Indonesian Journal of Electrical Engineering and Computer Science
Vol. 27, No. 1, July 2022, pp. 214∼221
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v27.i1.pp214-221

End-to-end multiple modals deep learning system for hand posture recognition

Huong-Giang Doan[1], Ngoc-Trung Nguyen[2]
[1]Department of Control and Automation, Electric Power University, Hanoi, Vietnam
[2]Department of Research Management and International Cooperation, Electric Power University, Hanoi, Vietnam

Article Info

Article history:

Received Oct 28, 2021
Revised Apr 19, 2022
Accepted Jun 02, 2022

Keywords:

Deep learning
End-to-end system
Hand posture recognition
Human-machine interaction
Multi-modality

ABSTRACT

Multi-modal or multi-view dataset that was captured from various resources (e.g. RGB and Depth) of a subject at the same time. Combination between different cues has still faced to many challenges as unique data and complementary information. In adition, the proposed method for multiple modalities recognition consists of discrete blocks, such as: extract features for separative data flows, combine of features, and classify gestures. To address the challenges, we proposed two novel end-to-end hand posture recognition frameworks, which are integrated all steps into a convolution neuronal network (CNN) system from capturing various types of cues (RGB and Depth images) to classify hand gesture labels. Both frameworks use the Resnet50 backbone that was pretrained by ImageNet dataset. We proposed a novel end-to-end multi-modal frameworks, which are named attention convolution module (ACM) and gated concatenation module (GCM). Both of them are deployed, evaluated and compared on various multi-modalities hand posture datasets. Experimental results show that our proposed method outperforms with others state-of-the-art techniques (SOTA) methods.

Corresponding Author:

Huong-Giang Doan
Department of Control and Automation, Electric Power University
235 Hoang Quoc Viet, Hanoi, Vietnam
Email: giangdth@epu.edu.vn

1. INTRODUCTION

Hand gesture has been become a natural way for human-machine interaction. In the literature, reviews in [1]-[3] have focused on many aspects of the recognition approaches. These reviews show that hand gesture recognition has still faced to many challenges as non-rigid hand, complex background, real-time system and hand recognition accuracy. Many relevant approaches have been proposed using both hand craft-based features: hand shape as speeded up robust feature (SURF) [4], kernel descriptor (KDES) [5]) or data driven-based approaches with deep neuronal networks in [6]-[8]. Performances of the deep learning approaches are out-performance the hand craft-based features ones in most relevant tasks of hand such as detection, pose estimation, and gesture recognition. The main idea of state-of-the-art neuronal networks (i.e., based on convolution neuronal networks (CNN)) to extract robust features [9]-[11]. In relevant works, performances of hand gestures recognition reach very impressive results in constrained (or lab-based) environments. However, in many practical applications, recognizing hand gesture needs to be considered in the unconstrained environments, multi-modalities, end-to-end system. The end-to-end multiple modalities will be main obstacles of the current deep neuronal networks. In these condition, current hand gesture recognition becomes a new bottleneck of

the pre-trained model. Therefore, investigating performances and tackle issues of the current state-of-the-art techniques will be a crucial step in the pipeline for gesture-based interface applications.

It is noticed that although to recognize hand-in-wild gestures, a number of other techniques have been proposed using different types of sensor such as radar [12], sonar [13], physical-motion [14]. While radar is often used to detect objects with electromagnetic wave beams but it is only efficient with larger objects without detail in fast movements. Physical sensors can capture from detail to global movement of the hand, but it is quite difficult to explain the achieved results. Vision-based sensors or cameras capture rich information about objects then are widely utilized for many applications. Therefore, recognizing the hand gestures collected from the vision-based sensors should be carefully considered in unconstrained environments. Some of complex algorithms used were hybrid approach with Depth maps (HAGR-D) [15], other modality of hand gesture dataset as surface electromyography hand gesture [16] or spatio-pemporal convolution neuronal networks [17], [18] are out of the scope in this study.

