MANIFOLD SPACE ON MULTIVIEWS FOR DYNAMIC HAND GESTURE RECOGNITION

KHÔNG GIAN ĐA TẠP CỦA CỬ CHỈ ĐỘNG BÀN TAY TRÊN CÁC GÓC NHÌN KHÁC NHAU

Huong Giang Doan

Electric Power University

Ngày nhận bài: 15/03/2019, Ngày chấp nhận đăng: 28/03/2019, Phản biện: TS. Nguyễn Thị Thanh Tân

Tóm tắt:

Recently, a number of methods for dynamic hand gesture recognition has been proposed. However, deployment of such methods in a practical application still has to face with many challenges due to the variation of view point, complex background or subject style. In this work, we deeply investigate performance of hand designed features to represent manifolds for a specific case of hand gestures and evaluate how robust it is to above variations. To this end, we adopt an concatenate features from different viewpoints to obtain very competitive accuracy. To evaluate the robustness of the method, we design carefully a multi-view dataset that composes of five dynamic hand gestures in indoor environment with complex background. Experiments with single or cross view on this dataset show that background and viewpoint has strong impact on recognition robustness. In addition, the proposed method's performances are mostly increased by multi-features combination that its results are compared with Convolution Neuronal Network method, respectively. This analysis helps to make recommendation for deploying the method in real situation.

Từ khóa:

Manifold representation, Dynamic Hand Gesture Recognition, Spatial and Temporal Features, Human-Machine Interaction.

Abstract:

Gần đây, có nhiều giải pháp nhận dạng cử chỉ động của bàn tay người đã được đề xuất. Tuy nhiên, việc triển khai trong các ứng dụng thực tế vẫn còn phải đối mặt với nhiều thách thức như sự thay đổi về hướng nhìn của máy quay, điều kiện nền phức tạp hoặc đối tượng điều khiển. Trong nghiên cứu này, chúng tôi đánh giá hiệu quả của không gian đa tạp biểu diễn cho các cử chỉ động của bàn tay đối với sự thay đổi hướng nhìn của máy quay. Hơn nữa, kết quả còn được đánh giá với sự kết hợp các đặc trưng của cùng một cử chỉ trên nhiều góc nhìn khác nhau. Chúng tôi xây dựng một cơ sở dữ liệu gồm năm cử chỉ động của bàn tay trên nhiều góc nhìn và thu thập trong môi trường trong phòng, với điều kiện nền phức tạp. Các thử nhiệm được đánh giá trên từng góc nhìn cũng như đánh giá chéo giữa các góc nhìn. Ngoài ra, kết quả còn cho thất sự hiệu quả khi kết hợp thông tin thu được trên nhiều luồng thông tin tại cùng một thời điểm, ngay cả so với những giải pháp sử dụng mạng nơ ron tiên tiến hiện nay. Kết quả phân tích trong nội dung của bài báo cung cấp những thông tin hữu ích giúp cho triển khai ứng dụng điều khiển sử dụng cử chỉ động của bàn tay trong thực tế.

Keywords:

Biểu diễn đa tạp, nhận dạng cử chỉ động, các đặc trưng không gian và thời gian, tương tác người máy.

1. INTRODUCTION

In recent years, hand gesture recognition has gained a great attention of researchers thanks to its potential applications such as sign language translation, human computer interactions [1][2][3], robotics, virtual reality [4][5], autonomous vehicles [3]. Particularly, Convolutional Neuronal Networks (CNNs) [7] have been emerged as a promising technique to resolve many issues of the gesture recognition. Although utilizing CNNs has obtained impressive results [6][8], or multiview hand gesture information[18][19][20]. Moreover, there exists still many challenges that should be carefully carried out before applying it in reality. Firstly, hand is of low spatial resolution in image. However, it has high degree of freedom that leads to large variation in hand pose. Secondly, different subjects usually exhibit different styles with different duration when performing the same gesture (this problem is identified as phase variation). Thirdly, hand gesture recognition methods need to be robust to changes in viewpoint. Finally, a good hand gesture recognizer needs to effectively handle complex background and varying illumination conditions.

Motived by these challenges, in this paper, we comprehensively analyze critical factors which affect to performance of a dynamic hand gesture recognition through conducting a series of experiments and evaluations. The manifold space's performances are examined under different conditions such as view-point's variations, muti-modality combinations and combination features strategy. Through these quantitative measurements, the important limitations of deploying manifold space representation could be revealed. Results of these evaluations also suggest that only by overcoming these limitations, one could make the methods being able to be applied in real situation.

