https://doi.org/10.2352/ISSN.2470-1173.2018.18.3DIPM-461

Skeleton-based Dynamic Hand Gesture Recognition using 3D Depth Data

Dan Zhao, Yue Liu, Guangchuan Li, School of Optoelectronics, Beijing Institute of Technology, Beijing, China, 100081*

Abstract

Hand gesture recognition is a crucial but challenging task in the field of Virtual Reality (VR) and Human Computer Interaction (HCI). In this paper, a skeleton-based dynamic hand gesture recognition approach is proposed, in which the skeleton structure of the hand captured by 3D depth sensor is firstly exploited and the spatiotemporal multi-fused features that concatenate four skeleton hand shape features and one hand direction feature are extracted. Then the hand shape features are encoded by Fisher Vector obtained from a Gaussian Mixture Model (GMM). To add the temporal information, hand shape Fisher Vector and hand direction feature are represented by a Temporal Pyramid (TP) to obtain the final feature vectors to be fed into a linear SVM classifier to recognize. The proposed approach is evaluated on a challenging dataset containing eight gestures performed by ten participants. Compared with the state-of-the-art dynamic hand gesture recognition methods, the proposed method shows a relative high recognition accuracy of 90.0%.

Keywords: hand gesture recognition, skeleton-based, Gaussian Mixture Model, Fisher Vector, SVM

Introduction

With the rapid development of HCI, as one of the most effective and natural way of interaction, dynamic hand gestures have gained more attentions in such applications of VR and HCI as medical care [1], education [2], interaction with robots [3] and entertainment [4]. However, how to recognize a specific dynamic hand gesture is a challenging task because of the variability of hand shape and the temporal information of each dynamic gesture.

Traditional dynamic hand gesture recognition is based on such flexure motion capture devices as data glove or video information captured by a monocular video sensor, however the additional glove and influential factors in hand tracking such as various hand appearances and illuminations limit the application of these recognition methods. During the past five years, the progress of commercial 3D depth sensing technologies has brought many innovative depth sensors, such as Leap Motion and Microsoft Kinect sensor [5], which may provide 3D hand skeleton structure and depth data of the scene. The adoption of 3D depth sensors in dynamic hand gesture recognition may effectively compensate the shortcomings of the above problems. Compared with the traditional recognition method, 3D depth is insensitive to light change, thus enhances the performance under light change environment, the different size of hands can also be easily normalized. Unlike Kinect sensor which provides the full-body depth, Leap Motion focuses on the accurate 3D hand data and the output of the Leap Motion is the skeleton data which consists of fingertips positions, palm center position, palm direction, palm normal and other relevant points positions as shown in Figure 1. No extra computational work is required to obtain such information. Moreover, such virtues as the highly precise localization in the effective range, small in size, and ease to be built into some VR applications make it a kind of specific somatosensory equipment for hand gesture recognition.

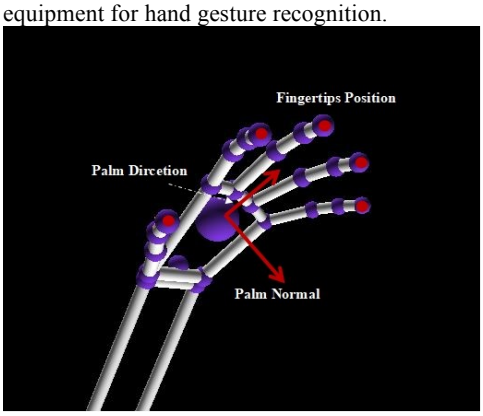

Figure 1. Illustrations of Palm direction, Palm normal and Fingertips positions.

Related Work

Hand gesture recognition has become a popular topic in recent years. Gestures can be divided into static gestures and dynamic gestures. In the existing static gesture recognition systems, most algorithms use the depth image information to obtain the hand silhouettes and extract features from the hand region [6-8]. Except the research based on depth information, some researches extract features based on skeleton information which mainly focus on the recognition of the sign language gestures lexicon. The authors of Ref. [9] first extracted the fingertip features and distance features from the Leap Motion and Kinect sensor respectively and combined them to recognize American Sign Language (ASL), then used SVM and Random Forest (RF) for classification. Mohandes et al. [10] developed a system for Arabic alphabet sign recognition using two Leap Motions, in which the two Leap sensors are placed perpendicular to each other to acquire the sign data. They exploited 12 features returned by each Leap Motion and fused them through Dempster-Shafer (DS) Theory.

When recognizing dynamic hand gestures, due to the lack of publicly available dynamic hand gesture datasets for benchmarking, most approaches focus on recognizing the motion trajectories including the digits and alphabet. The authors of Ref. [11] proposed a fast recognition method for motion trajectory of the digits and the alphabet. Such steps as data filtering, data quantization, feature vectors zero-padding and experimental data extension are paid close attention to. In addition, a Polynomial Kernel Function was chosen as the kernel of SVM, thus high recognition accuracy and fast recognition speed are obtained. Other approaches such as [12] and [13] selected the coordinates of the fingertip positions and finger angles as the trajectory features to recognize the numbers from 0 to 9. These two approaches exploit Geometric Template Matching

method and Radial Basis Function (RBF) neural networks for classification.

