IMPROVING HAND POSTURE RECOGNITION PERFORMANCE USING MULTI-MODALITIES

NÂNG CAO HIỆU QUẢ NHẬN DẠNG HÌNH TRẠNG BÀN TAY KẾT HỢP NHIỀU LUỒNG DỮ LIỆU

Doan Huong Giang

Electric Power University

Ngày nhận bài: 26/02/2021, Ngày chấp nhận đăng: 16/03/2021, Phản biện: PGS.TS. Ngô Quốc Tạo

Abstract:

Hand gesture recognition has been researched for a long time. However, performance of such methods in practical application still has to face with many challenges due to the variation of hand pose, hand shape, viewpoints, complex background, light illumination or subject style. In this work, we deeply investigate hand representations on various extractors from independent data (such as RGB image and Depth image). To this end, we adopt an concatenate features from different modalities to obtain very competitive accuracy. To evaluate the robustness of the method, two datasets are used: The first one, a self-captured dataset that composes of six hand gestures in indoor environment with complex background. The second one, a published dataset which has 10 hand gestures. Experiments with RGB and/or Depth images on two datasets show that combination of information flows has strong impact on recognition results. Additionally, the CNN method's performances are mostly increased by multi-features combination of which results are compared with hand-craft-based feature extractors, respectively. The proposed method suggests a feasible and robust solution addressing technical issues in developing HCI application using the hand posture recognition.

Keywords:

Electronic Home Appliances, Deep Learning, Machine Learning, Hand Poseture/Gesture Recognition, Human Machine Interaction, Multi-modalities, Late Fusion, Early Fusion.

Tóm tắt:

Nhận dạng cử chỉ tay đã được nghiên cứu trong thời gian vừa qua. Tuy nhiên, đây vẫn là một mảng nghiên cứu còn phải đối mặt với nhiều thách thức nếu muốn triển khai trong thực tế do: tồn tại nhiều hình trạng bàn tay khác nhau, hình dáng của cùng một hình trạng, góc nhìn khác nhau, điều kiện nền phức tạp, điều kiện chiếu sáng, mỗi người có cách thức thực hiện khác nhau. Bài báo này sẽ nghiên cứu cách biểu diễn bàn tay sử dụng các bộ phân lớp khác nhau trên các luồng thông tin (ảnh màu RGB và ảnh độ sâu Depth). Sau đó, các đặc trưng được kết hợp với nhau để nâng cao hiệu quả của quá trình nhận dạng. Các thử nghiệm được thực hiện trên các bộ sơ sở dữ liệu (CSDL) khác nhau gồm bộ CSDL tự thu thập và bộ CSDL được công bố trên mạng cho cộng đồng nghiên cứu. Ngoài ra, tác giả cũng sử dụng mạng nơron nhân tạo để thử nghiệm và so sánh với các giải pháp sử dụng các bộ trích chọn đặc trưng tự thiết kế. Kết quả cho thấy giải pháp sử dụng mạng

nơron đạt kết quả tốt hơn so, trong đó giải pháp đề xuất kết hợp các luồng thông tin trên tất cả các bộ phân lớp đạt hiệu quả tốt hơn so với sử dụng từng luồng thông tin riêng biệt. Các kết quả này cho thấy, giải pháp đề xuất khả thi khi triển khai ứng dụng trong tương tác giữa người và thiết bị sử dụng cử chỉ của bàn tay.

Từ khóa:

Thiết bị điện tử gia dụng, học sâu, học máy, nhận dạng cử chỉ bàn tay, tương tác người - máy, đa thể thức, kết hợp muộn, kết hợp sớm.

1. INTRODUCTION

Hand gesture recognition has been become an attractive field in computer vision [5][11][17][19][20] because of huge range of it applications such as Human-Machine-Interaction (HCI) [15], entertainment, virtual reality [18][21], autonomous vehicles [15], and so on. Moreover, its system performance (accuracy, time cost,…) has been face to many challenges due to various appearances of hand poses, non-rigid objects, different scales, too many degrees of freedoms, illumination, complex of background. Thanks to the development of new and low-cost depth sensors, new opportunities for posture recognition have emerged. Microsoft Xbox-Kinect is a successful commercial product that provides both RGB and Depth information for recognizing hand gestures to control game consoles [12]. Combination these data could be considered to improve recognition results. Particularly, Convolutional Neuronal Networks (CNNs) [14][16] have been emerged as a promising technique to resolve many issues of the posture/gesture recognition. Although utilizing CNNs has obtained impressive results [13][15], there exists still many challenges that should be carefully carried out before applying it in reality.