In the literature works, while recent deep neuronal networks reach the mature level for hand relevant tasks such as hand detection, segmentation, and tracking; deployments of these advantaged techniques in practical applications seem existing a big gap. Therefore, investigation of performances of gesture recognition and denoting common issues based on current state-of-the-art techniques would be an initial step towards development of a standard framework for gesture-based interface application. In this study, we propose new end-to-end frameworks using multiple cues for hand gesture recognition. Then, we examine performances of our methods for the hand gestures recognition via a series of public hand gesture datasets. Consequently, we report following questions: i) when does training of a new end-to-end convolution system saturate?; ii) what is the best end-to-end framework using multiple modals? If these state-of-the-art could be only suitable for different common gestures datasets?; iii) how is a gap time costs and memory usages between end-to-end models.

2. PROPOSED METHOD

Our proposed end-to-end CNN frameworks are illustrated in Figure 1. This framework consists of four methods. Firstly, two single modal flows: i) RGB hand gesture recognition and ii) Depth hand gesture recognition. These single models are utilized to compare with combination modalities; Secondly, two end-to-end multi-modal flows: iii) attention convolution module (ACM) and iv) gated concatenation module (GCM). These models combine both RGB image and Depth image of the same hand gesture. Given RGB image and Depth image, data are synchronously putted into the $Resnet50_{RGB}$ (blue blocks) and $Resnet50_{Depth}$ (brown blocks) networks to extract features at FC layers. The pretrained Resnet50 model [6] is applied and transferred learning for RGB and Depth hand gesture task. The ACM method (yellow block) utilizes attention convolutions and the GCM framework (green block) uses convolution layers to combine two features. These proposed frameworks are presented in detail in section 2.2 and section 2.3.

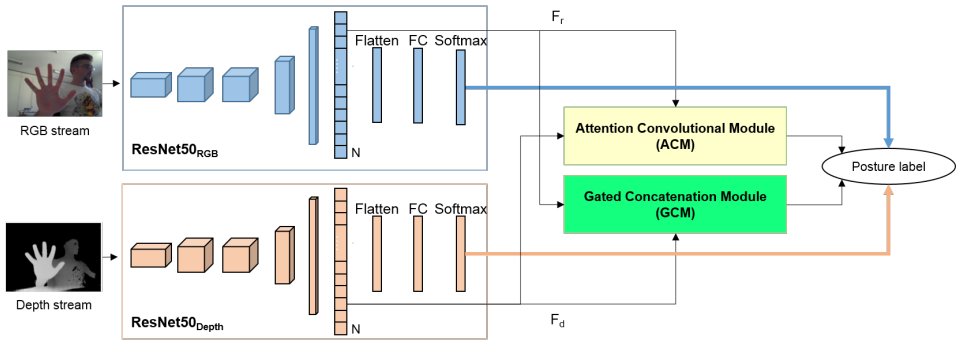


Figure 1. Propose end-to-end CNN strategies for hand posture recognition

2.1. Single hand gesture recognition

We use the pretrained Resnet50 model [6] which was trained on ImageNet dataset. This is very big dataset but it is not hand gesture. In this work, the Resnet50 model is fine-turned by hand gesture datasets with two modalities. The setup of Resnet50 models are presented in detail in the 2^{nd} column of Table 1.

Experimental results of two single frameworks and the discretely combination method - distributed

control system (DCS) framework, which were deployed in our previous works. Moreover, the DCS multi-modal method is not end-to-end method. It composed a serial of discrete steps, such as: Resnet50 extractors, normalization, concatenation and support vector machines (SVM) classifier. The apart of results of these methods are presented in this paper to compare with our proposed end-to-end multi-modal frameworks as presented in detail in section 3.