However, both hand pose and hand movement are used to represent the gesture during HCI. In order to recognize various dynamic hand gestures, some researchers began to collect their own datasets in their researches. In the study of [14], the authors collected a novel dataset from Leap Motion which contains 12 dynamic hand gestures. The data are used to train a 3D recognition model based on Convolutional Neural Networks(CNN), which can recognize 2D projections of the 3D space. Instead of recognizing dynamic gestures in daily life, authors from [15] created a comprehensive dataset for medical applications, their dataset contains 11 gestures such as click, zoom in/out, move left/right and so on. Arithmetic mean, Standard deviation, covariance and root-mean-square features are extracted from the skeleton information and then SVM is used for classification. Ref. [16] presented a Dynamic Hand Gesture 14-28 (DHG) dataset, which provided sequences of hand skeleton in addition to the depth image. There are 14 gestures performed by 10 participants in two ways in their dataset. For each frame, authors extracted 9 descriptors with the name of Shape of Connected Joints as feature and trained a linear SVM classifier for recognition.

In this paper, we propose a new dynamic hand gesture dataset, which is designed to cover several daily interactive gestures collected by Leap Motion. The robust spatiotemporal multi-fused features are exploited to represent each dynamic hand gesture, which is concatenated by two kinds of features as hand shape features dealing with changes among hand shapes and hand direction features dealing with local motion of palm center. The hand shape features are used to train a Gaussian Mixture Model and the Fisher Vector is adopted to represent the final hand shape features. Considering the temporal information, a temporal pyramid is used to encode each gesture sequence. The classification process is performed by a linear SVM. We discuss the effects of each feature and model parameters on the experimental results. We also demonstrate a dynamic hand gesture interactive application in a virtual decorative scene using our dataset.

Method

Dataset

We built a Dynamic Interactive Hand Gesture (LMDI) dataset using Leap Motion which provides sequences of hand skeleton. In the LMDI dataset 8 gestures are collected including Grab, Expand, Tap, Pinch, Swipe Right, Swipe Left, Clockwise and Counter Clockwise as shown in **Figure 2**. Each gesture is performed 10 times by 10 participants (5 males and 5 female), resulting in 800 sequences. All the participants are right-handed. The hand skeletons were captured at 50 frames per second for such hand information as palm position, palm direction, palm normal and fingertip positions.

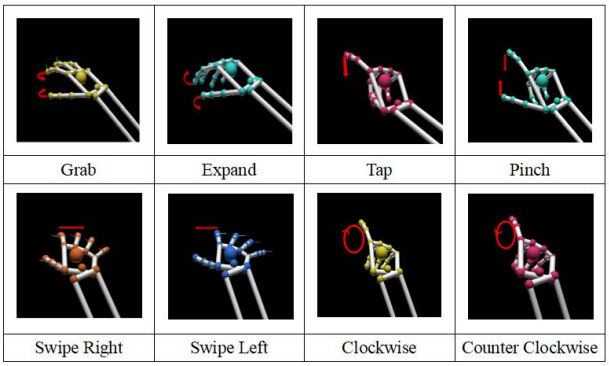

Figure 2. Eight hand gestures from LMDI. Left to right, top to bottom: Grab, Expand, Tap, Pinch, Swipe Right, Swipe Left, Clockwise and Counter Clockwise. The blue arrows are the motion trajectories of the fingers.

Feature Extraction

In order to represent a hand gesture, the features can be divided into two types, the first type of hand shape features mainly captures the hand shape variation based on fingertips and palm center, and the second type of hand direction features focuses on describing the movement of hand location in 3D space. Moreover， the temporal nature of dynamic gestures is encoded using a temporal pyramid and the classification process is completed by a SVM classifier. Figure 3 shows a general overview of the proposed approach.

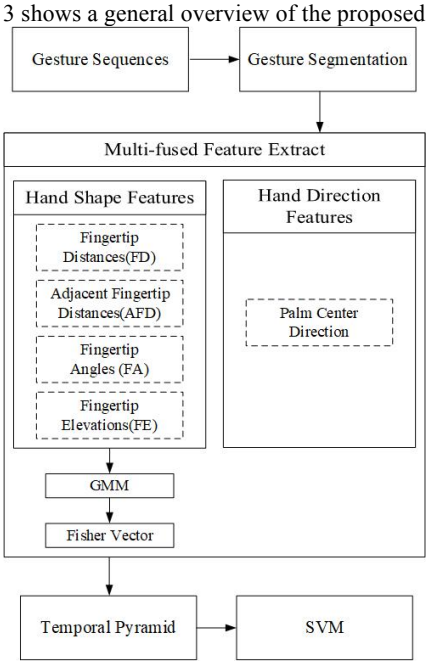

Figure 3. Overview of the proposed gesture recognition system.

Hand Shape Features

To represent the hand shape using a full skeleton, we extract 4 different sub-features based on hand joints to form hand shape features, including fingertip distances, adjacent fingertip distances, fingertip angles and fingertip elevations.

The fingertip distances (FD) feature measures the Euclidean distance between the i**th** fingertips and the palm center at the t**th** frame as:

$$FD_i(t) = \|F_i - C\|/M\,,\ \ i = 1,\cdots,5 \tag{1}$$

where F_i represents the 3D position of each fingertip. Vector C is the palm center location in the 3D space. In order to make hand shape features relatively invariant to hand geometric transformations, we normalize the features in the interval $[0,\ 1]$ by dividing scale factor M, which is the distance between the palm center and the middle fingertip, which can be computed with the palm opening completely before participants performs gestures. FD is a vector of 5 dimensions.

