modeling the skeleton as a graph structure, with nodes representing joints and edges representing the connections between them. However GCN-based approach suffers from its heavy computation. Therefore, DDNet - a light weight model has been introduced for skeleton-based activity and gesture recognition [28].

For continuous recognition, besides the task of classifying which hand gesture is being performed, this task must determine the starting and ending times of each gesture. Noted that isolated hand gesture recognition is the first step in the framework handling continuous hand gesture recognition. For this , sliding window is one of the popular technique for video segmentation. Isolated recognition methods are applied into each video segmentation. After that, a post-processing mechanism is utilized to filter and aggregate the obtained results of the isolated recognition to accurately locate the time duration of each hand gesture [6,13,17,24]. Recently, large language models (LLM) such as ChatGPT [3], LLaMA [23] have become relatively popular and have been widely used for dealing with numerous human language. These models are pre-trained over various human language to learn common language-related characteristics. As a consequence, large language models can be effectively and efficiently applied to even new human languages. Inspired by this, Qu et al. [18] proposed a novel framework for action recognition in which skeleton sequences treated as the description sentences and LLM is used to capture some useful characteristics for action recognition. This is a novel approach for action recognition, the relation-

ship between consecutive frames is regarded as the connection between words in a sentence.

Based on the analysis above, this study introduces a novel framework for continuous gesture recognition that integrates text descriptions with skeleton information. This approach aims to overcome the limitations of isolated gesture recognition by managing multiple hand gestures in real time and leveraging natural language descriptions for enhanced interpretation and classification. The subsequent section will provide a comprehensive explanation of this framework and its components.

3 Proposed method for hand gesture recognition

The proposed method for continuous gesture recognition, illustrated in Fig. 1, consists of two phases: training phase and inference phase. In the training phase, a dual encoder model TDDNet is trained from pairs of segmented samples and their corresponding text description. In the inference phase, a sliding window method is used to allow the trained model to stride through a video and make predictions on each step. A threshold-based mechanism is used to aggregate predictions from all steps to a sequence of gesture.

3.1 Skeleton Data Preprocessing

Since the input data consists of RGB frames, we need to extract skeleton joints from these frames. It is important to note that any human pose estimation model can be applied. In our study, we utilize an off-the-shelf model, Google's Mediapipe, for this purpose.

Given an image of arbitrary resolution, the Mediapipe returns a $J \times D$ matrix, with $J = 21$ being the number of extracted joints for each hand and $D = 3$ being the spatial dimension of the joint. Because some gestures in our self-collected dataset requires the involvement of two hands simultaneously, resulting in two matrices of shape 21×3. Because some gestures involve arm movements over a long trajectory, we select six additional joints from the arm and shoulder, represented by a 6×3 matrix, to enhance the recognition accuracy. This matrix captures critical motion data, improving the system's ability to interpret and classify complex gestures effectively. Noted that, face and foot landmarks are neglected since they do not contribute to the meaning of the gestures. For every input frame, we perform pose estimation and receive two 21×3 matrices and a 6×3 matrix. These matrices are concatenated to form a 48×3 matrix. For frames where the pose estimators fail to extract any skeleton joints, zeros are replaced to guarantee the shape of the final matrix.

3.2 TDDNet for isolated hand gesture recognition

As analyzed in the related works, current approaches proposed for skeleton-based gesture recognition do not fully exploit the semantic relationships between

