P-ISSN 1859-3585 | E-ISSN 2615-9615

MULTIVIEWS DYNAMIC HAND GESTURE RECOGNITION AND CANONICAL CORRELATION ANALYSIS-BASED RECOGNITION

NHẬN DẠNG CỬ CHỈ ĐỘNG CỦA BÀN TAY ĐA HƯỚNG NHÌN VÀ NHẬN DẠNG VỚI KỸ THUẬT PHÂN TÍCH THÀNH PHẦN TƯƠNG QUAN

Doan Thi Huong Giang

ABSTRACT

Nowaday, there have been many approaches to resolve the problems of hand gesture recognition. Deployment of such methods in practical applications still face to many issues such as in change of viewpoints, non-rigid hand shape, various scales, complex background and small hand regions. In this paper, these problems are considered of feature extractions on different view points as well as shared correlation space between two views. In the framework, we implemented hand-crafted feature for hand gesture representation on a private view. Then, a canonical correlation analysis method (CCA) based techniques [1] is then applied to build a common correlation space from pairs of views. The performance of the proposed framework is evaluated on a multi-view dataset with five dynamic hand gestures.

Keywords: *Dynamic hand gesture recognition, multivew hand gesture, cross-view recognition, canonical correlation analysis.*

TÓM TẮT

Ngày nay, có nhiều hướng tiếp cận nhằm giải quyết bài toán nhận dạng cử chỉ động của bàn tay người được đã đề xuất. Triển khai những đề xuất trong các ứng dụng thực tế vẫn phải đối mặt với nhiều thách thức như sự thay đổi của hướng nhìn, thay đổi kích thước, ảnh hưởng của điều kiện nền, độ phân giải của vùng bàn tay quá nhỏ so với toàn bộ khung hình. Trong bài báo này, những vấn đề về bài toán nhận dạng cử chỉ tay được xem xét trên các đặc trưng biểu diễn đa tạp trên từng hướng nhìn, trên nhiều hướng nhìn khác nhau cũng như trên không gian biểu diễn chung kết hợp thông tin từ các hướng. Không gian biểu diễn chuyển đổi giữa các góc nhìn được tạo ra dựa trên dữ liệu từ các hướng nhìn khác nhau sử dụng kỹ thuật phân tích các thành phần tương quan CCA. Hiệu quả của giải pháp đề xuất được đánh giá trên bộ cơ sở dữ liệu với năm cử chỉ bàn tay.

Từ khóa: *Nhận dạng cử chỉ động, các cử chỉ đa hướng nhìn, nhận dạng chéo, phân tích thành phần tương quan.*

Faculty of Control and Automation, Electric Power University
Email: giangdth@epu.edu.vn
Received: 01 June 2019
Revised: 11 July 2019
Accepted: 15 August 2019

1. INTRODUCTION

Hand gestures have been becoming one of the natural method for Human Computer Interaction (HCI) [2, 3, 4]. Many techniques for hand gesture recognition have been proposed and developed, for example sign language recognition [3, 5], home appliance controls [6] and so on. Hand gesture recognition researches and hand pose estimation frameworks are introduced in a recent survey [7, 8]. Moreover, the some challenges as view-point changing or cluttered background [8, 9], low-resolution of hand regions are still remaining is existing challenges [9, 10]. In addition, when deploys practical applications as home appliance system [6, 9, 11] that requires not only natural way but also robustness systems. In some case, interaction systems require some constrains of end-user's interaction such as they rise their hand to the camera with the fix direction [4, 10, 12]. Almost proposed methods resolve with a common viewpoint. Different viewpoints result in different hand poses [13, 19], hand appearances and complex background and light condition. This degrades dramatically the performance of pre-trained models. Therefore, proposing robust methods for recognizing hand gestures from unknown viewpoint [8] is pursued in this work.

Our focus in this paper is evaluated the performance of cross-view on multiview dynamic hand gestures and analyzing how to improve entire evaluation results. A dynamic hand gesture recognition framework is proposed with handcrafted features using manifold technique. Then canonical correlation analysis (CCA) is employed that builds a linear transpose space, uses learning linear transforms between two views.

