A Fast Automatic Holoscopic 3D Micro-gesture Recognition System for Immersive Applications

Rui Qin[1], Yi Liu[1], Mohammad Rafiq Swash[1], Maozhen Li[1], Hongying Meng[1,*], Tao Lei[2], and Tong Chen[3]

[1] Department of Electronic and Computer Engineering, Brunel University London, UK
[*] *Hongying.Meng@brunel.ac.uk*,
[2] Shaanxi University of Science and Technology, Xi'an, China
[3] Southwest University, Chongqing, China

Abstract. Immersive technology attempts to emulate a physical world through the means of a digital or simulated world. Micro-gestures are small variation actions on human hands defined by user that is one of the most convenient human action in immersive technology. Holoscopic 3D imaging uses bionics technology to capture spatial image in the pattern of fly's eye and it has fruitful 3D cubic information compared to 2D images that can be used for high accurate micro-gesture controller systems.In this paper, a new micro-gesture recognition system based on holoscopic 3D imaging system is proposed for immersive applications. It is built on fast pre-processing, dynamic image feature extraction and a non-linear Support Vector Machine classifier. It is evaluated on the public Holoscopic Micro 3D Gesture (HoMG) dataset outperforming all the existing state-of-the-art methods on the same dataset.

Keywords: Holoscopic 3D imaging, Micro-gesture recognition, LPQ-TOP, Support Vector Machine.

1 Introduction

Immersive application has been developed fast in recent year as the Augmented Reality (AR) and Virtual Reality (VR) applications and equipment have been developed and utilized widely in our daily life. Most immersive applications are constituted in perception, interaction and software. Interaction applications include gesture interaction, brain-computer interface and speech recognition. Software usually includes Artificial Intelligence (AI) technology and visual world applications to achieve immersive missions.

Holoscopic 3D (H3D) imaging is an image capture technology first proposed by Aggoun et al. [1] according to the theory of fly's eye on bionics by Wu et al. [12]. It utilised a mathematical image process method called integral imaging by Lippmann et al. [6] in 1908.

The H3D camera, which mimics the fly's eyes to capture and repeat the scenes, is used to get holosocopic 3D micro-gesture more clearly. The H3D sensor

provides RGB and depth information in high-resolution images and full HD videos. H3D imaging system is a novel potential technique which can satisfy the higher demand for user interactive experience. Precision 3D micro-gesture data can make use of the wide view coverage of the H3D camera to capture accurate finger movement [7].

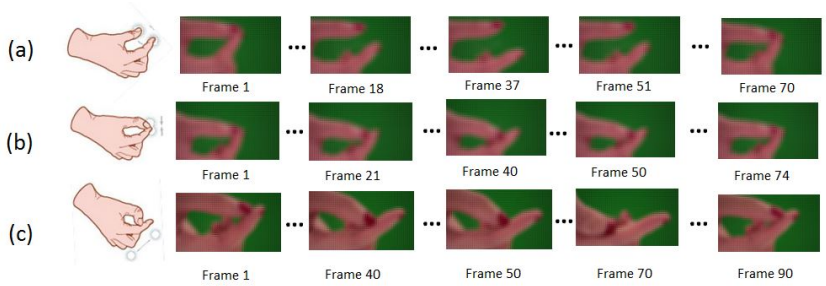

Fig. 1. Video sequences of three H3D micro-gestures (a) button, (b) dial, (c) slide.

Micro-gesture interaction is a significant method in human computer interaction area, which follows the new trend of enhancing the user experience introducing new forms of sensing, perception, interaction, and comprehension [2].

In this paper, we will explore a fast micro-gesture recognition system based on H3D imaging. It involves three types of the ubiquitous micro-gesture to assist augmented reality manipulation. The three types of simple micro-gestures are combined with the environmental and wearable computing paradigms which are responsible for the control of selection, confirmation and adjustment to enhance the interaction experience of information exhibition as shown in Figure 1 .

2 Related Works

Holoscopic 3D imaging based micro-gesture recognition research is progressing slowly in last few decades due to lacking of the data. Recently, HoMG dataset is collected and published [7]. Then the international Holoscopic 3D Micro-Gesture Recognition (HoMGR) challenge workshop was held in 2018 [7]. It speeds up the research in this particular area.

HoMGR dataset has two parts: 1). video subset where each micro-gesture is recorded as a video, some frames of which are shown in the Figure 1; 2). image subset where each micro-gesture is recorded as an image. In theory, video based system should produce better results as a complete micro-gesture normally has more frames (images).

In video based micro-gesture recognition competition, Zhang et al. [13] proposed the method using the Convolutional Neural Network (CNN) on each frame and making a fusion for the whole video, and Sharma et al. [10] proposed a more direct method based on Long Short-Term Memory (LSTM) model.