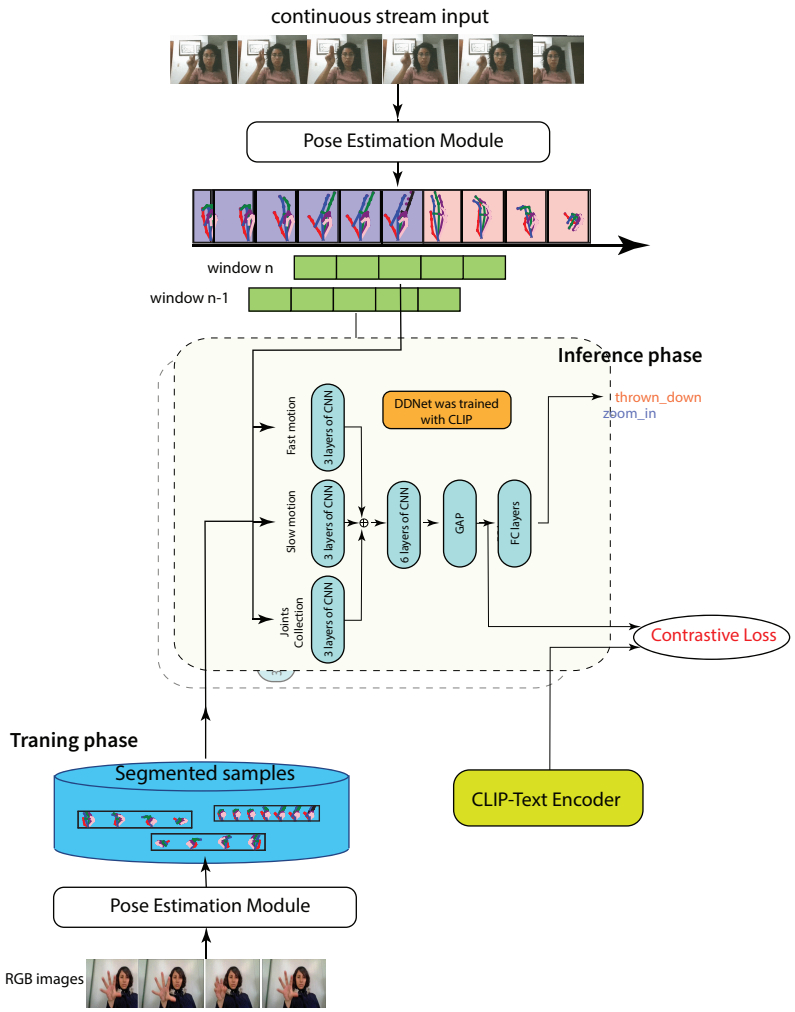


Fig. 1: Proposed method for continuous gesture recognition

actions. For instance, in some gestures, the primary distinction lies solely in the shape or movement of the hands. This information may be ignored by visual models due to the low resolution of the hands. However, in such cases, action descriptions could reveal these details. Therefore, recent research has attempted to incorporate action descriptions in human action recognition [25].

In our study, to leverage the information from text description, a dual encoder method TDDNet (illustrated in Fig. 2) that comprises of two encoders: one for skeleton and one for text is proposed. The skeleton encoder extracts features from human skeletons, and the text encoder provides auxiliary information from text descriptions. The text encoder helps the skeleton encoder learn effectively during the training stage, without affecting the inference time of the skeleton encoder. For skeleton encoder, we employ DD-Net [28], a lightweight 1D-CNN network for action recognition thanks to its superior performance obtained for action recognition. The text encoder relies on the model of CLIP [19] that takes a gesture description as input and results a text embedding feature.

We formulate the loss function in our proposed model as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}}\ (E_s(\mathcal{S})) + \lambda \mathcal{L}_{\text{con}}\ (E_s(\mathcal{S}), E_t(\mathcal{T})) \tag{1}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss, $\mathcal{L}_{\text{con}}$ is the contrastive loss between the features generated from the text encoder and the skeleton encoder. The contrastive loss is similar to the one used in the original paper by Radford et al. [19]. The weights λ is chosen through experimentation. $E_s(\mathcal{S})$ and $E_t(\mathcal{T})$ refer to the skeleton encoder model and the natural language encoder model whereas $\mathcal{S}$ and $\mathcal{T}$ are skeleton sequence and text description, respectively.

