I.J. Image, Graphics and Signal Processing, 2018, 9, 17-26
Published Online September 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2018.09.03

Spatial-Temporal Shape and Motion Features for Dynamic Hand Gesture Recognition in Depth Video

Vo Hoai Viet
University of Science, Ho Chi Minh City, 700000, Viet Nam
Email: vhviet@fit.hcmus.edu.vn

Nguyen Thanh Thien Phuc
University of Science, Ho Chi Minh City, 700000, Viet Nam
Email: ntthienphuc6195@gmail.com

Pham Minh Hoang
University of Science, Ho Chi Minh City, 700000, Viet Nam
Email: pmhoang@fit.hcmus.edu.vn

Liu Kim Nghia
University of Science, Ho Chi Minh City, 700000, Viet Nam
Email: liukimnghia@gmail.com

Received: 02 March 2018; Accepted: 22 June 2018; Published: 08 September 2018

Abstract—Human-Computer Interaction (HCI) is one of the most interesting and challenging research topics in computer vision community. Among different HCI methods, hand gesture is the natural way of human-computer interaction and is focused on by many researchers. It allows the human to use their hand movements to interact with machine easily and conveniently. With the birth of depth sensors, many new techniques have been developed and gained a lot of achievements. In this work, we propose a set of features extracted from depth maps for dynamic hand gesture recognition. We extract HOG2 for shape and appearance of hand in gesture representation. Moreover, to capture the movement of the hands, we propose a new feature named HOF2, which is extracted based on optical flow algorithm. These spatial-temporal descriptors are easy to comprehend and implement but perform very well in multi-class classification. They also have a low computational cost, so it is suitable for real-time recognition systems. Furthermore, we applied Robust PCA to reduce feature's dimension to build robust and compact gesture descriptors. The robust results are evaluated by cross-validation scheme using a SVM classifier, which shows good outcome on challenging MSR Hand Gestures Dataset and VIVA Challenge Dataset with 95.51% and 55.95% in accuracy, respectively.

***Index Terms*—**HCI, dynamic hand gesture, depth sequences, HOG2, HOF2

I. Introduction

Human-Computer Interaction (HCI) has become one of the most researched areas recent years. HCI plays an important role in many interactive systems, which human being can interact with machines by using hardware devices, gestures or voice.

In early studies, the hardware-based methods were considered for gesture recognition approach due to accurate hand position deliverability with high speed and hand model reconstruction in the 3D coordinate system. The devices like data gloves, leap-motion [1, 2, 3] can collect directly hand gesture data such as shape, movement, position and curve of finger joints. However, users are forced to wear expensive devices to perform gestures and it requires high techniques to handle. This makes it difficult to apply dynamic hand gesture in real applications. On the other hand, the vision-based methods have achieved popularity in many fields. It helps human to be able to interact with computers by natural gestures without extra device requirements. The vision-based approaches can be divided into two types: static-gesture based methods and dynamic-gesture based methods. Static hand gesture refers to single image of a hand posture which corresponds to a command. This kind of gesture is simple in performance and costs less computational power to extract feature. Otherwise, the dynamic hand gesture is a sequence of hand postures, happening within a certain time and performing a command. the dynamic hand gesture is more complex, more computation power requiring and confuse in some cases. Besides, some of the dynamic hand gestures express different meanings but have nearly same representation. The meaning of dynamic hand gesture usually depends on the culture, duration of performance and single handed or double handed representation. In this study, we focus on examining the dynamic hand gesture recognition.

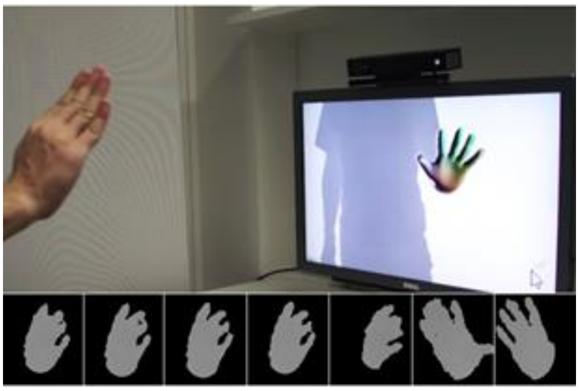


Fig.1. Some depth frames are captured by Microsoft Kinect

There are two main ways to collect hand information for hand gesture recognition: using RGB camera and depth camera (depth sensor). Color image captured by RGB camera can express completely and clearly the shape of hand but too sensitive to illumination, occlusion, heterogenous background and 3D information loss. Having too many colors in background is also a drawback in hand segmentation section. On the other hand, depth sensors, having become cheaper recent years (as Microsoft Kinect), have been used in many efficient recognizing techniques. Captured by depth sensor, depth image is strong in against illumination and occlusion problems and easy in 3D hand gesture modelling, although the quality of depth image is influenced by a variety of noises, quality of the depth sensor or capturing distance limitation.

In this paper, we develop a real-time, vision-based dynamic hand gesture recognition framework, using depth maps with robust performance. We mainly consider developing spatial-temporal features, which represent completely changes of shape and motion flow of hand gesture on each frame and on the whole video. The features after being extracted are reduced dimensions by Robust PCA and then are classified by a linear SVM model. Feature standardization technique is applied to increase the speed of convergence and recognizing rate of the system.

The remainder of the paper is organized as follows. In section II, we present related work. Our proposed approach is explained in session III. Then, the experimental results and discussion are shown in Section IV. Finally, we draw conclusions.

