15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Dynamic Hand Gesture Recognition from Cyclical Hand Pattern

Huong-Giang Doan
MICA HUST
Hanoi - VietNam
huonggiang80dl@gmail.com

Hai Vu
MICA HUST
Hanoi - VietNam
hai.vu@mica.edu.vn

Thanh-Hai Tran
MICA HUST
Hanoi - VietNam
thanh-hai.tran@mica.edu.vn

Abstract

In this paper, we tackle advantages of cyclical movement patterns of hand gestures. The cyclical patterns are defined as closed-form which hand moves away from a rest position, follows one or more of a series of the movement of hand shapes and returns to its rest position. Due to the cyclical pattern characteristic, phase of gestures are supportive cues for deploying robust recognition schemes. We conduct a spatial-temporal representation of the hand gestures which takes into account both hand shapes and its movements during a gesture. The phase alignment then is deployed in the conducted space. The proposed scheme ensures inter-period phase continuity as well as normalizes length of the hand gestures. Three different datasets of dynamic hand gestures consisting of non-cyclical and cyclical patterns are examined. Evaluation results confirm that the best accuracy rate achieves at 96% for cyclical pattern that is significantly higher than results for typical gestures. The proposed method suggests a feasible and robust solution addressing technical issues in developing human-computer interaction applications such as using hand gestures to control home appliance devices.

1 Introduction

Dynamic hand gesture is a natural communication between human and computer. Particularly, they are more popular in controlling home appliances, thanks to recent advances of intelligent computing, smart devices. Many research works [1, 5] have been intensively proposed for automatically detecting and classifying dynamic hand gestures. However, this topic is still challenging because of: (1) a large diversity in how people perform gestures, making detection and classification difficult; (2) real-time requirements for critical tasks such as hand detection, tracking, and gesture recognition; (3) spotting gestures from data stream. To achieve robust systems, many constraints, or strict conditions are deployed in real applications/systems. For instance, the most common ways are utilizing gloves, attaching makers/motion sensors on hands, palms, or simplifying background/scenes of the applications. Obviously, these solutions are not feasible for applications which are deployed with more flexible contexts such as controlling items around the house (e.g., lights, doors, air conditioners, and so on).

In this paper, we consider and tackle cyclical movement patterns of hand gestures whether they are supportive and natural cues for deploying a recognition system. Movements of hand gestures are organized as excursions [16], in which the gesticulating limb moves away from a rest position, engages in one or more of a series of hand movement patterns, then is returned to its rest position. Intuitively, cyclical movements are discriminative form comparing with typical gestures. Although a number of dynamic hand gestures datasets (e.g., MSRGesture3D [2], R3DCNN [18], Chalearn [1]) are published. It is not clearly observing the advantages of the cyclical gestures (some examples are shown in Fig. 1); or do they bring benefits comparing with non-cyclical ones. Another question is that do such gestures ensure the naturalness to end-user?

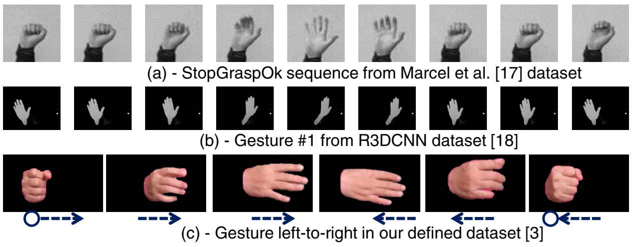
(a) - StopGraspOk sequence from Marcel et al. [17] dataset
(b) - Gesture #1 from R3DCNN dataset [18]
(c) - Gesture left-to-right in our defined dataset [3]

Figure 1. Examples of cyclical gestures selected from public datasets

To match a probe and gallery dynamic gestures, a critical task is to register phrase of them. With typical patterns, this registration leads to align/match each pair of hand shapes (e.g., utilizing Dynamic Time Warping (DTW) algorithms). However, the cyclical gestures form a closed-pattern of hand moves if the gestures are projected into an associated space, which conveys both spatial and temporal features. Thanks to the periodicity of the gestures, we can solve the phase synchronization with whole sequence of frames. To this end, we firstly represent hand gesture sequences in a spatial-temporal space whose dimensions are conducted from the most important features extracted from hand shape (spatial) and the hand trajectory (temporal). The hand shapes are exploited through an isometric feature mapping algorithm (ISOMAP [4]). While dominant trajectories of the hand is extracted by connecting keypoints tracked using KLT (Kanade-Lucas-Tomasi) technique. We then deploy an interpolation scheme on each dimension to reconstruct a new image sequence with the pre-determined number of frames. This registration scheme takes into account the inter-period phase continuity in the conducted space. The support vector machine (SVM) technique is utilized to assign gesture label of the interpolated sequence. We evaluate performance of the proposed approaches by comparing on different public datasets. The achieved performance with cyclical movement patterns are very competitive. Moreover, other technical issues such as gestures spotting could be resolved thanks to the periodic patterns.

