New Cyclical Pattern and Temporal-Spatial Representation for Robust Dynamic Hand Gesture Recognition

Huong-Giang DOAN[1,2] Hai VU[1] Thanh-Hai TRAN[1]
[1] International Research Institute MICA, HUST-CNRS/UMI-2954-GRENOBLE INP
and Hanoi University of Science & Technology, Vietnam
[2] Industrial Vocational College Hanoi, VietNam

***Abstract*— This paper tackles a new prototype of dynamic hand gestures and its advantages to apply in controlling smart home appliances. The proposed gestures convey cyclical patterns of hand shapes as well hand movements. Thanks to the periodicity of defined gestures, on one hand, some common technical issues that may appear when deploying the application (e.g., spotting gestures from a video stream) are addressed. On the other hand, they are supportive features for deploying robust recognition schemes. To this end, we propose a novel hand representation in temporal-spatial space. Particularly, the phase continuity of the gesture's trajectory is taken into account underlying the conducted space. This scheme obtains very competitive results with the best accuracy rate is of 96%. We deploy the proposed techniques to control home-appliances such as lamps, fans. Such systems have been evaluated in both lab-based environment and real exhibitions. In the future, the proposed gestures will be evaluated in term of naturalness of end-users and/or robustness of the systems.**

I. INTRODUCTION

Home-automation products have been widely used in smart homes thanks to recent advances in intelligent computing, smart devices, and new communication protocols. To maximize user-ability, we intend to deploy a human-computer interaction method, which allows users to use their hand gestures to perform conventional operations for controlling home appliances. To this end, we propose a new prototype of hand gestures and also deploy a real-time gesture recognition system.

The performance of dynamic hand gestures recognition strongly depends on the type of dataset used in relevant works. There were many self-defined dynamic hand gestures datasets such as [1], [2], [3], [4]. Many other works proposed hand gestures datasets that have been collected and widely published for different purposes: MSRGesture3D dataset for evaluating human action recognition [5], [6]; Cambridge-Gesture dataset for evaluating hand detection [7], [8]. In this paper, we consider and tackle cyclical hand-gestures where hand shapes are cyclical patterns and their trajectories (hand movements) are in a closed-form. Intuitively, cyclical gestures are discriminative form comparing with common ones.

Although much intensive works in the dynamic hand gesture recognitions [9], [10], [11], [12], deploying such techniques in real practical applications faces many technical issues such as real-time requirement and and complex movement of hands, arms, face, and body. Thanks to the periodicity of the defined gestures, technical issues such as spotting gestures from a video stream become more feasible and the phase normalization with the whole sequence of frames is more tractable. To obtain this, we firstly represent hand gesture sequences in a spatial-temporal feature space. The hand shapes are exploited through an isometric feature mapping algorithm (ISOMAP [13]). The dominant trajectories of the hand are extracted by connecting keypoints tracked using KLT (Kanade-Lucas-Tomasi) technique [14]. We then deploy an interpolation scheme on each dimension to reconstruct the phase-normalized image sequence with a pre-determined number of frames. This registration scheme takes into account the inter-period phase continuity in the conducted space. A support vector machine (SVM) classifier [15] is utilized to assign gesture label for the interpolated image sequence. We evaluate the performance of the proposed approaches on different public datasets with various scenarios to confirm the robustness of the proposed method. The achieved performance is very competitive.

II. PROPOSED METHOD FOR DYNAMIC HAND GESTURE RECOGNITION

A. Designing an unique dataset of dynamic hand gestures and their characteristics

To control a device, the user stands in front of a Kinect sensor [16] in the valid range from 1.2 to 4.0 meter. A gesture command is implemented through three phases: preparation; performing; relaxing. At preparation phase, the user stays immobile. At performing phase the user raises his/her hand (e.g., right hand) and moves the hand according to a pre-defined trajectory. Simultaneously, while moving hand, the hand shape also changes following three states: *initial* state, *implementing* state and *ending* state. Hand shapes changes follow a cyclic pattern from fist shape at *initial* state to open shape at *implementing* state, and fist shape again at *ending* state (Fig. 1). In this study, we design five commands

Fig. 1. In each row, changes of the hand shape during a gesture performing. From left-to-right, hand-shapes of the completed gesture chance in a cyclical pattern (closed-opened-closed).

which are the most common used to control home appliances: Turn on/Turn off; Next; Back; Increase; Decrease. Actually, the number of commands is quite limited. However, there is no limitation to design new gestures based on the same methodology. The proposed scheme are discriminated from existing hand gestures in both characteristics: hand shape and direction of hand movement. Hand shapes represent a cyclic pattern of a gesture, whereas hand movements represent the meaning of a gesture.