II. Related Work

In static hand gesture recognition, just one single image contains whole information of shape and appearance of a hand gesture. Most methods in this case introduced features which represent the geometric properties of hand such as hand palm, center point of the convex hull, number of fingers or hand contour. Many approaches used familiar descriptors to represent hand gesture like the methods using SIFT proposed by M. Hamissi et al [6, 8], RANSAC proposed by Z. Zhang et al [5] and Z. Ren, J. Meng et al [13], Super-pixel for hand shape representation proposed by C. Wang et al [4]. Some new algorithms introduced new efficient ways to extract hand characteristics. The method proposed by F. Dominio et al [7] divided detected hand region into many parts based on an estimated plane and Principal Component Analysis (PCA) output, then extracted features from these split hand parts. Meanwhile, C. Chan and S. Mirfakharaei et al [9] introduced Fourier Shape Descriptor to extract hand feature. De Gu [11] proposed a new method for finger tracking in hand gesture recognition. An approach proposed by Yi Li et al [10] provided hand convex hull detection using Graham Scan and S. Poularakis et al [12] introduced a method supporting fingers detection and analysis, using Minimum Enclosing Ellipsoid (MEE). These above studies focused on creating a good descriptor for shape and appearance of a hand, an important characteristic of gesture.

In dynamic hand gesture, the information of gesture is contained in a sequence of images, so besides extracting gesture characteristics in each frame, the recognition approaches must perform the changes of gesture movement. After extracting spatial descriptors in each frame, temporal properties of hand gestures are created by common methods as using hand detecting and tracking [15, 21, 22] or using Dynamic Time Warping (DTW) [14, 18], spending expensive cost to find the optimal alignment path or Temporal Pyramid [19, 20], connecting with spatial location.

The approach, proposed by H. Liang et al [15], used hand and arm data to detect hand fingers and hand palm at each frame by applying Geodesis Distance and then fingers are labeled by GPS Point to track. Using Skeleton data extracting by Microsoft Kinect, T. Osunkoya, C. Chern et al [21] tracked hand and defined gesture actions for mouse control and PowerPoint presentation. In handwriting recognition, T. Murata and J. Shin et al [22] defined "Click" and "Wave" hand gestures and recognized handwriting letters and numbers using DP Matching and Inter-stroke Information features.

Furthermore, many approaches combined different descriptors to represent hand gesture. A. Kurakin, Z. Zhang et al [16] introduced a hand shape representation forming from Cell Occupancy Descriptor and Silhouette Descriptor, then the proposed feature was reduced dimension by Principal Component Analysis (PCA) and a classifier was built by Action graph, including 5 encoding schemes: Action-Specific Viterbi Decoding (ASVD), Uni-Gram and Bi-Gram Global Viterbi Decoding (UGVD and BGVD), Uni-Gram and Bi-Gram Maximum Likelihood Decoding (UMLD and BMLD). The approach proposed by D. Smedt [18] introduced new descriptors using skeleton data as Shape of Connected Joints (SoCJ), Histogram of Hand Direction (HoHD) and Histogram of Wrist Rotation (HoWR). Also based on skeleton data, H. Takimoto, J. Lee, A. Kanagawa et al [17] proposed Hand Movement Descriptor and SURF to detect hand shape and applied Hidden Markov Model (HMM) to classify. Using depth maps, Diego G. Santos et al [19] proposed CIPBR algorithm and a hybrid classifier - HAGR-D model, which composed of DTW and HMM. After detecting and tracking hand position, Ryan et al [20] applied K-Curvature Algorithm to recognize finger, then the feature was classified by DTW. Eshed Ohn-Bar, Mohan M. Trivedi et al [23] introduced new spatial-temporal descriptors such as Angular Skeletal Feature, JAS-Joint Angles Pairwise Feature and HOG2 by using skeleton data and depth maps, eventually the proposed features were evaluated in a bag of words scheme, using a linear SVM. To develop automatic driven car system, Eshed Ohn-Bar, Mohan M. Trivedi et al [24] proposed a new in-car hand gesture recognition framework, using HOG2 feature extracted from RGBD data. The proposed method was examined on MSR Hand Gesture 3D Dataset and VIVA Challenge Dataset in real-time recognition system and was evaluated by cross-validation scheme which is based on a linear SVM classifier.

Not only using hand-craft features, some studies applied deep learning and gained high accuracy in dynamic hand gesture recognition. P. Molchanov, S. Gupta, K. Kim and J. Kautz et al [39] proposed 3D Convolutional Neural Networks consisting of HRN (High-resolution Network) and LRN (Low-resolution Network) on VIVA Challenge Dataset, a challenge natural hand gesture and achieved significant results.

Because of the complexity and diversity of dynamic hand gesture, especially gestures representing natural actions, the accuracy of recognizing methods is in range of over 70% and requires to be improved more.

III. Proposed Methodology

We propose a dynamic hand gesture recognition framework with three main stages as Fig. 2:

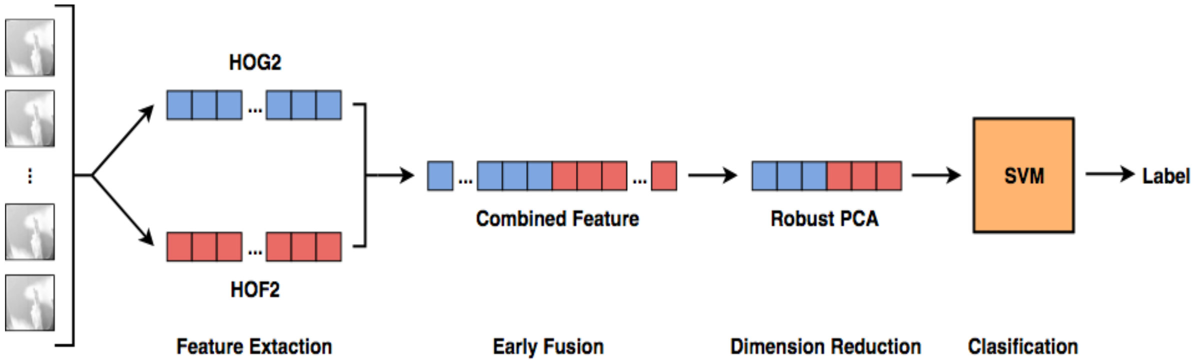


Fig.2. Our proposed framework for dynamic hand gesture recognition system

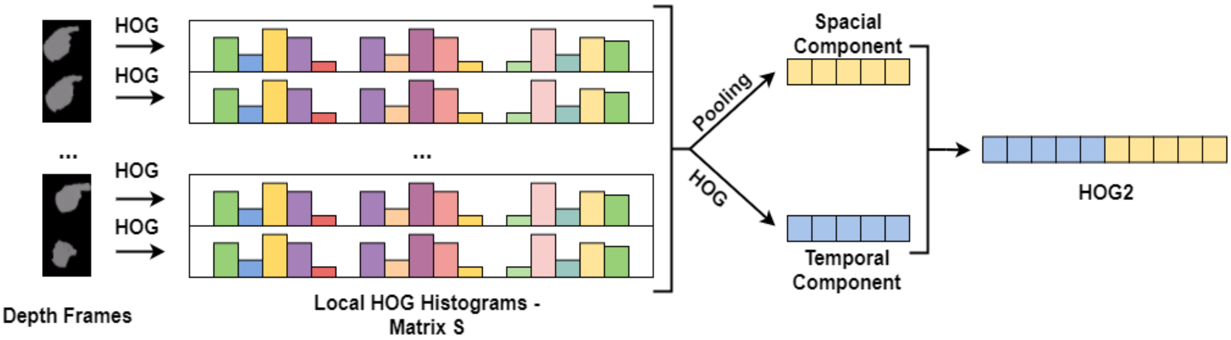


Fig.3. Flowchart of HOG2 extraction

Feature extraction: Significant characteristics of dynamic hand gesture are posture and movement. The main idea is that extracting features which represent the variation of hand shape and movement in each frame (spatial term) and whole video (temporal term). In this step, from depth map sequences, we implement new spatial-temporal descriptors as HOG2, HOF2 and a combined feature. During extracting feature process, we apply one of two different kinds of pooling techniques, max pooling and average pooling, to obtain a robust and compact gesture representation.

Feature dimensional reduction: To avoid the worst case when extracted feature collection is affected by large and sparse noise, we apply Robust PCA to reduce the dimension of the feature. After that, the feature values are normalized by feature standardization.

Multiclass classification: we use a linear SVM model introduced by Singer and Crammer [28, 29] and validate the model using leave-one-subject-out cross-validation scheme.

A. *Shape and Appearance Feature*

The characteristic which is often used in hand gesture recognition is hand shape and appearance. In this work, we extract HOG2 Feature [23] to represent the changes of hand shape and appearance. The HOG2 feature is formed from two components in spatial and temporal term (shown in Fig. 3.).

In spatial term, we extract the first component of HOG2 from each frame of the depth video.

Assume a depth video has K frames. Let $I(x,y)$ as a depth frame with size of $m\times n$. The gradient matrix G_x and G_y are calculated on $I(x,y)$ by using a $1D$ mark $[-1,0,1]$ to obtain a magnitude matrix G and a quantized orientation matrix θ. We also denote B as the number of bins of the extracted histogram.

First, we divide $I(\mathrm{x},\mathrm{y})$ into $M\times N$ blocks, with each block overlaps 50% together.

In the s^{th} divided block, with $s\in\{1,\ldots,M\times N\}$, we compute an orientation histogram h^s with B bins. The q^{th} bin of histogram h^s is denoted as:

$$h^s(q)=\sum_{x,y}G^s_{x,y}\bullet\mathbf{1}\left[\Theta^s(x,y)=\theta\right] \tag{1}$$

Where:

- $\theta\in\left\{-\pi+\frac{2\pi}{B}:\frac{2\pi}{B}:\pi\right\}$.
- $\mathbf{1}$ is indicator function.
- $q\in\{1,\ldots,B\}$.

After that, the histogram h^s is normalized by L2-norm:

$$h^s=\frac{h^s}{\sqrt{\left\|h^s\right\|_2+\varepsilon}} \tag{2}$$

By overlapping 50 percent of each block, we can collect the local information and the correlation of blocks within a frame.

Next, we create a HOG descriptor h_k for the frame k, with $k\in\{1,\ldots,K\}$, by concatenating the histograms of blocks within the frame t.

Denote S as a $2D$ matrix, S is formed by combining HOG histograms extracted from frames. The values in S correspond the hand shape in video.

On the matrix S, we apply pooling technique to compute a compact feature for the considered depth video. The pooling technique also help to avoid over-fitting problem in the recognition step. One of two kinds of pooling techniques (max pooling and average pooling are computed follow as formula (3) and (4)) is calculated to get the first spatial component h_S of the HOG2 Feature. In this work, we adopted max pooling in our experiments.

Average Pooling:

$$h_S(j)=\text{mean}\,S(j) \tag{3}$$

Max Pooling:

$$h_S(j)=\arg\max\ S(j) \tag{4}$$

Where:

- $h_S(j)$ is the j^{th} component of h_S, $j\in d$.
- $S(j)$ is the j^{th} column of S.
- d is the dimension of h_S.

Since each row in the matrix S is HOG histogram of a frame, in temporal term, we calculate the derivative along rows of matrix S to perform the change of hand appearance in the video. Therefore, we apply HOG algorithm again on the matrix S to extract the second temporal component h_T of HOG2.

$$h_T=\text{HOG}(S)=\text{HOG}\left(\begin{bmatrix}h_1\\ \vdots\\ h_K\end{bmatrix}\right) \tag{5}$$

The final HOG2 Feature is created by combining h_S and h_T together. Eventually, HOG2 feature is normalized by L2-norm, L1-sqrt or L2-Hys [26].

$$HOG2=\left[h_T;h_S\right] \tag{6}$$

The final feature is named HOG2 feature because HOG algorithm is applied twice. In our case, the size of HOG block (M and N) and B bins of local histogram are fixed in two times we apply HOG. Thus, the size of HOG2 feature is $1\times(2\times M\times N\times B)$.