The remaining of this paper is organized as follows: Section 2 describes our proposed approach. The experiments and results are analyzed in Section 3. Section 4 concludes this paper and recommends some future works.

2. PROPOSED METHOD

The main flow-work for hand pose recognition from RGB and Depth modalities consists of a series of the cascaded steps as shown in Fig. 1. Given RGB-Depth images, hand region regions are detected, extracted, and recognized. The steps are presented in detail as the next sections following:

2.1. Pre-processing data

RGB images (I_{RGB}) and Depth images (I_D) are captured by the Kinect camera version 1 (640×480 pixels). Because coordinate of pixels are reflected. It must be calibrated as presented detail in our previous research [7].

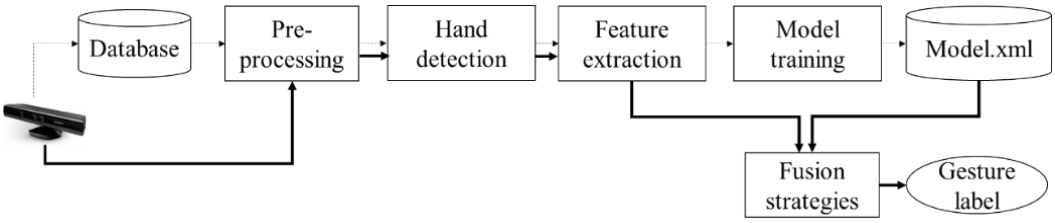


Fig. 1. Propose framework for hand posture recognition

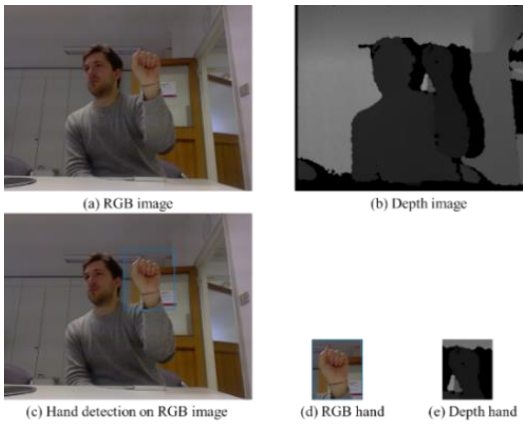


Fig. 2. Hand region detection

2.2. Hand detection

This step aims to have coordinate of hand region in image. Haar like cascade is used. That is an object detection algorithm in machine learning. This algorithm is proposed by [1]. It is composed by four stages: Haar features selection, Integral image creating, Adaboost training and Cascade classifiers. In this paper, we used pre-train Haar cascade model that is trained through all those steps and authors used a large hand dataset. The xml file of the pretrained model is published at [2]. We use three type models of hand parts such as: Palm.xml, Wrist.xml and Hand.xml.

Given an input RGB image I_{RGB} that is passed through three Haar like cascade models to detect RGB hand region I_{Hand}^{RGB} in a RGB image (Fig. 2(c,d)) as presented in following (1) equation:

$$I_{Hand}^{RGB} = F_{cropt}^{union}\left(F_{Palm}^{RGB}\left(I_{RGB}\right), F_{Wrist}^{RGB}\left(I_{RGB}\right), F_{Hand}^{RGB}\left(I_{RGB}\right)\right) \quad (1)$$

Next, coordinate of RGB hand region is marked on Depth image I_{Depth}. Depth hand region I_{Hand}^{Depth} (Fig. 2(e)) as illustrated in following (2) equation:

$$I_{Hand}^{Depth} = F_{cropt}^{\text{mark}}\left(I_{Hand}^{RGB}, I_{Depth}\right) \quad (2)$$

2.3. Hand posture representation

In this section, series of digit hand images I_{Hand}^{RGB} and I_{Hand}^{Depth} are then represented by both hand craft-based methods and deep learning-based method as Sec. 2.3.1. Then, digit hands are recognized as presented detail in the following Sec. 2.3.2.