B. Real-time dynamic hand gesture spotting

A dynamic hand gesture is a sequence of consecutive hand postures varying in time. Therefore, it is necessary to determine the starting and ending times of a hand gesture before recognizing it. In this study, all pre-defined gesture commands have the same hand shape at starting and ending times. Moreover, hand shapes of a gesture follow a cyclical pattern. We then rely on these properties for gesture spotting.

Before spotting a hand gesture, we conducted some pre-processing such as depth and RGB calibration (Fig. 2(b)), human body detection (Fig. 2(c)), hand detection using Gaussian Mixture Model (GMM) [17] (Fig. 2(d)), skin color pruning for hand region segmentation. Detail of these techniques were presented in our previous work [18].

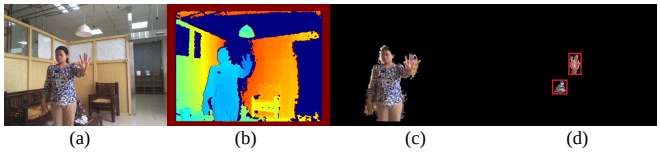

Fig. 2. Hand detection and segmentation procedures.(a) RGB image; (b) Depth image; (c) Extracted human body; (d) Hand candidates.

Given a sequence of segmented hand regions, we define a function $f_x(t)$ representing the evolution of hand region's area x_t extracted at time t. Figure 3 shows an example of $f_x(t)$ (the red curve) computed from 389 frames of several consecutive hand gestures performed by one person. The below graphic presents a zoomed cutoff segment containing three gestures with manual spotting. We observe that for each gesture, area of hand region increases then decreases.

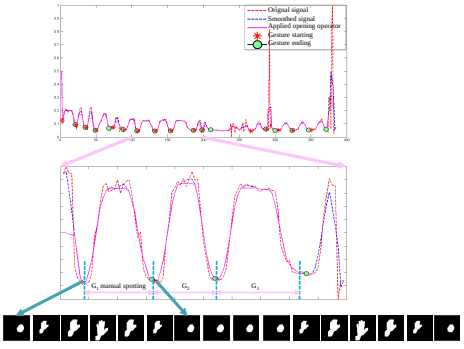

Fig. 3. First row: representation of signal $f_x(t)$. Second row: zoom of a segment of $f_x(t)$. Third row: variation of hand shape during the implementation of a hand gesture.

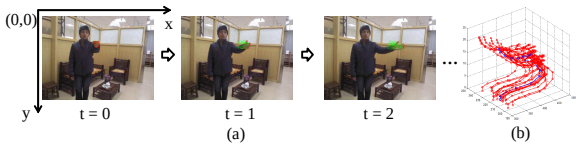

Fig. 4. An example of KLT-based trajectory. (a) Optical flow; (b) Trajectory.

This is suitable to the variation in hand shape from closure to open then to closure state. The highest value is found at the moment when hand spreads. Two lowest boundary values correspond to the closed hand posture. This remains the same for all other gestures. Based on this observation, we determine starting and ending times of a gesture as follows. First, we smooth $f_x(t)$ by a Gaussian filter to remove noise and dummy local peaks. Then we apply a morphological operator (opening operator) on the smoothed signal $f_{sx}(t)$. The obtained signal is $O(f_{sx}(t))$.

The final step is searching for local extrema of $O(f_{sx}(t))$. A gesture candidate exists in the interval determined by two consecutive local minimums with a local maximum at the middle. Once the starting and ending times of gesture candidates are determined, we will annotate them and store in the database for further processing.

C. Robust dynamic hand gesture recognition

Spatial-temporal feature extraction for gesture representation: Given a sequence of M frames of a spotted gesture, every frame, we extract spatial and temporal features then concatenate them to build the final representation of the gesture. 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. 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 [19], [14] (Fig. 4(a-b)). Once spatial and temporal features are extracted, they are combined to completely represent dynamic hand gestures. Figure 5(a)-(e) illustrates new representations in 3-D space of five different hand gestures. As shown, the separation between five gestures are very discriminative. Fig. 5(f) confirms inter-class variances when whole dataset is projected in the proposed space.