In addition, we are highly motivated by the fact that variation of view-points and complex background are real situations, particularly when we would like to deploy hand gesture recognition techniques automatic controlling home appliances using hand gestures. These factors ensure that strict constraints in common systems such as controlling's directions of end-users or context’s background are eliminated. They play important roles for a practical system which should be maximizing natural feeling of end-user. To do this, we design carefully a multi-view dataset of dynamic hand gestures in home environment with complex background. The experimental results show that the change of viewpoint.

Finally, other factors such as cropping hand region variations, length of a hand gesture sequence that could impact the hand gesture recognition’s performances are analyzed. As a consequent, we show that hand region crop strategy and view-points although has been proved to be very efficient for hand gesture recognition.

The remaining of this paper is organized as follows: Sec. 2 describes our proposed approach. The experiments and results are

analyzed in Sec. 3. Sec. 4 concludes this paper and proposes some future works.

2. PROPOSED METHOD FOR HAND GESTURE RECOGNITION

2.1. Multiview dataset

Our dataset consists of five dynamic hand gestures which corresponds to controlling commands of electronic home appliances: ON/OFF, UP, DOWN, LEFT and RIGHT. Each gesture is combination between the hand movement in the corresponding direction and the changing of the hand shape. For each gesture, hand starts from one position with close posture, it opens gradually at half cycle of movement then closes gradually to end at the same position and posture as describe in [15]. Fig. 1 illustrates the movement of hand and changes of postures during gesture implementation.

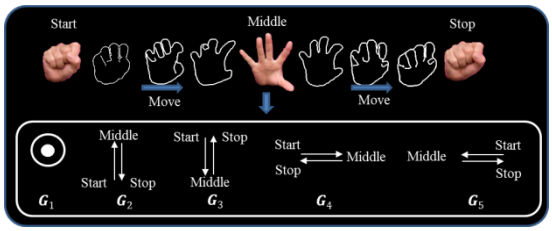

Figure 1. Five defined dynamic hand gestures

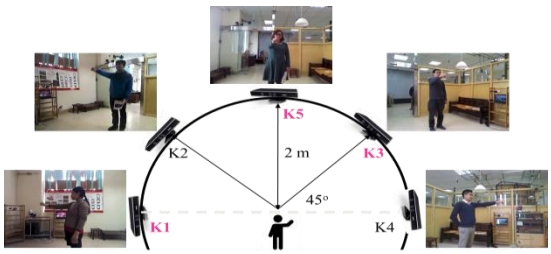

Figure 2. Setup environment of different viewpoints

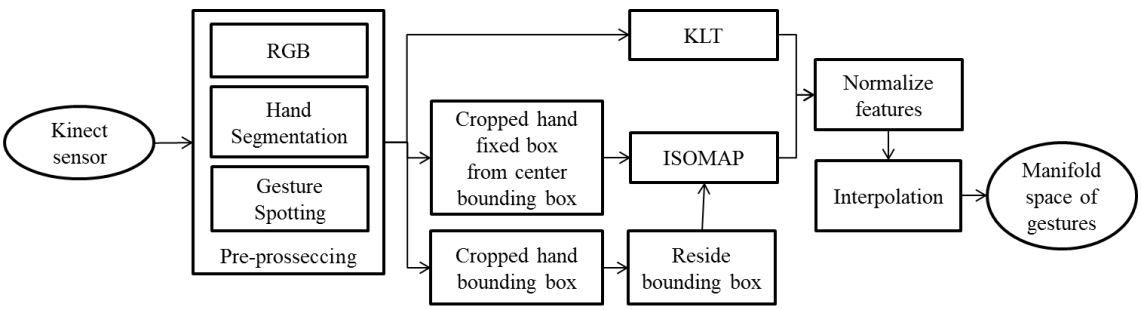

Figure 3. Pre-processing of hand gesture recognition

Five Kinect sensors K_1, K_2, K_3, K_4, K_5 are setup at five various positions in a simulation room of 4mx4m with a complex background (Fig. 2). This dataset MICA1 is collected in a lab-based environment of the MICA institution with indoor lighting condition, office background. A Kinect sensor is fixed on a tripod at the height of 1.8m. The Kinect sensor captures data at 30 fps with depth, color images which are calibrated between depth images and color images. This work aims to capture hand gestures under multiple different viewpoints at the same time. Subjects are invited to stand at a nearly fixed position in front of five cameras at an approximate distance of 2 meters. Five participants (3 males and 2 females) are voluntary to perform gestures (Pi; (i=1...5)}). Each subject implements one gesture from three to six times. Totally, the dataset contains 375