The rest of paper is organized as follows. Section 2 summaries hand gestures recognition techniques. Section 3 describes the proposed method. Section 4 reports the experimental results. Finally, Sec. 5 concludes works and suggests research directions.

2 Related works

There are uncountable solutions for developing a vision-based hand posture/gesture recognition system in the literature. Readers can refer good surveys such as [5, 6]. For detecting and recognizing hand gestures from a video stream, most of the related works have to deal with common issues such as the complexity of hand shapes, a variation of gesture trajectories, cluttered background, light conditions, changing velocity, and missed phases in the dynamic gestures. For dynamic gestures, the phase synchronization issue has been particularly interested in many relevant works. For example in [8], the authors proposed to use DTW algorithms. This technique is adopted from time series analysis domain to align a pair of hand shapes. Additionally, Hidden Markov Model and its variant in [7] are preferred to solve state issues what appear in the dynamic hand gestures. Recently, the authors in [18] perform simultaneous detection and classification of dynamic hand gestures with a recurrent three-dimensional convolutional neural network from multi-modal data. The method in[18] achieved very promising results comparing with state-of-the-arts. However, it requires not only depth, but also IR, motion flow, and RGB data, which may raise other critical issues such as data synchronization, computational time, fusions of them.

A temporal and/or spatial feature space has been considered and conducted in many topics of dynamic action recognition [10]. The advantages of the spatial-temporal space are not only to represent but also to address the temporal misalignment issues. The most common ways are based on interpolation idea in the conducted space. These approaches usually try to generate a high frame-rate video from a single/multiple low frame-rate video [9, 10]. Other approaches [11] enhanced both spatial-temporal resolution from two sequences that ones is high resolution and low frame rate, the other is low resolution and high frame rate. Particularly, a phase estimation approach is proposed in [12] dealing with low frame-rate videos for human gait recognition. Their techniques aim to create a periodic temporal super resolution. For the dynamic hand gestures, although temporally phases (e.g., generally, consisting of preparation, nucleus, and retraction) are noticed in [16, 18], their affects to recognition rate have been not investigated. In this work, the phase alignment is considered in context of cyclical-pattern movements. We deploy this technique by inspiring from temporal resolution technique. The interpolated video sequences are normalized in terms of length of the sequence. The proposed technique is evaluated by comparing recognition rate of the cyclical patterns and typical ones.

3 Proposed method

We propose a framework for hand gesture recognition which composes of three main components: hand gesture extraction, hand gesture representation and recognition, as shown in Fig. 2. To extract a hand gesture from video stream, we rely on the techniques presented in [13]. In this section, we focus on representing hand gestures (yellow/middle panel in Fig 2). We utilize a manifold learning technique to present phase shapes. The hand trajectories are reconstructed using a conventional KLT trackers [14]. We then propose an interpolation scheme which maximize inter-period phase continuity, or periodic pattern of image sequence is taken into account. In the following, we will present in detail sub-steps of the proposed hand gesture representation. For the gesture recognition, we utilize various existing classification techniques.

