Dynamic Hand Gesture Recognition using Cyclical Patterns of Hand Movement and Its Applications

Huong-Giang Doan[1,2], Hai Vu[1], Thanh-Hai Tran[1]*
[1]*Hanoi University of Science and Technology - No. 1, Dai Co Viet Str., Hai Ba Trung, Ha Noi, Viet Nam*
[2]*Industrial Vocational College Ha Noi, VietNam*
Received: March 29, 2017; accepted: June 9, 2017

Abstract

This paper tackles a new prototype of dynamic hand gestures and its advantages to apply to controlling smart home appliances. The proposed gestures convey cyclical patterns of hand shapes as well as hand movements. Thanks to the periodicity of defined gestures, on one hand, common technical issues that 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 a robust recognition scheme. To this end, we propose a novel hand representation in a temporal-spatial space. Particularly, the phase continuity of the gesture's trajectory is taken into account underlying the temporal-spatial space. This scheme obtains very promising results with the best accuracy rate is 96%. The proposed techniques are deployed to control home appliances such as lamps, fans. These systems have been evaluated in both lab-based environment and real exhibitions. In the future, the proposed method will be evaluated in term of the naturalness of end-users and/or robustness of the systems.

Keywords: Human Computer Interaction, Dynamic Hand Gesture Recognition, Spatial-Temporal Features

1. 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 to control home appliance devices such as bulbs/lamps, fans.

In relevant works, the performance of dynamic hand gestures recognition strongly depends on the type of dataset used. 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 styles comparing with common ones.

In the literature, many relevant works in the [9], [10], [11], [12] have deployed real practical applications using dynamic hand gestures. Such system faces many technical issues such as real-time requirement and complex movement of hands, arms, face, and body. In this study, 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 these, 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 key-points tracked using KLT (Kanade-Lucas-Tomasi) technique [14]. We then deploy an interpolation scheme on each dimension to reconstruct the phase-normalized image sequence. This interpolation 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.

*Corresponding author: Tel.: (+84) 97.656.0526
Email: thanh-hai.tran@mica.edu.vn

2. Proposed system

In this section, we present how the specific characteristics of the proposed hand gesture set will be utilized for solving the critical issues of an HCI application (e.g., in this study, it is a lighting control system). Fig. 1 shows the proposed framework. There are four main blocks: three first blocks compose steps for extracting and spotting a hand region from an image sequence; two next blocks present our proposed recognition scheme which consists of two phases: dynamic hand gesture representation and recognition. Once dynamic hand gesture is recognized, lighting control is a straightforward implementation.

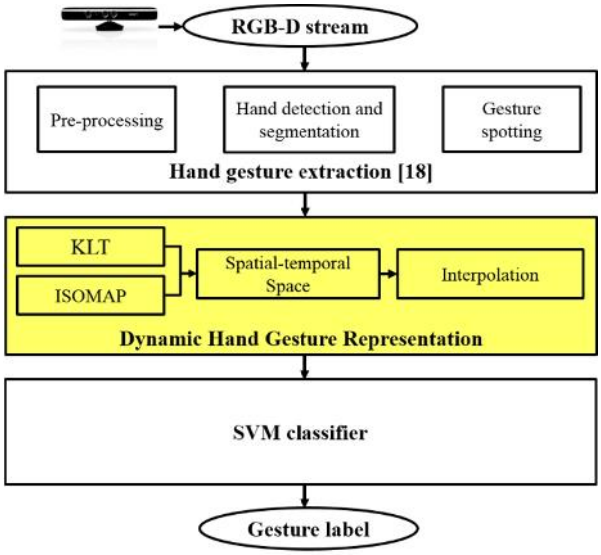

Fig. 1. The proposed framework for the dynamic hand gesture recognition.

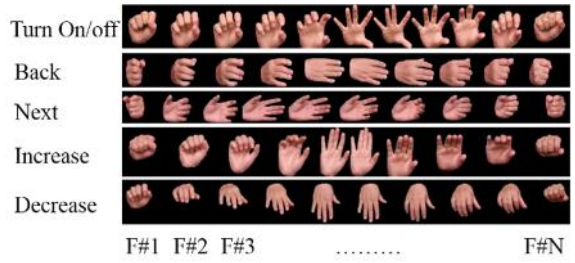

Fig. 2. In each row, changes of the hand shape during a gesture performing. From left-to-right, hand-shapes of the completed gesture chance in a cyclical pattern (closed-opened-closed).