In [13], a deep learning method of CNNs using ResNet has been applied. With the powerful CNN model and the novel fusion method, the recognition accuracy is 82.0%, which is the state of the art performance on the HoMGR dataset. This method considers a video as several images added together and ignores the dynamic characteristics of videos. However, without dynamic information, essentially, it is more like a image-based recognition method.

In Sharma et al. [10] study, Gated Recurrent Unit (GRU) and LSTM are utilised in video-based dataset. The recognition accuracy of video-based result is 69.17%. Although with dynamic recognition deep learning method, the recognition accuracy is worse than that of [13]. Inspired by these two methods, this paper aims to build a system based on dynamic feature extraction and direct classifiers for fast implementation.

3 System Overview

The overview of the proposed system is shown in Figure 2. The first step is to do some pre-processing on the videos. In order to reduce the huge amount of the data for a fast recognition system, the resolution of each frame was reduced and the colour images were converted into grayscale images.

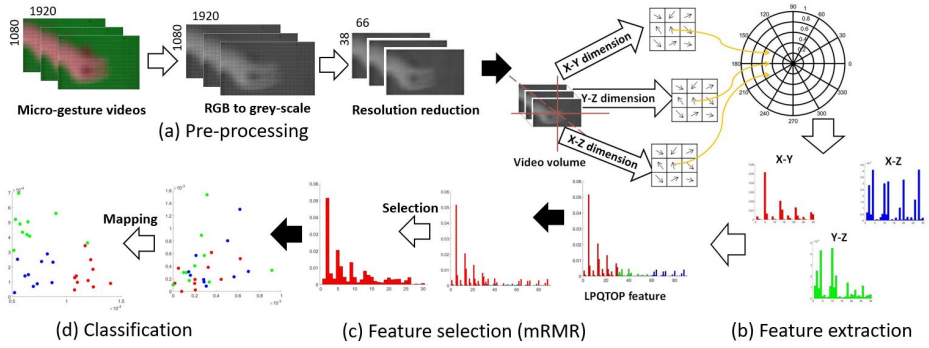

Fig. 2. Overview of automatic 3D micro gesture recognition system. (a) pre-processing, (b) LPQTOP feature extraction, (c) mRMR feature selection, (d) non-linear SVM classification.

The second step is feature extraction. In the system proposed in [7], both Local Binary Patterns from Three Orthogonal Planes (LBPTOP) [14] and Local Phase Quantisation from Three Orthogonal Planes (LPQTOP) [5] were used. According to previous experiments, LPQTOP has been proved to be a better solution that will be used in our system. The third step is feature selection. Feature selection can not only remove redundant information from feature, but also reduce the computing cost in the next step. The final step is the classification. There are lots of classifiers available, we chose non-linear Support Vector Machine (SVM) [9] classifier as it suits the characteristics of the features.

3.1 Pre-processing

As the resolution of the videos is very high (e.g. 1080×1920), direct use of the original video will need huge amounts of computing resources. It is natural to consider reducing the resolution. As the RGB image would not fit the feature extraction methods of LBPTOP and LPQTOP. The RGB images were transformed to grayscale images.

After that, the resolution was reduced to 38×66 according to the regulation of array of lenslets. Every pixel in 38×66 video comes from a max pooling in each lenslet of original high resolution video.

3.2 Feature extraction

LPQ is a very popular feature that was proposed by Chan et al. [8] and it has been used in facial recognition [11]. The LPQ characterizes texture or appearance by using sign-based, magnitude-based and orientation-based differences. It includes four stages: (1) three kinds of information (local sign, magnitude and orientation patterns) are extracted from the image, in which a local orientation pattern is realised by using orientation estimation and quantification; (2) three separate codebooks (O, S and M, respectively) are learned by using vector quantisation; (3) the sign, magnitude and orientation patterns are mapped into their corresponding codebook by using lookup table (LUT); and (4) three histograms are concatenated into one vector. The inference stage consists of all stages except the second. The schematic of LPQ is shown in (b) of Figure 2. LBPTOP is a combined feature of computing LBP on three different directions of the video volume. It was firstly proposed by Jiang et al. [5].

3.3 Feature selection

After feature extraction, usually a feature selection or reduction would be applied, especially when the number of feature vectors is larger than that of samples. Minimum Redundancy Maximum Relevance (mRMR) [3] is an algorithm used in feature selection frequently, which could narrow down the relevance of the features and identify characteristics of them.

