Improving Dynamic Hand Gesture Recognition on Multi-views with Multi-modalities

Huong-Giang Doan and Van-Toi Nguyen

***Abstract*—Hand gesture recognition topic has been researched for many recent decades because it could be used in many fields as sign language, virtual game, human-robot interaction, entertainment and so on. However, this problem has been faced to many challenges such as combination of multi information in a temporal flow in order to understand the meaning of human hand gesture. In recent times, thanks to the advances in hardware technologies such as readily available 3D cameras, Kinect sensors, and etc. The impressive performance of cutting-edge techniques in computer vision, which is known as: manifold learning, deep learning techniques and/or the presentation of various multimodal fusion strategies. There have been many improvements in exploiting of features from multimodal data to effectively solve human hand gesture recognition tasks. Therefore, this paper focuses on solving the problem of dynamic hand gesture recognition in our daily life. We consider methods for extracting features of different data sources (RGB images and depth images) based on both manifold learning and deep learning technique. For RGB information, a manifold technique is performed to extract spatial feature that is then composed with temporal feature extracted by KLT technique. Among many deep learning architectures proposed in the literature that achieved good results in detecting human activities, I studied and proposed a simple convolutional neural network to extract feature of depth motion map. This technique extracts hand features from depth information which combines spatial and temporal aspects. Besides that, fusion algorithms are deployed to unite with those extracted features and enhance the accuracy of a final dynamic hand gesture results. Evaluation results confirm that the best accuracy rate achieves at 84.7% that is significantly higher than results from previous works (at 78.4%). The proposed method suggests a feasible solution addressing technical issues in using multimodality and multi-viewpoint of hand gestures.**

***Index Terms*—Hand gesture recognition, convolutional neuron network, RGB-Depth images, multi-modalities, multi-viewpoints, depth motion map, manifold learning.**

I. Introduction

Interaction using hand gesture is a natural and friendly way in Human Computer Interaction (HCI) [1]-[3]. In addition, using vision cues brings an independent way for end-users. Therefore, this hand gesture recognition topic has been attracted the many researches in computer vision field in recent decades [4]-[8]. Although, this topic has still faced to many challenges such as complex background, occlusion, blur, light condition, various scales, different views directions, non-rigid hand objects, etc. Kinect sensor brings new approach for research community because it provides not only RGB images but also Depth information at the same time. Moreover, deploiting and utilizing those cues flows is father researched in order to obtain better efficient recognition accuracy and on different viewpoints.

The problem of dynamic hand gesture recognition with both spatial and temporal cues on RGB and Depth images is considered in this paper. For this task, to improve efficient of hand gesture recognition system, we study and deploy on two types of above image flows. Firstly, with RGB flow, for spatial representation, manifold learning could be applied on segmented RGB hand images to exploit the lower dimension of data with ISOMAP technique [9] while movement of hand is extracted by KLT algorithm [10], [11]. Those features of RGB hand are concatenated together. Phase synchronization algorithm is applied as presented detail in our previous research [12]. Secondly, with depth flow, DMM [4] technique is implemented to obtain hand movements of depth images sequence. New convolutional neural network architecture is considered and proposed to extract depth dynamic hand feature. Different from the existing CNN technique, our proposed CNN architecture is quite simple, suitable and efficient for our input DMM images. Thirdly, from the various multimodal presentations, utilization strategies of data aim to perform early and late fusion of those extracted features as well as obtain middle outputs to enhance the accuracy of a final dynamic hand gesture classification. Experimental results show advantages of fusing multi-modal data for better hand gesture recognition.

Finally, we evaluated experiments of the proposed framework on two datasets MSRGesture3D [13] and MICA4 dataset [14]. Both two datasets consists twelve dynamic hand gestures. MSRGesture3D is used to show the effect of combination between DMM method and proposed CNN architecture on Depth cues. Specially, MICA4 dataset is multi-views dynamic hand gesture that captured both RGB and Depth images at the same time as presented in detail at [5]. This dataset composes five fixed Kinect sensors [15] and captured in indoor environment at various instance time and different invited subjects.

II. Related Works

Hand gesture recognition could be considered on many approaches such as data modalities, dependent and/or

Manuscript received August 7, 2019; revised October 13, 2019. This work was supported in part by Vietnam National Foundation for Science and Technology Development (NAFOSTED).