In the training process, for each gesture, a corresponding description is given. As the dataset used in our experiment is collected in the context of human-robot interaction, a description that explain how to perform the gesture is already available. For example, the text descriptions for two gestures Pickpart and Report illustrated in Fig. 4 are *"A person places one arm in front of their chest, vertical to the body, with palms facing out in front of the body, fingers clustered in a claw shape"* and *"A person raises one arm in front of their chest, vertical to the body axis, with fingers spread out and coming together at one point, the tips of the hands facing the front of the body."*, respectively.

3.3 Continuous hand gesture recognition

After having finished training on the isolated task, the best model will be saved and used in the continuous task. Similar to previous works [6,15], we use a sliding window mechanism as the data loader to the model. For each step, the size of the sliding window is fixed and chosen via the isolated evaluation. This sequence of frames is fed to the trained model to provide a prediction, just like a single forward pass in the isolated recognition. After each step, this window strides forward by one frame to get another sequence of frames. The sliding window mechanism is described as in Fig. 3

Given a sliding window with size T and an input video of N frame, the data fed to the model at step n will be the sequence of $(n-T+1)^{th}$ to n^{th} frames.

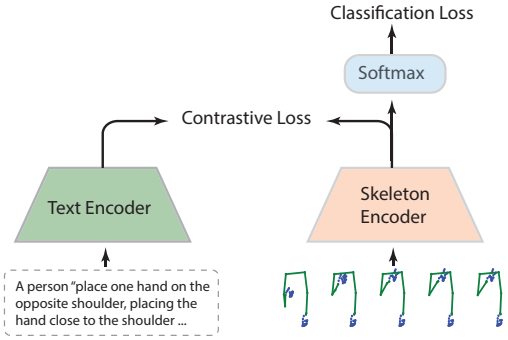


Fig. 2: TDDNet for isolated gesture recognition with dual encoder

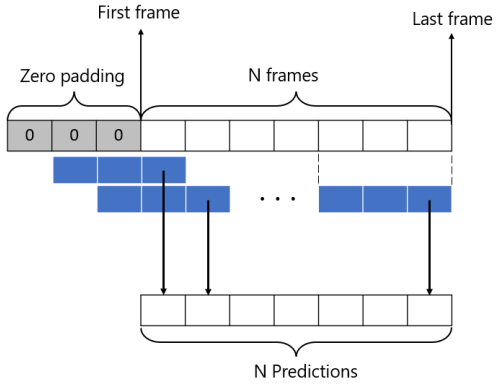


Fig. 3: Sliding window mechanism for continuous recognition

Using this approach, the first $T-1$ steps will not have enough frames to fit in the sliding window. Therefore, frames that have the indices being smaller than one will be replaced by zeros, or in other words they are zero-padded.

The sliding window shifts forward frame by frame. At each step, the model produces a prediction based on the current position of window. The predicted label of the current window is assigned to the middle frame of the window.

4 Experiments and Results

4.1 Dataset

To evaluate the proposed method, we use our self-collected dataset. The dataset was gathered as part of our project, which aims to develop gesture-based human-COBOT interaction. It consists of 19 gestures, designed according to [2], and was collected from 50 subjects, resulting in a total of 300 videos. Each subject performed 19 different gestures continuously, with random breaks between gestures lasting two to five seconds. In addition to these gestures, "no gesture" sequences (labeled with ID 0) were also collected. Table 1 provides statistical information about the gesture sequences in our dataset. We observe that the duration of gesture classes varies significantly. Moreover, some gestures are quite similar, as illustrated in Fig. 4, which poses challenges for gesture recognition methods. For evaluation, we define a subject-independent protocol, where sequences from subjects with odd IDs are used for training, while sequences from the remaining subjects are reserved for testing.

Table 1: Statistic information of 19 gesture classes in our dataset.

