DYNAMIC HAND GESTURE RECOGNITION USING DEPTH DATA

NHẬN DẠNG CỬ CHỈ ĐỘNG CỦA BÀN TAY SỬ DỤNG DỮ LIỆU ẢNH ĐỘ SÂU

Doan Thi Huong Giang, Bui Thi Duyen

Electric Power University

Ngày nhận bài: 05/07/2019, Ngày chấp nhận đăng: 24/04/2020, Phản biện: TS. Nguyễn Thị Thanh Tân

Abstract:

Recently, hand gesture recognition has been becomce a attractive field in computer vision. Which consists some main step such as: hand detection, hand segmentation, spotting gesture, feature extraction and classification. There are many state-of-the-art methods has been proposed while have almost ultilized RGB images. Moreover, almost recent method employed RGB images for these consequence states dynamic hand gesture recognition. Such modality still has to face with many challenges due to the light condition, motion blur, complex background, low resolution and so on. In this paper, we propose a new framework for deeply evaluate efficient of Depth information for dynamic hand gesture recogniton. In addition, the suitable frames number of depth images in a gestures are evaluated to obtain very competitive accuracy.

Keywords:

Dynamic hand gesture recognition, depth motion map, human-computer interaction.

Tóm tắt:

Gần đây, nhận dạng cử chỉ động của bàn tay trở thành một chủ đề hấp dẫn trong xử lý ảnh. Bài toán nhận dạng cử chỉ động của bàn tay bao gồm các bước chính như: phát hiện tay, trích trọn vùng bàn tay trong ảnh, phân đoạn chuỗi cử chỉ tay, trích trọn đặc trưng của chuỗi cử chỉ động và nhận dạng. Đã có nhiều giải pháp đề xuất cho bài toán nhận dạng cử chỉ tay trong đó hầu hết là sử dụng ảnh màu. Tuy nhiên, hầu hết chúng vẫn phải đối mặt với các thách thức như điều kiện chiếu sáng, nhòe, phông nền phức tạp, độ phân giải thấp,… Trong bài báo này, chúng tôi đề xuất một giải pháp phân tích sự hiệu quả của thông tin ảnh độ sâu trong bài toán nhận dạng cử chỉ động của bàn tay. Ngoài ra, chúng tôi còn đánh giá số lượng các khung hình phù hợp cho mỗi cử chỉ động để đạt hiệu quả tốt nhất.

Từ khóa:

Nhận dạng cử chỉ động, bản đồ chuyển động của độ sâu, tương tác người - máy.

1. INTRODUCTION

In recent years, hand gesture recognition has become a great attention of researchers thanks to its potential applications such as sign language translation, human computer interactions [3][4][5][6] robotics, virtual reality [4] [5], autonomous vehicles [3]. In many last proposed methods, community researchers are concentrated on RGB

images. Which are sensitive with light condition as well as motion blur. Such methods have been proposed for hand gesture recognition such as [2] [4] [5] [15]. In [2], authors firstly used RGB images on both entire background and segmented hand and. The KDES descriptor and SVM classifier is then used to recognize hand gestures. Authors in [5] proposed a dynamic hand gesture method with KLT and ISOMAP combination for RGB gesture representation. Authors in [15] deploy convolutional neuron network (CNN) on RGB sequence to recognize dynamic hand gestures. Recently, Kinect sensor of Microsoft company [10] has bring a new approach for researchers in computer vision which provided both RGB and Depth information at the same time. The depth maps could provide shape and motion information in order to distinguish human getures/actions. This depth information has been motivated for recent researches work to explore gesture recognition based on depth maps such as [6] [8] [11] [16]. Hand posture recognition method is proposed by using a Bag-of-3D-Points [16] for sampling 3D points from depth maps. An action graph was then employed to model the sampled 3D points to perform action recognition. However, this research require an expensive computations because the sampled 3D points of each frame generated a considerable for entire data. [8] ultilized DMM and HOG descriptor for action representation. Moreover, this method requires a threshold to calculate depth map. In [2], KDES despriptor is quite efficient for hand posture recognition on RGB images which has motivated for our research. We must be try an aproach with non-threshold to create DMM images and KDES method for dynamic hand gesture representation.