2.3.1. Handcraft-based method

In this part, some state-of-the-art descriptors are used to extract features for hand pose such as: SIFT[7], SURF[8], HOG[9] and KDES[10]. By using those corresponding descriptors, the number of

most important key-points of hand í detected. Hand gesture is then presented by feature vectors such as: $F^{(2)} = F_{Sift}$; $F^{(3)} = F_{Surf}$; $F^{(4)} = F_{HOG}$ and $F^{(5)} = F_{KDES}$. Which is presented detail in equations (3), (4), (5) and (6) (while $N_1 = 256$; $N_2 = N_3 = N_4 = 1024$) following:

$$F^{(1)}_{RGB/Depth} = F_{SIRF} = [K^{(1)}_1 \quad ... \quad K^{(1)}_{N_1}]^T \qquad (3)$$

$$F^{(2)}_{RGB/Depth} = F_{SURF} = [K^{(2)}_1 \quad ... \quad K^{(2)}_{N_2}]^T \qquad (4)$$

$$F^{(3)}_{RGB/Depth} = F_{HOG} = [K^{(3)}_1 \quad ... \quad K^{(3)}_{N_3}]^T \qquad (5)$$

$$F^{(4)}_{RGB/Depth} = F_{KDES} = [K^{(4)}_1 \quad ... \quad K^{(4)}_{N_4}]^T \qquad (6)$$

2.3.2. Deep learning method

Recently, deep learning has been widely used in computer vision in various tasks as feature extraction, recognition, identification. In this research, Resnet50 model is utilized to extract feature of human hand. This convolutional neural network composes of 5 Conv layers and FC layer. The architecture of pertained Resnet50 network is illustrated in the following Fig. 7:

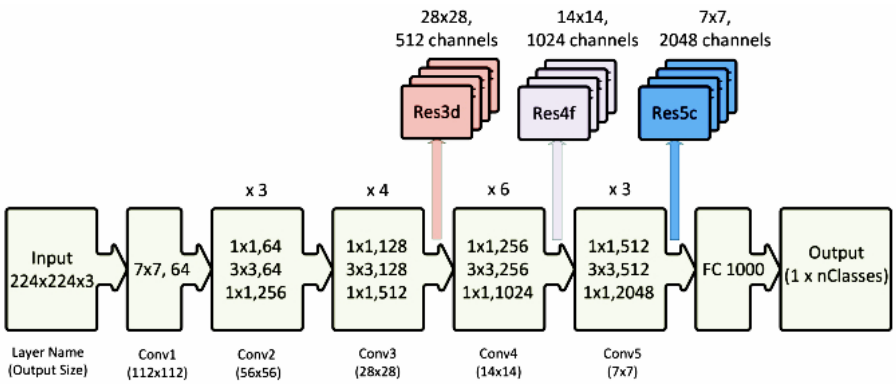


Fig. 3. The Resnet50 architecture

The cropped images I^{RGB}_{Hand} and I^{Depth}_{Hand} are different sizes that are then resized to (224×224pixels). These same dimension hand images are utilized as inputs of this Resnet50 convolutional neuron network. Then, this network is used as feature extractor. The dimension of output feature is taken at last FC layer of network and the feature size is presented by $F^{(6)} = F_{Resnet50}(1xN_5)(N_5 = 1000)$ as following (7) equation:

$$F^{(5)}_{RGB/Depth} = F_{Resnet50} = [K^{(5)}_1 \quad ... \quad K^{(5)}_{N_5}]^T \qquad (7)$$

2.4. Hand gesture classification

The features $F^{(1)};...;F^{(6)}$ (extracted from RGB and Depth information) are utilized as the inputs of the classification strategies *late fusion* and *early fusion* as presented in the following sections:

2.4.1. Early fusion strategy

In this part, modalities of hand posture representation that are extracted from RGB and Depth hand images (as presented in previous section) by five

descriptors. RGB and Depth features of the same extractors are combined together. Both of features are normalized in following equation:

$$F_{multi} = \left\{ \| \, F^{i}_{RGB}, F^{i}_{Depth} \, \| ; i = (1,2,3,4,5) \right\} \quad (8)$$

Next, F_{multi} features are inputs of the SVM classifier [6] that is utilized as multi-class SVM classifier. The output of multi-class SVM will be one number value among {0, 1, 2, …, N} with N is number ID of gestures in dataset.