The mRMR use mutual information (MI) to achieve the minimum redundancy between features and maximum relevance between features and target. Using I to present MI, f_i presents the ith feature and the number of features are L. So the MI can be defined by the formula in 1. [4]

$$I(f_i, f_j) = \iint p(f_i, f_j)(x, y) log \frac{p(f_i, f_j)(x, y)}{p(f_i)(x)p(f_j)(y)} dxdy \quad (i, j = 1, 2, ..., L; i \neq j) \tag{1}$$

where $p(f_i, f_j)$ is joint probability distribution, $p(f_i)$ and $p(f_j)$ are marginal probability distribution. $\Omega = \Omega_S \cup \Omega_T$ stands for the whole features while $\Omega = \Omega_S$ presents selected features conclude m features and Ω_T presents target features conclude n features. The relevance D of a feature f_i with its target

c can be calculated by equation 2 and redundancy R of the feature f_j in Ω_T can be calculated by equation 3.

$$D(f_i) = I(f_i, c), \quad (f_i \in \Omega, \quad i = 1, 2, ..., m) \tag{2}$$

$$R(f_j) = \frac{1}{m} \sum_{f_i \in \Omega_S} I(f_j, f_i), \quad (f_j \in \Omega_T, \quad j = 1, 2, ..., n) \tag{3}$$

To achieve the maximum relevance and minimum redundancy, the equation 2 and 3 can be combined as equation 4. And the mRMR would be achieved by finding the maximum of equation 4 based on the suitable Ω_T.

$$\max_{\Omega_T}[I(f_j, c) - \frac{1}{m} \sum_{f_i \in \Omega_S} I(f_j, f_i)], \quad (i = 1, 2, ..., m; j = 1, 2, ..., n; f_j \in \Omega_T) \tag{4}$$

3.4 Classification

SVM has been approved as a very popular classifier in recent years. However, linear SVM classifier is not working well if the feature space is complex.

Different from linear SVM, non-linear SVM would map features into another suitable spatial coordinates and it would be easy to divide these samples from two or more classes. [9]The basic idea for non-linear SVM is to project feature into a higher dimensional space and then apply optimal hyper-plane algorithm in the new space.

4 Evaluation

4.1 HoMG database

This dataset was collected using a H3D camera by Liu [7] at Brunel University London. Three micro-gestures were captured, i.e. Button, Dial and Slider. In all, 50 subjects' micro-gestures have been collected including 33 males and 17 females. In the final HoMG database, 40 subjects have been selected.

The experiment has been done under subject independent setting. All subjects was divided into 3 groups, i.e. training set, development set and test set. In the baseline experiment, only training set and development set were used. The examples of video-based micro-gesture recording in HoMGR dataset are shown in Figure 1. The detailed information about the HoMGR dataset can be found in the baseline paper [7].

4.2 Experimental results

In the task of micro-gesture recognition, the focus is on extracting dynamic features. We have investigated both LBPTOP and LPQTOP methods for feature extraction.

Table 1. First results on training and development set for LBPTOP and LPQTOP features under different setting using a linear SVM classifier. (all means all gestures, including close gestures (C.G.) and far gestures (F.G.), statistical method includes leave one out (L.O.T) and 10 cross-validation (10 C.V.))

Feature	SVM (%)	k-NN (%)	Sub. Dis.	Setting
LBPTOP all	68.9	50.6	72.5	L.O.T.
LPQTOP all	78.9	51.9	**81.1**	L.O.T.
LBPTOP all	71.4	52.4	72.5	10 C.V.
LPQTOP all	78.6	51.2	79.7	10 C.V.
LBPTOP C.G.	59.7	36.1	55.0	10 C.V.
LPQTOP C.G.	74.4	37.5	68.3	10 C.V.
LBPTOP F.G.	60.8	35.0	56.1	10 C.V.
LPQTOP F.G.	66.7	42.5	70.0	10 C.V.

The accuracy in Table 1 is calculated by equation 5. Where in the equation, TP, TN, FP, FN stand for true positive, true negative, false positive and false negative respectively.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

In the first attempt, training set and development set are used together for building the models in which there are 30 subjects.The 30 subjects have been divided into 10 groups for 10-fold cross validation (C.V.) . The other way for the 30 subjects is leaving one out (L.O.T). Both experiments are subject independent.

In the next attempt, we follow the exact setting of the HoMGR dataset, where the dataset has been divided into 3 groups. According to the results in Table 1, LPQTOP is a better feature as it has the higher performance than LBPTOP. In addition, mRMR feature selection method was also added into the system.

Table 2 shows results of different features and options. From this table, it is clearly shown that the LPQTOP feature is better than LBPTOP, also approved in the first attempt. In addition, non-linear SVM classifier is much better than linear SVM. With the addition of mRMR, the performance was improved further.

The system is running on MATLAB 2018a with Window 10. The CPU is Core i7-7500U and RAM is 16GB. The Table 2 also shows the execution time of each system on testing one video sample from pre-processing to the final recognition result. The videos have different length, the average number of frames is 339 which is about 14 second under a 25 frame rate. It can been seen that one video clip can be processed within half minute that is good for most applications.