ID	Label	No.Samples	Ave duration(in frames)	ID	Label	No.Samples	Ave duration(in frames)
1	Start	188	94.86	11	PickPart	162	57.15
2	Stop	154	94.63	12	DepositPart	158	51.08
3	Slower	154	91.15	13	Report	138	55.41
4	Faster	154	82.64	14	Ok	153	46.57
5	Done	154	89.22	15	Again	154	47.96
6	FollowMe	120	92.67	16	Help	154	45.77
7	Lift	154	88.2	17	Joystick	154	50.41
8	Home	154	57.24	18	Identification	156	53.51
9	Interaction	151	49.16	19	Change	152	49.23
10	Look	157	48.18				

For our method, we use the skeleton data instead of the original RGB frames. An off-the-shelf model, Mediapipe, is employed to extract skeleton sequences from the RGB video sequences. Then, we process skeleton sequence as explained in Sec. 3.1. The skeleton file is a tensor of shape $T \times J \times D$, where T is the number of frames, $J = 48$ is the number of skeleton joints, and $D = 3$ indicates the 3D joints.

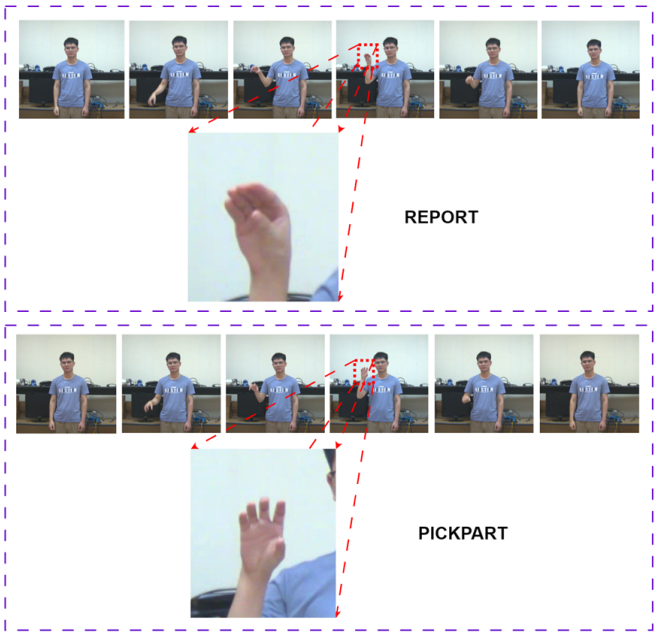


Fig. 4: Examples sequences of Report and PickPart gestures in our dataset.

4.2 Evaluation metrics

Continuous hand gesture involves two individual tasks, different sets of metrics are needed for each task to evaluate the efficiency of the model.

Isolated recognition refers to the evaluation of individual instances of gestures, making it be similar to a classification task. Therefore, traditional metrics that is Accuracy is employed for evaluation.

While isolated recognition is performed at the instance level, continuous recognition is evaluated at the video level, where the input consists of multiple gesture instances. To assess the performance of continuous recognition, we use frame-wise accuracy.

Frame-wise accuracy is the average accuracy of every frame of a video. Specifically, given a video with T frames, the prediction for T frames being $P = \{P_1, P_2, \ldots, P_T\}$ and the corresponding ground-truth being $G = \{G_1, G_2, \ldots, G_T\}$ the frame-wise accuracy can be formulated as:

$$\text{Frame-wise accuracy} = \frac{1}{T} \sum_{n=1}^{T} S\left(P_n, G_n\right) \tag{2}$$

with

$$S\left(P_n, G_n\right) = \begin{cases} 1 & \text{if } P_n = G_n \\ 0 & \text{if } P_n \neq G_n \end{cases} \tag{3}$$

The accuracy for a set consists of N videos is calculated by the average Frame-wise accuracy of all videos.