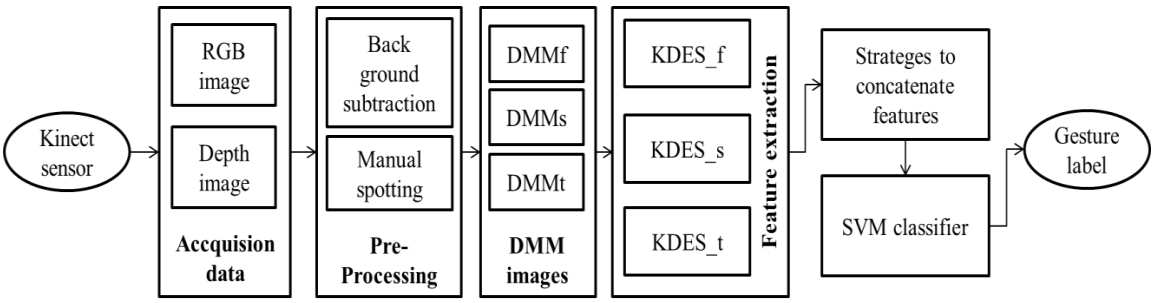


Figure 1. Proposed framework for dynamic hand gesture recognition

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

In this section, The main flow-work for dynamic hand gesture recognition from RGB-Depth images consists of a series of the cascaded steps as shown in Fig. 1 following. By using a fixed the Kinect

sensor, a RGB image and a Depth image are concurrently wrapped at the same time. Then, hand gestures are processed, extracted and recognitized. The steps are presented in detail at the next sections.

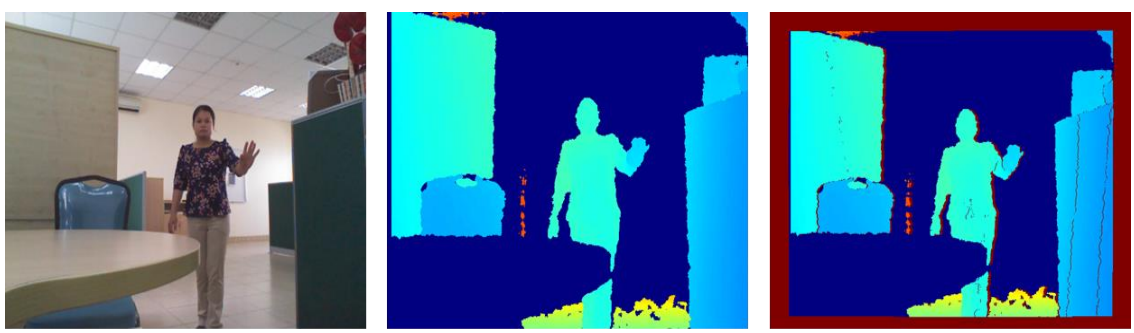
2.1. Accquision and Pre-processing data

Depth (I_D) and RGB (I_{RGB}) images from the Kinect sensor are not measured from the same coordinates. In our previous research, this problem was considered and resolved as presented in [1]. That we utilized calibration method of Microsoft to repair the depth images and RGB images. The result showed in Fig. 2a and Fig. 2b is original Depth and RGB image, Fig.2c is calibration depth image. Because Kinect sensor and background are immobile in scense. Moreover, subjects stand at the fixed position when implement dynamic hand gestures. Calibrated depth is used for the background subtraction because the depth data is less sensitive with illumination. Among numerous techniques of the background subtractions, we adopt Gaussian Mixture Model (GMM) [7] as presented detail in our other work [2]. Firstly, noise and background model with parameters $(\boldsymbol{\mu_p}, \boldsymbol{\eta_p}, \boldsymbol{\sigma_p})$ are calculated from n depth frame through each pixel p on temporal dimension of $\boldsymbol{s_p} = [\boldsymbol{I_{D1}}, \boldsymbol{I_{D2}}, \dots, \boldsymbol{I_{Dn}}]$. Then, each depth image $(\boldsymbol{I_D})$ is given from the Kinect sensor is recalculated by quotion (1) following:

$$\boldsymbol{H} = \begin{cases} \boldsymbol{\mu_p} & (\boldsymbol{\eta_p\ is\ noise})\ \boldsymbol{and}\ (\boldsymbol{invalid\ pixel}) \\ \boldsymbol{I_D} & \boldsymbol{otherwise} \end{cases} \quad (1)$$

The result showed in Fig. 3a is calibrated depth image, Fig.3b is result of human depth image (H).

Given depth human continuous sequence, we then implemented manual spotting in order to divide continuous frames into meaning gestures and manual label it. Depth human gesture consists different number of postures as shown in Fig. 4. There three dynamic hand gestures are implementd by the same subject in three times but phase of gestures are not the same. This problem is quite challenge for synchrolization of dynamic hand gestures before gesture recognization.



(a) RGB image (b) Orignal Depth image (c) Calibration Depth image

Figure 2. Combination of RGB and Depth images for human detection

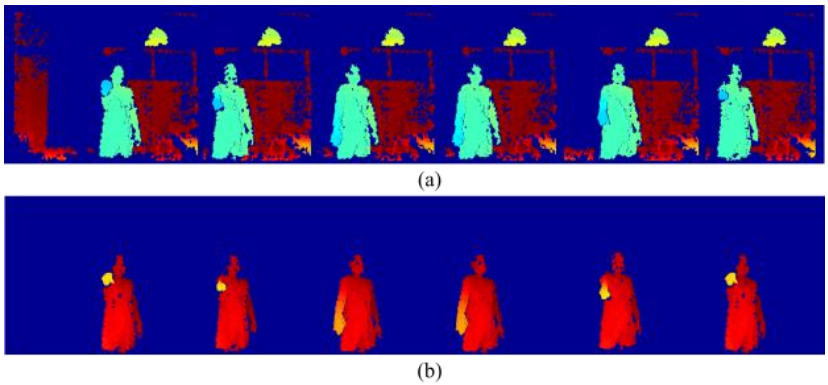


Figure 3. Manual spotting for hand gestures

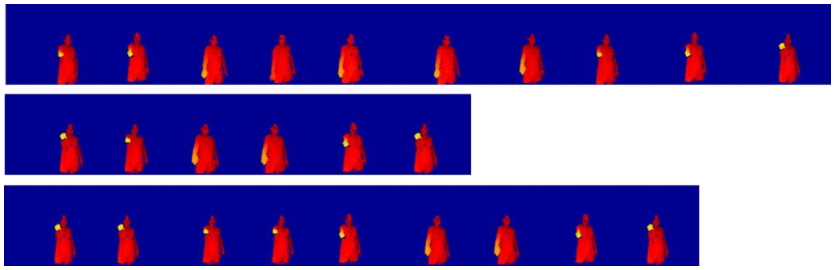


Figure 4. Different number of postures in dynamic hand gestures

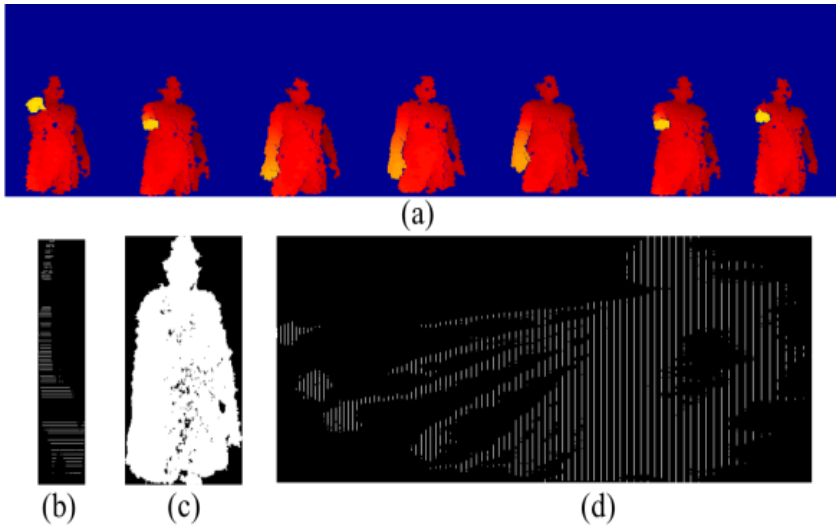


Figure 5. Three projected view using depth motion map for each dynamic hand gesture

Fig. 2b is original Depth and RGB image, Fig. 2c is calibration depth image. Because Kinect sensor and background are immobile in scense. Moreover, subjects stand at the fixed position when implement dynamic hand gestures. Calibrated depth is used for the background subtraction because the depth data is less sensitive with illumination. Among numerous techniques of the

background subtractions, we adopt Gaussian Mixture Model (GMM) [7] as presented detail in our other work [2]. Firstly, noise and background model with parameters $(\boldsymbol{\mu_p}, \boldsymbol{\eta_p}, \boldsymbol{\sigma_p})$ are calculated from n depth frame through each pixel p on temporal dimension of $\boldsymbol{s_p} = [\boldsymbol{I_{D1}}, \boldsymbol{I_{D2}}, \dots, \boldsymbol{I_{Dn}}]$. Then, each depth image $(\boldsymbol{I_D})$ is given from the Kinect sensor is recalculated by quotion (1) following:

$$\boldsymbol{H} = \begin{cases} \boldsymbol{\mu_p} & (\boldsymbol{\eta_p\ is\ noise})\ \boldsymbol{and}\ (\boldsymbol{invalid\ pixel}) \\ \boldsymbol{I_D} & \boldsymbol{otherwise} \end{cases} \quad (1)$$

The result showed in Fig. 3a is calibrated depth image, Fig.3b is result of human depth image (H).

Given depth human continuous sequence, we then implemented manual spotting in order to divide continuous frames into meaning gestures and manual label it. Depth human gesture consists different number of postures as shown in Fig. 4. There three dynamic hand gestures are implementd by the same subject in three times but phase of gestures are not the same. This problem is quite challenge for synchrolization of dynamic hand gestures before gesture recognization.

2.2. Depth motion map representation

First, N humand depth images of dynamic hand gesture $\boldsymbol{G_k}$ $([\boldsymbol{H_{Gk}^1}, \boldsymbol{H_{Gk}^2}, \dots . \boldsymbol{H_{Gk}^N}])$ are projected into three orthogonal Cartesian planes: top, side and bottom views as presented in [8]. The dynamic hand gesture composes a volumn that contains images following time series. Therefore, 3D depth frame generates three 2D maps according to front, side, and top views $(\boldsymbol{D_f^i}, \boldsymbol{D_s^i}, \boldsymbol{D_t^i})$. In this work, the motion energies are calculated without a threshold as in [8] to have projected map between two consecutetive maps. The binary map of motion energy indicates motion regions or where movement happens in each temporal interval. It provides a strong information of the gestures. Then, we stack the motion energy through entire image sequences to generate the depth motion map $\boldsymbol{DMM_g}$ for each projection view of dynamic hand gesture as equation (2), (3) and (4) following:

$$\boldsymbol{DMM_f} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N-1}} \left| \boldsymbol{D_f^{i+1}} - \boldsymbol{D_f^i} \right| \quad (2)$$

$$\boldsymbol{DMM_s} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N-1}} \left| \boldsymbol{D_s^{i+1}} - \boldsymbol{D_s^i} \right| \quad (3)$$

$$\boldsymbol{DMM_t} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N-1}} \left| \boldsymbol{D_t^{i+1}} - \boldsymbol{D_t^i} \right| \quad (4)$$