A dataset of dynamic hand gestures is used in this paper that captured from different viewpoints. Thanks to the proposed frame-work and the defined dataset, performances of the gestures recognition from different views are deeply investigated. Consequently, developing a practical application is feasible.

The remaining of this paper is organized as follows: Sec. 2 describes the 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. 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. 1.

Hand segmentation and gesture spotting: Firstly, continuous sequences of RGB images are captured from five Kinect sensors. Then, original video clip and the corresponding segmented one annotated manually. Finally, we just apply an interactive segmentation tool to manually detect hand from images as presented in detail at [13].

Spatial and Temporal feature extraction for dynamic hand gesture representation: Given dynamic hand gestures is manually spotted and labeled. To extract a hand gesture from video stream, we rely on the techniques presented in detail at [14]. For representing hand gestures, we utilize a manifold learning technique to present phase shapes. On one hand, The hand trajectories are reconstructed using a conventional KLT trackers [15, 16] as proposed in [14]. On the other hand, The spatial features of a frame is computed though manifold learning technique ISOMAP [8] by taking the three most representative components of this manifold space as presented in our previous works [14, 17].

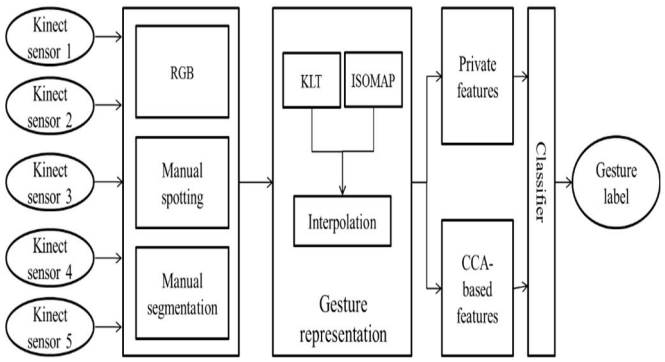

Figure 1. Proposed dynamic hand gesture recognition

Given a set of N segmented postures X = {Xi, i=1,...,N}, after compute the corresponding coordinate vectors Y = {Yi ∈ 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. A Temporal feature of hand gesture then is represented by: $\mathbf{Y_i} = \{(\mathbf{Y_{i,1}} \quad \mathbf{Y_{i,2}} \quad \mathbf{Y_{i,3}})\}$] which is chosen to extract three most significant dimensions of 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 (x_i, y_i). In [17], we have combined a hand posture Pi and spatial features Yi as eq. 1 following:

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

Manifold spaces on multiviews: In our previous researches [17], 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 each view, postures are capture from each Kinect sensor that is 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:

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

We then used an interpolation scheme which maximize inter-period phase continuity on each viewpoint, or periodic pattern of image sequence is taken into account as in [17, 18].

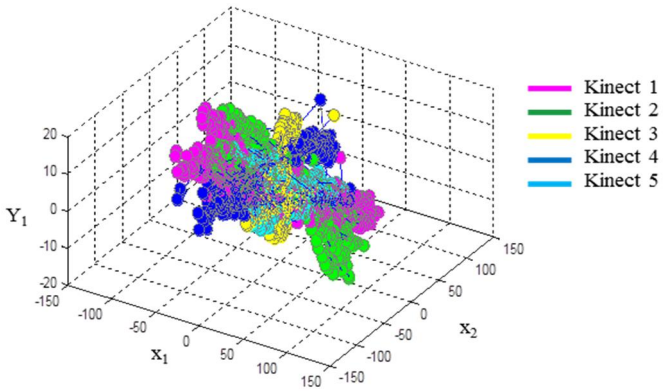

Figure 2. Manifold space of the gesture G_2 on five difference view-points

Figure 2 shows separations of the same gesture G_2 from five difference views of five Kinect sensors (K_1,K_2,…,K_5). This figure confirms inter-class variances when whole dataset is projected in the manifold space. In particularly, the patterns of the same hand gesture are presented on five views which are distinguished with others. while its manifold space is similar trajectory. The G_2 dynamic hand gestures of Kinect sensor K_1 presented in magenta; K_2 is showed on blue color; K_3 is illustrated on yellow color; K_4 is cyan color; and K_5 is green curves respectively. Features vector then are recognized on two cases by SVM classifier [18] as showed in Fig. 1. On the first one, gesture is evaluated on each view. On the other hand, features are evaluated on cross view. Figure 2 shows that hand gestures are distinguished in exter-class and they are converged in inter-class.