B. Movement Feature

Since hand motion is an important information of dynamic gesture recognition, we introduce a feature based on computing Histogram of Optical Flow (HOF) to represent the changes of hand motion flow. Like HOG2, our new movement feature, known as HOF2, also has two components in the spatial and temporal term (illustrated in Fig. 4.).

Assume a depth video has K frames. Let $I(x,y)$ as a depth frame with size of $m\times n$.

On two consecutive frames, we apply F ärneback Dense Optical Flow Estimation Algorithm [27] to extract an Optical Flow Image $I_{OF}(x,y)$.

$$I_{OF}(x,y)=\text{OF}\left(I_{k'-1}(x,y),I_{k'}(x,y)\right) \tag{7}$$

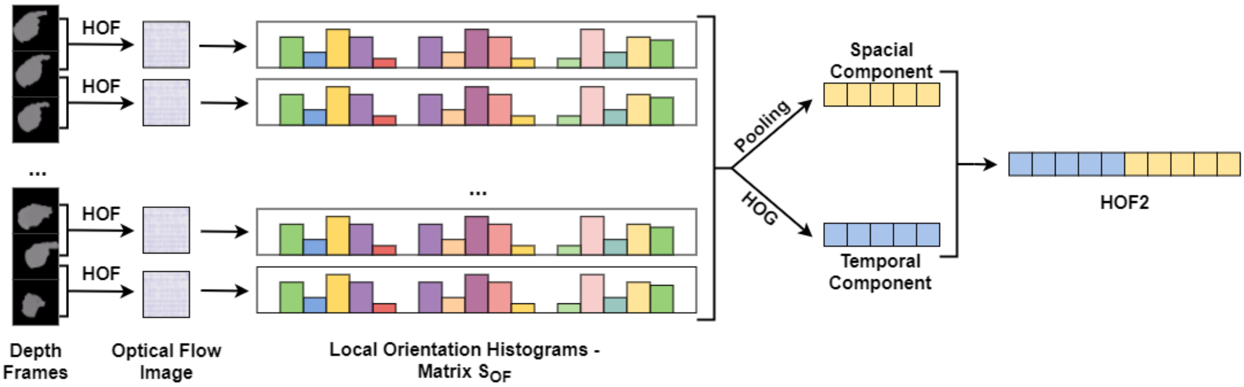


Fig.4. Flowchart of HOF2 extraction

Where:

- $I_{OF}(x, y)$ is Optical Flow Image.
- $k' \in \{2, \ldots, K\}$.
- OF is the Optical Flow Estimation Method.

After that, $I_{OF}(x, y)$ is divided into $M \times N$ blocks with each block overlaps 50% each other. In a local block, a magnitude matrix G^s and an orientation matrix θ^s with $s \in \{1, \ldots, M \times N\}$ are calculated to build a B-bin Local Orientation Histogram h_{OF}^s . Next, we create Optical Flow Histogram h_k for a frame k (with $k \in \{1, \ldots, K\}$) by concatenating Local Orientation Histograms.

Then, we collect Optical Flow Histograms of video frames and create an Orientation Histogram Matrix S_{OF} . The values of S_{OF} show hand movement in a depth video.

On the matrix S_{OF} , we apply pooling techniques (mentioned in the previous part) to get the first component h_{OFS} of HOF2.

In temporal term, the HOG Algorithm (mentioned above) is applied again on S_{OF} to extract the second component h_{OFT} of HOF2.

$$h_{OFT} = \text{HOG}(S_{OF}) = \text{HOG}\left(\begin{bmatrix} h_1 \\ \vdots \\ h_{K-1} \end{bmatrix}\right) \quad (8)$$

Eventually, the final HOF2 feature is formed when we concatenate h_{OFS} and h_{OFT} . The HOF2 feature, then, is normalized by L2-norm, L1-sqrt or L2-Hys [26].

$$HOF2 = [h_{OFT}; h_{OFS}] \quad (9)$$

The HOF2 extraction process is similar the HOG2 extraction method, so the final extracted feature is named HOF2. In this case, the size of local block (M and N) and B bins of histogram feature are fixed, so the size of HOF2 is $1 \times (2 \times M \times N \times B)$.

C. Dynamic Hand Gesture Representation

In the previous parts, we introduced the ways to extract two spatial-temporal histogram features. In this part, we create a compact feature for dynamic gesture representation by early fusion two mentioned descriptors HOG2 and HOF2.

$$h_{fusion} = [h_{HOF2}; h_{HOG2}] \quad (10)$$

The order of histogram components in the fusion feature is not important.

D. Dimensional Reduction and Feature Standardization

In this step, we reduce the dimension of extracted feature by using Robust Principal Component Analysis (Robust PCA).

Although classical PCA is also used to reduce dimension, it cannot deal completely with missing data and large noise distribution (shown Fig. 5). In Fig.5-a, the samples (cross symbols) on a subspace are influenced by small Gaussian noise. The estimated component of classical PCA (arrow) is very close to the true subspace despite all samples being noisy. Meanwhile, in fig. 5-b, samples (cross symbols) on a subspace are influenced by sparse, large errors. The estimated component of classical PCA (arrow) is quite far away from true subspace when even almost all samples are affected by noise.

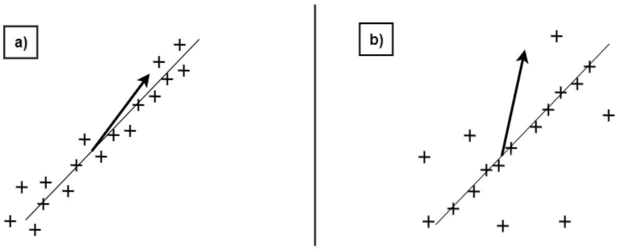


Fig.5. Illustration of subspace of noise data: a) subspace in Gaussian noise, b) subspace in large noise

After feature extraction, we are not sure whether the extracted feature matrix is affected by large noise or not. Therefore, Robust PCA is the best solution which proposes an optimization solution by minimizing rank of true value matrix of extracted feature matrix as below:

$$\min_{X,E} rank\left(X\right)+\gamma\left\|E\right\|_0 \text{ subject to } D=X+E \qquad (11)$$

The equation (11) can be solved by Convex Optimization [25, 32].