Considering the interaction between adjacent fingertips, a 4-dimension adjacent fingertip distance feature (AFD) is computed. AFD feature denotes the Euclidean distance between adjacent fingertips at the tth frame as:

$$AFD_i(t) = \|F_i - F_{i+1}\|/M\ \ \ i = 1,\cdots,4 \tag{2}$$

The fingertip angle (FA) feature measures the angles corresponding to the orientation of the projected fingertip on the palm plane with respect to the hand direction at the tth frame as:

$$FA_i(t) = \angle\left(F_i^P - C, H\right)/\pi\,,\ \ i = 1,\cdots,5 \tag{3}$$

where F_i^P denotes the projection of each fingertip on the palm plane identified by the palm normal vector N. Vector H is the hand direction. The FA feature is normalized with π. FA is a vector of 5 dimensions.

The last hand shape feature is fingertip elevations (FE), which represents the fingertip's elevation from palm plane. The feature at the tth frame is:

$$FE_i(t) = sgn\left(\left(F_i - F_i^P\right)\cdot N\right)\left\|F_i - F_i^P\right\|/M\ \ \ i = 1,\cdots,5 \tag{4}$$

and the sign operator describes the fingertip belonging to which of the two space defined by the palm plane. The dimension of the FE feature is 5.

We compute four features on each frame and obtain a combined hand shape feature of 19 dimensions by concatenating the four features. **Figure4** shows an example of four sub-features.

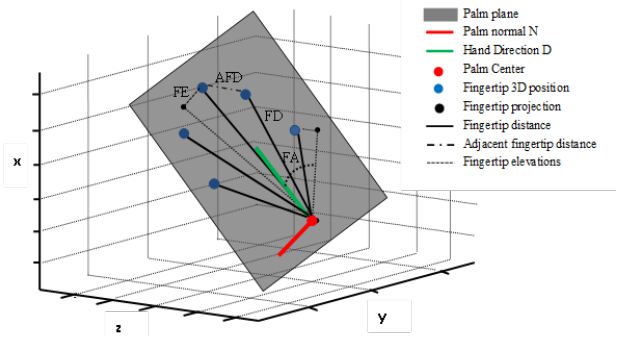

Figure 4. An example of 4 sub-features.

Hand Direction Features

Because some gestures are actually defined only by the way of hand movement (such as swipe right and swipe left), it is not sufficient to use only hand shape features to discriminate these gestures. To overcome this shortcoming, we compute a direction vector to describe the movement of hand in 3D space.

The direction feature using the position of palm center along the gesture sequence can be written as:

$$DirectionF = palm^t - palm^{t-Step},\ step + 1 \le t \le Num \tag{5}$$

where $palm^t$ is the palm position at tth frame. $Step$ is a constant chosen through experiment and Num is the number of frames for a specific hand gesture sequence. For direction features, spherical coordinate is adopted to represent each direction vector as shown in Ref. [16]. The 3D space is divided into $n_\theta m_\varphi$, which n_θ and m_φ are the parts of space θ and space φ respectively. Each direction vector is then localized at a unique interval.

Fisher Vector Representation and Gaussian Mixture Model

Fisher Vector has been widely used in image classification [17] and human action recognition [18] to incorporate the additional information about the distribution of the features and it performs well even with simple linear classifiers.

Let $F_{seq} = \{x_i, i = 1,\ldots,N\}$ represent the set of four sub-features with D dimensions extracted from the gesture sequences. The generation process of sub-features can be modeled by a probability density function ψ_λ, where λ denotes the function parameters. Hand shape Fisher Vector can be described by the derivative of each log-likelihood of the ψ_λ with respect to λ as:

$$G_\lambda^{F_{seq}} = \frac{1}{N}\nabla_\lambda log\psi_\lambda\left(F_{seq}\right) \tag{6}$$

Usually ψ_λ is chosen as a K-component Gaussian Mixture Model[19]: $\psi_\lambda(x) = \sum_{k=1}^K \omega_k\,\psi_k(x)$ with the parameters $\lambda = \{\omega_k, \mu_k, \Sigma_k\ \ k = 1,\cdots,K\}$, where ω_k,μ_k and Σ_k are respectively the weight, mean and covariance matrix of GMM. Note that the covariance matrices are diagonal and denoted by the variance vector σ_k^2 [20]. After using the sub-features to train a GMM, the FV can be computed as:

$$\begin{aligned} G_{\mu,k}^{F_{seq}} &= \frac{1}{N\sqrt{\omega_k}}\sum_{i=1}^N \gamma_i(k)\left(\frac{x_i-\mu_k}{\sigma_k}\right) \\ G_{\sigma,k}^{F_{seq}} &= \frac{1}{N\sqrt{2\omega_k}}\sum_{i=1}^N \gamma_i(k)\left(\frac{(x_i-\mu_k)^2}{\sigma_k^2}\right) \end{aligned} \tag{7}$$

where $\frac{1}{N}$ is the normalization term and $\gamma_i(k)$ is the soft assignment of the features x_i to Gaussian k. The final 2KD-dimensional FV is the concatenation between $G_{\boldsymbol{\mu},k}^{F_{seq}}$ and $G_{\sigma,k}^{F_{seq}}$. Final FV is also normalized with a power and $L2$ normalization to increase its discriminability [21].

Temporal Pyramid and Classification

Both hand shape features and hand direction features describe the variation of hand shape and direction for a specific hand gesture sequence without taking into account the order of different gestures.

To add the temporal information, we exploit a 3-layer TP which divides the gesture sequence into 3-level sub-sequences [22]. The principle of TP is shown in **Figure 5**. For each sub-sequence, we re-concatenate two types of features.