Table 1. The detail setup of CNN architectures

Parameter	Resnet50	Attention conv module	Gated concatenation module
Learning rate	$5*10^{-5}$	10^{-5}	$5*10^{-5}$
Batch size	32 images	16 images	16 images
Optimizer	Adam	Adam	Adam
Loss function	Cross entropy	Cross entropy	Cross entropy
Transfer learning	All layer Resnet	All layer Resnet + attention Conv	All layer Resnet + Gated concatenation conv
Input image	224x224 pixels	224x224 pixels	224x224 pixels
Output feature	2048	2048	2048

2.2. Attention module for end-to-end hand gesture recognition framework

In this work, Resnet50 model [6] is backbone in this our proposed method. The 2D CNN models are transferred learning by hand gesture datasets. The Resnet50 models are independently retrained and updated parameters following RGB dataset and Depth dataset, respectively. The setup of the ACM architecture is presented in the 3^{rd} column of Table 1. The first end-to-end multi-modal framework is illustrated in below Figure 2.

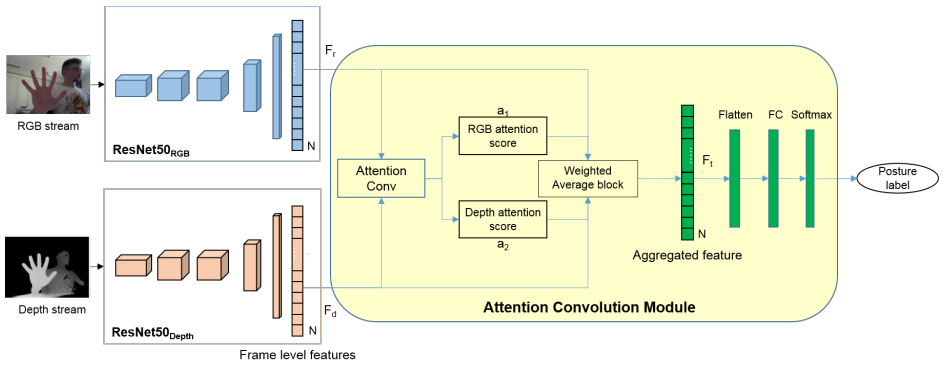


Figure 2. Aggregated multi-modalities framework with attention convolution module

Both retrained 2D RGB CNN model and 2D Depth CNN model are used as the feature extractors for image level features. RGB image I_r and Depth image I_d are the inputs of these 2D CNN extractors. The outputs of $Resnet50_{RGB}$ and $Resnet50_{Depth}$ extractors are taken from FC6 layers with ($F_r[1, ..., N]$ and $F_d[1, ..., N]$, N = 2048). Both features are normalized into $F_{ra}^{L2}[1, ..., N]$ and $F_{da}^{L2}[1, ..., N]$ as (5).

$$F_{ra}^{L2} = ||F_r||; F_{da}^{L2} = ||F_d|| \tag{1}$$

In the next step, an attention conv layer (2, N, 1) is applied for images-level features to generate attention scores. The attention conv utilizes both F_r and F_d as inputs that output of this block are attention scores $a_i, i = (1, M), M = 2$; i=1 corresponds to $a_1 = a_r$ and i=2 corresponds to $a_2 = a_d$, respectively. The attention scores are calculated by $Sigmoid$ function and L_1 normalization function [19] as presented in (2).

$$a_i = \frac{\sigma^{x_i}}{\sum_{n=1}^{M} \sigma^{x_i}} = \frac{\frac{1}{1-e^{x_i}}}{\sum_{n=1}^{M} \frac{1}{1-e^{x_i}}} \tag{2}$$

The attention conv trains and generates attention factors following roles of features. It helps to present affects of feature vectors through attention scores. The attention weights are applied for both RGB and Depth features to obtain $F_t[1, ..., N]$ feature. The aggregated feature is built based on single features and efficient scores that is presented in detail in (3).

$$F_t = \frac{1}{M}\sum_{i=1}^{M} a_i f_i \tag{3}$$

In this work, the lost function of Resnet50 models are exploited for RGB CNN model and Depth CNN model. In addition, Softmax cross-entropy loss function is also utilized to train the attention networks, and classify hand posture. Given predict result of hand gesture $\bar{p}_i$, ground truth is p_i, loss function is calculated as illustrated in (4).

$$L_{softmax} = \frac{1}{K}\sum_{i=1}^{K} p_i log\bar{p}_i \tag{4}$$

2.3. Gated concatenation module for end-to-end hand gesture recognition framework

As presentation in the previous section 2.2, this framework (Figure 3) also utilizes the Resnet50 backbone to extract single modality RGB image I_r and Depth image I_d. Two frame feature levels F_r and F_d are inputs of the gated concatenation module, we utilize concatenation Conv to train and combine two features F_r and F_d. Then, aggregated feature is used to classify hand gesture labels. The setup of 2D CNN gated connection architecture is presented in detail in the 4^{rd} column of Table 1.