Next, feature standardization is applied to obtain a more robust feature matrix before feeding the classifier.

$$\mathrm{X}=\frac{x-\overline{x}}{\sigma} \qquad (12)$$

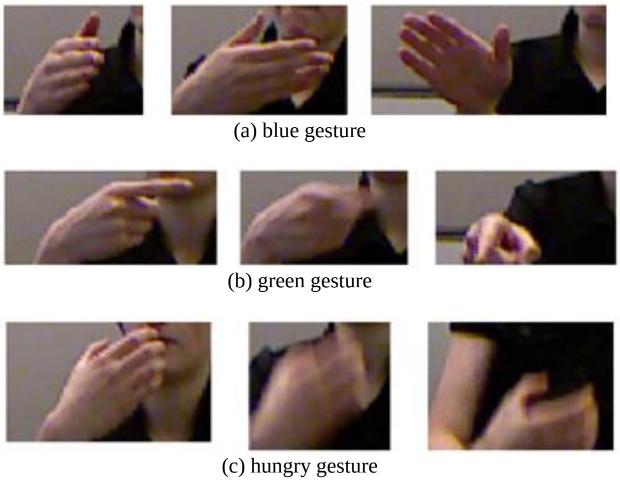
Where:

- X is the normalized value.
- x is the original value.
- $\overline{x}$ is the mean value.
- σ is the standard deviation.

E. Classification by Support Vector Machine

Support Vector Machine (SVM) [28,29] is originally designed for binary classification. To extend SVM for multiclass classification, there are two ways: one way is that constructing and combining several binary classifiers and another way is that considering all data in an optimization function. The methods extending from binary SVM for multiclass classification meet the drawback – the size of the quadratic optimization is proportional to the number of categories. The consequence in this case is that the quadratic problem is hard to solve and difficult to store.

Therefore, Crammer and Singer [28] proposed a linear SVM model, using one optimization solution for multiclass classification. Crammer and Singer's SVM model is a version of SVM model proposed by Weston, Watkins, and Vapnik. The method gives a solution of following optimized problem [29]:



(a) blue gesture

(b) green gesture

(c) hungry gesture

Fig.6. Some samples in MSR Hand Gesture 3D Dataset

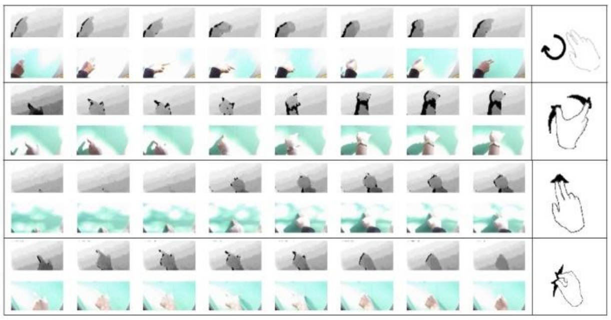


Fig.7. Some samples in VIVA Challenge 3D Dataset

$$\min_{f_1,\ldots,f_N\in\mathcal{H},\zeta\in\Re^{\ell}}\left(\sum_{i=1}^{N}\left\|f_i\right\|_K^2+C\sum_{i=1}^{\ell}\sum_{j\neq y_i}\zeta_i\right)$$
$$\text{subject to: } \begin{array}{c} f_{y_i}\left(x_i\right)\geq f_j\left(x_j\right)+1-\zeta_i \\ \zeta_i\geq 0 \end{array} \qquad (13)$$

To solve formula (13), the Lagrange dual function is solved by a complex but effective dual decomposition algorithm. In this paper, we apply Crammer and Singer's linear SVM model, which is implemented in LibLinear library.

IV. Experiments

We evaluate our methodology on two benchmark datasets and compare our results to the state-of-the-art methods to demonstrate the feasibility of the proposed method.

A. Dataset and Experimental Settings

The MSR-Hand Gesture 3D dataset, which is introduced in [23, 24] is a depth-only dynamic hand gesture dataset. This dataset contains 12 American sign language gestures: "bath-room", "blue", "finish", "green", "hungry", "milk", "past", "pig", "store", "where", "j", "z".

Although the MSR Hand Gesture 3D Dataset is segmented very well with zero-value background but it meets noise and self-occlusion issues.

In another experiment, the VIVA Challenge 3D Dataset [24, 30] is a depth dataset captured by Microsoft Kinect for natural human activity study. This dataset contains 19 hand gesture for occupant-vehicle interaction.

The VIVA Dataset is very challenging in many aspects:

- Performance of a hand gesture can be different when it is performed by different people.
- Some video samples have too few frames so that it is hard to extract robust feature.
- Videos are influenced by illumination changes and shadow artifacts.

The results of classification are evaluated by leave-one-subject-out cross-validation approach. The dataset is split into k subsets, with $k-1$ subsets for training and the rest one for testing. The final accuracy is mean value of k evaluation times.

$$acc = \frac{1}{k}\sum_{i=1}^{k} acc_i \quad (14)$$

Where:

- acc_i is the accuracy rate on the i^{th} evaluation.

B. Experimental Results

In this work, we evaluate the performance of the proposed methods on the two challenging dynamic hand gesture datasets. Then we compare our results to the state-of-the-art methods to demonstrate the superiority of the proposed method.