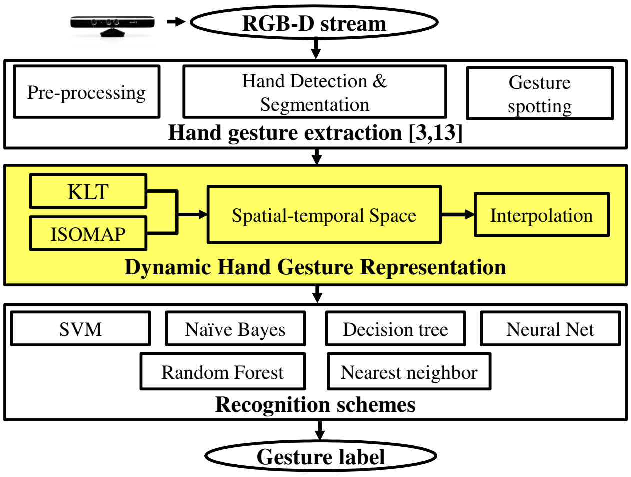

Figure 2. The proposed framework

3.1 Temporal features extraction

The temporal features of a frame are two coordinates (x, y) of the average trajectory of the hand during gesture implementation. This trajectory is computed by averaging all trajectories extracted using KLT tracker [14] (Fig. 3(a-b)). This work was presented in detail at our previous work [3]. The temporal features (Tr_N^G) extracted from an image sequence of a hand gesture as: $Tr_N^G = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$

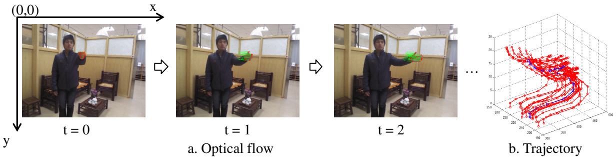

Figure 3. An example of KLT-based trajectory.

3.2 Spatial features extraction

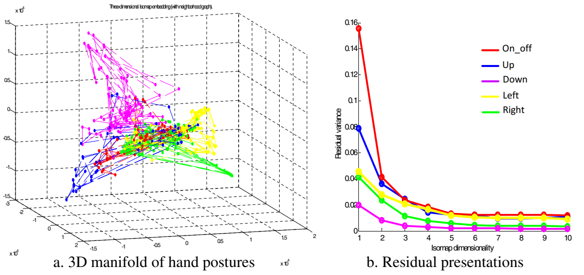

Figure 4. Distribution of dynamic hand gestures in the low-dimension.

The spatial features of a frame is computed though manifold learning technique ISOMAP [4] by taking the three most representative components of this manifold space as presented in our previous work [20]. Given a set of N segmented postures $\boldsymbol{X} = \{X_i, i = 1, ..., N\}$, after compute the corresponding coordinate vectors $\boldsymbol{Y} = \{Y_i \in R^d, i = 1, ..., N\}$ 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 R_d 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 D_d. 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. 4(a) illustrates 3-D manifolds of five different hand gestures, Fig. 4(b) show the residual error). A hand gesture then is represented as: $Y_i = \{(Y_{i,1}, Y_{i,2}, Y_{i,3})\}$

In [20], we have presented a hand posture by temporal and spatial features $P_i = (Tr_i, Y_i) = (x_i, y_i, Y_{i,1}, Y_{i,2}, Y_{i,3})$. Figure 5(a)-(e) illustrates new representations in 3-D space of five hand gestures. In comparison with Fig. 4, separation between five gestures are clearer than that is presented in Fig. 4. The main reason is that the extracted temporal features are embedded in this new space. Fig. 5(f) confirms interclass variances when whole dataset is projected in the proposed space. In particularly, cyclic patterns of the hand gestures are presented as closed-circles.

3.3 Phase alignment based on interpolation

By utilizing the conducted space, comparison between two gestures could be straightforward implementation. However, inter-period phase would be discarded. In other words, periodic pattern of image sequence has been omitted. To overcome this issue, we deploy an interpolation scheme so that hand gesture sequences have same length, and maximize inter-period phase continuity. The proposed scheme is based on piecewise interpolation and similarity measurement between two adjacent points in the proposed hand gesture space. Supposing M is the desired length for each gesture, given $\boldsymbol{G}^{TS} = \{P_1, P_2, ..., P_N\}$ at time instances ($t_1, t_2, \ldots, t_N$) respectively, a distance vector of $\boldsymbol{G}^{TS}$ is calculated by $\boldsymbol{D_{inter}} = \{d_i; (i = 1, ..., N-1)\}$ where is Euclidean distance between two consecutive postures P_i and P_{i+1}.

In case $N < M$, we find a maximal distance from vector $\boldsymbol{D_{inter}}$ ($d_{max} = \max(D_{inter})$). This furthest point is the first priority to do the interpolation. Then the interpolated point is inserted between them. The length of the new sequence after inserting is $N + 1$. This procedure is iterated until the sequence length reaches M postures.