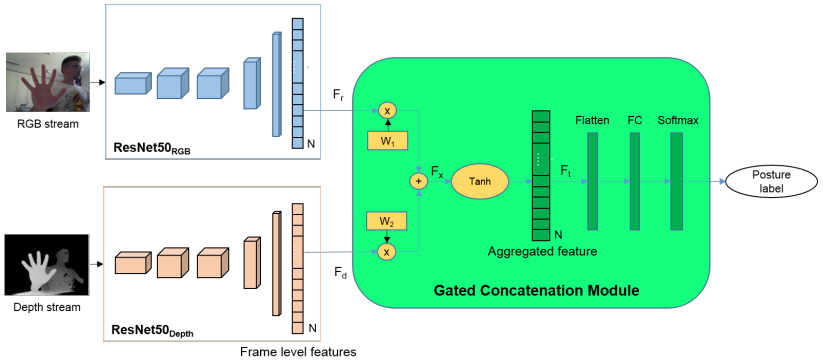


Figure 3. Aggregated multi-modalities framework with gated concatenation module

Given 2048-D feature vectors of RGB stream $F_r[1, ..., N]$ and Depth stream $F_d[1, ..., N]$ (N=2048), they are L_2 normalized into F_r^{L2} and F_d^{L2} as presented in (5).

$$(F_{rg}^{L2}; F_{dg}^{L2}) = (||F_r; F_d||) \tag{5}$$

Then, two conv layers with weights w_1, w_2 and $Tanh$ function are utilized to combine a unimodal feature $Fx[1, .., N]$ and $Ft[1, .., N]$ as illustrated in (6) and (7).

$$F_x = F_{rg}^{L2} * w_1 + F_{dg}^{L2} * w_2 \tag{6}$$

$$F_t = tanh(F_x) = \frac{e^{F_x} - e^{F_x}}{e^{F_x} + e^{F_x}} \tag{7}$$

Then, F_t aggregated feature is input of classification block that utilizes cross-entropy loss function as presented in (4). Output is a corresponding hand posture label [1,...,K]. Where K is number of hand gesture class in a dataset.

3. RESULTS AND DISCUSSION

In this paper, we follow "One-leave-subject-out" protocol [20] in order to evaluate efficiency of evaluations. It is mean that only one subject is utilized for testing and remaining people will be used for training model. Experiments are rolled for every subjects in a dataset to ensure that every people could be tested. Hand

gestures of a subject are used for testing that does not appear on training phase. This evaluation protocol is implemented for five datasets with multiple modalities (RGB image and Depth image), such as: KinectLeap [21], HKU [22], Zen3D [23], ASLFinger [24] and LaRED [25] datasets. In this research, we conduct experiments to evaluate: i) transfer learning processes of various end-to-end frameworks; ii) efficiency of various end-to-end multiple modalities (combination of RGB image and Depth image) frameworks versus single modality (either RGB image or Depth image); iii) time cost and memory usages of different end-to-end frameworks. The evaluation codes are written in Python; deep learning framework is Pytorch and run on a work-station with NVIDIA GPU 11G.