We describe dynamic hand gesture as a fusion of two characteristics: i) changes in shape and appearance of hands; ii) change in movement of hands. These characteristics are extracted from HOG2 and HOF2. From experimental results in Fig. 7. and 8, we argue that no one single category of feature can deal with all types of dynamic hand gesture datasets equally well. So, it is quite necessary and useful to combine different categories of features to improve the dynamic hand gesture recognition performance.

We also study the impact of parameters and algorithm in our framework on the performance of the system as follows:

- The parameters of HOG2 and HOF2 features.
- The dimensional reduction using RobustPCA.
- Feature standardization.

We evaluate the performance of HOG2, HOF2, and the fusion feature (see Fig. 8 and 9).

For HOG2 descriptor, the accuracy depends on the size of block (M and N) and B bins of the histogram while for HOF2 descriptor, the accuracy also depends on Optical flow estimation parameters.

The best result of our proposed feature - HOF2 is achieved when we choose $poly\_n = 9$, $poly\_signma = 1.1$, $flags = 0$ for Optical Flow Estimation.

On MSR Hand Gesture 3D Dataset, the best result gained with HOG2 (93.61%) if $M = N = 8, B = 8$; HOF2 (90.39%) if $M = N = 8, B = 8$ or $B = 18$; the fusion feature (95.51%) if $M = N = 8, B = 8$.

On CVRR-HANDS 3D Dataset, the best result is obtained with HOG2 (53.11%) if $M = N = 4, B = 18$; HOF2 (28.96%) if $M = N = 4, B = 8$; the fusion feature if $M = N = 4, B = 18$.

When applying Robust PCA, we obtain two matrices: low-rank matrix (the true value matrix) and sparse matrix. The results gained by low-rank matrix are better than the results from the sparse matrix and quite similar or better than results which are obtained by not using Robust PCA (see Fig. 10 and 11).

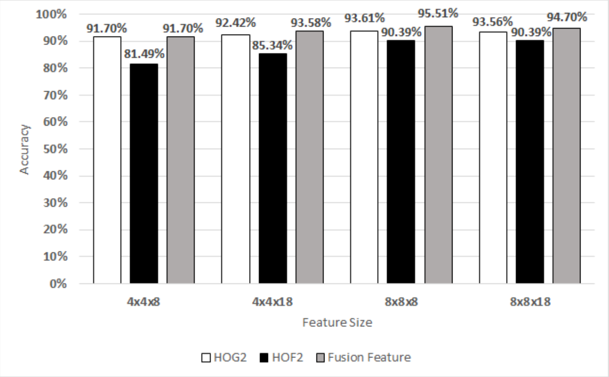


Fig.8. The experimental results on MSR Hand Gesture 3D Dataset

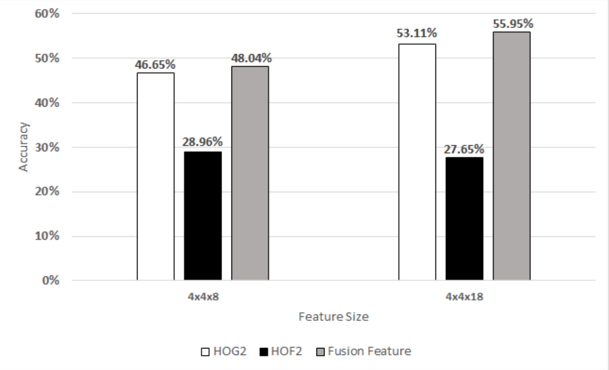


Fig.9. The experimental results on VIVA Challenge Dataset

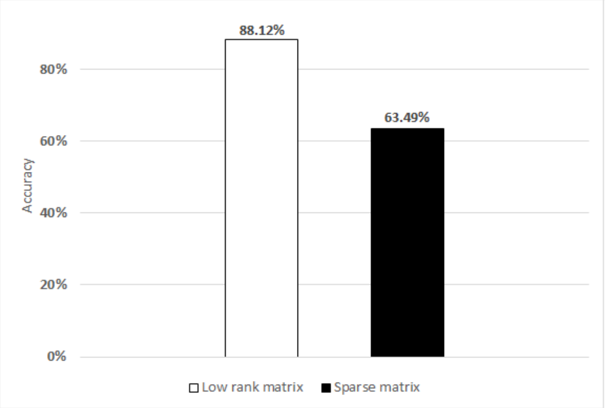


Fig.10. The experimental results of Robust PCA on MSR Hand Gesture3D dataset

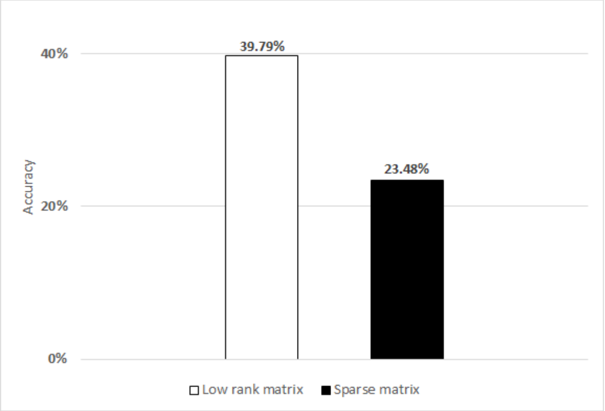


Fig.11. The experimental results of Robust PCA on VIVA dataset

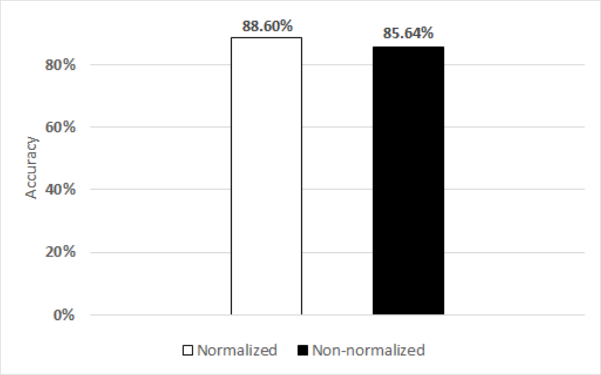


Fig.12. The experimental results of feature standardization on MSR Hand Gesture3D Dataset