Huong-Giang Doan is with the Control and Automation Faculty at Electrical Power University Hanoi, Vietnam (e-mail: giangdth@epu.edu.vn).

Van-Toi Nguyen is with the Faculty of Information Technology, Posts and Telecommunications Institute of Technology, Vietnam (e-mail: toinv@ptit.edu.vn)

doi: 10.18178/ijmlc.2019.9.6.875

independent equipment, static and/or dynamic hand gesture processing algorithm, etc. This problem has been researched for a long time because of its many practical applications such as sign language [8], [16], entertainment [3], and human-machine interaction [1], [2]. As hand gesture recognition is seen by information flows such as RGB image [5], [8], [17], Depth image [4], [6], optical flow [17], and so on. However, this issue is a difficult and challenging task as combination of signals, representation techniques as well as classifiers. In [12], authors utilized RGB image from the Kinect sensor to recognize hand gesture from various distance. Because of low resolution and small hand shape, thus those effect on final results. Only depth image was used in [4] with DMM projection and HOG descriptor. In [1], authors proposed method using multi-modalities with complex algorithm for online hand gesture recognition. Their system requires a high configuration computer, and RGB and Depth multi-views dataset is captured by two camera but this dataset does not designed to evaluate the impact of viewpoints. In [2], [16], RGB-D images from the Kinect sensor are used to improve the hand gesture accuracy results. However, a real application is hard to deploy and integrate.

This paper focuses on solving the problem of dynamic hand gesture recognition in dual data flows as RGB and Depth images. In recent times, thanks to the advances in hardware technologies such as readily available 3D cameras [19], [20], Kinect sensors [15], and so on. The impressive performance of cutting-edge combination presentation of various multimodal fusion methods [5]; there have been many improvements in exploiting features derived from multimodal to effectively solve hand gesture recognition tasks. Therefore, in this research, we study methods for extracting features of different data sources based on manifold learning [9] and deep learning technique [5]. Among deep learning architectures proposed in the literature that achieved good results in detecting hand gesture recognition [5], [8], [13], convolutional neural network is studied and self-designed network architecture (2D-CNN) is proposed. This self-designed network is built to extract depth features. It is simple and mining depth data. Besides that, a fusion algorithm of feature flows is performed aims to perform early and late fusion to enhance the final accuracy for the hand gesture classification.

Finally, the proposed framework is evaluated on two datasets: MSRGesture3D dataset is collected from ten subjects, each subject implemented twelve gestures in three times and depth data is captured by one fixed Kinect sensor at near distance at 0.8 to 1m (one viewpoint). MICA4 dataset performed by sub-dataset with 6 subjects with self-designed 12 dynamic hand gestures under multiple viewpoints; Experimental results shows advantages of fusing multi modalities for better hand gesture recognition.

The remaining of this paper is organized as follows: Section II presents some approaches that use the manifold learning and the deep learning technique for the hand gesture recognition problem. Section III describes our proposed approach. The experiments and results are analyzed in Section IV. Section V concludes this paper and proposes some future works.

III. Approach for Multi-modalities Hand Gesture Recognition

A. The Proposed Framework

Because hand gestures are captured from the Kinect sensor [15] that gives both RGB images and Depth images at the same time. Therefore, additional information about RGB hand pose, depth images could help to improve hand recognition accuracy. In this work, our proposed framework for hand gesture inherits our previous work on RGB feature extraction [12]. This work extended the original manifold learning architecture by extra-flow depth method where one stream is RGB (as presented detail in [12]) and one stream utilize depth motion map and CNN method. The framework for hand gesture recognition is illustrated detail in Fig. 1. It consists of following main components:

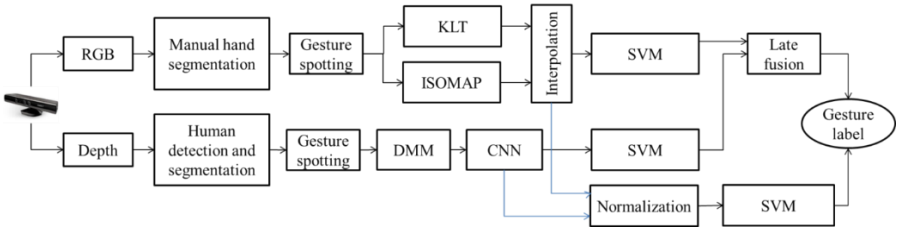


Fig. 1. General proposed framework for multi-modalities hand gesture recognition.

a) Original RGB images of dynamic hand gesture G_3

b) Segmented RGB dynamic hand gesture G_3

Fig. 2. RGB modality of a dynamic hand gesture.