3.1. Transfer learning results

A glance at the below Figure 4 that Figure 4(a) and Figure 4(b) show loss and accuracy values of the ACM framework while Figure 4(c) and Figure 4(d) show loss and accuracy values of the GCM framework. These works are evaluated on various multi-modal datasets, such as: KinectLeap [21], HKU [22], Zen3D [23], ASLFinger[24] and LaRED[25] datasets. Retraining processes of these models are saturated after 20 to 75 epochs depending on datasets and CNN architecture. As shown, the transfer learning model of the ACM framework Figure 4(a) and Figure 4(b) are saturated after 30 epochs for HKU dataset and Zend3D dataset; remaining datasets require upto 70 epochs. Particularly, the end-to-end GCM method obtains both accuracy and loss values of which are converge and stability after 20 epochs when we train the CNN models with small hand gesture datasets.

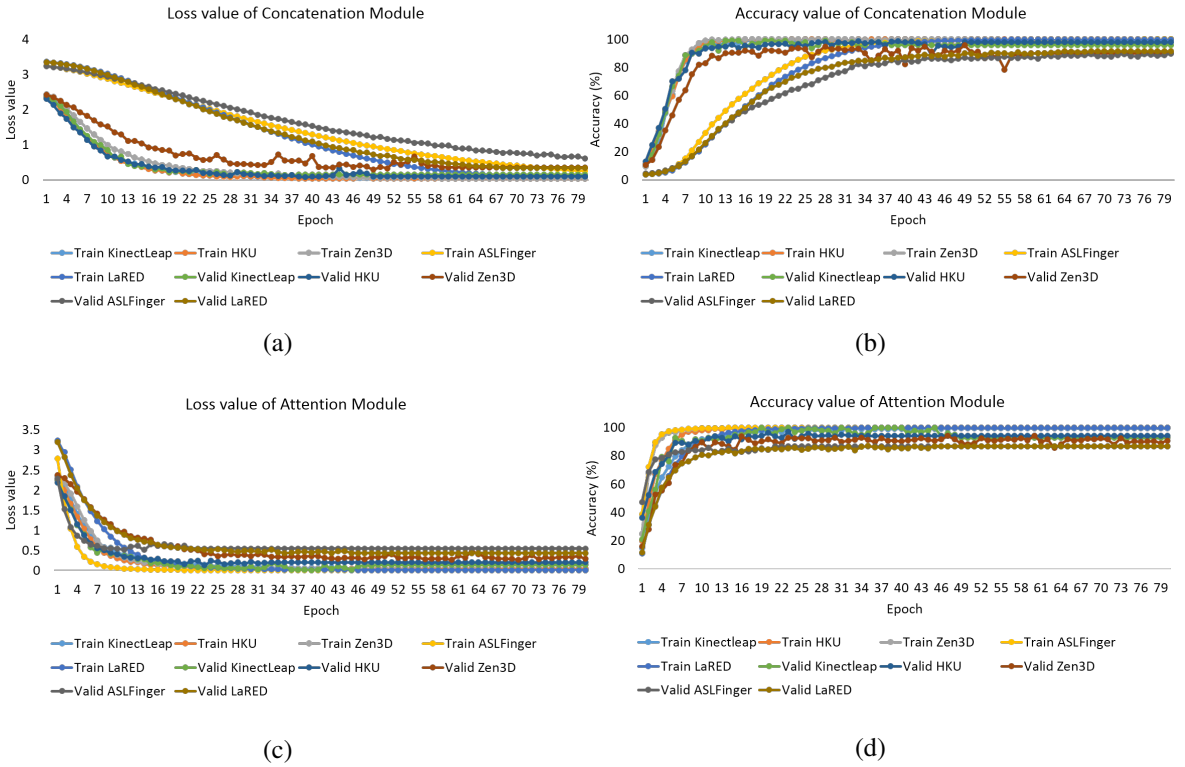
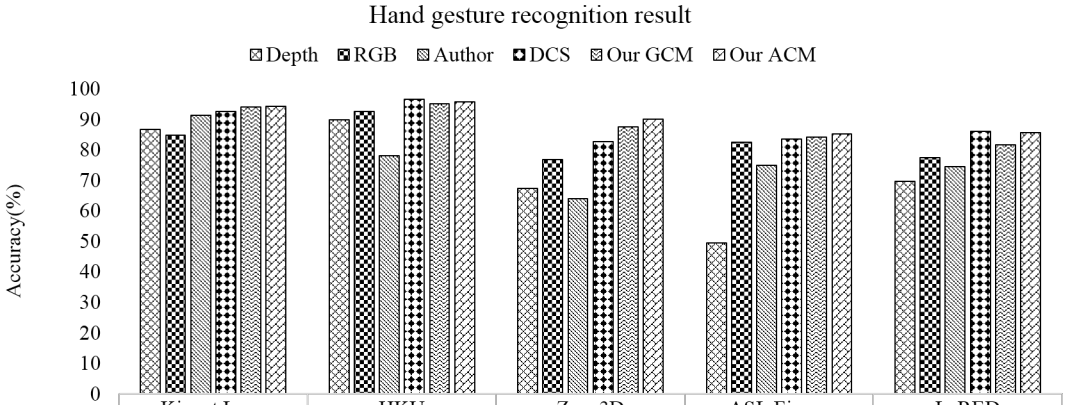