In case $N > M$, we find a minimal value of vector $\boldsymbol{D_{inter}}$ ($d_{min} = \min(D_{inter})$) between two nearest points, supposing P_i, P_{i+1}. We then eliminate one from these two points as follows (1):

$$P_{removed}=\begin{cases} P_i & [(d_{i-1}<d_{i+1})\&(i\neq N-1)]or[(i=1)] \\ P_{i+1} & [(d_{i-1}>d_{i+1})\&(i\neq 1)]or[(i=N-1)] \end{cases} \tag{1}$$

Gesture recognition is then performed using different classification methods such as SVM, Naive Bayes, Decision tree, Random Forest, Neural Net, Nearest Neighbor. An optimal classifier is selected based on the performance evaluation results.

4 EXPERIMENTAL RESULTS

4.1 Evaluations on different dataset

The performance of the proposed method is evaluated on three different datasets: $MSRGesture3D$ [2]; and a subset of R3DCNN dataset [18], and the dataset prepared by ourself. MSRGesture3D consists of 12 dynamic American Sign Language gestures, is implemented by 10 people. The second dataset consists of 15 dynamic gestures. This dataset consists of cyclical hand movements collected from 25 gestures of R3DCNN dataset. To intensively evaluate impacts of cyclical movements, we construct the third one. Eight volunteers (4 males and 4 females) are invited to perform five pre-defined gestures at various positions in a lab-experimental room. We adjust the distances from the subjects to a Kinect sensor [15], as well as set different directions. Totally, 13 positions are recorded. Each position therefore consists of 120 dynamic hand gestures (8 subjects × 5 gestures × 3 times). The total number of collected videos is 1560 sequences. The proposed framework is warped by a C++ program on a PC Core i5 3.10GHz CPU, 4GB RAM. Kinect sensor captures data at 30 fps.

For each dataset, 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. For each validation, based on the confusion matrix, precision and recall measurements are averagely calculated. The evaluation results are shown in Table 1. For MSRGesture3D dataset, the state-of-the-art method achieved ups to 92.45% in [17] and 94.72% in [19]. Obviously, with recall rate of 92.03%, results of the proposed method is comparable. An interesting point is that with the second dataset, the recall rate achieved far from that was reported in [18] (83.6% for depth data). It is noticed that original result in [18] was evaluated on the full dataset with 25 gestures. With the third dataset, although number of gestures are only five, but this is more challenge because the proposed method is evaluated from various positions.

Table 1. Performance of the proposed method (%) on three different datasets

Dataset	Precision	Recall
MSRGesture3D	94.5 ± 3.1	92.03 ± 5.1
R3DCNN subset	91.0 ± 4.7	87.5 ± 4.2
Our dataset	96.1 ± 3.2	96.9 ± 2.1

4.2 Impacts of the phase alignment

We continuously evaluate the performance with different 13 positions with 3 recognition schemes: DTW-based in [3]; a CNN (Convolution Neuron Networks) features combining SVM and the proposed method. While DTW attempts a pair of hand shape alignment, CNN is a must-to-try technique, the proposed method dedicates to resolve phase alignment in cyclical movements. The evaluation results are shown in Fig. 6.

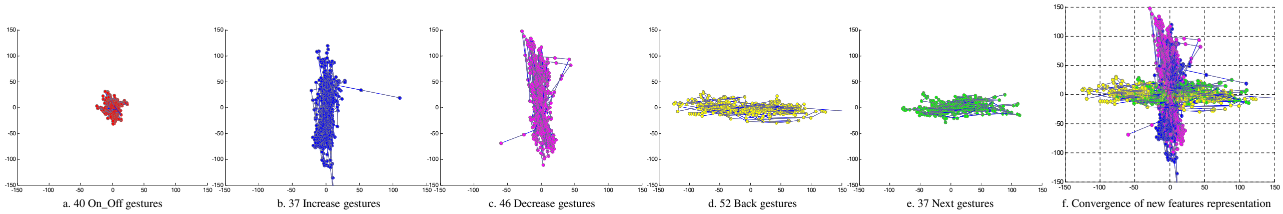

Figure 5. Five dynamic hand gestures in the 3D dimension.

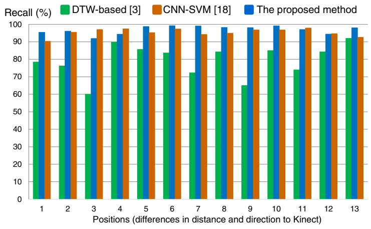

Figure 6. Comparison results between the proposed method vs. others at thirteen positions.

