A Novel Pseudo Viewpoint based Holoscopic 3D Micro-gesture Recognition

Yi Liu
Brunel University London
London, UK
liuyi61april@gmail.com

Shuang Yang
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
shuang.yang@ict.ac.cn

Hongying Meng
Brunel University London
London, UK
hongying.meng@brunel.ac.uk

Mohammad Rafiq Swash
Brunel University London
London, UK
Rafiq.Swash@brunel.ac.uk

Shiguang Shan
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
sgshan@ict.ac.cn

ABSTRACT

Recently, video-based micro-gesture recognition with the data captured by holoscopic 3D (H3D) sensors is getting more and more attention, mainly because of their particular advantages to use a single aperture camera to embed the 3D information in 2D images. However, it is not easy to use the embedded 3D information in an efficient manner due to the special imaging principles of H3D sensors. In this paper, an efficient Pseudo View Points (PVP) based method is proposed to introduce the embedded 3D information in H3D images into a new micro-gesture recognition framework. Specifically, we obtain several pseudo view points based frames by composing all the pixels at the same position in each elemental image(EI) in the original H3D frames. This is a very efficient and robust step, and could mimic the real view points so as to represent the 3D information in the frames. Then, a new recognition framework based on 3D DenseNet and Bi-GRU networks is proposed to learn the dynamic patterns of different micro-gestures based on the representation of the pseudo view points. Finally, we perform a thorough comparison on the related benchmark, which demonstrates the effectiveness of our method and also reports a new state of the art performance.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; **3D imaging**; **Supervised learning by classification**.

KEYWORDS

Holoscopic 3D imaging, pseudo view points, micro-gesture recognition

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425249

ACM Reference Format:
Yi Liu, Shuang Yang, Hongying Meng, Mohammad Rafiq Swash, and Shiguang Shan. 2020. A Novel Pseudo Viewpoint based Holoscopic 3D Micro-gesture Recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425249

1 INTRODUCTION

Hand gestures have been widely used for natural human-computer interaction (HCI) over past decades. As an effective manner to capture the various types of both normal and micro gesture, holoscopic 3D (H3D) camera has attracted increasingly attention in recent years [12][6][8][3][10][5]. One particular advantage of H3D cameras, compared with most existing sensors such as Kinect and Leap Motion, is that 3D objects could be captured in 2D format by the special microlens-array of the camera sensor. Each microlens is responsible for recording the object from a particular view and all these microlens together are able to record the object from several different views in 3D scenes. Therefore, H3D sensors could reflect the 3D information in the final 2D images including the depth, angle, and so on. One sample captured by a H3D sensor is shown in Figure 1.

Due to the above advantages, H3D imaging system is able to capture widely viewing and rich information by a single aperture camera. However, the recorded depth and position information are all latent information, which can not be used directly. Thus, various pre-processing procedures are required to effectively extract images from different viewpoints and so as to display the 3D depth well. Due to the different purposes of the task, the previous work mainly used the geometric distortion correction, detection and Gaussian filters to perform viewpoint extraction. After a series of the above fine-tuning steps, the recorded images could be displayed in a very high quality. Inspired by these appealing displayed results, we try to introduce the viewpoints extraction for the recognition task for the first time. However, the traditional viewpoint extraction process involves several mannual settings and complicated details, which makes it not flexible enough for the final recognition task.

In this paper, we propose a novel pseudo view point based representation for micro-gesture recognition, together with a new recognition framework. Based on the principle of the integral imaging that depth information could be extracted from the omnidirectional holoscopic 3D images[2], we simulate this process to obtain pseudo view point based representations for recognition. Specifically, we learn the target gesture sequence by extracting information from several Pseudo View Points (PVP) of the H3D images automatically. The representation of each pseudo view point mimics the true observation of the sequence from a specific real view point. All the pseudo view points together could provide a relative complete representation of the target gestures and would be used as an initial representation of the gestures. Then a new CNN-BiGRU based framework is introduced to accomplish the final recognition task. The results on the HoMG dataset clearly show the effectiveness of our method and we also report a new state of the art performance.