2.1. Designing a 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 predefined trajectory. Simultaneously, while moving the hand, the hand shape will be changed following three states. There changes are underlying a cyclic pattern/closed-form in which the hand shape is closed at initial state then opened at the middle state, and closed again in the state, as shown in Fig. 2. In this study, we design five commands which are the most commonly used to control home appliances: Turn on/ off; Next; Back; Increase; Decrease. Although the number of commands is quite limited, there is no limitation to design new gestures based on the same concepts. The proposed gestures are discriminated from existing ones in both characteristics: hand shape and direction of hand movement. Hand shapes represent a cyclical pattern of a gesture, whereas hand movements represent the meaning of a gesture. Before spotting a hand gesture, we implemented some pre-processing procedures such as depth and RGB calibration (Fig. 3(b)), human body detection (Fig. 3(c)), hand detection using Gaussian Mixture Model (GMM) [17] (Fig. 3(d)), skin color pruning for hand region segmentation. Details of these techniques were presented in our previous work [18].

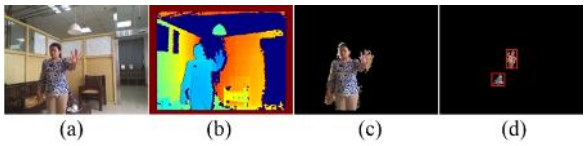

Fig. 3. Hand detection and segmentation procedures. (a) RGB image; (b) Depth image; (c) Extracted human body; (d) Hand candidates.

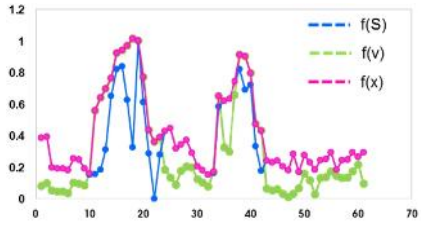

Fig. 4. Representation of signal $f(S), f(v), a \quad f(x)$.

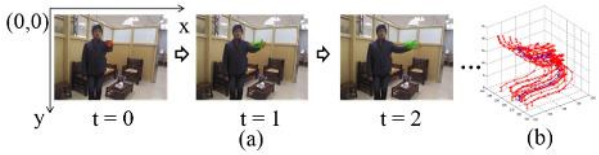

Fig. 5. An example of KLT-based trajectory. (a) Optical flow extracted from consecutive frames; (b)

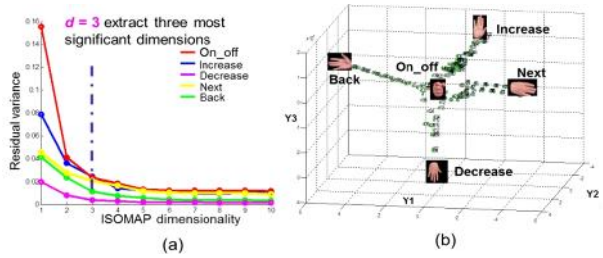

Fig. 6. The spatial feature extraction ISOMAP-based. (a) Residual representation; (b) Three most significant dimensions of ISOMAP.

The estimated trajectory of the gesture, and ending times of a hand gesture before recognizing it. In this study, we rely on the cyclical pattern of hand gestures for gesture spotting implementations that are combined between the area convolution of hand region as presented in our previous work [26] $f(S)$ and velocity of hand movement $f(v)$. Which is f(x) as the following (1):

$$f(x) = (\|f(S)\|) \cup (\|f(v)\|) \quad (1)$$

In Fig. 2, the blue curve illustrates area convolution of hand regions, the green curve illustrates velocity of hand movement and the pink curve is a combination of these signals. The pre-defined gestures consist of the identical hand shapes and hand movements at starting and ending times. We then applied method as presented in [26] to search two consecutive local minimums values on correspond to the closed form of hand shapes from the $f(x)$ signal. Once the starting and ending times of a gesture are determined by these local minimums. We will annotate them and store in the database for further processing.