The blue columns present results the proposed method. Obviously, the proposed method is over-performed others at various positions. Main reasons are that it ensures the inter-period phase continuity. This evaluation also confirmed its robustness and tolerance with changing of subject positions and/or different hand directions.

5 CONCLUSION

This paper tackled cyclical patterns of hand movements during performing a gesture. Due to closed-form of gestures, we taken into account phase alignment. This characteristic is deployed in a spatial-temporal feature space. To resolve the phase alignment, the interpolation method to normalize length of hand gestures is deployed so that the inter-phase continuity is maximal. The experimental results confirmed that closed-form pattern of dynamic gestures can be archived higher performance comparing with typical gestures. The proposed algorithm is more robust and tolerant with changing of subject positions and/or different hand directions. Moreover, the cyclical pattern also suggests a solution to overcome other technical issues for developing human-computer interaction applications. It shows the feasibility to deploy real applications to control home appliance devices.

Acknowledgement

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number FWO.102.2013.08.

References

[1] S. Escalera, J. Gonz‘alez, et al., “Multi-modal gesture recognition challenge 2013: Dataset and results,” *Proc. of ICMI 2013.*

[2] A. Kurakin, Z. Zhang, Z. Liu, “A Real-Time System for Dynamic Hand Gesture Recognition with a Depth Sensor, ” *Proc. of EUSIPCO, 2012.*

[3] H. Doan, H. Vu, and T. Tran, “Recognition of hand gestures from cyclic hand movements using spatial-temporal features,” in *Proc. of SoICT 2015.*

[4] J. B. Tenenbaum, et al. “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[5] X. Zabulis, H. Baltzakis, and A. Argyros, *Vision-based Hand Gesture Recognition for Human Computer Interaction.* Lawrence Erlbaum Associates, 2009.

[6] S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” *Artif. Intel. Rev.*, vol. 43, pp. 1–54, 2015.

[7] M. Elmezain, A. Al-Hamadi, and C. Michaelis, “Real-Time Capable System for Hand Gesture Recognition Using HMM in Stereo Color Image Sequences,” *WSCG*, vol. 16, pp. 65–72, 2008.

[8] K. Barczewska and A. Drozd, “Comparison of methods for hand gesture recognition based on Dynamic Time Warping algorithm,” *FedCSIS*, pp. 207–210, 2013.

[9] M. Shimano, T. Okabe, I. Sato et al., “Video temporal super-resolution based on self-similarity,” in *Proc. of ACCV 2010.*

[10] E. Shechtman, Y. Caspi, and M. Irani, “Space-Time Super-Resolution,” vol. 27, no. 4, pp. 531–545, 2005.

[11] K. Watanabe, Y. Iwai, et al., “Video synthesis with high spatio-temporal resolution using motion compensation and spectral fusion,” *IEICE Transactions*, vol. 89-D, no. 7, pp. 2186–2196, 2006.

[12] Y. Makihara, A. Mori, and Y. Yagi, “Periodic Temporal Super Resolution Based on Phase Registration and Manifold Reconstruction,” *CVA*, vol. 3, no. 1, 2011.

[13] H.-G. Doan, V.-T. Nguyen, et al., “A combination of user-guide scheme and kernel descriptor on RGB-D data for robust and realtime hand posture recognition,” *EAAI*, vol. 49, Mar. 2016.

[14] J.Shi and C.Tomasi, “Good features to track,” in *Proc. IJCAI*, 1994, pp. 593–600.

[15] “http://www.microsoft.com/en-us/kinectforwindows.”

[16] A. Kendon,“Current issues in the study of gesture,” in *The biological foundations of gestures: motor and semiotic aspects*, Lawrence Erlbaum Ass., 1986

[17] O. Oreifej and Z. Liu, “HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences,” *Proc. of CVPR 2013.*

[18] P. Molchanov,X. Yang,et al.,“ Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks,” in *Proc. of the CVPR 2016* .

[19] X. Yang and Y. Tian, “Super Normal Vector for Action Recognition Using Depth Sequences,” *Proc. of CVPR 2014.*

[20] H. Doan, H. Vu, and T. Tran, “Phase Synchronization in a Manifold Space for Recognizing Dynamic Hand Gestures from Periodic Image Sequence,” in *Proc. of RIVF 2017.*