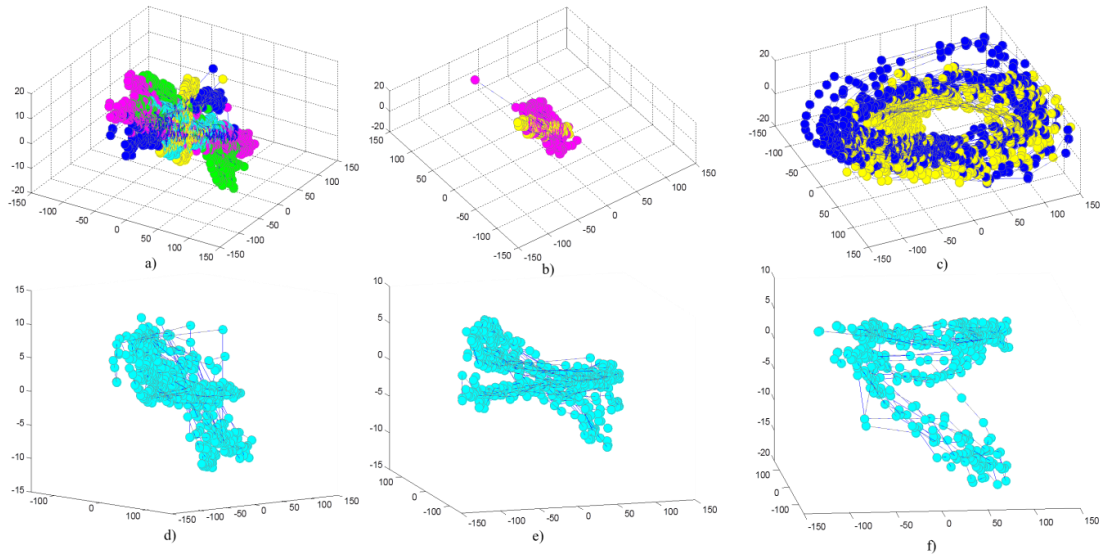
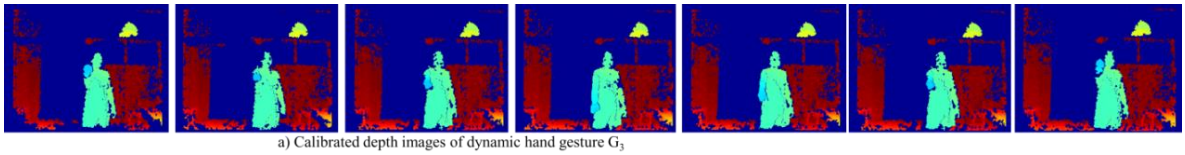
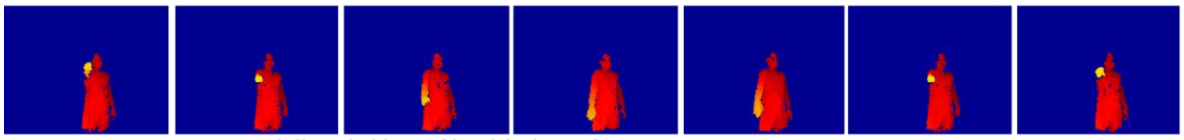


Fig. 3. RGB hand gesture representation in new space.



a) Calibrated depth images of dynamic hand gesture G_3



b) Human depth images of dynamic hand gesture G_3

Fig.4. Depth modality of a dynamic hand gesture.

B. Hand Gesture Extraction with RGB modality

Original RGB images (I_{RGB}) in dynamic hand gestures are directly captured from the Kinect sensors. Then, continuous image sequence is manually spotted as illustrated in Fig. 2a. Next, RGB hand is manual segmented as showed in Fig. 2b. Dynamic hand gesture is contemporaneously performed at all viewpoints. To extract a hand gesture from RGB video stream, we rely on the techniques presented in our previous paper [12]. On the first hand, we utilize ISOMAP algorithm [9] to present hand shapes of dynamic hand gesture. On the other hand, the hand trajectories are reconstructed using a conventional KLT trackers [10], [16] as performed in [5]. Each hand posture P_i in a dynamic hand gesture is normalized and composed by two hand trajectory elements (Tr_i = [x_i, y_i]) and three hand shape elements (Y_i=[$Y_{i,1}$, $Y_{i,2}$,$Y_{i,3}$]) as equation (1) following:

$$\boldsymbol{P_i} = (\boldsymbol{Tr_i}, \boldsymbol{Y_i}) = \left(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{Y_{i,1}}, \boldsymbol{Y_{i,2}}, \boldsymbol{Y_{i,3}}\right) \quad (1)$$