4.3 Experimental Results

Isolated recognition results Table 2 summarizes the results of our ablation study using different models from CLIP [19]. In the original CLIP implementation, the image encoder leverages two distinct architectures: one based on a ResNet and the other on a Vision Transformer (ViT). The results in Table 2 evaluate the performance of these architectures and their impact on isolated gesture recognition. The best performance, 85.50%, is achieved when using the ResNet architecture in CLIP, while employing the ViT model results in only a slight variation in performance. It is worth noting that RN50x4 and RN50x16 are variants of ResNet-50, scaled up by factors of 4 and 16, respectively, following the principles of EfficientNet [22] scaling rule. This scaling strategy allow the network to handle more complex tasks by improving its capacity to learn more useful feature representations. Therefore, in the remaining experiments, we utilizes ResNet architecture in CLIP model.

Table 2: Ablation study of different models used for text encoder.

Model name	Accuracy (%)
DDNet+RN50	85.20
DDNet+RN101	84.60
DDNet+RN50 × 4	**85.50**
DDNet+RN50 × 16	84.80
DDNet+ViT-B/16	82.80
DDNet+ViT-L/14	84.40

The performance of the proposed TDDNet model, in comparison to the baseline model, is shown in Table 3. Notably, the proposed method, leveraging a dual encoder architecture, outperforms the baseline DDNet model by 1.3%. This improvement is significant, especially considering the challenging nature of the dataset, which follows a subject-independent testing protocol. The challenge arises from the high variability in how different subjects perform the same gesture and the substantial inter-similarity between different gestures.

Table 3: Comparison with the baseline method.

Model name	Accuracy (%)
DDNet (baseline) [7]	84.20
Proposed method TDDNet	**85.50**

Figure 5 and 6 present the confusion matrices for both the baseline model (DDNet) and the proposed model. Both models demonstrate good classification

performance across most gesture classes. The proposed method outperforms the baseline method in majority classes except "Look" and "Pickpart". Although the results are promising, we can observe that for certain gestures, such as "Look" "PickPart" "Report" "Identification" and "Change", the performance of both methods is still low due to their similar execution patterns. Figure 7 illustrates two gestures with almost identical movements, where the key distinguishing feature is the positioning of the fingers, which makes the models more susceptible to misclassification.

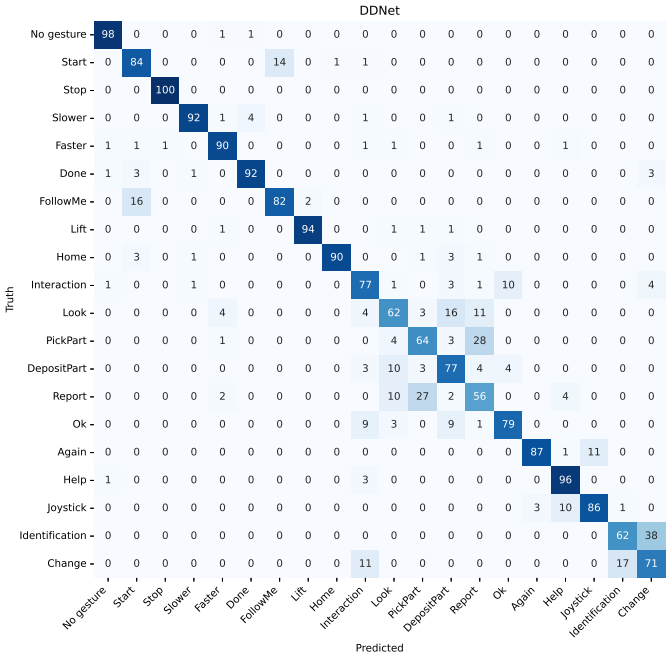


Fig. 5: Confusion matrix obtained by the baseline model DDNet

Figure 8 shows the t-SNE visualization of the gesture class distribution for both the baseline and proposed models. The proposed model exhibits better feature clustering, with gestures of the same class forming tighter clusters, indicating improved feature representation.