Phase normalization 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 (named phase-normalized image sequence), 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. Details of this technique were presented in our previous work [20]. Figure 6 presents some results of the interpolation procedure with the same length

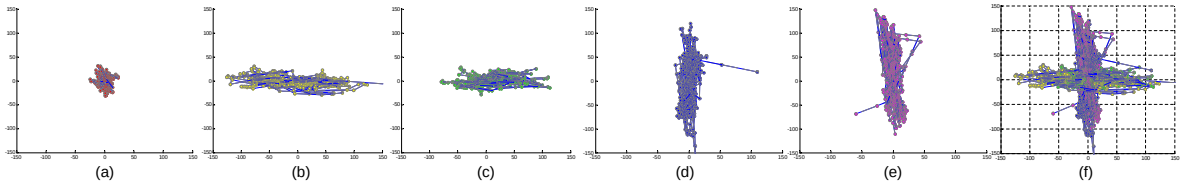

Fig. 5. Distribution of dynamic hand gestures in the low-dimension. (a) 40 On_Off gestures; (b) 52 Back gestures; (c) 37 Next gestures; (d) 37 Increase gestures; (e) 46 Decrease gestures; (f) Convergence of new features representation.

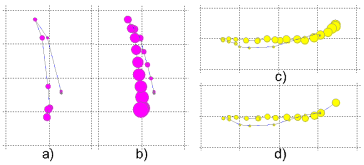

Fig. 6. Interpolation of dynamic hand gestures. a) Original gesture Decrease (9 frames); c) Original gesture Back (30 frames); b,d) corresponding interpolated hand gestures (20 frames).

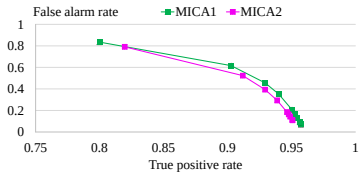

Fig. 7. Performances of the dynamic gesture spotting on two datasets MICA1 and MICA2

of sequence M is equal to 20. The number of postures in Fig. 6 (a) is equal to 10. In Fig. 6 (c), the number of postures is up to 28. Fig. 6 (b),(d) are two interpolated hand gestures after applying the interpolation procedure.

After interpolated, all dynamic hand gestures are represented by features vector of the same length. Gesture recognition is performed using SVM classifier [15]. The input of this classifier is feature vectors extracted from interpolated sequence.

III. EXPERIMENTAL RESULTS

A. Evaluating performance of the gesture spotting algorithm

We evaluate our algorithm of gesture spotting on our two datasets $MICA1$ (16 video of 16 subjects) and $MICA2$ (33 video of 33 subjects). These datasets are available at "http://mica.edu.vn/perso/Doan-Thi-Huong-Giang/MICADynamicHandGestureSet/". Each video in these datasets includes fifteen pre-defined gestures and some undefined gestures performed by one person. For quantitative evaluation, we use Jaccard Index JI[21]. A true positive (TP) is detected when $JI \geq \theta$ where θ is a pre-defined threshold. Otherwise, it is considered as an insertion (False Positive - FP). Fig. 7 illustrates quantitative spotting results in term of true positive rate and false alarm rate with θ varying from 0.1 to 0.9. When θ increases, the true positive rate lightly reduces from 0.96 to 0.8 (on the MICA1 dataset) or from 0.95 to 0.82 (on the MICA2 dataset). That shows our algorithm performs very well on true samples. However, the false alarm rate increases significantly from 0.07 to 0.83 (on the MICA1 dataset) or 0.11 to 0.79 (on the MICA2 dataset). We propose to choose $\theta = 0.5$ that gives the best trade-off between the true positive rate and false alarm rate for testing the whole system of recognition.

B. Evaluating recognition performances

The performance of the proposed method is evaluated on two benchmark datasets: $MSRGesture3D$ [22]; and a subset of $R3DCNN$ dataset [23]. To intensively evaluate impacts of cyclical movements, we construct the third one. Eight volunteers (4 males and 4 females) are invited to perform three times five pre-defined gestures at various thirteen positions and orientations in a lab-experimental room. Each position consists of 120 dynamic hand gestures.

TABLE I

PERFORMANCE OF THE PROPOSED METHOD ON THREE DATASETS

Dataset	Precision (%)	Recall (%)
MSRGesture3D	94.5 $\pm$ 3.1	92.03 $\pm$ 5.1
R3DCNN subset	91.0 $\pm$ 4.7	87.5 $\pm$ 4.2
Our dataset	96.1 $\pm$ 3.2	96.9 $\pm$ 2.1

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 I. For MSRGesture3D dataset, the state-of-the-art method achieved ups to 92.45% in [24] and 94.72% in [25]. 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 [23] (83.6% for depth data). It is noticed that original result in [23] was evaluated on the full dataset with 25 gestures. With the third dataset, although number of gestures are only five, but this is more challenging because the proposed method is evaluated from various positions.

C. Impacts of the phase alignment

We continuously evaluate the performance with different 13 positions with 3 recognition schemes: DTW-based in [26]; a CNN (Convolution Neuron Networks) features combining SVM [27] 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

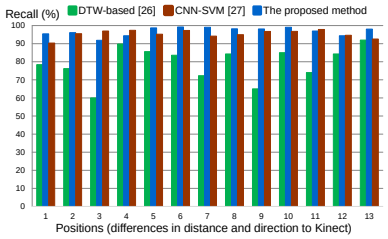

Fig. 8. Comparison results between the proposed method vs. others at thirteen positions.

shown in Fig. 8. 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.