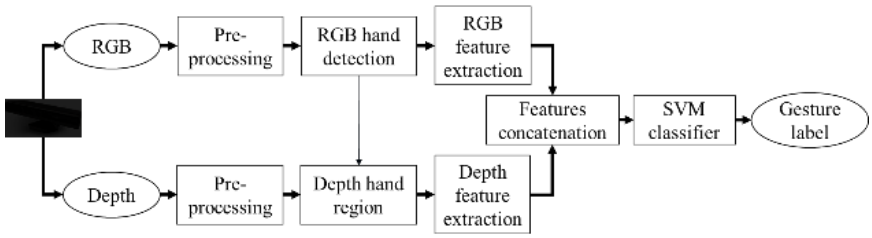


Fig. 4. Early fusion strategy

2.4.1. Late fusion strategy

Differ from early fusion method, in this case, exploiting features derived from multimodal data are independently used as input of the separate SVM classifiers [6] as illustrated in the following Fig. 5. Then, at decision layer, output scores of classifiers are decision vectors (D_{RGB} and D_{Depth}) that are combined to obtain the final results. This method requires one more classifier as well as long time cost than early fusion. Furthermore, it is more flexible and easy to expand model. Additionally, it allows to use the classifier that is best suitable for the each modality.

The results of fusion strategies are presented detail in the following Sec. 3.

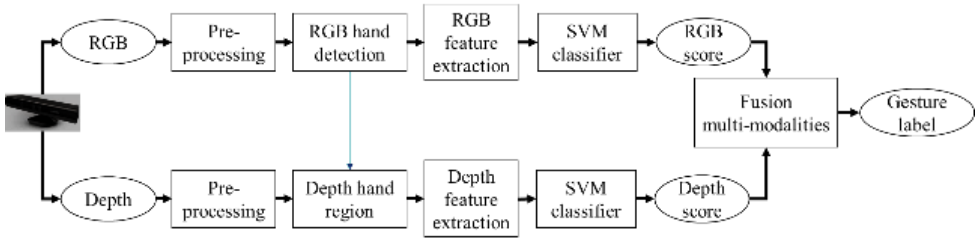


Fig. 5. Late fusion strategy

3. EXPRIMENTAL RESULTS

The proposed framework is warped by Python program on a PC Core i6 4.2 GHz CPU, 8GB RAM, NVIDIA 8G GPU. We evaluate performance of the hand gesture recognition on EPUHandPose2 dataset and KinectLeap [19] dataset. In entire evaluations, we follow Leave-p-out-cross-validation method as presented detail in [11], with p equals 1. It means that gestures of one subject are utilized for testing and the remaining subjects are

utilized for training. Three evaluations are considered such as: (1) How is the better between Hand-craft and Deep learning feature method; (2) Comparison of accuracy recognition rate between kernel SVM classifiers; (3) Compare accuracy of fusion strategies. The detail evaluations are presented as following sub-sections:

3.1. Evaluation of hand recognition rate on different feature representations

In this evaluation, we test the accuracy rate of various Hand-craft feature representation with SVM classifier. EPUHandPose2 dataset is used in this work. The accuracy is evaluated at independent modalities RGB and Depth. Look at the Fig. 8, it is apparent that the Resnet-based descriptor obtains the best percentage at both flows, 95.3% for RGB and 90.5 for Depth. While SIRF descriptor is lowest accuracy in overall that is stood at 34.5% and 28.3%, respectively. In Hand-craft extractors, KDES feature obtains the best evaluations (at 90.3% and 78.2%) that is dramatically higher than remain extractors. Moreover, accuracy of the hand craft-based methods (SIFT, SIRF, HOG, and KDES extractors) are far smaller than the deep learning-based approach. Therefore, in this paper, the Resnet model and KDES model will be utilized in the next experiments.