2.3. Robust dynamic hand gesture recognition

Spatial-temporal feature extraction for gesture representation: Given a sequence consisting of L frames of a spotted gesture, we extract spatial and temporal features of every frame then concatenate them to build the final representation of the gesture. The spatial features are computed through manifold learning technique ISOMAP [13] by taking the three most representative components of this manifold space as shown in Fig. 6. The temporal features 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. 5(a-b)). Fig. 7 illustrates a representation in 3-D space of five different hand gestures. As shown, the separations of five gestures are very discriminative. It expresses inter-class variances when the whole dataset is projected in the proposed space [20].

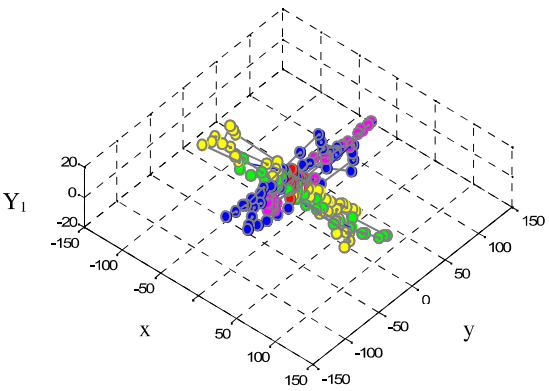

Fig. 7. Distribution of gestures in the low-dimension

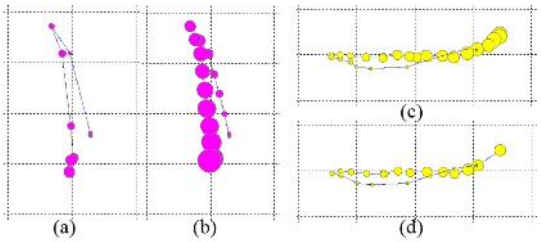

Fig. 8. Interpolation of dynamic hand gestures. a) Original gesture Decrease (9 frames); c) Original gesture Back (30 frames); b, d) corresponding interpolated hand gestures (20 frames).

Phase normalization based on interpolation: By utilizing the spatial-temporal space, the comparison between two gestures could be straightforward implementation by using DTW (Dynamic Time Warping) algorithms. However, DTW techniques discard inter-period phase. In other words, due to locally comparing hand shapes of two gestures (e.g., one from a gallery, one is probe gesture), the inter-period phase is ignored. Thanks to a periodic pattern of the image sequence, we deploy an interpolation scheme so that hand gesture sequences have the same length, and therefore maximize inter-period phase continuity. The proposed scheme is based on piecewise interpolation and similarity measurement between two adjacent points in the proposed spatial-temporal space. Details of this techniques were presented in our previous works [20]. Fig. 8 presents some results of the interpolation procedure so that length of interpolated sequence is equal to a predetermined value M. (For instance, M is set to 20 frames). The frame numbers of a gesture in Fig. 8 (a) equals to 10. Fig. 8 (c) consists of 28 frames. Fig. 8 (b), (d) are two interpolated gestures after applying the interpolation procedure. In [20], we adjusted M and obtain the recognition accuracy rates at M equals 18 with our datasets. After applying phase normalizing scheme, all dynamic hand gestures are represented by feature vectors of the same length. Gesture recognition is performed using a SVM

classifier [15]. The input of this classifier is the feature vectors extracted from interpolated sequences.

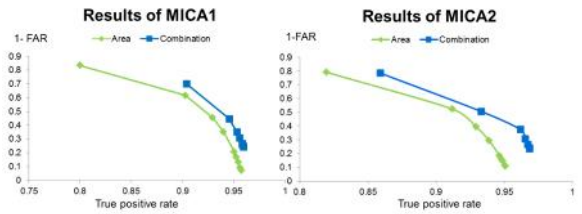

Fig. 9. Performances of the dynamic gesture spotting on two datasets MICA1 and MICA2.