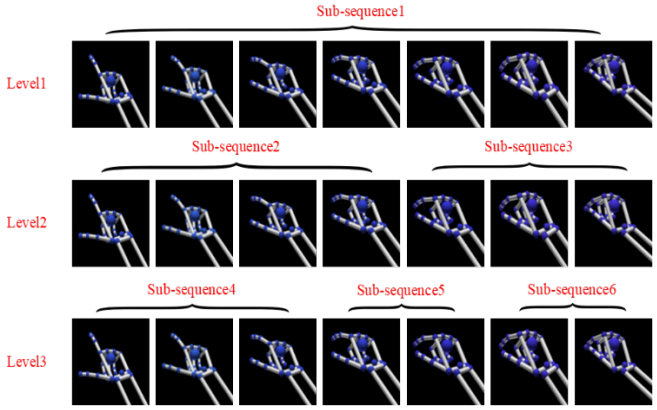

Figure 5. A 3-layer TP representation on a gesture example (sample of gesture Pinch from our dataset). The complete features are the concatenation of six sub-sequences.

The classification scheme exploits a multi-class SVM classifier based on the one-vs-one approach [23]. In particular a set of $G(G-1)/2$ binary SVM classifiers is used where G is the number of different gestures and a linear kernel has been selected. The gesture with the maximum number of votes is selected as the final output label.

Experimental Results

To evaluate the effectiveness of the proposed method, we first compare the results for each type of features separately. Then, the effects of different GMM and TP parameters on the recognition results are studied and a comparison with other researches is presented. Finally, an interactive demonstration experiment is implemented using the proposed dynamic hand gesture recognition approach. In the experiment, hand shape features with size $D = 19$ per frame are used to train a GMM of $K = 20$ clusters, resulting in 760-elemetnt FVs. For the hand direction features, n_θ and m_φ are 6 and 8 respectively. Finally, we choose 3 layers TP, leading to a 4848-dimensional vector feature for each sequence. We test the proposed dynamic hand gesture recognition method on our LMDI dataset with *cross-validation* strategy.

Experimental Results with Different Features

Two types of features are adopted, i.e. hand shape feature and hand direction features which focus on describing the variation of hand shape and hand movement in 3D space. Table 1 shows the recognition accuracy of the proposed method obtained by using each type of features independently and jointly. It can be seen from Table 1 that when only hand shape features or hand direction features are adopted, the accuracy is lower than adopting them jointly.

Table1: Recognition accuracy of two features for LMDI dataset

Feature set	Accuracy
Shape Features	70.0%
Direction Features	87.5%
Shape Features + Direction Features	90.0%

Figure 6 shows that the recognition accuracy of only using hand shape features to recognize such gestures as *Grab*, *Expand*, *Tap* and *Pinch* is higher than the recognition result of such gestures as *Swipe Right*, *Swipe Left*, *Clockwise* and *Counter Clockwise* and the only adoption of the direction features will generate completely opposite results, which shows that the hand shape features is the most effective features for the dynamic gestures with large variations in the hand shape.

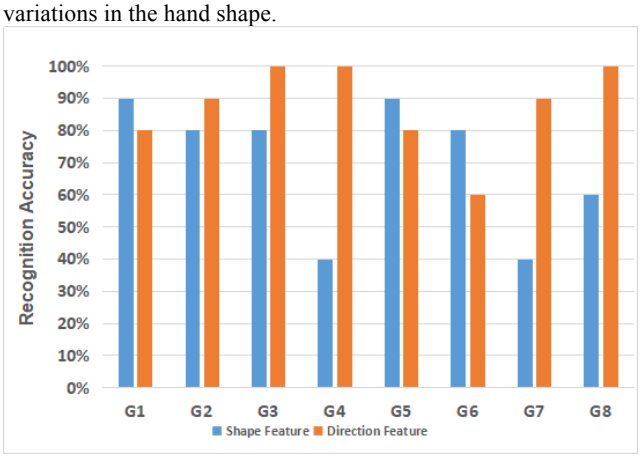

Figure 6. The accuracy of each gesture is classified by two features respectively. G1 to G8 are gestures: Grab, Expand, Swipe Right, Swipe Left, Tap, Pinch, Clockwise and Counter Clockwise.

On the other hand, the accuracy is also high when recognizing the gesture *Expand* using direction features only, because the highly precise Leap Motion can capture small change of direction during the acquisition of dynamic gestures. The further proof of the validity of the hand shape is shown in **Figure 7**.

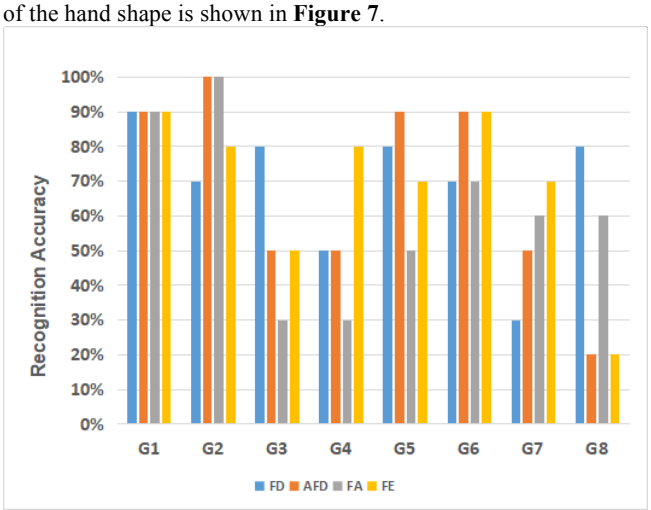

Figure 7. The accuracy of each gesture is classified by four sub-features form hand shape features. G1 to G8 are gestures: Grab, Expand, Swipe Right, Swipe Left, Tap, Pinch, Clockwise and Counter Clockwise.