Figure 4. Loss and accuracy values of two end-to-end ACM and GCM frameworks on various datasets (a) loss values of GCM, (b) accuracy values of GCM, (c) loss values of ACM, and (d) accuracy values of ACM

These results show that the fine-tune scheme deploys a CNN transfer learning for backbone model and training for convolution layer of framework successfully based on complexity of datasets and CNN architecture. For the same dataset, the ACM framework's training processes are extremely faster than the GCM framework. In addition, applying the transfer learning model on testing data, we realized that most of these models outperform recognition rates, reported in the original works, as given in below section 3.2.

3.2. End-to-end hand gesture recognition

In this part, we examine hand gesture recognition with combining of RGB and Depth features as presented in section 2. We evaluate the efficiency of multi-modalities with our proposed GCM and ACM architectures. These networks are evaluated on five aforementioned RGB and Depth datasets. The "Leave-one-subject-out" protocol is utilized in this experimental as presented in detail in our research [5]. The evaluation results are illustrated in Figure 5 with following dominant observations: i) hand gesture recognition results are improved when both RGB and Depth cues are combined. The combination of RGB and Depth achieves far higher accuracy on five datasets and using three various end-to-end CNN architectures as DCS, GCM and ACM networks; ii) the ACM framework obtains the best results over almost datasets. For the Zen3D dataset, ACM accounts recognition accuracy upto 89.98% while DCS is 82.73% and GCM is 87.42%. For ASLFinger dataset, the ACM method also obtains the highest results at 85.18% where the DCS method and the GCM method only achieve 83.51% and 84.1%, respectively. With LaRED, the DCS method has the highest results at 86.11% and the ACM method is sight lower than the GCM method at 85.54% and result of the GCM method is far smaller remaining methods at 81.68%; iii) the DCS method is non end-to-end method while accuracy results are far lower than remaining multi-modality method; iv) both the GCM method and the DCM method are not only end-to-end CNN framework but also highest accuracy. Nevertheless, these architectures could be considered with complementary aspections as presented in detail in section 3.3.



	Kinect Leap	HKU	Zenz3D	ASL Finger	LaRED
Depth	86.57	89.75	67.35	49.34	69.69
RGB	84.86	92.6	76.74	82.41	77.51
Author	91.28	78	64	74.8	74.55
DCS	92.5	96.5	82.73	83.51	86.11
Our GCM	94.04	95.15	87.42	84.1	81.68
Our ACM	94.2	95.75	89.98	85.18	85.54

Figure 5. Hand posture recognition accuracy of the end-to-end CNN methods

3.3. Time cost and memory usage

In term of computational cost, Table 2 reports the time cost and memory usage of three end-to-end hand gesture models on five multi-modal datasets. The ACM architecture requires the highest computational cost and memory usage. This results indicate that: i) the GCM model is a real-time system which requires small memory size. Memory capacitance of the GCM model is around 180 MB. While the ACM model requires the largest memory, it is around 278 MB; ii) the ACM model requires extremely high time cost while it model size is far larger than the GVM model on all datasets. The GCM network time cost is a little higher than the DCS architecture's one. By combining both accuracy rate and computational cost, the GCM framework could be recommended to develop end-to-end hand gesture-based interfaces. It is robust in term of gesture recognition, and also achieves the real-time performances.