2.2. Learning view-invariant representation for cross-view recognition

As mentioned previously, private features of the same gesture are very different at different viewpoints. They should

be represented in another common space to be converged. There exists a number of techniques to build the viewpoint invariant representation. In this paper, we will deploy a variant of Canonical correlation analysis method (CCA [1]). However, most of multi-view discriminant analysis in the literature as well as in [1] were exploited for still images. To the best of our knowledge, our work is the first one to build cross corelation space for video sequences. We will see how such techniques could help to improve cross-view recognition overral.

Canonical Correlation Analysis method (CCA) [1]: a method of correlating linear relationships between two multidimensional variables. CCA can be seen as the problem of finding basis vectors for two sets of variables such that the correlations between the projections of the variables onto these basis vectors are mutually maximized.

Hand gestures consist c classes (c = 5) which are observed from v views (v = 5), the number of hand gestures from the jth view of the ith class is n_{ij}. G is defined as (4) quotient following:

$$\boldsymbol{G} = \left\{\boldsymbol{G_{TS}^{ijk}} | \boldsymbol{i} = (\boldsymbol{1}, \dots, \boldsymbol{c}); \boldsymbol{j} = (\boldsymbol{1}, .., \boldsymbol{v}); \boldsymbol{k} = \left(\boldsymbol{1}, .., \boldsymbol{n_{ij}}\right)\right\} \quad (4)$$

Given gestures from two views: $\boldsymbol{G_{TS}^{ijk}}$ and $\boldsymbol{G_{TS}^{i(j+1)k}}$; $\boldsymbol{j} = (\boldsymbol{1}, .., \boldsymbol{v})$ which $\boldsymbol{G_{TS}^{ijk}} \in \boldsymbol{R^{dj}}$ is the k^{th} gesture from the j^{th} view of the i^{th} class, *dj* is the dimensions of data at the j^{th} view. The Canonical Correlation Analysis method tries to determine a set of v linear transformations to project all gestures from each view j = (1,..,v) to another view. The projection results of G on the view j^{th} on $j+1^{th}$ is denoted by (5) quotient following:

$$\boldsymbol{Y} = \begin{Bmatrix} \boldsymbol{y_{ijk}} = \boldsymbol{w_j^T} * \boldsymbol{G_{TS}^{ijk}} | \boldsymbol{i} = (\boldsymbol{1}, \dots, \boldsymbol{c}); \boldsymbol{j} = (\boldsymbol{1}, .., \boldsymbol{v}); \\ \boldsymbol{k} = \left(\boldsymbol{1}, .., \boldsymbol{n_{ij}}\right) \end{Bmatrix} \quad (5)$$

Canonical correlation analysis seeks vectors w_j and w_{j+1} that $\boldsymbol{w_j^T} * \boldsymbol{G_{TS}^{ijk}}$ and $\boldsymbol{w_{j+1}^T} * \boldsymbol{G_{TS}^{i(j+1)k}}$ maximine correlation. Then one seeks vectors maximizing the same correlation subject to the constraint that they are to be uncorrelated with the first pair of canonical variables; this gives the second pair of canonical variables. This procedure may be continued up to the last case. The objective is formulated by a quotient (6) following:

$$\boldsymbol{argmax\ Corr}\left(\boldsymbol{w_j^T} * \boldsymbol{G_{TS}^{ijk}}, \boldsymbol{w_{j+1}^T} * \boldsymbol{G_{TS}^{i(j+1)k}}\right) \quad (6)$$