Dynamic hand gesture consists of N postures (G^k=[P_1^k, P_2^k,…., P_N^k]) as equation (2) following:

$$\boldsymbol{G_{RGB}^k} = \begin{bmatrix} \boldsymbol{x_1^k} & \boldsymbol{x_2^k} & \dots & \boldsymbol{x_N^k} \\ \boldsymbol{y_1^k} & \boldsymbol{y_2^k} & \dots & \boldsymbol{y_N^k} \\ \boldsymbol{Y_{1,1}^k} & \boldsymbol{Y_{2,1}^k} & \dots & \boldsymbol{Y_{N,1}^k} \\ \boldsymbol{Y_{1,2}^k} & \boldsymbol{Y_{2,2}^k} & \dots & \boldsymbol{Y_{N,2}^k} \\ \boldsymbol{Y_{1,3}^k} & \boldsymbol{Y_{2,3}^k} & \dots & \boldsymbol{Y_{N,3}^k} \end{bmatrix} \quad (2)$$

We then use an interpolation scheme which maximizes inter-period phase continuity, or periodic pattern of image sequence is taken into account as presented in detail [5]. In [12], all dynamic hand gestures (with different length) are interpolated to M frames as showed in equation (3) following:

$$\boldsymbol{G_{RGB}^k} = \begin{bmatrix} \boldsymbol{x_1^k} & \boldsymbol{x_2^k} & \dots & \boldsymbol{x_M^k} \\ \boldsymbol{y_1^k} & \boldsymbol{y_2^k} & \dots & \boldsymbol{y_M^k} \\ \boldsymbol{Y_{1,1}^k} & \boldsymbol{Y_{2,1}^k} & \dots & \boldsymbol{Y_{M,1}^k} \\ \boldsymbol{Y_{1,2}^k} & \boldsymbol{Y_{2,2}^k} & \dots & \boldsymbol{Y_{M,2}^k} \\ \boldsymbol{Y_{1,3}^k} & \boldsymbol{Y_{2,3}^k} & \dots & \boldsymbol{Y_{M,3}^k} \end{bmatrix} \quad (3)$$

Fig. 3(a)-(f) illustrates new representations in 3-D space of twelve different hand gestures of MICA4 dataset at frontal view. In Fig. 3a, first five hand gestures are showed on the same figure that gesture representations are distinguished with other types. Those dynamic hand gestures were introduced detail in [5] and [14]. In addition, they are separated with remain gesture as showed in Fig. 3(b)-(f). In Fig. 3(b) illustrates 22G_6 and 26G_7 gestures corresponding yellow and purple color. Fig. 7 shows 26 G_8 and 27 G_9 gestures are the same circle curves but its start and stop points as well as directions are inverse together. These twelve dynamic hand gestures were new designed and collected and presented detail in [5] and [14]. Gestures in the same type are converged and distinguished with others.

C. Hand Gesture Extraction with Depth Modality

In pre-processing depth information, original depth images are captured (I_D) at the same time with RGB images (Fig. 2(a)). Moreover, their coordinate of two images is athwart. In this work, we applied calibration method as in

our previous research [19] to adjust original depth images following original RGB images. The result of calibrated depth images I'_D are illustrated in Fig. 4(a). Then, model background is trained and applied as in our other research [17] that aims to remove background $I_{bg}(\mu_{bg}, \sigma_{gb}, \eta_{bg})$ and remain only human depth (H) from calibrated depth images as equation (4) following. Result is showed in Fig. 4(b).