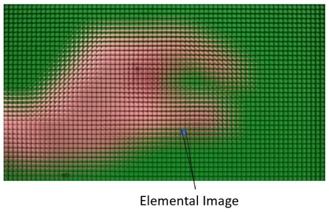

Figure 1: One Sample of Holoscopic 3D micro-gesture image and its Elemental Image (EI).

2 RELATED WORK

H3D imaging system is also called integral imaging, presents the real-world 3D scenes in 2D format by a micro-lens array in the aperture [1]. Each image is composed of several small, 2D, low-resolution images, which are named "elemental images" (EIs) as shown in Figure 1. Each micro-lens in the camera is responsible for recording the scene from a particular view, and different micro-lens would lead to different EIs. The difference among different EIs from these micro-lens could reflect the 3D information of the captured object.

In the past few years, many researchers have performed several attempts to extract the 3D information embedded in H3D images for a high display quality. For example, Wu et al. [9] introduced Hough Transform to correct the geometric distortion to extract valid viewpoint based images. Swash et al. [7] proposed to handle the nonlinear distortions. And some other works proposed to introduce third-party tools, such as Photoshop [2], to solve the distortions. Besides used as a device for display, the H3D camera has been introduced to capture data for micro-gesture recognition in recent years. In [10], LPQTOP [4] is used to extract the features in H3D images and then combined with Support Vector Machine (SVM) to recognize the micro-gestures. Later, Sharma et al. [6] introduced LSTM and GRU with the original H3D images as input to finish the recognition. Zhang et al. [12] introduced image-based subset features together with the video-based data and combined several convolutional neural networks to obtain the final recognition result. Recently, Qin et al. [5] proposed a method by combining LPQTOP features and a non-liner SVM classifier for recognition.

Different from the above recognition methods using the H3D images directly and ignoring the embedded 3D information, we propose a new method to extract the embedded 3D information for recognition. Instead of using a series of complicated procedures to extract accurate view points, we mimic and simplify the viewpoint extraction process based on the intrinsic principle of H3D imaging. At the same time, we also introduce a new CNN-BiGRUs framework for effective video-based micro-gesture recognition.

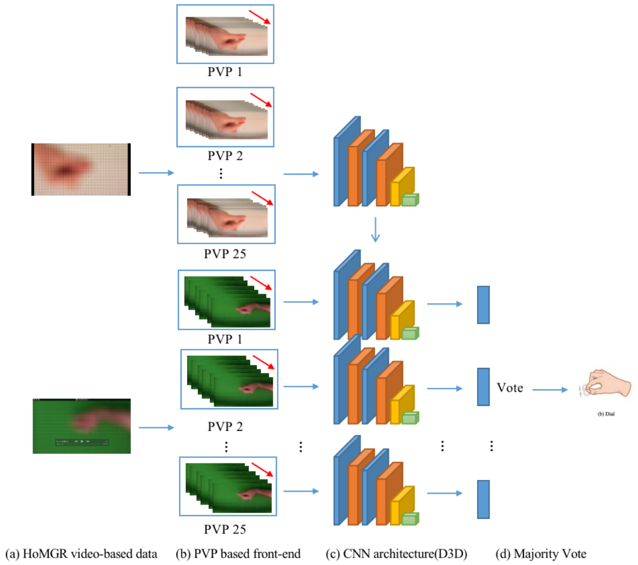

Figure 2: The video-based H3D micro-gesture recognition pipeline: (a) Sample video data of HoMG, (b) Front-end which consists the PVP extraction and pre-processing, and (c) the back-end which consist the deep network models and majority voting.