3. EXPRIMENTIAL RESULTS

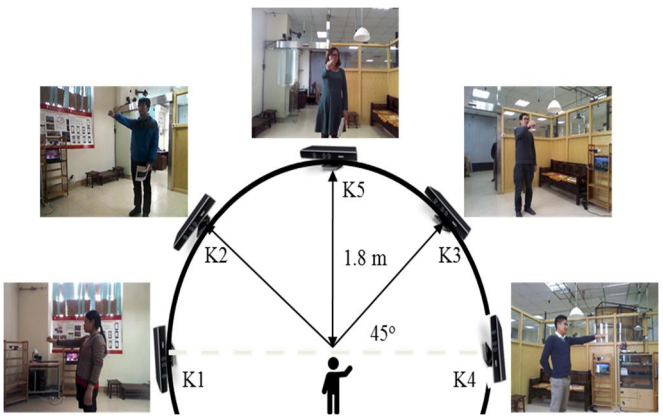

Figure 3. Environment setup of difference view-points

To evaluate the proposed framework, we utilize a multi-view dataset which is collected from multiple camera viewpoints (five Kinect sensors: K_1, K_2, K_3, K_4, K_5) in indoor environment with complex background as showed in Figure 3. Detail about this dataset is presented in other previous work [13].

The average accuracy is firstly computed to evaluate performance for two techniques with variation of viewpoints on both single and cross view. The canonical correlation analysis (CCA) is then applied to project all dynamic hand gestures from each pair of viewpoints.

Preparation of the training and testing data in this paper is described in detail at [14, 17]. That uses leave-one-subject-out cross-validation. 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 evaluated on different viewpoints with the training set. The evaluation metric used in this paper is presented in eq. (7) following:

$$\boldsymbol{accuracy} = \frac{\sum \boldsymbol{Corrects}}{\boldsymbol{Total}}\% \quad (7)$$

3.1. Evaluation hand gesture recognition on multi views

Table 1 shows the dynamic hand gesture recognition results of different numbers of classes which manifold features are extracted as described in detail at our previous research [16]. As that could be seen from the Tab. 1 that the proposed method gives the best results on all single views (K_1, K_2, K_3, K_4, K_5). In which the highest value belongs to single view with 99.36% and the smallest value at 81.31%.

Table 1. Cross-view hand gesture recognition with hand-craft feature of five gesture classes

	K_1	K_2	K_3	K_4	K_5
K_1	**81.31**	59.6	58.62	47.89	41.38
K_2	66.72	**92.68**	89.56	58.46	53.45
K_3	73.86	76.27	**99.36**	88.18	76.4
K_4	63.85	72.82	96.55	**98.52**	76.03
K_5	42.93	45.86	62.52	77.02	**90.48**
Single view			**92.47%**		
Cross view			**66.39%**		

Table 1 shows the detail cross-view results between five Kinect sensors these are setup as Fig. 2. A glance at the Tab. 2 provided evident reveals that:

- **Single view** gives more competitive performance than **cross-view**. The average value is 92.47% that is higher than other cases, 71.61% respectively. This is apparent that orient of hand to Kinect sensor directly affects on the gesture recognition result.

- **Single view** gives quite good results on all of five Kinect sensors while K_2, K_3 and K_4 are best results at the front views, with 92.68%, 99.36% and 98.52% respectively. The **cross-view** of K_1 gives the worst results which fluctuate at somewhere from 41.38% to 59.6% only, and the **cross-view**

K_5 obtains from 42.93% to 77.02%. These results are because the hands are occluded or out of camera field of view, or because the hand movement is not discriminative enough.

3.2. Evaluation hand gesture on shared space learning

Table 2 presents results when hand craft feature is projected from the Kinect sensor to other shared spaces [1]. Overall, the accuracy in cross view of five Kinect sensors are experienced a balance results over the period shown. Specially, some results dramatically increase from 41.38% to 52.84% accounted for pair between K_1 and K_5, and from 42.93% to 58.27% with pair between K_5 and K_1, respectively.