$$H = f(I'_D, I_{bg}) \quad (4)$$

Then, N human depth images of dynamic hand gesture $G_D^k = (\left[\mathrm{H}_1^{\mathrm{k}}, \mathrm{H}_2^{\mathrm{k}}, \ldots . \mathrm{H}_{\mathrm{N}}^{\mathrm{k}}\right])$ are projected into three orthogonal Cartesian planes: top, side and bottom views as presented in [1]. The dynamic hand gesture composes a column that contains images following time series. Therefore, 3D depth frame generates three 2D maps according to front, side, and top views (D_f^i, D_s^i, D_t^i). In this work, the motion energies are calculated without a threshold as in [1] to have projected map between two consecutive maps. The binary map of motion energy indicates motion regions or where movement happens in each temporal interval. It provides strong information of the dynamic hand gestures. Then, we stack the motion energy through entire image sequences to generate the depth motion map $\mathrm{DMM_g}$ for each projection view of dynamic hand gesture as equation (5), (6) and (7) following:

$$DMM_f = \sum_{i=1}^{N-1} \left| D_f^{i+1} - D_f^i \right| \quad (5)$$

$$DMM_s = \sum_{i=1}^{N-1} \left| D_s^{i+1} - D_s^i \right| \quad (6)$$

$$DMM_t = \sum_{i=1}^{N-1} \left| D_t^{i+1} - D_t^i \right| \quad (7)$$

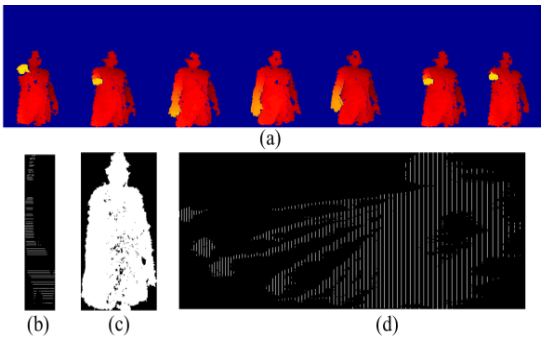


Fig. 5. Three projected views using depth motion maps for each dynamic hand gesture.

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

Recently, deep learning (CNN) has been widely used in computer vision in various tasks such as feature extraction, recognition. E.g. deep neural networks for dynamic hand gesture recognition [1], [8], [18]. In our previous research [17], authors try to capture simultaneously spatial and temporal information by applying a 3D-CNN on the whole video sequence. In this method, designed-network is employed with trained parameters of model. That requires high and complex configuration of computer. Moreover, in almost proposed method applied for RGB image or RGB sequences. This paper try to implement a simple designed CNN, this deep architecture tested on two general hand gesture datasets that is suitable for depth hand gesture recognition with a relatively low spatial of hand resolution. Moreover, it is evaluated various viewpoint in order to extract depth features.

In this section, we will introduce the self-design convolutional neural network that will be utilized to extract depth feature from above DMM images. Self-design network composes of 8 convolutional layers, 4 max pooling and 2 fully connected layer followed by a soft-max output layer. The architecture of designed network is illustrated in Fig. 6. In this network, the convolutional operation is 2D convolution which represents both spatial and temporal information from depth images. They are composed onto the DMM images.

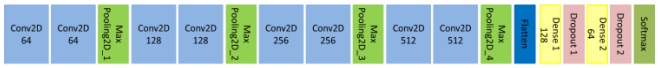


Fig. 6. Proposed CNN architecture for depth hand gesture extractor.

Given an input video $W \times H \times L$ where W, H are weight and height of frame, L is the video length. Dynamic hand gestures are implemented by different subjects and instance time that has different number of postures in a gesture. A dynamic hand gesture is projected into three DMM images $((D_f^i, D_s^i, D_t^i)$; those images then convert to the same size (200×200pixels) so dimension of DMM image is $W \times H$; W = 200 and H = 200), to extract features, DMM images are utilized as inputs of designed convolutional neuron network. These DMM images will be passed to CNN to extract depth extract features. These features of a gesture are concatenated together. In this work, this designed network is used as feature extractor. The dimension of output feature are [64×3, 1] as equation (8) following:

$$G_D^k = [D_1^k \quad D_2^k \ldots D_{191}^k \quad D_{192}^k]^T \quad (8)$$

D. RGB-Depth Dynamic Hand Gesture Representation and Classification

Given RGB and Depth feature of dynamic hand gesture as equation (3) and (4) above, they is firstly normalized to the same scale between RGB data (G_{RGB}^k) and Depth data (G_D^k). It is then concatenated to present feature of dynamic hand gesture $G^k = [G_{RGB}^k, G_D^k]$ as equation (9) following:

$$G^k = [\boldsymbol{x_1^k}, \boldsymbol{y_1^k}, \boldsymbol{Y_{1,1}^k}, \boldsymbol{Y_{1,2}^k}, \boldsymbol{Y_{1,3}^k}, \ldots, \boldsymbol{x_M^k}, \boldsymbol{y_M^k}, \boldsymbol{Y_{M,1}^k}, \boldsymbol{Y_{M,2}^k}, \boldsymbol{Y_{M,3}^k}, D_1^k, \ldots, D_{194}^k]^T \quad (9)$$