(5 views × 5 gestures × 5 subjects × (3 to 6 times)) dynamic hand gestures with frame resolution is set to 640×480. Each gesture's length varies from 50 to 126 frames (depending on the speed of gesture implementation as well as different users) as present in Tab. 1. Where the G_1 has the smallest frame numbers that is only from 33 to 66 frames fer a gesture. While other gestures fluctuated at somewhere approximately 60 to 120 frames per a gesture. This leads to a different number of frames to be processed and create large challenges for phase synchronization between different classes and gestures. In this work, only the three views K1, K3 and K5 were used because of their discriminants on view points. In addition, in each view, only videos taken from 5 subjects will be spotted and annotated with different numbers of hand gestures. This work requires large number of manual hand segmentation therefore they are sampled three frames on continuous images sequences: (1) All views have the same number of gestures with others. (2) In each view, the number of gestures of G_3 is highest at 33 gestures, G_1 and G_4 have the same number (26 gestures) while the number of G_2 and G_5 are 22, 23 gestures, respectively. These dataset will used to divide to train and test as presented in Sec. 3.

The dataset was synthesized at MICA institute, five dynamic hand gestures performed by five different subjects under five different viewpoints. Fig. 2 shows the information of five different views used in the dataset. However, only gestures in three views K_1, K_3 and K_5 were used in this paper. Tab. 1 shows the numbers of videos for each gesture: with average frame numbers of gesture as show in Tab. 1 following:

Table 1. Average frame numbers in a gesture

Subject	P_1	P_2	P_3	P_4	P_5
G_1	**49.2**	51	33	54	66.3
G_2	61.7	115	49.7	104.7	**126.2**
G_3	55.8	98.7	**118.5**	106.5	103.3
G_4	70.2	101.7	69	**108.8**	107.2
G_5	59.5	83	72.7	92.7	**102.5**

2.2. Manifold representation space

We propose a framework for hand gesture representation which composes of three main components: hand segmentation and gesture spotting, hand gesture representation, as shown in Fig. 3.

Hand segmentation and gesture spotting: Given continuous sequences of RGB images that are captured from Kinect senssors. Hands are segmented from background before spotted to gestures. Any algorithm of hand segmentation can be applied, from the simplest one basing on skin to more advanced techniques such as instance segmentation of Mask R-CNN [16]. In this work, we just apply an interactive segmentation tool[1] to manually detect hand from image. This precise segmentation helps to avoid any additional effect of automatic segmentation algorithm that could lead to wrong conclusion. Fig. 4 illustrates an original video clip and the corresponding segmented one annotated manually.

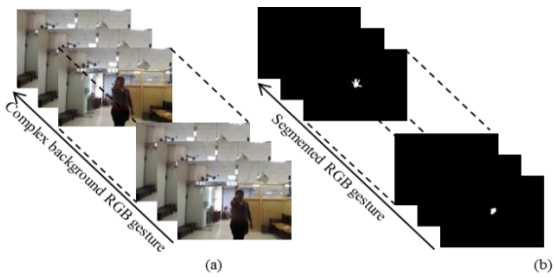

Figure 4. Hand segmentation and gesture spotting. (a) Original video clips; (b) The corresponding segmented video clip

Given dynamic hand gesture that is manually spotted by hand. To extract a hand gesture from video stream, we rely on the techniques presented in [11]. For representing hand gestures, we utilize a manifold learning technique to present phase shapes. The hand trajectories are reconstructed using a conventional KLT trackers [8] as proposed in [11]. We then used an interpolation scheme which maximize inter-period phase continuity, or periodic pattern of image sequence is taken into account.