D. Deployment in a practical application of lamp controlling

We have deployed the proposed techniques in a demo of controlling lamps and testing in with different people in both lab based environment and exhibitions. The participant expressed their strong interest in using hand gestures to control devices. This show a big potential and feasibility to deploy such techniques in reality (Fig. 9).

Fig. 9. Illustration of an user controlling lamps using hand gestures.

IV. CONCLUSIONS AND FUTURE WORKS

This paper described phase-normalized image sequence for increasing the performance of the dynamic hand gesture. Due to temporal variations of hand movements, we took into account normalizing phase during the completed gesture. The proposed technique is deployed in a spatial-temporal feature space. To resolve the phase normalization, the interpolation method to normalize length of hand gestures is deployed so that the inter-phase continuity is maximal. The experimental results confirmed that the proposed technique is able to achieve higher performance comparing with conventional methods with common public datasets. Moreover, the experimental results also confirmed that the proposed algorithm is more robust and tolerant with changing of subject positions and/or different hand directions. The proposed technique, therefore, suggests a feasible solution overcome technical issues for developing human-computer interaction applications.

REFERENCES

[1] S. Marcel, O. Bernier, J.-E. Viallet, and D. Collobert, "Hand gesture recognition using input-output hidden markov models," in *FG*, 2000, pp. 456–461.

[2] Z. Ren, J. Yuan, and Z. Zhang, "Robust Hand Gesture Recognition Based on Finger-Earth Movers Distance with a Commodity Depth Camera," in *ACM International Conference on Multimedia*, 2011.

[3] Y. Song, D. Demirdjian, and R. Davis, "Tracking body and hands for gesture recognition: Natops aircraft handling signals database," in *FG*, 2011, pp. 500–506.

[4] A. I. Maqueda, c. del Blanco, and F. G. Jaureguizar, "Human-computer interaction based on visual recognition using volumegrams of local binary patterns," in *ICCE*, 2015, pp. 583–584.

[5] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth," in *EUSIPCO*, 2012, pp. 1975–1979.

[6] Y.-T. Li and J. P. Wachs, "Hierarchical elastic graph matching for hand gesture recognition," in *ICPR*. Springer, 2012, pp. 308–315.

[7] D. Kim, J. Song, and D. Kim, "Simultaneous Gesture Segmentation and Recognition Based on Forward Spotting Accumlative HMMs," *Journal of Pattern Recognition Society*, vol. 40, pp. 1–4, 2007.

[8] T.-K. Kim and R. Cipolla, "Canonical Correlation Analysis of Video Volume Tensors for Action Categorization and Detection," in *TPAMI*, 2009, pp. 1415–1428.

[9] I. Bayer and T. Silbermann, "A multi modal approach to gesture recognition from audio and video data," *ICMI*, pp. 461 – 466, 2013.

[10] X. Chen and M. Koskela, "Online rgb-d gesture recognition with extreme learning machines," in *ICMI*, 2013, pp. 467–474.

[11] A. El-Sawah, C. Joslin, and N. Georganas, "A dynamic gesture interface for virtual environments based on hidden markov models," in *IEEE International Workshops on HAVE*, 2005, pp. 109–114.

[12] S. Escalera, J. Gonzalez, X. Baro, M. Reyes, O. Lopes, I. Guyon, V. Athitsos, and H. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," *ACM on ICMI*, pp. 445 – 452, 2013.

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

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

[15] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[16] "http://www.microsoft.com/en-us/kinectforwindows."

[17] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *CVPR*, vol. 2, 1999, pp. 246–252.

[18] H.-G. Doan, H. Vu, T.-H. Tran, and E. Castelli, "Improvements of rgb-d hand posture recognition using an user-guide scheme," in *CIS-RAM*, 2015, pp. 24–29.

[19] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, 1981, pp. 674–679.

[20] 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 *RIVF*, 2016, pp. 163–168.

[21] K. McGuinness and N. E. O Connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognition*, vol. 43, no. 2, pp. 434–444, Feb. 2010.

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

[23] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," in *CVPR*, 2016, pp. 4207–4215.

[24] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *CVPR*, 2013, pp. 716–723.

[25] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *CVPR*, 2014, pp. 804–811.

[26] H. G. Doan, H. Vu, and T. H. Tran, "Recognition of hand gestures from cyclic hand movements using spatial-temporal features," in *SoICT*. ACM, 2015, pp. 260–267.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.