3 PSEUDO VIEWPOINT BASED H3D MICRO-GESTURE RECOGNITION

Fig. 2 gives the whole framework of our recognition pipeline, which contains three parts. Firstly, we represent each H3D frame in the input video $\mathbf{X_i}(\mathbf{i}=\mathbf{1},\mathbf{2},...,\mathbf{N})$ based on several Pseudo View Points (PVP), where N is the total number of videos. Each original H3D frame would lead to several different PVP based frames, and so each input video $\mathbf{X_i}$ will be transformed to several PVP based videos $\mathbf{X_i^{(1)}},\mathbf{X_i^{(2)}},...,\mathbf{X_i^{(P)}}$ where $\mathbf{X_i^{(p)}}$ denotes the p-th PVP based representation of video $\mathbf{X_i}$. All the PVP videos transformed from the same input video could represent different views of the same gesture in the original video. Based on this point, all the PVP based videos $\mathbf{X_i^{(p)}}$ can be used as training data to perform the recognition. When coming to the test process, each test video would also generate several PVP based videos and the model would output a prediction for each PVP. Then the majority voting is applied to get the final prediction of the input test video.

3.1 The Pseudo View Point (PVP) based Front-end

As the first step, we perform pseudo view point based image extraction from the original video frames, to generate the corresponding PVP based sequences $\mathbf{X_i^{(1)}},\mathbf{X_i^{(2)}},...,\mathbf{X_i^{(P)}}$ of each input video $\mathbf{X_i}(\mathbf{i}=\mathbf{1},\mathbf{2},...,\mathbf{N})$.

Compared with traditional viewpoint methods from correcting distortion to extracting different depth planes, the extraction of

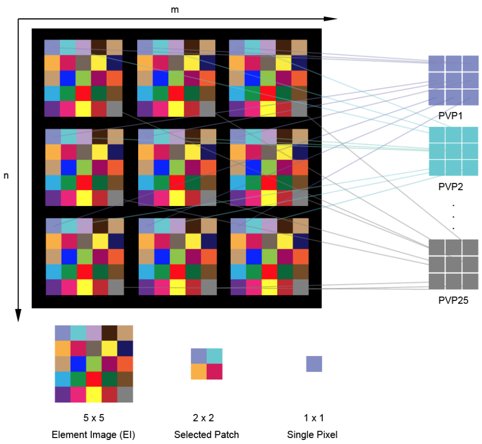

Figure 3: An example of the pseudo view point based image extraction.

PVP is convenient, efficient and can retain more informative pixels. In this work, we simplify this process by temporarily ignoring the distortions in the corner and the edges. Therefore, we ignore the distortions in the corner and the in-complete edges and leave the problem of resisting these residual noises to the subsequent recognition framework.

Specifically, for a H3D image of $H_{original}$ (1920) $\times$ $W_{original}$ 1080 pixels, we resize each image to $H \times W$ to make the size divisible by the size $h \times w$ of EI. Therefore, the number of EIs could be approximately $m \times n$. Figure 3 is an example to extract PVP based images from $m \times n$ EIs ($m = n = 3, h = w = 3$ in the figure3). The reconstructed PVP-based image consists of pixels at the same position in each EI. We select one pixel at a time and shift one-pixel horizontally or vertically. Therefore, the PVPs' resolution in the picture is $m \times n = 9$ pixels, and there are 25 PVP in total. In our experiments, $H_{original} = 1920, W_{original} = 1080, H = 1866, W = 999, h = 26, w = 26, m = 73, n = 41, selectedpatch = 5$. The final reconstructed PVP image is related to the size of the selected patch and the number of EI. We sample several local patches in each EI, and then obtains the final PVP frame by putting all the patches from all EIs together in a single frame. To make a fair comparison with the traditional methods, we extracted 16 PVPs because the traditional method only can extract 16 VPs. Then, PVP25 extraction can provide more depth information, and can also provide more features for training.

3.2 A New CNN-BiGRUs based Back-end

There are already a few methods that have begun to introduce deep learning methods to solve the problem of micro-gesture recognition. Whereas, these methods do not use the embedded depth information from H3D, which enables deep network better training and recognition. This paper is to simplify the method of extracting depth information used in the H3D imaging system of previous work and adjust it to be a front-end suitable for network training. In this paper, the D3D network is selected to extract features of the micro-gesture sequence. Compared with general problems, micro-gesture identification requires both spatial and temporal fine-grained information, then to correctly identify the corresponding micro-gesture and classify it.