Table 2. Cross-view hand gesture recognition with canonical correlation analysis method

	K1	K2	K3	K4	K5
K1		**63.18**	56.72	**55.40**	**52.84**
K2	**67.32**		73.86	**61.95**	53.52
K3	72.70	75.97		76.36	75.56
K4	61.89	67.13	76.67		68. 90
K5	**58.27**	**53.44**	**66.46**	**78.22**	

4. DISCUSSION AND CONCLUSION

In this paper, the hand gesture recognition in the different view points is firstly deployed. The hand gesture recognition with the canonical correlation analysis method is then evaluated. Results show that the single view results are higher than cross view results with some main conclusions following: i) Hand craft feature is obtained highest performance with frontal view, it is still good when view point deviates in the range of 45^0 and drastically reduced when the viewpoint deviates from 90^0 to 135^0. The recommendation is to learn dense viewpoints so that testing view point could avoid huge difference compared to learnt views; ii) The common share space is applied that the cross view recognition results impacted on performance of the manifold recognition method. It is recommended to project to the share space between difference view points of the same human hand gesture in order to combine multi-view information that help to obtain higher recognition accuracy overall.

REFERENCES

[1]. Hotelling, H., 1936. *Relations Between Two Sets of Variates*. Biometrika. 28 (3–4): 321–377.

[2]. D. Shukla, Ö. Erkent and J. Piater, 2016. *A multi-view hand gesture RGB-D dataset for human-robot interaction scenarios*. ROMAN 2016, USA, pp. 1084-1091.

[3]. Haiying Guan, Jae Sik Chang, Longbin Chen, R. S. Feris and M. Turk, 2006. *Multi-view Appearance-based 3D Hand Pose Estimation*. CVPRW 2006, pp. 154-154.

[4]. K. He, G. Gkioxari, P. Dollar, R. Girshick, 2017. *Mask R-CNN*. In Proceedings of the ICCV 2017, pp. 2980–2988.

[5]. P. Jangyodsuk, C. Conly, and V. Athitsos, 2014. *Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features*. PETRAE 2014, pages 50:1–50:6.

[6]. J. Do, H. Jang, S. Jung, J. Jung, and B. Z, 2005. *Soft remote control system in the intelligent sweet home*. IRS 2005, pp. 3984–3989.

[7]. T. Simon, H. Joo, I. Matthews, and Y. Sheikh, 2017. *Hand keypoint detection in single images using multiview bootstrapping*. CVPR 2017, pp. 1145 - 1153.

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

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

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

[11]. Q. Chen, A. El-Sawah, C. Joslin, N. D. Georganas, 2005. *A dynamic gesture interface for virtual environments based on hidden markov models*. HAVE 2005, pp. 109-114.

[12]. B. D. Lucas and T. Kanade, 1981. *An iterative image registration technique with an application to stereo vision.* The 7th International Joint Conference on Artificial Intelligence, Vol. 2, USA, pp. 674-679.

[13]. Dang-Manh Truong, Huong-Giang Doan, Thanh-Hai Tran, Hai Vu, Thi-Lan Le, 2019. *Robustness analysis of 3D convolutional neural network for human hand gesture recognition*. IJMLC, Vol.9(2), pp. 135-142.

[14]. 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.* RIVF 2016, pp. 163 - 168.

[15]. J. S. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, 2018. *Depth-based hand pose estimation: methods, data, and challenges*. International Journal of Computer Vision, Vol. 126(11), pp. 1180–1198.

[16]. J. Shi and C. Tomasi, 1994. *Good features to track*. CVPR 1994, USA, pp. 593-600.

[17]. Huong-Giang Doan, Hai Vu, and Thanh-Hai Tran, 2017. *Dynamic hand gesture recognition from cyclical hand pattern*. MVA 2017, pp. 84-87.

[18]. C. 1. C. Burges, 1997. *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery Journal, vol. 43, pp. 1-43, 1997.

[19]. Poon, Geoffrey & Chung Kwan, Kin & Pang, Wai-Man, 2018. *Real time Multiview Bimanual Gesture Recognition*. SIPROCESS 2018.

THÔNG TIN TÁC GIẢ

Đoàn Thị Hương Giang

Khoa Điều khiển và Tự động hóa, Trường Đại học Điện lực