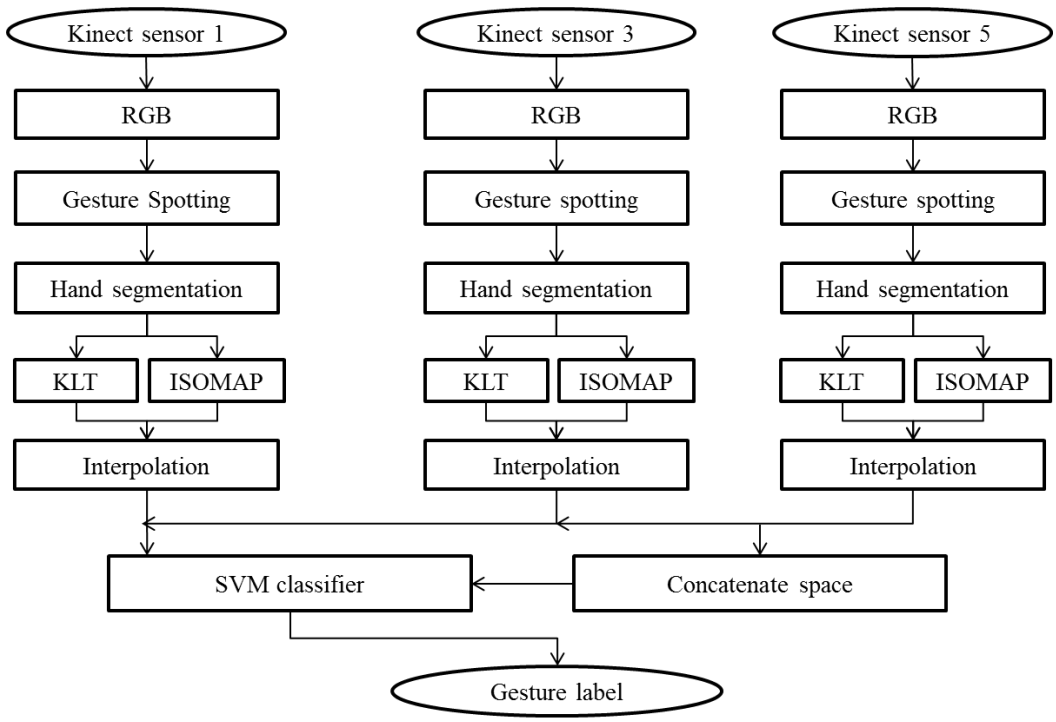

Figure 5. The proposed framework of hand gesture recognition

The spatial features of a frame is computed though manifold learning technique ISOMAP [13] by taking the three most representative components of this manifold space as presented in our previous works [11], [15]. Moreover, in [11], [15], we cropped hand regions around bounding boxes of hands in a gesture. Then, all of them are resided to the same size before using as inputs of ISOMAP technique as show in Fig. 3. That should be changed characteristics of hand shapes. In this work, we take hand region from center of bounding boxes with the same size. These cropped hand regions is not converted and directly

applied ISOMAP technique. The affects of these works are compared in Sec. 4. In both two methods, given a set of N segmented postures X = {Xi, i=1,...,N}, after compute the corresponding coordinate vectors Y = {Yi Є Rd, i = 1,...,N} in the d-dimensional manifold space (d << D), where D is dimension of original data X. To determine the dimension d of ISOMAP space, the residual variance Rd is used to evaluate the error of dimensionality reduction between the geodesic distance matrix G and the Euclidean distance matrix in the d-dimensional space Dd. Based on such evaluations, three first components (d = 3) in the manifold space are extracted as spatial features of each hand shape (e.g. Fig. 6 (a) illustrates 3-D manifolds of five different hand gestures. A Temporal feature of hand gesture then is represented as: $\mathbf{Y_i} = \{(\mathbf{Y_{i,1}} \quad \mathbf{Y_{i,2}} \quad \mathbf{Y_{i,3}})\}]$. Which is chosen to extract three most significant dimensions for hand posture representations. Three first components in the manifold space are extracted as spatial features of each hand shape/posture. Each posture Pi has coordinates Tr_i that are trajectory composes of K good feature points of a posture and then all of them are averaged by (xi, yi). In [15], we have combinated a hand posture Pi and spatial features Yi as eq. 1 following:

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

2.3. Manifold spaces on multiviews

In our previous researches [15], we only evaluated discriminant of each gesture with others on one view. In this paper, we investigate the difference of same gesture from different views on both separation spaces and concatenate hand gesture space as show in Fig. 4.