This feature as presented in (5) that is utilized as input of multi-class SVM classifier [10]. The output of multi-class SVM will be one value among {1,2,…,12} corresponding to the gesture elements in both two datasets.

IV. Experiment

The proposed framework is warped by a C++ program and a Matlab program on a PC Core i5 3.1 GHz CPU, 4GB

RAM. We evaluate performance of the hand gesture recognition on three two datasets: MSRGesture3D dataset [13]; MICA4 dataset [5], [14]. In entire evaluation, we follow Leave-p-out-cross-validation method as presented detail in [5], with p equals 1. It means that gestures of one subject are utilized for testing and the remaining subjects are utilized for training. We conduct four evaluations: (1) The performance of the proposed depth feature extraction method DMM-CNN; (2) When viewpoints is changed, the accuracy rate of the hand gesture recognition with depth images; (3) efficient of hand gesture recognition system when RGB feature and Depth feature are combined; (4) Compare accuracy of fusion strategies. The detail evaluations are presented as following sub-sections.

A. Evaluate Feature Extractors on Depth Modality

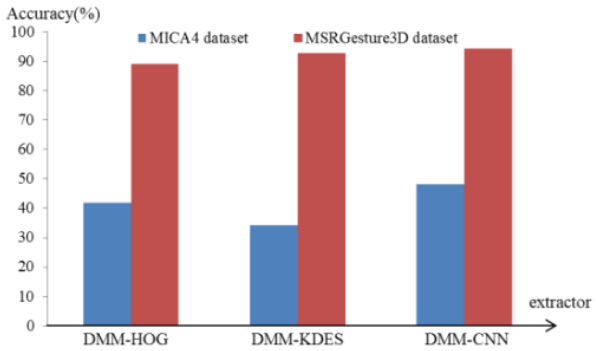


Fig. 7. Accuracy of hand gesture recognition from depth data.

Fig. 8. Example of a depth dynamic hand gesture in MSRGesture3D.

Fig. 7 shows that feature is extracted by DMM-CNN method obtains the best results on both published dataset (MRGGesture3D) and self-designed dataset (MICA4 dataset) at 94.47% and 48.18% respectively. While DMM-HOG [4] is archived the lowest results at only 89.17% and 41.91%. This result stimulated for our later evaluations in this paper. In addition, although both of two datasets have the same number of gesture but MSRGesture3D dataset gives the higher accuracy than MICA4 dataset. This may be that MSRGesture3D dataset was captured at the nearer distance, thus, depth hand has better quality than our dataset as illustrated in Fig. 8 (a)-(b). Moreover, MSRGesture3D dataset remained only depth hand region in images while our dataset contains full depth human as illustrated detail in Fig. 4(b).

B. Evaluate for Multi-views Depth Hand Gesture Recognition

Fig. 9 shows accuracy of dynamic hand gesture recognition of five Kinect sensors on three different feature extraction methods: DMM-HOG [4], DMM-KDES [17] and our proposed method DMM-CNN. A glance at the Fig. 9, it shows that DMM-CNN obtains the highest values on all five Kinect sensors with average value about 48.1±7%. Inverse, while DMM-HOG gives the smallest on overall with 41.9±9% and DMM-KDES method is 43.1±9%.

Look at the trends in the Fig. 9 over five Kinect sensors, it is apparent that the uptake of the different feature extractors increase dramatically entire views. In addition, accuracy of the frontal view (K_3) always obtains the best results on all methods with 51.1%, 55.2% and 56.2% for DMM-HOG, DMM-KDES and DMM-CNN respectively. The only little smaller monitory than accuracy of the Kinect sensor 1 that is values of two side views (K_1 and K_5) while the best results still belongs to our proposed feature extraction method (DMM-CNN) at 50.1% and 54.8% respectively. This average figure is just 35.1% and 33.6% on K_2 and K_4 (view directions are 45^0 and 135^0).