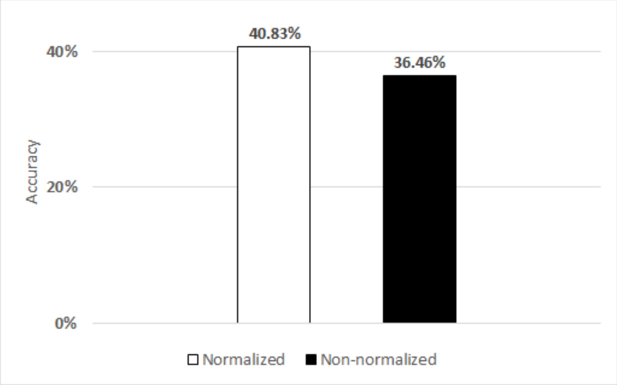


Fig.13. The experimental results of feature standardization on VIVA Challenge Dataset

From experimental results, the performance obtained by applying feature standardization in final step is more robust than results gained by not applying (shown in Fig. 12 and 13).

We also compare our experimental results with state-of-the-art results on MSR Hand Gesture3D and VIVA dataset in tables 1 and 2, respectively.

Table 1. Compare with the state of the art method on MSR Hand Gesture3D dataset

Method	Accuracy (%)
HOF2	**90.39**
Ohn-Bar [23]	92.64
HOG2	**93.17**
Our Fusion feature from HOF2-HOG2	**95.51**

Table 2. Compare with the state of the art method on VIVA dataset

Method	Accuracy (%)
HOF2	**28.96**
Ohn-Bar [23]	46.9
HOG2	**53.11**
Our Fusion feature from HOF2-HOG2	**55.95**
CNN [33]	77.5

On MSR Hand Gesture3D, our recognition rate is 95.51%, more than the current best rate by 2.87%. On VIVA, our recognition rate is 55.95%, more than the current best rate by 9.05% in the same methodology. We also compare with deep learning method so that we have a good overview for selecting the best method in the real-world application with particular cases. In [33], the authors try capturing shape and motion properties of multi-resolution by using CNN 3D. The approach has a complexity in the computational process and take time for training phase. Moreover, it needs a lot of labeled data that is difficult to have in practical applications. The experimental results also confirm that the shape and appearance characteristics are the most important for dynamic gesture recognition system. Besides, we also see that movement features contribute to the performance the system. This means that we can use handcraft or learning features in dynamic gesture recognition but gesture descriptors must have an ability to describe both shape and appearance and movement features in order to obtain the robust dynamic hand gesture recognition system.

V. Conclusion

In this paper, we have presented an efficient method for dynamic hand gesture recognition system. These key problems of this work can be summarized as: i) the changes in shape and appearance of hand by time; ii) the change in movement of hand by time. For the first problem, we used HOG2 that represented fully shape and appearance characteristics of dynamic hand gestures in the spatial and temporal term. To solve the second problem, we proposed a new feature is called HOF2 that is extracted from optical flow using Gunner Farneback algorithm. We adopted early fusion technique to yield the final feature vector for a dynamic gesture. And Robust PCA was used to create a robust feature for dynamic gesture representation. In classification phase, SVM method was used to identify the label gesture that video belongs to. We have done experiments on two challenging benchmark datasets such as MSR Hand Gestures and VIVA Challenge Dataset with 95.51% and 55.95% in accuracy, respectively. The experimental results also have shown that our proposed approach significantly outperforms of the state-of-the-art methods on both datasets. This demonstrates the good performance of our proposed approach. We also evaluated on the different parameters on feature extraction to show more clearly how the features impact on the performance of the system. Moreover, our approach has low computational cost and is easy to comprehend. This makes our approach suitable for real-time hand gesture recognition systems and robotics.

In the future work, we will fuse more cues to further improve the accuracy of our proposed fusion framework in feature-level as well as learning features.

References

[1] Arjunlal, and Minu Lalitha Madhavu, “A survey on hand gesture recognition and hand tracking”, *International Journal of Scientific Engineering and Applied Science*, 2016.

[2] Alexander Cardona López, “Hand Recognition Using Depth Cameras”, *TECCIENCIA*, 2015.

[3] Arpita Ray Sarkar, G. Ganyal, and S. Majumder, “Hand Gesture Recognition Systems: A survey”, *International Journal of Computer Application,* vol 71, no.15, 2013.

[4] Chong Wang, Zhong Liu, and Shing-Chow Chan, “Superpixel-Based Hand Gesture Recognition with Kinect Depth Camera”, *IEEE Transactions on Multimedia,* vol 17, pp 29 - 39, 2015.

[5] Zhou Ren, Junsong Yuan, Jingjing Meng, and Zhengyou Zhang, “Robust Part-Based Hand Gesture Recognition Using Kinect Sensor”, *IEEE Transactions on Multimedia,* vol 15, no 5, pp 1110 - 1120, 2013.

[6] Minoo Hamissi, and Karim Faez, “Real-time Hand Gesture Recognition Based on the Depth Map for Human Robot Interaction”, *International Journal of Electrical and Computer Engineering (IJECE)*, vol 3, no 6, 2013.

[7] F. Dominio, M. Donadeo, G. Marin, P. Zanuttigh, and G. M. Cortelazzo, “Hand gesture recognition with depth data”, *4th IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream,* pp 9-16, 2013.

[8] Hasan Mahmud, Md. Kamrul Hasan, and Abdullah-Al-Tariq, “Hand Gesture Recognition Using SIFT Features on Depth Image”, *The 9th International Conference on Advances in Computer-Human Interactions (ACHI),* pp 59-64, 2016.

[9] Cliff Chan, Seyed Sepehr Mirfakharaei, “Hand Gesture Recognition using Kinect”, *Boston University*, 2013.