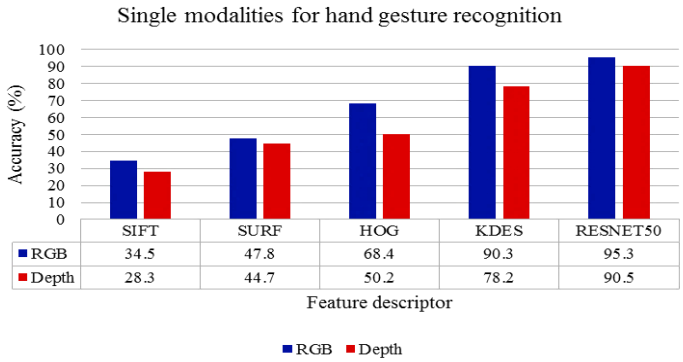


Fig. 6. Accuracy with the different feature representations

3.2. Comparison between Kernels of SVM classifiers

In this Section, two evaluations are performed on four Kernels of SVM classifiers: Linear, Sigmoid, RBF and Poly. Two datasets are used in this evaluation as: EPUHandPose2 dataset and KinectLeap[19] dataset.

In the first case, the best KDES feature extractor is used. A glance at the Table. 1, it is clear that, the RBF kernel obtains the best accuracy on all modalities in both two datasets, at 90.3% and 78.2% for RGB and Depth of EPU dataset while KinectLeap dataset are 78.4% and 60.2% respectively.

In the second cases, we try to evaluate on deep learning-based feature. Given the Table. 2 that shows a comparison of recognition accuracy of four various

Kernels at two datasets. It is evident from the information provided that RBF Kernel has by far the highest percentage in almost cases (at 97.2%, 92.6% for EPU dataset and 93.2% and 90.6% for KinectLeap dataset).

Table 1. Hand craft-based feature on various Kernels of SVM

Dataset / Kernel SVM	EPU dataset		KinectLeap	
	RGB	Depth	RGB	Depth
Linear	85.1	72.2	73.4	58.6
Sigmoid	15.6	11.2	13.6	10.8
Rbf	90.3	78.2	78.4	60.2
Poly	86.9	78.3	71.3	60.1

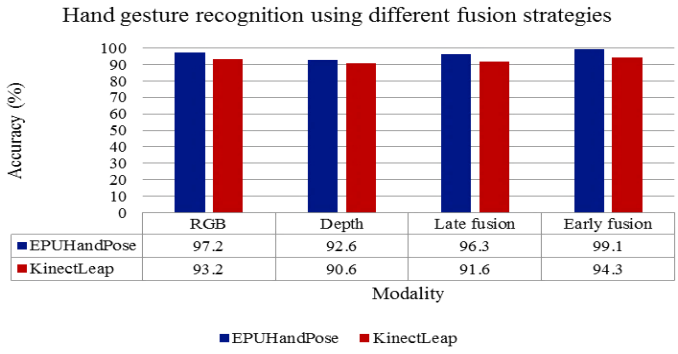
Table 2. Deep learning-based feature on various Kernels of SVM

Dataset / Kernel SVM	EPU dataset		KinectLeap	
	RGB	Depth	RGB	Depth
Linear	95.3	90.5	91.6	89.7
Sigmoid	25.7	23.4	27.8	15.4
Rbf	97.2	92.6	93.2	90.6
Poly	86.9	78.3	88.9	79.3

Moreover, results of Linear Kernal are slightly lower than RBF method at 95.3%, 96,5% for EPU dataset and 91.6%, 89.7% for KinectLeap dataset. This evaluation illustrates that, the deep learning method is more efficient than hand craft-based method over the period show. In addition, the highest hand-craft-based methods (KDES-SVM) are far lower than CNN method. Also note worth is fact that Resnet50 model is more likely to deploy a real application.

3.3. Hand gesture recognition using various fusion strategies

A glance at the figure provided reveals hand pose recognition accuracy of fusion methods (late fusion and early fusion) of two cues (RGB and Depth flows) during the period shown. It could be seen from the Fig. 7 that, combination of both RGB and Depth information obtains higher accuracy than independent cue representation. Additionally, the early fusion method obtains the best hand gesture recognition accuracy, with the highest value at 99.1% on EPUHandPose2 dataset and 94.3% for KinectLeap dataset. While late fusion approach accounted by far on both two EPUHandPose2 and KinectLeap datasets at 96.3% and 91.6%, respectively.