Our method performs well on the majority of the gestures with the exception that the recognition accuracy of one or two gestures is not always optimal as shown in the red confusion matrix in **Figure 8**. For example, it can be noticed that the accuracy of all gestures except *Pinch* have reached an accuracy of more than 80%, which means that the features we exploited can represent the variation of

hand shape and position perfectly. Gesture *Pinch* is sometimes confused with gesture *Grab* and gesture *Tap*, which is most probably due to the similarity of the three gestures, especially gesture *Tap* and gesture *Pinch*. The main difference between them is the number of fingers used in the whole gesture.

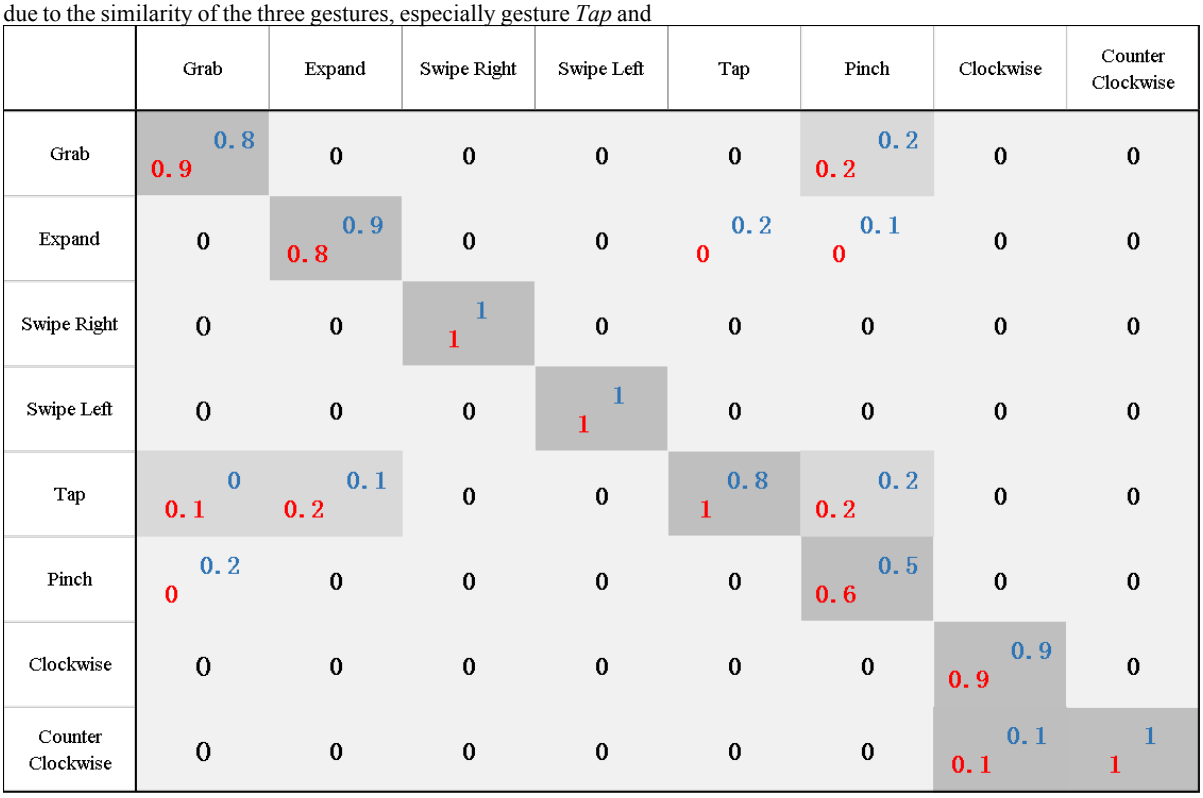

Figure 8. The confusion matrix for LMDI dataset. The red values represent the model parameters with 20 clusters and 3 layers, blue values represent the model parameters with 20 clusters and 2 layers.

Experimental Results with Different Parameters

To evaluate the contribution of GMM and TP parameters on the experimental results, we performed a series of experiments to determine the best combination of numbers of GMM cluster and TP layer. The different setting with corresponding results is listed in Table 2. The best accuracy of 90% for LMDI dataset occurs at 20 clusters and 3 layers. The experimental results show that the hand shape features can be well fitted by 20 single Gaussian models. As for TP layer, we find that adding more temporal information will increase the recognition accuracy. The confusion matrix for LMDI dataset with 20 clusters and 2 layers (blue values) is also shown in **Figure 8**, which shows that with the increase of temporal information, the accuracy of gesture *Tap* and gesture *Pinch* is increased by 20% and 10% respectively. However, adding more levels to TP not only allows the temporal information to be more precise, but also increases the size of the final features. For example, the dimension of the final features at layer 3 is 4848, but at layer 4 the final size is 8080. The large increase of dimension not only increases the computing time, but also increases the influence of each independent dimension information, and decreases the recognition accuracy. Moreover, if the level is reduced, the model cannot represent the local time information of gestures well, especially for the gestures which are similar to the hand shape such as *Pinch* and *Tap*.