Continuous gesture recognition results The results for continuous gesture recognition are shown in Table 4, while Figure 9 illustrates an example of recognition for continuous videos from subject ID 20. It can be observed that

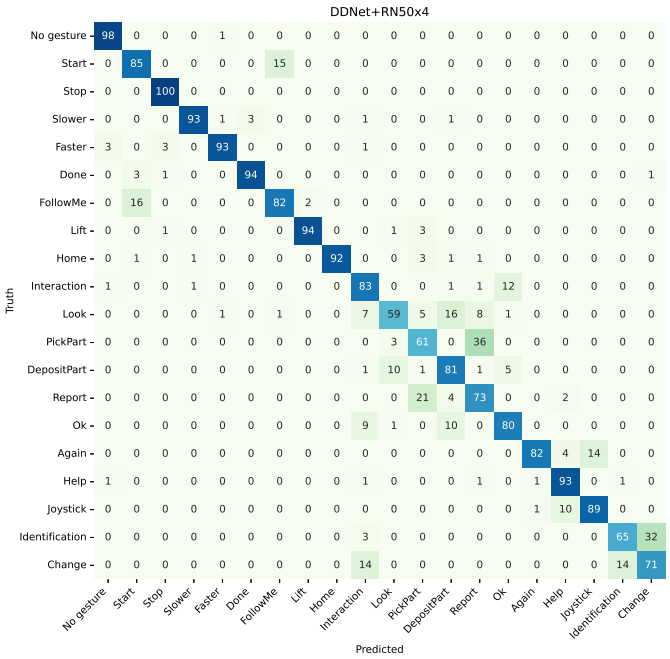


Fig. 6: Confusion matrix obtained by the proposed method TDDNet.

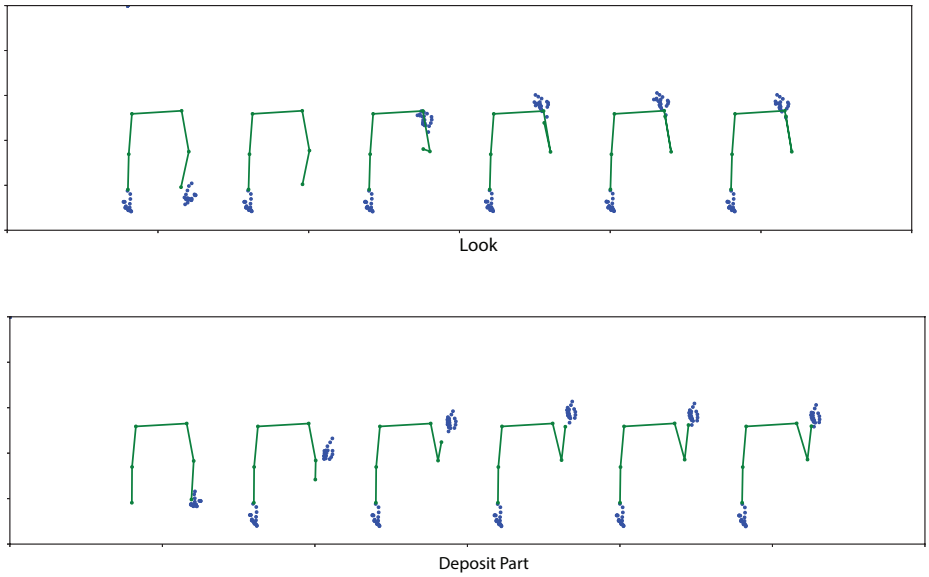
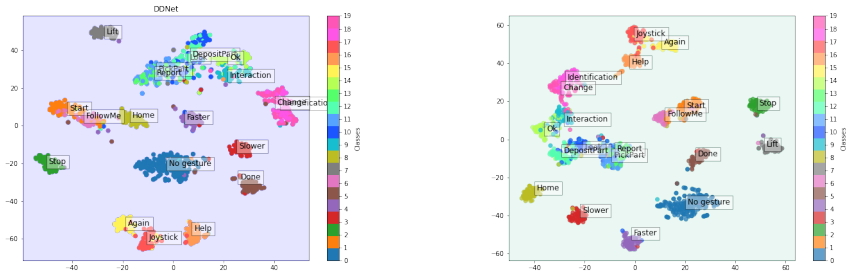


Fig. 7: Skeleton sequence of "Deposit Part" and "Look" gesture



(a) DDNet **(b)** Proposed model

Fig. 8: The t-SNE visualization of gesture classes distribution when using the baseline model DDNet (a) and the proposed method (b).

the proposed method outperforms also the baseline model for continuous recognition. Although the improvement with 0.6% in term of frame wise accuracy is not significant, however, from the visualization in Fig. 9, the proposed method does not results many fragments.