N is number of frames in a dynamic hand gesture. $\boldsymbol{DMM_g} = (\boldsymbol{DMM_f}; \boldsymbol{DMM_s}; \boldsymbol{DMM_t})$ contains binary maps of motion energy. Which present appearance/shape motion of hand gesture in temporal. which characterize the accumulated motion distribution and intensity of this action. The $\boldsymbol{DMM_g}$ representation encodes the 4D information of body shape and motion in three projected planes, meanwhile significantly reduces considerable data of depth sequences to just three 2D maps. Figure 5 illustrate $\boldsymbol{DMM}$ images in three views of dynamic hand gesture. Fig. 5a shows human depth images in dynamic hand gesture and Fig. 5b,c,d is bottom, frontal and side DMM images of dynamic hand gesture, respectively.

2.3. Feature extraction and classification

Given three $\boldsymbol{DMM_g}$ of dynamic hand gesture, difference from [8], authors concatenate three feature vectors that are extracted by HOG method. In this paper, we ultilize KDES descriptor as presented in [2] for feature extraction in frontial, side and top projected views. $\boldsymbol{DMM_g}$ images of depth motion map of hand gesture is presented by kernels [2] which follows consequence steps: pixel feature extraction, patch feature extraction and DMM image feature extraction. In addition, in this paper, we use adaptive patch size and pyramid structure in [2] to extract feature vectors. Each gesture composes of three features $\boldsymbol{F_f}$, $\boldsymbol{F_t}$ and $\boldsymbol{F_s}$ with each feature vector size is [1x4096]. Next, we implement the strategy to concatenate above feature vectors in order to create the feature vector representations for a hand gestures F (size of F is [1x(4096x3)]) as quotion (5) following:

$$\boldsymbol{F} = \left[\boldsymbol{F_f}, \boldsymbol{F_t}, \boldsymbol{F_s}\right] \quad (5)$$

Finally, we use Multi-class SVM classiffer [9] with the input is feature vector of dynamic hand gesture and output is label of gesture. The accuracy rate is the ratio between the numbers of true positives rate per total number of hand gestures used in testing.

3. EXPRIMENTIAL RESULTS

We evaluate performance of the hand gesture recognition on two datasets: MSRGesture3D [14] and the sub-dataset MICA [15]. This datataset is captured by five Kinect sensors that are fixed on a tripod at the height of 1.8m. Kinect sensors are collected in a lab-based environment of the MICA institution with indoor lighting condition, office background. The Kinect sensor captures data at 30 fps with depth, color images. Six users are invited to implement 3 to 5 times for five dynamic hand gestures. Five dynamic hand gestures are presented detail in our previous researche [5][15]. In entire evaluation, we follow Leave-p-out-cross-validation method, with p equals 1. It means that gestures of one subject are utilized for testing and the remaining subjects are utilized for training. In this paper, three evaluations are conducted: (1) The performance of the proposed method when the number of frame is changed, (2) The accuracy rate of the hand gesture recognition system and (3) The performance of other datasets.

3.1. Influence of resolution with hand gesture recognition rate

In this evaluation, we test the accuracy rate with various values of the number frames of dynamic hand gestures. This number of frame is changed from 15 to 55 frames for each gesture. The accuracy rates are illustrated in Fig.5, that show results on MICA dataset [15] with Kinect sensor 3. As shown, if this value is small, hand gesture recognition result is degraded. Performance are saturated when the number of frame is equal to 30 frames per one dynamic gesture. In next evaluations, this number of frames should be ultilized for other exprimentials.