Table 2. The time cost and memory usage of end-to-end CNN architectures

Dataset	DCS		GCM framework		ACM framework	
	Time cost (ms)	Memory usage (MB)	Time cost (ms)	Memory usage (MB)	Time cost (ms)	Memory usage (MB)
KinectLeap	24.58	180.02	35.71	180.32	55.71	278.05
HKU	9.42	180.02	14.75	180.32	20.45	278.05
Zen3D	15.27	180.02	20.37	180.33	35.15	278.06
ASLFinger	6.51	180.24	8.81	180.43	15.65	278.16
LaRED	3.48	180.24	9.6	180.45	12.73	278.18

4. CONCLUSION

This paper presents the new end-to-end CNN architectures for hand gesture recognition, which are evaluated on five various public datasets. Among the evaluations, the ACM framework is the best option because it achieved the trade-off between accuracy rate and the computational cost. Evaluation results on the common datasets also denoted that combination of multiple modalities (e.g. RGB image, Depth image) achieves better performances. This proposed method opens critical research directions which require further investigation on data augmentation, multi-view gesture analysis for development of a gesture-based interface in practical applications is feasible.

ACKNOWLEDGEMENT

This research is funded by Electric Power University (EPU) under grant project title "Research on multi-modal and multi-view hand gesture recognition combines sensors and image information".

REFERENCES

[1] M. M. Hasan and P. K. Mishra, "Hand gesture modeling and recognition using geometric features: a review," *Canadian Journal on Image Processing and Computer Vision*, vol. 3, pp. 12-26, 2012.

[2] J. Suarez and R. R. Murphy, "Hand gesture recognition with Depth images: A review," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012, pp. 411-417, doi: 10.1109/ROMAN.2012.6343787.

[3] P. Pisharady, P. Vadakkepat, and A. Loh, "Attention based detection and recognition of hand postures against complex backgrounds," *Int J Comput Vis*, vol. 101, no. 3, pp. 403-419, 2013, doi: 10.1007/s11263-012-0560-5.

[4] W. X. H. Tang, H. Liu and N. Sebe, "Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion," *NeuroComputing*, vol. 331, pp. 0925-2312, 2019, doi: 10.1016/j.neucom.2018.11.038.

[5] H.-G. Doan, V.-T. Nguyen, H. Vu, and T.-H. Tran, "A combination of user-guide scheme and kernel descriptor on rgb-d data for robust and realtime hand posture recognition," *Eng. Appl. Artif. Intell.*, vol. 49, no. C, pp. 103-113, 2016, doi: 10.1016/j.engappai.2015.11.010.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[7] A. Howard, *et al*., "Searching for mobilenetv3," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1314-1324, doi: 10.1109/ICCV.2019.00140.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.

[9] N. M. H. Aseel Ghazi Mahmoud, Ahmed Mudheher Hasan, "Convolutional neural networks framework for human hand gesture recognition," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2223-2230, 2021, doi: 10.11591/eei.v10i4.2926.

[10] A. Dixit and T. Kasbe, "Multi-feature based automatic facial expression recognition using deep convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 3, pp. 1406-1419, 2022, doi: 10.11591/ijeecs.v25.i3.pp1406-1419.

[11] A. E. Minarno, W. A. Kusuma, and Y. A. Kurniawan, "Human activity recognition for static and dynamic activity using convolutional neural network," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 19, no. 6, pp. 1857-1864, 2021, doi: 10.12928/telkomnika.v19i6.20994.

[12] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor," *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3278-3289, 2018, doi: 10.1109/JSEN.2018.2808688.

[13] M. Abavisani, H. Joze, and V. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1165-1174, doi: 10.1109/CVPR.2019.00126.