Table2: Recognition accuracy of GMM and TP under different parameters for the LMDI dataset

K	L	Accuracy	K	L	Accuracy
15	3	88.75%	20	1	86.25%
20	**3**	**90.0%**	20	2	86.25%
25	3	86.25%	20	3	90.0%
31	3	85.0%	20	4	86.25%

Comparison with Other Method

To evaluate the performance of the proposed method, we test our method on other datasets as well. Authors in [15] collected a new dataset containing 11 dynamic hand gestures with a Leap Motion sensor, which are *Click, Left rotation, Right rotation, Increase contrast, Decrease contrast, Zoom in, Zoom out, Move left, Move right, Previous* and *Next.* We use G1 to G11 to represent them. Each gesture was performed by 10 subjects for five times. The dataset includes 550 gesture sequences. Some of the gestures in the dataset are highly similar to each other which makes the recognition algorithm quite challenging.

During the experiment, one tenth of the gesture sequences are used for training and the rest for testing. Table 3 summarizes the comparison of the recognition accuracy of the proposed method with the method in [15]. In [15] the recognition accuracy on four features: Mean, Stander deviation, Covariance, Root mean square are tested separately and jointly, which reaches a recognition accuracy of 80.909%. We also test our two kinds of features respectively on the

dataset separately and jointly, which obtains a 4.55% improvement on the recognition accuracy as shown in Table 3. It is also found that each of our features has a better performance compared with [15].

More details about the recognition results on 11 dynamic hand gestures are shown in **Figure 9**. In the confusion matrix, the red values represent the recognition results of our method and the blue values represent the recognition results from [15]. It is found that almost all gestures can be recognized with an accuracy of 80% and some even with 100% with the proposed method. In particular, the recognition of G2 and G3 gestures by our algorithm is 100% and 80% respectively, which is 50% and 20% higher than that of [15]. Similar results can also be found in the recognition of G10 and G11 gestures, both of them have a more than 10% improvement. G2, G3, G10 and G11are four reciprocal gestures and our method is robust to such gestures. Both the proposed method and Ref. [15] show similar results on G4 to G9. It can also be found from the experimental results that both algorithms failed to recognize the challenging gesture G1 owing to the similarities between G1 and G5 and touching fingers.

Table3: The comparison of recognition accuracy of the proposed method with the method in [15]

Feature set	Accuracy
Mean(μ)	40.00%
Standard deviation(S)	53.1818%
Covariance(C)	40.9091%
Root mean square(RMS)	30.9091%
$\mu+S+C+RMS$	**80.9091%**
Shape Features	61.8182%
Direction Features	83.6364%
Shape Features+Direction Features	**85.4545%**

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11
G1	0.6 0.4	0.05 0.0	0.2 0.0	0	0.0 0.2	0	0	0	0	0	0
G2	0	0.5 1.0	0.0 0.2	0	0	0	0	0	0	0	0
G3	0.15 0.0	0.25 0.0	0.6 0.8	0	0	0	0	0	0.05 0.0	0	0
G4	0	0.05 0.0	0.1 0.0	1.0 1.0	0	0.0 0.2	0	0	0	0.2 0.0	0.05 0.0
G5	0	0	0	0	0.9 0.8	0	0	0	0	0	0
G6	0.2 0.4	0	0	0	0.05 0.0	1.0 0.8	0.0 0.2	0	0	0	0.05 0.0
G7	0	0	0	0	0	0	1.0 0.8	0	0	0	0
G8	0.05 0.0	0	0	0	0	0	0	1.0 0.8	0	0	0.05 0.0
G9	0	0.15 0.0	0.05 0.0	0	0	0	0	0	0.95 1.0	0.05 0.0	0
G10	0.0 0.2	0	0.05 0.0	0	0	0	0	0.0 0.2	0	0.75 1.0	0
G11	0	0	0	0	0.05 0.0	0	0	0	0	0	0.85 1.0

Figure 9. The confusion matrix for 11 dynamic hand gestures dataset. The red values represent our algorithm and the blue values represent the algorithm from [15].

Interactive Demonstration Experiment

In order to evaluate the performance of user's interaction with virtual objects using the proposed dynamic gesture recognition algorithm, an interactive demonstration experiment is designed and implemented.

The interactive demonstration experiment is implemented using a Leap Motion sensor connected via USB to a PC with Inter(R) Core(TM) i5-4570 CPU @ 3.20GHz and NVIDIA GeForce GTX 750. The software environment is Windows 10 operating system, 5.2.1f1 unity3D for windows and Leap Motion Core Asset v2.3.1.

There are three scenes in the gesture interaction experiment, in which scene one and scene two are implemented mainly through keyboard detection to control the switching between the scenes and acquire different users' hand size information. Then the user can interact using dynamic hand gesture in scene three. In the gesture interaction experiment, different functions are designed for 8 kinds of dynamic gestures that we designed. However, since in the actual test it is found that the proposed algorithm has poor recognition accuracy on *Expand* gesture, we only select other seven dynamic gestures and each gesture corresponds to different functions. The three scenes and the interactive results using dynamic gesture are:

Scene 1: Start interface. The user chooses whether to start interaction according to the prompt information on the interface. "Yes" will enter into Scene 2 and "No" will exit the system.