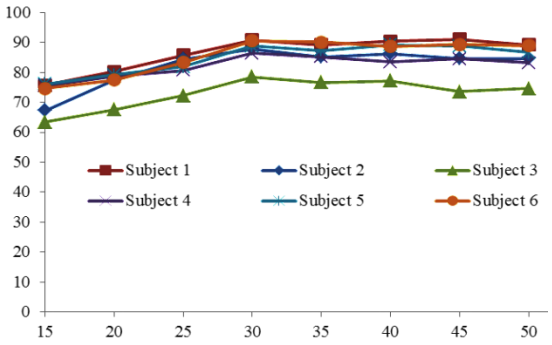


Figure 5. Evaluation with the different number of frames

3.2. Comparison of different methods

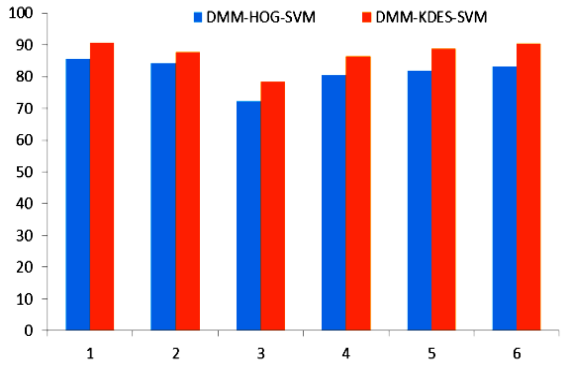


Figure 6. Evaluation with the different methods

Figure 6 shows the results of different schemes as described in other research [16]. As could be seen from the Fig. 8 that the combination between DMM and KDES method overall obtains the accuracy rate at $87.09 \pm \mathbf{4.1}\%$, is higher than $81.34 \pm \mathbf{4.4}\%$ with DMM and HOG descriptors. Averagely, the propose method gives the best results on all subjects with highest value at 91% for subject 1 and 6. The smallest accuracy belongs to subject 3 with 79%.

3.3. Comparison of different datasets

Table 1 presents the efficient of different hand gesture representation methods on different datasets. As could be seen from the Tab. 1 that the propose method obtains the best hand gesture recognition accuracy with the highest value at 92.89% on MSRGesture3D dataset. While method [8] brings only 89.17%. The same trength with MICA dataset, the better result belong to combination between DMM and KDES method with 78.09% that is far higher than 81.34% for DMM and HOG method[8].

Table 1. Evaluate accuracy on different datasets

	MICA[15]	MSRGesture3D[14]
DMM-HOG[8]	81.34%	89.17%
DMM-KDES	**87.09%**	**92.89%**

3.4. Depth data for dynamic hand gesture recognition on multiviews

Table 2 show the hand gesture recognition results on five Kinect sensor [15] (K1, K2,…K5) of MICA sub-dataset. This dataset contains dynamic hand gestures are captured by six subjects (S1,…S6). A glance at the Tab.2 reveals the difference values from five Kinect sensors with higest result belong to K3 and K5 at 87% and 88%, respectively. While the similarities are K1,K2 and K4 from 76% to 78%, respectively. As could be seen from the Tab. 2 that the propose method brings the best hand gesture recognition accuracy with the highest value at 100% for subject 1 on K5 and subject 5 on K1. In addition. Almost subjects on K5 give the high accuracy from the 93% to 96%. Avr results are mean values of six subjects on each Kinect sensor. These results show that best recognition result belong to Kinect sensor K5 while lowest

evaluations are K2 and K4.

Table 2. Evaluate accuracy on multi-views

	K1	K2	K3	K4	K5
S1	71.42	80.95	**90.71**	80.95	**100**
S2	70.12	62.5	87.75	84.37	**96.87**
S3	65.93	66.66	80.64	77.77	51.61
S4	**94.11**	86.36	88.34	64.54	**95.45**
S5	**100**	**95.83**	88.71	75.04	**95.83**
S6	74.59	76.41	86.23	73.32	**93.05**
Avr	**79.36**	**78.12**	**87.06**	**76.00**	**88.00**

4. DISCUSSION AND CONCLUSION

In this paper, an approach for human hand gesture recognition using depth imformation. Then we have deeply investigated the results of with suitable temporal resolution for the best dynamic hand gesture recognition using DMM-KDES method. Experiments were conducted on two datasets: self-designed dataset and published dataset. The evaluations lead to some following conclusions: i) Concerning depth imformation issue, the proposed method has obtained highest performance with both self-designed dataset and published dataset [14]. It is simple approach and avoid illumination with light condition. So one of recommendation is to combinate between depth and RGB data to obtain the higher accuracy of dynamic hand gesture recognition; ii) The extraction method of action region from DMM views has impact on performance of recognition method. Using KDES descriptor gives higher recognition accuracy.