3. Experimental results

3.1. Evaluating performance of the gesture spotting algorithm

We evaluate the gesture spotting technique on our two datasets MICA1 (16 videos of 16 subjects) and MICA2 (33 videos 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 subject. 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. 9 illustrates quantitative spotting results in term of true positive rate and false alarm rate with θ varying from 0.1 to 0.9 with the area convolution of the area and the combination between area signal and velocity of hand movement. When θ increases, the true positive rate slightly reduces from 0.96 to 0.8 with area signal, 0.9 to 0.96 with the combination (on the MICA1 dataset) or from 0.95 to 0.82 with area signal, 0.86 to 0.97 with the combination (on the MICA2 dataset). That shows our algorithm performs more effective with this combination of both our two datasets. However, the false alarm rate increases significantly from 0.21 to 0.76 (on the MICA1 dataset) or 0.23 to 0.79 (on the MICA2 dataset). We propose to choose $\theta = 0.75$ that gives the best trade-off between the true positive rate and false alarm rate for testing the whole system of recognition.

3.2. Evaluating performances of the representation spaces

We evaluate the gesture spotting technique on our datasets with different feature representations which are spatial, temporal and the combination of them. The evaluation results obtain the accuracy rate as shown in Tab. 1. A new representation space is the highest recognition result at 96.5%.

Table 1. The assessments of end-users on the proposed system.

ISOMAP	KLT	ISOMAP+KLT
59.02 ± 3.16	90.63 ± 0.94	96.5 ± 1.58

Table 2. Performance of the proposed method on three datasets.

Dataset	Precision (%)	Recall (%)
MSRGesture3D	94.5±3.1	92.03±5.1
R3DCNN subset	91.0±4.7	87.5±4.2
MICA3	**96.1±3.2**	**96.9±2.1**

Then, this combination is evaluated on four datasets and the results are compared with another method [26]. Fig. 10 shows that the proposed method is more effective than our previous method [26].

3.3. Evaluating performances of the recognition scheme

The proposed method is evaluated on three different datasets, in which consisting of two benchmark datasets: MSRGesture3D [22]; and a subset of R3DCNN dataset [23]. In our previous work [20], we evaluated on two our datasets which obtain the accuracy rate at 97.95±3.09% with MICA1 dataset and 94.95±4.65% with MICA2 dataset. Moreover, these datasets only captured at a fix position of end-users. To clearly confirm affects of the cyclical movements, we construct the third one, named MICA3. MICA3 dataset is constructed following setups: volunteers (4 males and 4 females) are invited to perform three times five pre-defined gestures at 13 positions in a lab-experimental room (As shown in Fig. 10, the various positions on the floor are marked). Therefore, each position consists of 120 dynamic hand gestures. 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 evaluation, based on the confusion matrix, precision and recall indexes are averagely calculated. The evaluation results are shown in Tab. 2. Although types of gestures are varying from three datasets, the cyclical gestures appear often in such datasets. Lowest performances are archived with R3DCNN, while highest performances are archived with MICA3. Comparing with recent works, for MSRGesture3D dataset, the sensitivity of state-of-the-art method achieved ups to 92.45% in [24]. With recall rate of 92.03%, the result of the proposed method is obviously comparable. For the second dataset, the recall rate achieved far from that was reported in [23] (83.6% for depth data). With the third dataset, this is more challenging because the proposed method is evaluated from various

positions/orientations from a subject to Kinect sensor, but the highest performances are achieved.

3.4. Impacts of the proposed phase normalization scheme

Using MICA3 Dataset, we evaluate the performances at different 13 positions with 3 recognition schemes: DTW-based in [26]; a CNN (Convolutional Neuron Networks) features combining SVM [27] and the proposed method. While DTW aligns locally a pair of hand shape alignment, CNN is a must-to-try machine learning technique. The proposed method dedicates to resolve phase-alignment for cyclical movements. The comparison results are shown in Fig. 11. Obviously, the proposed method is over-performed others at various positions, particularly, the proposed method significantly outperforms the DTW-based techniques. 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.