On one views, postures are capture from three Kinect sensors that are represented on both spatial and temporal as eq. 2 following:

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

In addition, a gesture is combined from n postures $\boldsymbol{G_{TS}^i} = [\boldsymbol{P_1^i} \quad \boldsymbol{P_2^i} \ldots \quad \boldsymbol{P_N^i}]$ as eq. 3 following:

$$\boldsymbol{G_{TS}^i} = \begin{bmatrix} \boldsymbol{x_1^i} & \boldsymbol{x_2^i} & \ldots & \boldsymbol{x_N^i} \\ \boldsymbol{y_1^i} & \boldsymbol{y_2^i} & \ldots & \boldsymbol{y_N^i} \\ \boldsymbol{Y_{1,1}^i} & \boldsymbol{Y_{2,1}^i} & \ldots & \boldsymbol{Y_{N,1}^i} \\ \boldsymbol{Y_{1,2}^i} & \boldsymbol{Y_{2,2}^i} & \ldots & \boldsymbol{Y_{N,2}^i} \\ \boldsymbol{Y_{1,3}^i} & \boldsymbol{Y_{2,3}^i} & \ldots & \boldsymbol{Y_{N,3}^i} \end{bmatrix} (\boldsymbol{i} = \mathbf{1}, \mathbf{3}, \mathbf{5}) \quad (3)$$

Separations the same gesture G2 from three views is presented in Fig. 5 following. This figure confirms inter-class variances when whole dataset is projected in the manifold space. In particularly, cyclic patterns of the same hand gesture are presented on three-views are distinguided with others while its manifold space is similar trajectory. The G2 dynamic hand gestures of frontal view K5 presented in red. Hand gestures on the Kinect sensor K3 are presented in magenta curves, and hand gestures on the Kinect sensor K1 are showed in green curves, respectively. Features vector then are recognized on two cases by SVM classifier[14] as showed in Fig. 5. On the first one, gesture is evaluated on each view and cross-view. On the other hand, features are concatenate together. Figure 6 following shows the five gesture

representations (G_1, G_2,…,G_5) on both two views frontal view - K_5 and 45 degree - K_3. This figure shows that five hand gestures are separated in exter-class and they are converged in inter-class.

2.4. Evaluation procedure

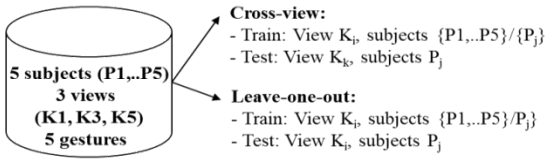

Figure 7. Evaluation procedure

In this paper, we use leave-one-subject-out cross-validation as described in [15] in order to prepare data for training and testing in our evaluations. Which each subject is used as the testing set and the others as the training set. The results are averaged from all iterations. With respect to cross-view, the testing set can be evaluate on different viewpoints with the training set. The evaluation metric used in this paper is presented in eq. (4) following:

$$accuracy = \frac{\sum Corrects}{Total}\% \quad (4)$$

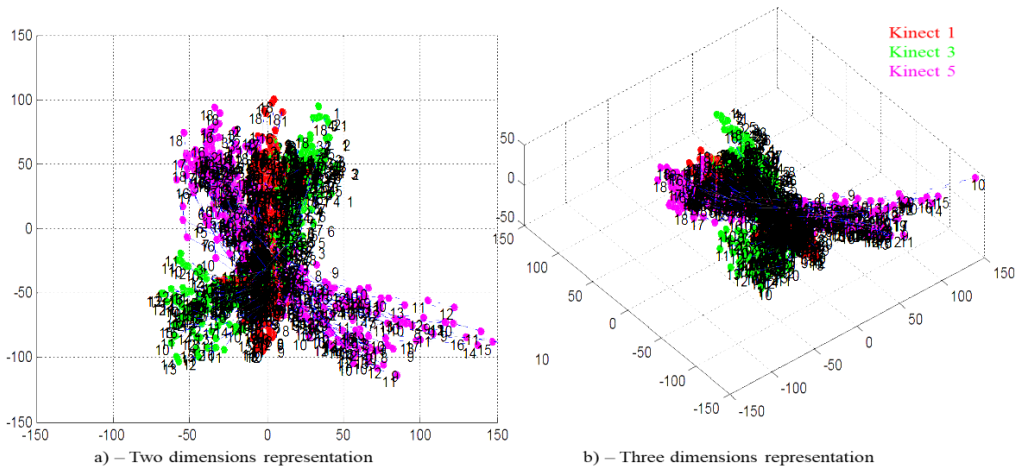

Figure 5. Discriminant manifold spaces of one type of hand gestures

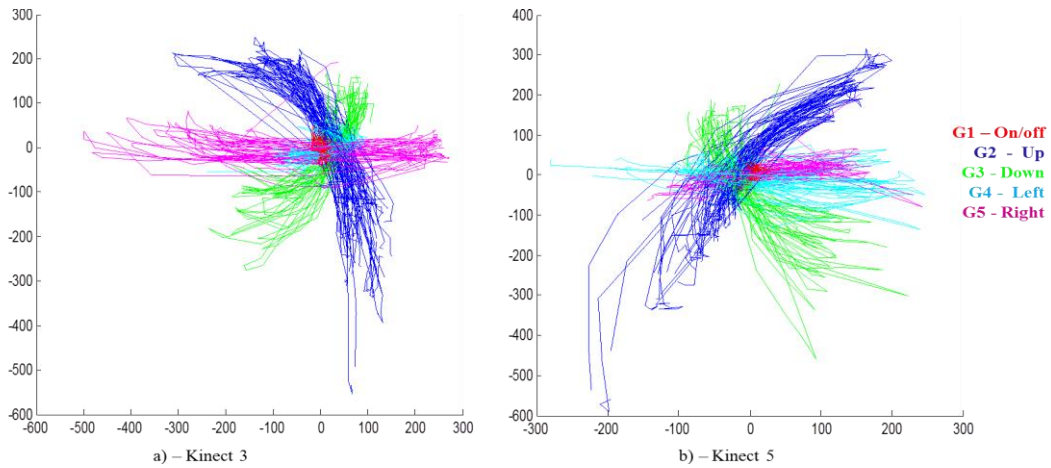

Figure 6. Discriminant manifold spaces of hand gestures between two views

3. EXPRIMENTIAL RESULTS

3.1. Cross-views evaluation

Table 2 shows the cross view results on two different cropped hand regions: (1) variable cropped hand regions, and (2) fixed cropped hand region. A glance at the Tab. 2 provided evident reveals that:

- Fixed cropped hand region gives more competitive performance than cropped hand regions. The average value is 78.64% that is higher than other case, 76.43% respectively. This is evident that cropped hand region directly affects on the gesture recognition result. We should focus on the fixed cropped hand in order to improve accuracy of the recognition system in our other researches.
- Single view gives quite good results on K_3 and K_5 that is best at the front views on all solutions, with 84.56%, 98.53% and 99.38% respectively. The view K_1 gives the worst results which fluctuate at some where from 42.06% to 84.56% only. These results is because the hands are occluded or out of camera field of view, or because the hand movement is not discriminative enough.
- Cross view has not strong impact on classification results, as could be seen from the comparison between single view and cross view results.

Table 2. Comparison of cross views with different cropped hand regions

	Variable bounding box			Fixed bounding box		
	K1	K3	K5	K1	K3	K5
K1	**81.58**	41.06	58.42	**84.56**	42.06	59.46
K3	59.22	**96.67**	95.38	65.15	**98.53**	98.33
K5	72.57	83.48	**98.21**	72.15	88.18	**99.38**
Avr	76.43			**78.64**		

3.2. Comparison of different methods

Figure 8 shows the results of different schemes as described in other our research [16]. As could be seen from the Fig. 8 that the proposed method gives the best results on all single views (K_1, K_3, K_5) with highest value at 99.38% on K5.

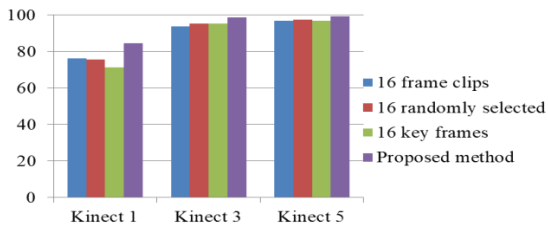

Figure 8. Evaluation with the different methods

3.3. Combination strategies of feature vectors

Table 3 shows the results of different concatenate schemes as described in Sec.2. As could be seen from the Tab. 3 that Kinect sensor K_5 (frontal view) gives the best results with highest value at 98.52%. While combination between Kinect sensor K_1 (180 degrees) and Kinect sensor 3 (45 degrees) is smallest results at 95.38%. Given results of combination from three view K_1, K_3 and K_5 as in Tab. 4 which shows confusion matrix of this concatenate strategy. Almost wrong recognition case belongs to dynamic hand gesture ON_OFF.