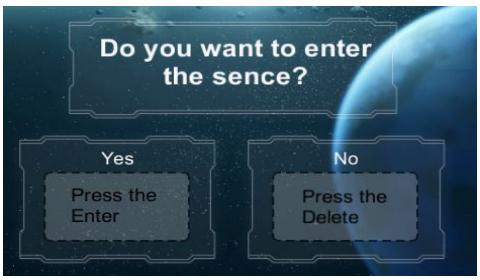

Figure 10. Scene 1: Star interface

Scene 2: Hand size information acquisition interface. In this scenario, users are asked to open their hand and put near the Leap Motion sensor with palm down to obtain the hand size for normalization. After obtaining 30 frames of hand information, the interactive experiment will enter into Scene 3.

Figure 11. Scene 2: Hand size information acquisition interface

Scene 3: Interactive interface. The interactive scenario is a virtual room design application. Users can select different interaction objects through the options on the menu and decorate the virtual room. Gesture interaction focuses on the interaction between the user and the UI menu. The interaction process by using different gestures is shown through **Figure 12** (a) to (e).

Conclusion

In this paper, we proposed an effective approach to recognize dynamic hand gestures using 3D hand skeleton data acquired from Leap Motion sensor. The multi-fused features can be divided into two types of spatiotemporal features such as hand shape feature to represent shape structure variation in the process of gesture completion and hand direction features to measure the change of palm center position in 3D space. Furthermore, a Fisher Vector representation encoded the hand shape features and TP method added the temporal information of each gesture. Experimental results tested on the two challenging datasets show that the proposed method achieved a relatively high recognition accuracy. A gesture interactive demonstration experiment has also been designed and users' interaction with virtual objects is realized in a virtual scene with dynamic hand gestures. In the future, we will focus on the segmentation of continuous dynamic hand gestures and try to apply our method for more complicated gestures including the dynamic hand gestures performed by two hands.

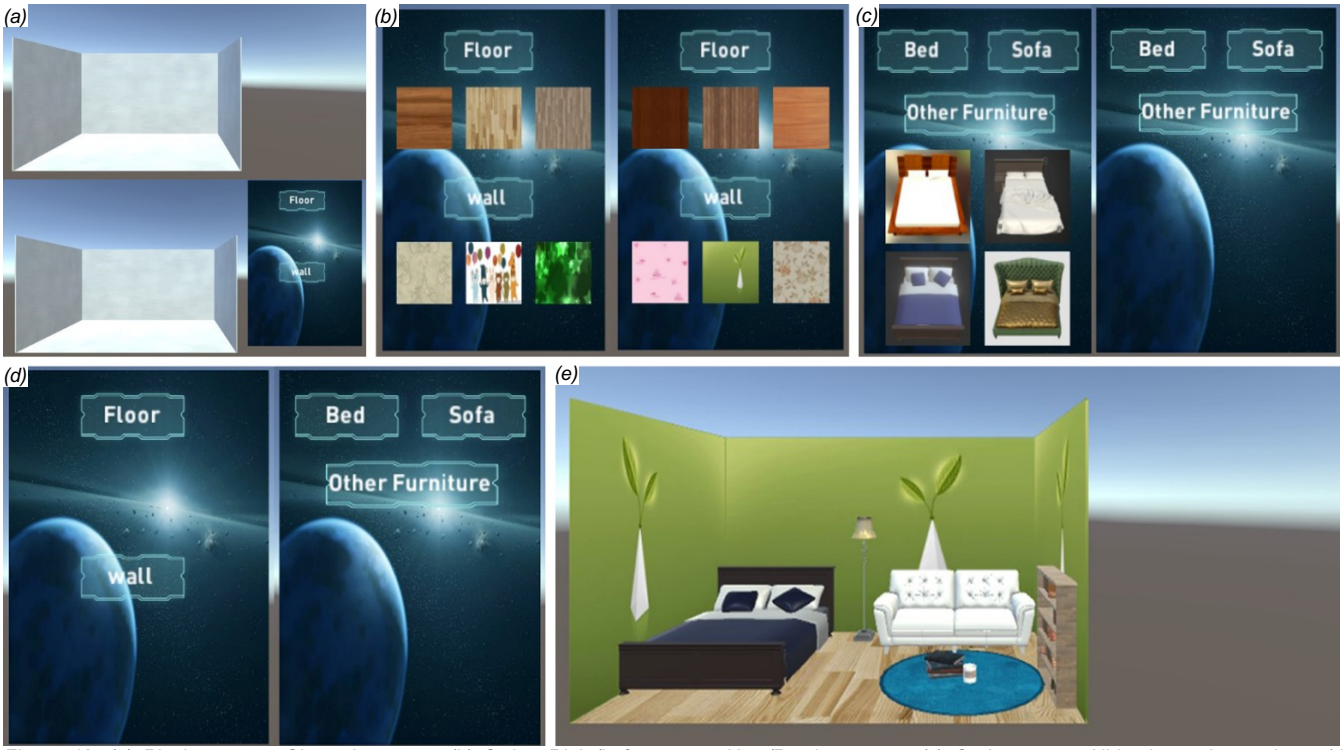

Figure 12. (a) Pinch gesture: Show the menu. (b) Swipe Right/Left gesture: Next/Previous page. (c) Grab gesture: Hide decoration options. (d) Clockwise/Counter Clockwise gesture: Next/Previous menu. (e) Tap gesture: Close the menu. After this, a decorated room is shown as (e).

References

[1] Vishal Nayakwadi, N. B. Pokale "Natural Hand Gestures Recognition System for Intelligent HCI: A Survey "International Journal of Computer Applications Technology and Research, vol.3, no.1, pp.10 - 19, 2013.T. Jones, "Sample Journal Article," Jour. Imaging Sci. and Technol., vol. 53, no. 1, pp. 1-5, 2009.