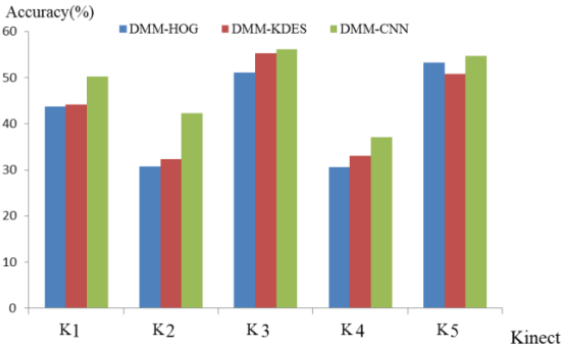


Fig. 9. Accuracy of hand gesture recognition from depth data.

C. Evaluate Impact of Modalities Combination for Multi-views Hand Gesture Recognition

TABLE I: MULTI-MODALITIES HAND GESTURE RECOGNITION (%)

	K_1	K_2	K_3	K_4	K_5	Avr
DMM-CNN	50.2	42.4	53.2	37.1	54.7	48.1±7.3
KLT-ISOMAP	63.7	70.9	81.2	74.2	61.2	70.2±7.2
Combination	**76.4**	**79.8**	**84.7**	**80.3**	**82.5**	**80.7±2.8**

Table I shows the dynamic hand gesture recognition accuracy of the two modalities and combination of them on five different Kinect sensors. It is evident that the percentages of features combination in five views are expected to the highest during the research period show at 80.7% respectively. Additionally, the figure for Depth modality is predicted to experience the smallest value at approximately 48.1%. The middle group is RGB modality with two features KLT and ISOMAP that stand at 70.2%. This result shows that, if modalities are utilized at single, the hand gesture accuracy is lower than when it is combined together.

Look at the Table I, it is clear from the information given that frontal view still obtains the best accuracy on not only two single modalities but also combination modalities with 53.2%, 81.2% and 84.7% respectively. While results of features concatenation are improved on the existence views that are fluctuated from 76.4% to 82.5% on K_1 to K_5.

D. Fusion Strategies of multi-Modalities for Dynamic Hand Gesture Classification

Table II shows results obtained with early fusion and late fusion of both RGB and Depth stream. Compared to using single stream (RGB or Depth flow), fusing both two streams help to improve the average accuracy about 1.6% while early fusion evaluation obtains lightly better than late fusion. In early fusion, highest accuracy has been obtained at the frontal view (K_3) with 84.7% and smallest value belongs to angle 0^0 – side view (K_1). This trend is the same with late fusion strategy at 83.7% and 74.3% respectively. Late fusion

method requires up to two classifiers that time cost and configuration of computer is more expensive than other one.

TABLE II: LATE AND EARLY FUSION FOR GESTURE RECOGNITION (%)

	K1	K2	K3	K4	K5	Avr
Early fusion (KLT-ISOMAP-DMM-CNN)	76.4	**79.8**	**84.7**	80.3	**82.5**	**80.74**
Late fusion (KLT-ISOMAP-DMM-CNN)	**74.3**	77.5	83.7	**81.1**	79.2	79.16

V. DISCUSSION AND CONCLUSION

In this paper, an approach for RGB-D hand gesture recognition using both RGB and Depth information in recognition phase. Then we have deeply investigated the results of with suitable spatial and temporal solution for the best dynamic hand gesture recognition that is combined from manifold learning method and deep learning method. Experiments were conducted on two datasets: self-designed dataset and published dataset. The evaluations lead to some following conclusions: i) concerning depth information issue, the proposed method has obtained highest performance with both self-designed dataset and published dataset. It is simple approach and avoids illumination with various light conditions. So one of recommendation is improve efficient representation of depth feature as well as more suitable convolutional neuron network of depth extractor; ii) Direction has impact on performance of proposed depth dynamic hand gesture that is good at the frontal view and two side view (0^0 and 180^0) while it obtains lower results at reflect directions (45^0 and 135^0); iii) combination between depth and RGB data that obtains the higher accuracy of dynamic hand gesture recognition; iv) depth segmented hand and high hand resolution gives better accuracy results with our proposed method.

These conclusions open some directions in our future works. Firstly, we will complete evaluation of all of twenty subjects in our MICA4 dataset. Secondly, we also test and compare with other state-of-the-art convolutional neuron network. Next, cross-view evaluation will be performed on the multi-view dataset (MICA4 dataset) using our RGB-Depth proposed representation method.