Accuracy with the fusion strategies

4. DISCUSSION AND CONCLUSION

In this research, an approach for hand pose recognition system that combines multi-modalities (RGB and Depth images). This paper has deeply investigated the results of some state-of-the-art feature extraction methods as SIRF, SURF, KDES, HOG and Deep learning method. Experimental try in order to test on various Kernels of SVM to choose best suitable model for different features. Experiments are conducted on our captured dataset and the published dataset. Furthermore, the evaluations lead to some following conclusions: i) Concerning both hand craft-based and CNN issues, the proposed method has obtained highest performance on both datasets. It is simple approach and obtains high accuracy system. So one of recommendation is to combine with others features such as optical flow and/or texture of hand as well as create larger training dataset to obtain the higher accuracy of hand posture recognition; ii) The proposed method will be evaluated on other published datasets.

5. ACKNOWLEDMENT

This research is funded by Electric Power University under the Project "Control home appliances using Computer Vision and Artificial Intelligent".

REFERENCES

[1] P.A. Viola and M.J. Jones, "Rapid object detection using a boosted cascade of simple features", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, I–I, 2001.

[2] https://github.com/Balaje/OpenCV/tree/master/haarcascades

[3] Caltech dataset: http://www.vision.caltech.edu/html-files/archive.html

[4] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In Procceding of ECCV, pp. 512-528,2019.

[5] Huong-Giang Doan and Van-Toi Nguyen, Improving Dynamic Hand Gesture Recognition on Multi-views with Multi-modalities, International Journal of Machine Learning and Computing, Vol. 9, No. 6, pp. 795-800, 2019.

[6] C.1.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Vol. 43, pp. 1-43, 1997.

[7] David G Lowe, Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image, David Lowe's patent for the SIFT algorithm, 2004.

[8] Bay, Herbert & Tuytelaars, Tinne & Van Gool, Luc. SURF: Speeded up robust features. The Proceedings of the 9th European conference on Computer Vision-ECCV, pp. 404-417. 2006.

[9] Navneet Dalal, Bill Triggs. Histograms of Oriented Gradients for Human Detection. International Conference on Computer Vision & Pattern Recognition (CVPR '05), United States. pp.886--893, 2005.

[10] Bo, Liefeng, Xiaofeng Ren, and Dieter Fox. Kernel descriptors for visual recognition. In Advances in neural information processing systems, pp. 244-252. 2010.

[11] Dang-Manh Truong, Huong-Giang Doan, Thanh-Hai Tran, Hai Vu, and Thi-Lan Le, Robustness Analysis of 3D Convolutional Neural Network for Human Hand Gesture Recognition, International Journal of Machine Learning and Computing (IJMLC), Vol. 9, No. 2, April 2019, pp.135-142.

[12] http://www.microsoft.com/en-us/kinectforwindows.

[13] F. Zhan, Hand Gesture Recognition with Convolution Neural Networks, 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), Los Angeles, CA, USA, pp. 295-298, 2019.

[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Curran Associates Inc., USA, 2012, pp. 1097–1105.

[15] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, 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.

[16] Kaiming He, Xiangyu Zhang and, Shaoqing Ren and Jian Sun, "Deep Residual Learning for Image Recognition", abs/1512.03385, CoRR 2015.

[17] Huong-Giang 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, Mar. 2016.

[18] Yang Li, Jin Huang, Feng Tian, Hong-An Wang, Guo-Zhong Dai, Gesture interaction in virtual reality, Virtual Reality & Intelligent Hardware, Volume 1, Issue 1, 2019, Pages 84-112.

[19] G. Marin, F. Dominio, P. Zanuttigh, "Hand gesture recognition with Leap Motion and Kinect devices", IEEE International Conference on Image Processing (ICIP), Paris, France, 2014.

[20] Ashish Sharma, Anmol Mittal, Savitoj Singh, Vasudev Awatramani, Hand Gesture Recognition using Image Processing and Feature Extraction Techniques, Procedia Computer Science,Volume 173, pp. 181-190, 2020.

[21] Marta Sylvia Del Rio Guerra, Jorge Martin-Gutierrez, Renata Acevedo and Sofía Salinas, Hand Gestures in Virtual and Augmented 3D, Environments for Down Syndrome Users, Applied Sciences, Vol.9, pp. 1-16, 2019.

Giới thiệu tác giả:

Doan Huong Giang, 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, Vietnam. She is a lecturer at Control and Automation faculty, Electric Power University, Ha Noi, Viet Nam. Her current research centers on human-machine interaction using image information, action recognition, manifold space representation for human action, computer vision.