The obtained experimental results were compared to all the existing results on video-based micro-gesture recognition as shown in Table 3. Clearly, our result is better than all state of the art methods. In general, deep learning methods can get better results as it make very deep models for the data. Sharma et al. [10] proved the effectiveness of deep learning. However, the performance was not

Table 2. The recognition accuracy and average execution time on a video clip of the HoMGR test set using different features and options. (s is short for second)

Method	Accuracy (%)	System execution time (s)
LBPTOP + Linear SVM	60.4	28.73
LPQTOP + Linear SVM	72.8	26.92
LPQTOP + non-Linear SVM	84.2	26.57
LPQTOP + non-Linear SVM + mRMR	**84.6**	27.98

Table 3. Performance comparison between all the results on HoMGR dataset.

Author	Method	Accuracy (%)
Sharma et al. [10]	LSTM	65.4
Sharma et al. [10]	Gated Recurrent Unit (GRU)	69.2
Zhang et al. [13]	Hybird NN	69.2
Zhang et al. [13]	ResNet	82.0
Zhang et al. [13]	Dense	82.0
Zhang et al. [13]	SE-ResNet	82.0
Ours	LPQTOP + non-Linear SVM + mRMR	**84.6**

as good as expected. Our system adds mRMR for feature reduction, and the accuracy reaches 84.6%.

5 Conclusion

This paper proposed an automatic recognition system for H3D micro-gestures. The proposed system utilised traditional feature extracted and classification methods rather than time consuming deep learning methods. But the final performance is better than deep learning methods. Also the computing load is lower than that of the deep learning methods.

The possible reasons for this are: Firstly, the dataset is a bit small, the deep model might not be able to be well trained. There are only 30 subjects, and each subject has 8 videos in total. Secondly, LPQTOP feature might capture the dynamics of the micro-gesture better than deep learning models. Thirdly, non-leaner SVM classification might capture the structure of the feature space better.

In the future work, more deep learning methods will be explored for the challenge with effective architecture and more feature extraction methods will be investigated to extract better 3D information from the images.

References

1. Amar Aggoun, Emmanuel Tsekleves, Mohammad Rafiq Swash, Dimitrios Zarpalas, Anastasios Dimou, Petros Daras, Paulo Nunes, and Luí Ducla Soares. Immersive 3d holoscopic video system. *IEEE MultiMedia*, 20(1):28–37, 2013.

2. Leonardo Angelini, Francesco Carrino, Stefano Carrino, Maurizio Caon, Denis Lalanne, Omar Abou Khaled, and Elena Mugellini. Opportunistic synergy: a classifier fusion engine for micro-gesture recognition. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 30–37, 2013.
3. Benjamin Auffarth, Maite López, and Jesús Cerquides. Comparison of redundancy and relevance measures for feature selection in tissue classification of ct images. In *Industrial Conference on Data Mining*, pages 248–262. Springer, 2010.
4. Yudong Cai, Tao Huang, Lele Hu, Xiaohe Shi, Lu Xie, and Yixue Li. Prediction of lysine ubiquitination with mrmr feature selection and analysis. *Amino acids*, 42:1387–95, 04 2011.
5. Bihan Jiang, Michel F Valstar, and Maja Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 314–321. IEEE, 2011.
6. Gabriel Lippmann. Epreuves reversibles donnant la sensation du relief. *J. Phys. Theor. Appl.*, 7(1):821–825, 1908.
7. Yi Liu, Hongying Meng, Mohammad Rafiq Swash, Yona Falinie A Gaus, and Rui Qin. Holoscopic 3d micro-gesture database for wearable device interaction. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 802–807. IEEE, 2018.
8. Ville Ojansivu, Esa Rahtu, and Janne Heikkila. Rotation invariant local phase quantization for blur insensitive texture analysis. In *19th International Conference on Pattern Recognition (ICPR 2008)*, pages 1–4. IEEE, 2008.
9. Stuart J Russell and Peter Norvig. Artificial intelligence: a modern approach (international edition). 2002.
10. Garima Sharma, Shreyank Jyoti, and Abhinav Dhall. Hybrid neural networks based approach for holoscopic micro-gesture recognition in images and videos. In *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 808–814. IEEE, 2018.
11. Muhammad Atif Tahir, Chi-Ho Chan, Josef Kittler, and Ahmed Bouridane. Face recognition using multi-scale local phase quantisation and linear regression classifier. In *18th IEEE International Conference on Image Processing (ICIP)*, pages 765–768. IEEE, 2011.
12. ChunHong Wu, Malcolm McCormick, Amar Aggoun, and SY Kung. Depth mapping of integral images through viewpoint image extraction with a hybrid disparity analysis algorithm. *Journal of Display technology*, 4(1):101–108, 2008.
13. Weizhe Zhang, Weidong Zhang, and Jie Shao. Classification of holoscopic 3d micro-gesture images and videos. In *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 815–818. IEEE, 2018.
14. Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.