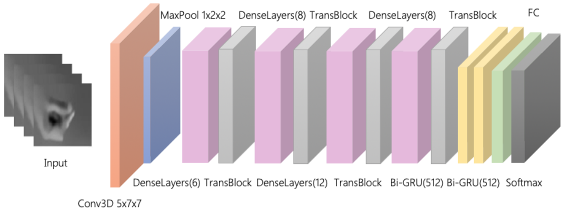
MaxPool 1x2x2
DenseLayers(8) TransBlock
DenseLayers(8)
TransBlock
FC
Input
DenseLayers(6) TransBlock DenseLayers(12) TransBlock Bi-GRU(512) Bi-GRU(512) Softmax
Conv3D 5x7x7

Figure 4: DenseNet in 3D version(D3D) architecture.

Here are two examples, such as b. Dial and c. Slider in Figure5. The two gestures belong to different actions. However, the difference is only the micro-movement on the fingertips, and it through the structure of the connecting frame in the DenseNet. The movement information in the sequence can be continuously enhanced to achieve better recognition. Here we refer to the structure design in [11] and use it to our main structure.

After going through the D3D network, we could obtain an intital representation of the input PVPs based seuqnece. To model the dynamic patterns in the sequence on the temporal dimension, we introduce a two-layer bi-GRU to take the output of the D3D network as input. Finally, the output at the last time step of the Bi-GRU is used as the final representation of the whole sequence. One fully-connected layer is followed to output the prediction probabilities after the softmax activations. For the whole network, we use the following Cross-Entropy loss (CE-loss) as the optimization goal for training. The the cross-entropy loss L_{CE} at each time step as follows:

$$L_{CE} = -\sum_{i=1}^{N}\sum_{p=1}^{P} y_i \log \hat{y}_i^{(p)}$$

where $\hat{y}_i^{(p)}$ is the prediction result of the model for the p-th PVP based representation of the i-th video.

4 EXPERIMENTS

4.1 Database

To evaluate our method, we perform a thorough comparison and analysis in the HoMG database, because it is the first public database, and it has many previous works that enable us to reference. There are 40 subjects in the dataset, each of whom performs three types of micro-gestures: button, slider, and dial respectively, as shown in Figure 5. The data involves both right-hand and left-hand samples in the set. All the samples are recorded with a static green or white background. There are 480 videos in total for training, 240 videos for validation, and 240 videos for testing [10]. There are 24 video recordings for each subject. 20 subjects are split to the training set and 10 subjects for the development and test set respectively.

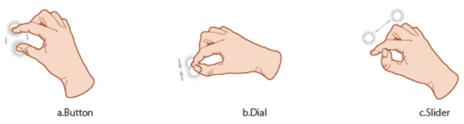

Figure 5: The three types of gesture.

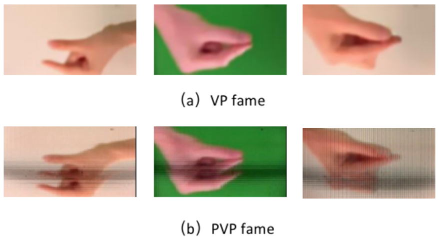

Figure 6: Pseudo viewpoint(PVP) frame and viewpoint frame. (a) obtained PVP frame, (b) obtained VP frame and the resolution is 340 by 185 pixels.

Table 1: Classification accuracy (%) comparison at development subset and testing subset respectively. "M.V." means "majority voting".