[14] W. Lu, Z. Tong, and J. Chu, "Dynamic hand gesture recognition with leap motion controller," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1188-1192, 2016, doi: 10.1109/LSP.2016.2590470.

[15] D. Santos, B. Fernandes, and B. Bezerra, "Hagr-d: A novel approach for gesture recognition with Depth maps," *Sensors*, vol. 15, no. 11, pp. 28646-28664, 2015, doi: 10.3390/s151128646.

[16] P. Kaczmarek, T. Mańkowski, and J. Tomczyński, "putEMG—A surface electromyography hand gesture recognition dataset," *Sensors*, vol. 19, no. 46, p. 3548, 2019, doi: 10.3390/s19163548.

[17] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 1-7, doi: 10.1109/CVPRW.2015.7301342.
[18] F. Armandika, E. C. Djamal, F. Nugraha, and F. Kasyidi, "Dynamic hand gesture recognition using temporal-stream convolutional neural networks," *2020 7th International Conference on Electrical Engineering, Computer Sciences and Informatics (EECSI)*, 2020, pp. 132-136, doi: 10.23919/EECSI50503.2020.9251902.
[19] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4694-4703, doi: 10.1109/CVPR.2017.499.
[20] D.-M. Truong, H.-G. Doan, T.-H. Tran, H. Vu, and T.-L. Le, "Robustness analysis of 3d convolutional neural network for human hand gesture recognition," *International Journal of Machine Learning and Computing*, vol. 9, no. 2, pp. 135-142, 2019, doi: 10.18178/ijmlc.2019.9.2.777.
[21] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with jointly calibrated leap motion and Depth sensor," *Multimedia Tools and Applications*, vol. 75, pp. 14991-15015, 2016, doi: 10.1007/s11042-015-2451-6.
[22] C. Wang, Z. Liu, and S.-C. Chan, "Superpixel-based hand gesture recognition with kinect Depth camera," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 29-39, Jan. 2015, doi: 10.1109/TMM.2014.2374357.
[23] A. Memo and P. Zanuttigh, "Head-mounted gesture controlled interface for human-computer interaction," *Multimedia Tools and Applications*, vol. 77, pp. 27-53, 2018, doi: 10.1007/s11042-016-4223-3.
[24] N. Pugeault and R. Bowden, "Spelling it out: Real-time asl fingerspelling recognition," in *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2011, pp. 1114-1119, doi: 10.1109/ICCVW.2011.6130290.
[25] Y.-S. Hsiao, J. Sanchez-Riera, T. Lim, K.-L. Hua, and W.-H. Cheng, "Lared: a large rgb-d extensible hand gesture dataset," in *Proceedings of the 5th ACM Multimedia Systems Conference*, 2014, pp. 53-58, doi: 10.1145/2557642.2563669.

BIOGRAPHIES OF AUTHORS

Huong-Giang Doan received B.E. degree in Instrumentation and Industrial Informatics in 2003, M.E. in Instrumentation and Automatic Control System in 2006 and Ph.D. in Control Engineering and Automation in 2017, all from Hanoi University of Science and Technology, Ha Noi, Vietnam. Her current researches are in fields of Human-Machine Interaction using image information, action recognition, deep learning, computer vision. She is a lecturer and researcher at Control and Automation Faculty, Electric Power University, Hoang Quoc Viet st., Ha Noi, Viet Nam. She can be contacted at Email: giangdth@epu.edu.vn.

Ngoc-Trung Nguyen received B.E degree in Power System in 2003, M.E in Electrical Engineering in 2006, all from Hanoi University of Science and Technology, Hanoi, Vietnam; received Ph.D in Electrical Engineering from University of Palermo, Palermo, Italy, in 2014. His current researches are in fields of Supervisory Control and Data Acquisition (SCADA), Smart-grid, Protection and Automation in Power System, Human Machine Interaction. He is lecturer and researcher at Department of Research Management and International Cooperation, Electric Power University, Hoang Quoc Viet st., Hanoi, Vietnam. He can be contacted at Email: trungnn@epu.edu.vn.