[10] Yi Li, “Hand gesture recognition using Kinect*”, Thesis, University of Louisville*, 2012.

[11] De Gu, “Fingertip Tracking and Hand Gesture Recognition by 3D Vision”, *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 6 (6), pp 5421-5424, 2015.

[12] S. Poularakis, and I. Katsavounidis, “Fingertip Detection and Hand Posture Recognition based on Depth Information”, *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),* 2014.

[13] Zhou Ren, Jingjing Meng, and Junsong Yuan, “Depth Camera Based Hand Gesture Recognition and its Applications in Human-Computer-Interaction”, 8th International Conference on Information”, *Communications and Signal Processing (ICICS),* 2011.

[14] Zahid Halim, and Ghulam Abbas, “A Kinect-Based Sign Language Hand Gesture Recognition System for Hearing and Speech Impaired: A Pilot Study of Pakistani Sign Language”, *Assistive Technology, the Official Journal of RESNA*, vol 27 no 1, pp 34-43 2014.

[15] Hui Liang, Junsong Yuan, and Daniel Thalmann, “3D Fingertip and Palm Tracking in Depth Image Sequences”, *20th ACM International Conference on Multimedia*, pp 785-788, 2012.

[16] A. Kurakin, Z. Zhang, and Z. Liu, “A Realtime System for Dynamic Hand Gesture Recognition with a Depth Sensor”, *20th European Signal Conference (EUSIPCO),* 2012.

[17] Hironori Takimoto, Jaemin Lee, and Akihiro Kanagawa, “A Robust Gesture Recognition Using Depth Data”, *International Journal of Machine Learning and Computing*, 2013.

[18] Quentin D. Smedt, Hazem Wannous, and J. P. Vandeborre, “Skeleton-based Dynamic hand gesture recognition”, *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016.

[19] Diego G. Santos, Bruno J. T. Fernandes, Byron L. D. Bezzerra, “HAGR-D: A Novel Approach for Gesture Recognition with Depth Maps”, *SENSORS-BASEL*, vol 15, no 11, 2015.

[20] Daniel James Ryan, “Finger and gesture recognition with Microsoft Kinect”, *Thesis, University of Stavanger*, 2012.

[21] Toyin Osunkoya, John-Chern Chern, “Gesture-Based Human-Computer-Interaction Using Kinect for Windows Mouse Control and PowerPoint Presentation”, *Chicago State University*, 2013.

[22] Tomoya Murata, Jungpil Shin, “Hand Gesture and Character Recognition Based on Kinect Sensor”, *International Journal of Distributed Sensor Network*, 2014.

[23] Eshed Ohn-Bar, and Mohan M. Trivedi, “Joint Angles Similarities and HOG2 for Action Recognition”, *IEEE Conference on Computer Vision and Pattern Recognition Workshops: Human Activity Understanding from 3D Data*, 2013.

[24] Eshed Ohn-Bar, and Mohan M. Trivedi, “Hand Gesture Recognition in Real-Time for Automotive Interfaces: A Multimodal Vision-based Approach and Evaluations”, *IEEE Transactions on Intelligent Transportation Systems*, vol 15, no 6, pp 2368 - 2377 , 2014.

[25] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright, “Robust Principal Component Analysis?”, Stanford University, 2009.

[26] Navneet Dalal, Bill Triggs, “Histograms of Oriented Gradients for Human Detection”, *Computer Vision and Pattern Recognition*, pp 886-893, 2005.

[27] Gunner Fernback, “Two-Frame Motion Estimation Based on Polynomial Expansion”, *SCIA*, pp 363-370, 2003.

[28] Koby Crammer, Yoram Singer, “On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines”, *Journal of Machine Learning Research*, vol 2, pp 265-292, 2001.

[29] Chih-Wei Hsu, and Chih-Jen Lin, “A Comparison of Methods for Multi-class Support Vector Machines”, *IEEE Computational Intelligence Society*, vol 13, no 2, pp 415 – 425, 2002.

[30] CVRR-HANDS 3D Dataset: http://cvrr.ucsd.edu/vivachallenge/index.php/hands/hand-detection/

[31] MSR 3D Dataset: http://www.uow.edi.au/~wanqing/#MSRAction3DDatasets

[32] Zhouchen Lin, Arvind Ganesh, John Wright, Leqin Wu, Minming Chen, Yi Ma, “Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix”, *Intl. Workshop on Comp. Adv. in Multi-Sensor Adapt. Processing, Aruba, Dutch Antilles* , 2009.

[33] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kaut, “Hand Gesture Recognition with 3D Convolutional Neural Networks”, *In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops,* 2015.

Authors' Profiles

Vo Hoai Viet is a Lecturer and Senior Researcher at the University of Science, VNU-HCMC, Vietnam from 2012. He is currently working toward the Ph.D degree in Computer Science at University of Science, VNU-HCMC, Vietnam. His research interests include Digital Image Processing, Programming Language, Computer Graphics, Computer vision, and Machine Learning.

Nguyen Thanh Thien Phuc graduated from the University of Science, VNU-HCMC, Vietnam in in 2017. His research interests include Image Processing and Computer Vision.

Pham Minh Hoang is a Lecturer and Senior Researcher at the University of Science, VNU-HCMC, Vietnam. His research interests include Image Processing, 3D Reconstruction, and Interactive Surface.

Liu Kim Nghia graduated from the University of Science, VNU-HCMC, Vietnam in 2017. His research interests include Image Processing, and Computer Vision.

How to cite this paper: Vo Hoai Viet, Nguyen Thanh Thien Phuc, Pham Minh Hoang, Liu Kim Nghia, " Spatial-Temporal Shape and Motion Features for Dynamic Hand Gesture Recognition in Depth Video", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.10, No.9, pp. 17-26, 2018.DOI: 10.5815/ijigsp.2018.09.03