[2] Yang, Dongseok, Jongkuk Lim, and Younggeun Choi. "Early childhood education by hand gesture recognition using a smartphone based robot." in 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, UK,2014.

[3] Shukla, Dadhichi, Ozgur Erkent, and Justus H. Piater. "A multi-view hand gesture RGB-D dataset for human-robot interaction scenarios." in 25th IEEE International Symposium on Robot and Human Interactive Communication, New York, USA, 2016.

[4] Krzysztof Pietroszek, "Watchcasting: Freehand 3D interaction with off-the-shelf smartwatch." in 2017 IEEE Symposium on 3D User Interfaces, Los Angeles, USA, 2017.

[5] Cheng, Hong, Lu Yang, and Zicheng Liu. "Survey on 3D Hand Gesture Recognition." in IEEE Transactions on Circuits and Systems for Video Technology, vol.26, no.9,2016.

[6] Suarez, Jesus, and Robin R. Murphy. "Hand gesture recognition with depth images: A review." in 21st IEEE International Symposium on Robot and Human Interactive Communication, Paris, France,2012

[7] Ferran Argelaguet, Mélanie Ducoffe, Anatole Lécuyer, and Rémi Gribonval. "Spatial and Rotation Invariant 3D Gesture Recognition Based on Sparse Representation." in 2017 IEEE Symposium on 3D User Interfaces, Los Angeles, USA, 2017.

[8] Plouffe, Guillaume, and Anamaria Cretu. "Static and Dynamic Hand Gesture Recognition in Depth Data Using Dynamic Time Warping." in IEEE Transactions on Instrumentation and Measurement, vol.65, no.2, pp.305-316,2016

[9] Marin, Giulio, Fabio Dominio, and Pietro Zanuttigh. "Hand gesture recognition with jointly calibrated Leap Motion and depth sensor." in Multimedia Tools and Applications, vol.75, no.22, pp.14991-15015,2016.

[10] Mohandes, Mohamed, S. Aliyu, and Mohamed Deriche. "Prototype Arabic Sign language recognition using multi-sensor data fusion of two leap motion controllers." in International Multi-Conference on Systems, Signals and Devices, Mahdia, Tunisia, 2015.

[11] Chen, Yanmei, et al. "Rapid recognition of dynamic hand gestures using leap motion." in 2015 IEEE International Conference on Information and Automation, Lijiang, China, 2015.

[12] Sharma, Jayash, Rajeev Gupta, and Vinay Kumar Pathak. "Numeral Gesture Recognition Using Leap Motion Sensor." in 2015 International Conference on Computational Intelligence and Communication Networks, Jabalpur, India, 2015.

[13] Qinghui Wang, Ying Wang, Fenglin Liu, Wei Zeng "Hand gesture recognition of Arabic numbers using leap motion via deterministic learning." in 36th Chinese Control Conference, Dalian, China,2017.

[14] McCartney, Robert, Jie Yuan, and Hans-Peter Bischof. "Gesture recognition with the leap motion controller." Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2015.

[15] Safa Ameur, Anouar Ben Khalifa, and Mohamed Salim Bouhlel. " A Comprehensive Leap Motion Database for Hand Gesture Recognition." in 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, Hammamet, Tunisia, 2016.

[16] De Smedt, Quentin, Hazem Wannous, and Jeanphilippe Vandeborre. "Skeleton-Based Dynamic Hand Gesture Recognition." in 2016 IEEE Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 2016.

[17] Perronnin, Florent, Jorge Sanchez, and Thomas Mensink. "Improving the fisher kernel for large-scale image classification." in 14th European Conference on Computer Vision, Amsterdam, Netherlands, 2016.

[18] Peng, Xiaojiang, et al. "Action Recognition with Stacked Fisher Vectors." in 13th European Conference on Computer Vision, Zürich, Switzerland, 2014.

[19] Dixit, Mandar, Nikhil Rasiwasia, and Nuno asconcelos. "Adapted Gaussian models for image classification." In 2011 IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA, 2011.

[20] Douglas Reynolds, "Gaussian Mixture Models." MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA

[21] Sanchez, Jorge, et al. "Image Classification with the Fisher Vector: Theory and Practice." International Journal of Computer Vision, vol.105, no.3, pp.222-245, 2013.

[22] Evangelidis, Georgios D., Gurkirt Singh, and Radu Horaud. "Skeletal Quads: Human Action Recognition Using Joint Quadruples." in 22nd International Conference on Pattern Recognition, Stockholm, Sweden,2014.

[23] Chang, Chihchung, and Chihjen Lin. "LIBSVM: A library for support vector machines." ACM Transactions on Intelligent Systems and Technology 2.3, 2011.

Author Biography

Dan Zhao received a Bachelor's degree in Optical engineering from Xidian University in 2016. She is currently a Master graduate student under the supervision of Prof. Yue Liu from Beijing Institute of technology. Her study interests are in dynamic hand gesture recognition and computer vision.

Yue Liu received his Ph.D. in Telecommunication and information System from Jilin University, China in 2000. He is currently a professor of Optical Engineering and software in Beijing Institute of technology, China. His research interests include computer vision, virtual and augmented reality, evaluation of 3D display devices etc.

Guangchuan Li graduated from Hefei University of Technology in 2015. From then on, he has been studying at Beijing Institute of technology applying for his Degree of doctor of engineering as a PhD candidate. His main research is about dynamic hand gesture recognition and object manipulation in virtual reality and augmented reality.