CONFLICT OF INTEREST

*.@epu.edu.vn; *.@mica.edu.vn; *.@ptit.edu.vn.

AUTHOR CONTRIBUTIONS

All works in this paper have been done by both two authors with the same role: Huong-Giang Doan; Van-Toi Nguyen.

REFERENCES

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

[2] D. Shukla, Ö. Erkent, and J. Piater, "A multi-view hand gesture RGB-D dataset for human-robot interaction scenarios," in *Proc. the 25th IEEE International Symposium on Robot and Human Interactive Communication*, USA, 2016, pp. 1084-1091.

[3] S.-de-U. Zelai, G.-Z. Begonya, and M. Z. Amaia, "Kinect-based virtual game for motor and cognitive rehabilitation: A pilot study for older adults," in *Proc. the 8th International Conference on Pervasive Computing Technologies for Healthcare*, 2014, pp. 262-265.

[4] X. D. Yang, C. Y. Zhang, and Y. L. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. the 20th ACM International Conference on Multimedia*, 2012, pp. 1057-1060.

[5] 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, April 2019.

[6] H. Takimoto, J. Lee, and A. Kanagawa, "A robust gesture recognition using depth data," *IJMLC*, vol. 3, no. 2, pp. 245-249, 2013.

[7] H. Y. Guan, J. S. Chang, L. B. Chen, R. S. Feris, and M. Turk, "Multi-view appearance-based 3D hand pose estimation," in *Proc.* the *2006 Conference on Computer Vision and Pattern Recognition Workshop*, USA, 2006, pp. 154-154.

[8] A. H. Vo, V.-H. Pham, and B. T. Nguyen, "Deep learning for Vietnamese sign language recognition in video sequence," *IJMLC*, vol. 9, no. 4, pp. 440-445, 2019.

[9] J. B. Tenenbaum, V. de Silva, and C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000.

[10] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. the 7th International Joint Conference on Artificial Intelligence*, vol. 2, San Francisco, CA, USA, 1981, pp. 674-679.

[11] J. Shi and C. Tomasi, "Good features to track," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, Ithaca, USA, 1994, pp. 593-600.

[12] H.-G. Doan, H. Vu, and T.-H. Tran, "Phase synchronization in a manifold space for recognizing dynamic hand gestures from periodic image sequence," in *Proc.* the 12*th IEEE-RIVF International Conference*, 2016, pp. 163-168.

[13] Research Microsoft. [Online]. Available: http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/

[14] MICA Dynamic Hand Gesture Set. [Online]. Available: http://mica.edu.vn/perso/Doan-Thi-Huong-Giang/MICADynamicHandGestureSet/

[15] Kinect for Windows. (2012). [Online]. Available: http://www.microsoft.com/enus/kinectforwindows

[16] D. H. Vo, H. H. Huynh, P. M. Doan, and J. Meunier, "Dynamic gesture classification for Vietnamese sign language recognition," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, 2017.

[17] 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," *Journal of Engineering Applications of Artificial Intelligence*, vol. 49, no. C, pp. 103-113, 2016.

[18] D. Shukla, Ö. Erkent and J. Piater, "A multi-view hand gesture RGB-D dataset for human-robot interaction scenarios," in *Proc. the 25th IEEE International Symposium on Robot and Human Interactive Communication*, USA, 2016, pp. 1084-1091.

[19] H.-G. Doan, H. Vu, and T.-H. Tran, 2014, "Ultilizing depth image from Kinect sensor: Error analysis and its application," in *Proc. the 7th Vietnamese Conference on FAIR*, Vietnam, 2014, pp. 216-222.

[20] M. V. den Bergh and L. V. Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," in *Proc. the IEEE Workshop on WACV*, 2011, pp. 66-72.

Huong-Giang Doan received the 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 with major in human computer, interaction using image information, all from Hanoi University of Science and Technology, Vietnam. She is working at Electric Power University, Ha Noi, Vietnam.

Van-Toi Nguyen received his BSc degree in informatics in 2001 from Hanoi National University of Education, MSc in computer science in 2005 from Thai Nguyen University, Vietnam. In 2016, he received his PhD degree in informatics and applications from The University of La Rochelle, France. He is working at Posts and Telecommunications Institute of Technology, Ha Noi, Vietnam.