3.5. Deployment in a practical application of lamp controlling

We have deployed the proposed techniques for controlling bulb/lamp. The proposed system is tested

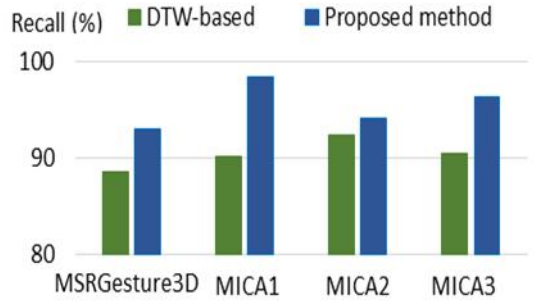

Fig. 10. Comparison results between the proposed method vs. other method

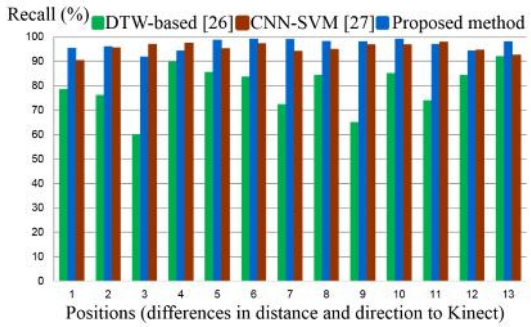

Fig. 11. Comparison results between the proposed method vs. others at thirteen positions.

Fig. 12. Illustration of a user controlling lamps using hand gestures.

Table 3. The assessments of end-users on the proposed system.

Survey	Lab-based	Real exhibition (MICA2)	Lab-based (MICA3)
Subject	16	35	8
Age	20 to 38	8 to 69	20 to 40
Male/Female	10/6	27/8	4/4
User's Feedback			
Natural	16	35	8
Memorial	15	35	8

with a number of end-users in the lab-based environment and technical exhibitions (Vietnam Techmart 2015). In Fig. 12 shows a demonstration of the system in the lab-based environment. In these evaluations, besides measuring the system's performance, we also asked end-users to answer some questions concerning the naturality and the memorability of the designed gesture dataset. The main purpose of this survey is to initially hear end user's feedback about the proposed system. As shown in Tab. 3, the user's feedback confirmed high usability and a promising technology. The participants expressed their strong interest in using hand gestures to control devices. This shows a big potential and feasible techniques to deploy real applications.

4. Conclusion

This paper described a new type of dynamic hand gestures and the robust recognition techniques. We focused on utilizing the cyclical pattern characteristics of the proposed hand gestures to solve critical issues when deploying a real application. While hand-shapes form a solution to spot a dynamic gesture, both hand-shapes and hand-movement are utilized to extract spatial and temporal features to deploy the recognition scheme. Particularly, we took into account normalizing length of the hand gestures via interpolation schemes. The proposed technique ensures that the inter-phase continuity of the gestures

is maximized. The experimental results confirmed that proposed techniques achieved higher performances comparing with conventional methods on public datasets. Moreover, deploying the proposed techniques is for controlling some home appliances are demonstrated. Initial evaluations of end-user shown a feasible and a natural way of human-computer interaction to control home appliances.

Acknowledgments

This research is funded by Hanoi University of Science Technology under grant number T2016-PC-189.

References

1. S. Marcel, O.Bernier, J.-E. Viallet, and D. Collobert, Hand gesture recognition using input-output hidden markov models, 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, 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, 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, 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, EUSIPCO, 2012, pp. 1975– 1979.
6. Y.-T. Li, and J. P. Wachs, Hierarchical elastic graph matching for hand gesture recognition, ICPR., 2012, pp. 308–315.
7. D. Kim, and J. Song, 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, 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, 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, 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, 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, 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, CVPR, vol. 2, 1999, pp. 246–252.
18. H.-G. Doan, H. Vu, T.-H. Tran, and E. Castelli, A combination of user-guide scheme and kernel descriptor on rgb-d data for robust and realtime hand posture recognition, EAAI, vol. 49, pp. 103–113, Mar. 2016.
19. B. D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, 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, 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/
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, CVPR, 2016, pp. 4207–4215.
24. O. Oreifej, and Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, CVPR, 2013, pp. 716–723.
25. X. Yang, and Y. Tian, Super normal vector for activity recognition using depth sequences, 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, SoICT, 2015, pp. 260–267.
27. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning spatial-temporal features with 3d convolutional networks, ICCV, 2015.