5. DISCUSSION AND CONCLUSION

In this paper, an approach for human hand gesture recognition using different views

in new manifold representation. Then we have deeply investigated the robustness of the method for hand gesture recognition. Experiments were conducted on a multi-view dataset that was carefully designed and constructed by ourselves. Different evaluations lead to some following conclusions: i) Concerning viewpoint issue, the proposed method has obtained highest performance with frontal view, it is still good when view point deviates in the range of 450 and reduced drastically when the viewpoint deviates from 900 to 1350. So one of recommendation is to learn dense viewpoints so that testing view point could avoid huge difference compared to learnt views; ii) Area of cropped hand region has impact on performance of recognition method. It is recommended to cut from the center to the edge of images before project them in to ISOMAP space; iii) using multi-view information obtains higher recognition accuracy.

Table 3. Multiviews dynamic hand gesture recognition with features combination

	Kinect 1-3	Kinect 1-5	Kinect 3-5	Kinect 1-3-5
Concatenate features-multiviews	**95.38**	**98.13**	**98.52**	**97.55**
Variable box-single view	72.43	77.77	79.08	76.43
Fixed box-single view	75.1	79.83	80.99	78.64

Table 4. Confusion matrix in concatenate space of Kinect 1,3,5

	G1	G2	G3	G4	G5
G1	**26**	0	0	0	0
G2	1	**21**	0	0	0
G3	0	0	**33**	0	0
G4	2	0	0	**24**	0
G5	0	0	0	0	**23**

These conclusions open some directions in future works. Firstly, we will complete our annotation and evaluation of all of five views and compare our methods with other existing ones. We also perform automatic hand segmentation and integrate into unified framework. Some adaption of the representation to face more with change of viewpoint also will be considered. One possibility is to learn more viewpoints and try to match the unknown gestures with the gestures having the most similar viewpoint in the training set. Another possibility is to extract invariant human pose features.

ACKNOWLEDGMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-17-1-4056.

TÀI LIỆU THAM KHẢO

[1] H. Doan, H. Vu, T. Tran, Dynamic hand gesture recognition from cyclical hand pattern, IAPR International Conference on Machine Vision Applications (MVA), 2017, pp. 97–100.

[2] M.M. Hasan and P.K. Mishra, Robust Gesture Recognition Using Gaussian Distribution for Features Fitting, IJMLC, Vol. 2, No. 3, 2012, pp. 266-273.

[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] V. Dissanayake, S. Herath, S. Rasnayaka, et al, Real-Time Gesture Prediction Using Mobile Sensor Data for VR Applications, IJMLC, Vol. 6, No. 3, June 2016, pp. 215-219.

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

[7] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, International Conference on Neural Information Processing Systems - Volume 1, 2012, pp. 1097–1105.

[8] B.D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, in Proceedings of the 7th International Joint Conference on Arti_cial Intelligence Volume 2, San Francisco, CA, USA, 1981, pp. 674-679.

[9] J. Shi and C. Tomasi, Good features to track, in IEEE Conference on Computer Vision and Pattern Recognition - CVPR'94, Ithaca, USA, 1994, pp. 593-600.

[10] Huong-Giang Doan, Hai Vu, and Thanh-Hai Tran. Recognition of hand gestures from cyclic hand movements using spatial-temporal features, in the proceeding of SoICT 2015, Vietnam, pp. 260-267.

[11] 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, Hanoi, VietNam, 2016.

[12] H.G. Doan, H. Vu, T.-H. Tran, and E. Castelli, Improvements of RGBD hand posture recognition using an user-guide scheme,in 2015 IEEE 7th International Conference on CIS and RAM, 2015, pp. 24-29.

[13] J.B. Tenenbaum, V. de Silva, and 1. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol. 290, no. 5500, pp. 2319-2323, 2000.

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

[15] Huong-Giang Doan, Hai Vu, and Thanh-Hai Tran. (2017). Dynamic hand gesture recognition from cyclical hand pattern, to appear in proceeding of The fifteenth IAPR International Conference on Machine Vision Applications (MVA2017), pp. 84-87 Nagoya, Japan, May 8-12, 2017.

[16] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, ICCV, 2017, pp. 2980–2988.

[17] Dang-Manh Truong, Huong-Giang Doan, Thanh-Hai Tran, Hai Vu, Thi-Lan Le , Robustness analysis of 3D convolutional neural network for human hand gesture recognition, ACMLC 2018, HoChiMinh, VietNam.

[18] 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.

[19] 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.

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

Giới thiệu tác giả:

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, 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.

.