Table 4: Results obtained by the proposed method for continuous gesture recognition.

Model name	**Frame wide accuracy (%)**
DDNet (baseline)	66.00
Proposed method	66.60

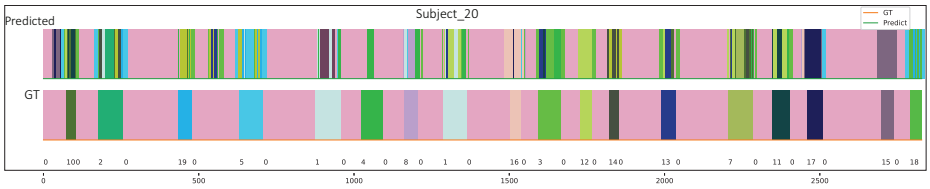


Fig. 9: Continuous recognition results for subject ID 20.

5 Conclusions and future works

In this paper, a lightweight deep learning model for continuous gesture recognition based on skeleton information was proposed. The model operates in two stages: isolated gesture recognition and a sliding window-based approach for continuous recognition. To improve performance in the isolated gesture recognition

step, we incorporate a text encoder model based on CLIP into the DDNet architecture. Experimental results on the self-collected dataset of 19 gestures demonstrate that our model improves isolated gesture recognition accuracy from 84.2% to 85.5%. For continuous gesture recognition, the model achieves a performance of 66.60%, compared to 66.00% achieved by the baseline model. Although the results for isolated gesture recognition are promising, the performance in continuous gesture recognition still requires improvement. Future work should explore action detection frameworks that allow for simultaneous action localization and recognition from a single representation to enhance continuous gesture recognition.

Acknowledgements This research was funded by the Vietnam Ministry of Education and Training under grant number B2024-GHA-11.

References

1. Abavisani, M., Joze, H.R.V., Patel, V.M.: Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1165–1174 (2019) 3
2. Barattini, P., Morand, C., Robertson, N.M.: A proposed gesture set for the control of industrial collaborative robots. In: 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication. pp. 132–137. IEEE (2012) 8
3. Brown, T.B.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020) 3
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299 (2017) 3
5. Dallel, M., Havard, V., Dupuis, Y., Baudry, D.: A sliding window based approach with majority voting for online human action recognition using spatial temporal graph convolutional neural networks. In: Proceedings of the 2022 7th International Conference on Machine Learning Technologies. p. 155–163. ICMLT '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3529399.3529425, https://doi.org/10.1145/3529399.3529425 2
6. Dallel, M., Havard, V., Dupuis, Y., Baudry, D.: A sliding window based approach with majority voting for online human action recognition using spatial temporal graph convolutional neural networks. In: 2022 7th International Conference on Machine Learning Technologies (ICMLT). pp. 155–163 (2022) 3, 6
7. Fan Yang, Sakriani Sakti, Y.W., Nakamura, S.: Make skeleton-based action recognition model smaller, faster and better. In: ACM International Conference on Multimedia in Asia (2019) 1, 2, 10
8. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7157–7173 (2022) 3