REFERENCES

[1] Huong-Giang Doan, Hai Vu, and Thanh-Hai Tran. (2014). Ultilizing Depth Image from Kinect sensor: Error Analysis and Its Application, in the proceeding of the 7th Vietnamese Conference on FAIR 2014, ThaiNguyen, VietNam, ISBN: 978-604-913-300-8, pp. 216-222, 2014.

[2] Huong-Giang Doan, Van-Toi Nguyen, Hai Vu, and Thanh-Hai Tran. (2016). A combination of user-guide scheme and kernel descriptor on rgb-d data for robust and realtime hand posture recognition, Journal of Engineering Applications of Artificial Intelligence (EAAI 2016 Journal), Elsevier, ISSN: 0952-1976, vol. 49, no. C, pp. 103-113, 2016.

[3] H. Takimoto, J. Lee, and A. Kanagawa, A Robust Gesture Recognition Using Depth Data, IJMLC, Vol. 3, No. 2, 2013, pp. 245-249.

[4] Q. Chen, A. El-Sawah, C. Joslin, N.D. Georganas, A dynamic gesture interface for virtual environments based on hidden markov models, IEEE International Workshop on Haptic Audio Visual Environments and their Applications, 2005, p. 109-114.

[5] Huong-Giang Doan, Hai Vu, and Thanh-Hai Tran. (2016). Phase Synchronization in a Manifold Space for Recognizing Dynamic Hand Gestures from Periodic Image Sequence, in the proceeding of the 12th IEEE-RIVF International Conference on Computing and Communication Technologies, pp. 163 - 168, 2016.

[6] P. Molchanov, S. Gupta, K. Kim, J. Kautz, Hand gesture recognition with 3d convolutional neural networks, CVPRW, 2015, pp. 1–7.

[7] C. Stauffer and W.E.L. Grimson, Adaptive background mixture models for real-time tracking, In the proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVRP 1999), Vol. 2, USA, 1999, pp. 246-252.

[8] Xiaodong Yang, Chenyang Zhang, and YingLi Tian, Recognizing Actions Using Depth Motion Maps-based Histograms of Oriented Gradients, In the proceedings of the 20th ACM International Conference on Multimedia, 2012, pp. 1057 - 1060.

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

[10] Microsoft Kinect for Windows, http://www.microsoft.com/enus/kinectforwindows., November 2013.

[11] D. Shukla, Ö. Erkent and J. Piater, "A multi-view hand gesture RGB-D dataset for human-robot interaction scenarios," *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, New York, NY, 2016, pp. 1084-1091.

[12] Haiying Guan, Jae Sik Chang, Longbin Chen, R. S. Feris and M. Turk, "Multi-view Appearance-based 3D Hand Pose Estimation," 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), New York, NY, USA, 2006, pp. 154-154.

[13] Poon, Geoffrey & Chung Kwan, Kin & Pang, Wai-Man. (2018). Real-time Multi-view Bimanual Gesture Recognition. 19-23. 10.1109/SIPROCESS.2018.8600529.

[14] http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/

[15] 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 2019), Vol. 9, No. 2, April 2019, pp.135-142.

[16] Li, W., Zhang, Z., and Liu, Z. 2010. Action Recognition based on A Bag of 3D Points. *IEEE Workshop on CVPR for Human Communicative Behavior Analysis.*

Biography:

Doan Thi 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.

Bui Thi Duyen received B.E. degree in Instrumentation and Industrial Informatics in 2004, M.E. in Automatic in 2007 and Ph.D. in Control engineering and Automation in 2020, 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 focus on measurement and control system, wireless sensor network, antenna and high-frequency circuit.

.