9. Gammulle, H., Ahmedt-Aristizabal, D., Denman, S., Tychsen-Smith, L., Petersson, L., Fookes, C.: Continuous human action recognition for human-machine interaction: A review. ACM Comput. Surv. **55**(13s) (jul 2023). https://doi.org/10.1145/3587931, https://doi.org/10.1145/3587931 1
10. Gao, Q., Liu, J., Ju, Z.: Hand gesture recognition using multimodal data fusion and multiscale parallel convolutional neural network for human-robot interaction. Expert Systems **38**(5), e12490 (2021) 3
11. Hu, Q., Gao, Q., Gao, H., Ju, Z.: Skeleton-based hand gesture recognition by using multi-input fusion lightweight network. In: International Conference on Intelligent Robotics and Applications. pp. 24–34. Springer (2022) 3
12. Kim, J.W., Choi, J.Y., Ha, E.J., Choi, J.H.: Human pose estimation using mediapipe pose and optimization method based on a humanoid model. Applied Sciences **13**(4), 2700 (2023) 3
13. Kwolek, B.: Continuous hand gesture recognition for human-robot collaborative assembly. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2000–2007 (2023) 3
14. Li, Y., Miao, Q., Qi, X., Ma, Z., Ouyang, W.: A spatiotemporal attention-based resc3d model for large-scale gesture recognition. Machine Vision and Applications **30**, 875–888 (2019) 3
15. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J.: 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. pp. 4207–4215 (2016) 6
16. Narayan, S., Mazumdar, A.P., Vipparthi, S.K.: Sbi-dhgr: Skeleton-based intelligent dynamic hand gestures recognition. Expert Systems with Applications **232**, 120735 (2023) 3
17. Nguyen, T.T., Nguyen, N.C., Ngo, D.K., Phan, V.L., Pham, M.H., Nguyen, D.A., Doan, M.H., Le, T.L.: A continuous real-time hand gesture recognition method based on skeleton. In: 2022 11th International Conference on Control, Automation and Information Sciences (ICCAIS). pp. 273–278. IEEE (2022) 2, 3
18. Qu, H., Cai, Y., Liu, J.: Llms are good action recognizers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18395–18406 (2024) 3
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:231591445 6, 10
20. Robinson, N., Tidd, B., Campbell, D., Kulić, D., Corke, P.: Robotic vision for human-robot interaction and collaboration: A survey and systematic review. ACM Transactions on Human-Robot Interaction **12**(1), 1–66 (2023) 1
21. Sincan, O.M., Keles, H.Y.: Using motion history images with 3d convolutional networks in isolated sign language recognition. IEEE Access **10**, 18608–18618 (2022) 3
22. Tan, M.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019) 10
23. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Open and efficient foundation language models. Preprint at arXiv. https://doi. org/10.48550/arXiv **2302** (2023) 3

24. Villani, V., Secchi, C., Lippi, M., Sabattini, L.: A general pipeline for online gesture recognition in human–robot interaction. IEEE Transactions on Human-Machine Systems **53**(2), 315–324 (2023) 3
25. Xiang, W., Li, C., Zhou, Y., Wang, B., Zhang, L.: Generative action description prompts for skeleton-based action recognition (2022). https://doi.org/10.48550/ARXIV.2208.05318, https://arxiv.org/abs/2208.05318 2, 6
26. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence **32** (01 2018). https://doi.org/10.1609/aaai.v32i1.12328 3
27. Yang, C.L., Li, W.T., Hsu, S.C.: Skeleton-based hand gesture recognition for assembly line operation. In: 2020 International Conference on Advanced Robotics and Intelligent Systems (ARIS). pp. 1–6. IEEE (2020) 3
28. Yang, F., Wu, Y., Sakti, S., Nakamura, S.: Make skeleton-based action recognition model smaller, faster and better. In: Proceedings of the 1st ACM International Conference on Multimedia in Asia. pp. 1–6 (2019) 2, 3, 6
29. Yu, J., Qin, M., Zhou, S.: Dynamic gesture recognition based on 2d convolutional neural network and feature fusion. Scientific Reports **12**(1), 4345 (2022) 3
30. Zhong, E., Del-Blanco, C.R., Berjón, D., Jaureguizar, F., García, N.: Real-time monocular skeleton-based hand gesture recognition using 3d-jointsformer. Sensors **23**(16), 7066 (2023) 3