	Dev	Dev+M.V.	Test	Test+M.V.
Original	48.3	-	60.8	-
VP16	87.1	88.1	75.5	76.7
PVP16	90.1	89.1	82.2	83.5
PVP25	91.4	94.7	84.	85.2

Table 2: Classification accuracy (%) comparison between the proposed method with all the existing methods on the testing subset of HoMG database. "M.V." means "majority voting" GRU means "Gated Recurrent Unit". "n.SVM" means "non-Linear SVM"

Author	Methods	Acc
Liu et al.[10]	LBPTOP+SVM	59.5
Liu et al.[10]	LPQTOP+SVM	66.7
Sharma et al.[6]	LSTM	65.4
Sharma et al.[6]	GRU	69.2
Zhang et al.[12]	ResNet152+M.V.	82.0
Zhang et al.[12]	DenseNet161+M.V.	82.0
Zhang et al.[12]	SeResNet50+M.V.	82.0
Qin et al.[5]	LPQTOP+n.SVM+mRMR	84.6
This work	PVP25+D3D+M.V.	**85.2**

4.2 Implementation Details

In our experiments, all the frames are resized to 122 × 122 pixels and then cropped to 112 × 112 pixels. All the PVP frames are converted to gray-scale and normalized with their mean and variance. All the frames are cropped randomly but kept in the same random position if the frames belong to the same sequence. When coming to test, each frame is cropped in the centre position. We fed different PVP based sequences as augmented data into the training process, then used each PVP based representation of the test data to obtain an accuracy for each PVP. Finally, the majority voting strategy is used to obtain the final prediction from all the predictions.

4.3 Results

A. Ablation Study

In this part, we perform a thorough ablation study from the following three aspects. Firstly, we compare our proposed PVP based method with the original frame based and VP based method in Table 1. Firstly, we input the sequence composed by the original H3D frames to the model and get a baseline performance, to verify the ability of our architecture for the micro-gesture recognition task. In this setting, the frames are as shown in Figure 1 and it's very challenging to discriminate different micro-gestures. However, our architecture still achieves a high performance of 48.3% and 60.8% on the development set and the test set respectively. This result could clearly demonstrate the effectiveness of our proposed architecture. Secondly, we performed experiments using the traditional VP based frames together with our CNN-GRU based architecture to compare with our PVP based method. As shown in Table 1, it can be seen that the performance based on VP16 is not as good as PVP, and even the recognition rate is much lower. We think that the reason may be that a lot of features are lost in the extraction of VP. Finally, We also compare and analyze the VP extracted by the traditional method with PVP16 and PVP25 extracted by the proposed method. It can be seen that the performance of PVP16 is higher than that of the VP16. The reason may be that the VP16 feature lost during extraction, and PVP25 retains more original information.

B. Comparison with the SOTA In this part, we compare with other related methods, including the state of the art methods in Table 2. In the table, [10] is the first one to perform recognition on this dataset. They employed LBPTOP and LPQTOP based features together with SVM based classifiers and achieved 59.5% and 66.7% respectively. Then Sharma's paper [6] is the first one to introduce the viewpoint image for recognition. They introduced the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) based networks and achieved 65.4% and 69.2% respectively. The work [12] employed three very deep network based models and perform majority voting over the predictions of different models. Finally, they achieved an accuracy of 82% accuracy. The existing sate of the the art performance is reported by [5], which combined dynamic image feature extraction and a non-linear Support Vector Machine (SVM) classifier to evaluate HoMG database. Notably our final result is 85.4%, which surpass all prior results shown in the table.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a new PVP based micro-gesture recognition method. By extracting the pseudo view point based frames in an efficient and effective manner, the computation burden and complicated process of the traditional view point extraction could be relieved. Besides, we introduce a new deep network based architecture to model the video based micro-gesture recognition problem. Comprehensive experiments and comparison are carried out and the results have clearly show the robustness and effectiveness of our method. In general, the proposed PVP method is a robust and efficient method to extract the embedded information in H3D images. It can be applied to the pre-processing of any H3D image based problems, and the exciting experimental results may bring some insights to the community over related topics.

REFERENCES

[1] A. Aggoun. 2006. Pre-Processing of Integral Images for 3-D Displays. *Journal of Display Technology* 2, 4 (Dec 2006), 393–400. https://doi.org/10.1109/JDT.2006.

884691

[2] O. A. Fatah. 2015. *Post-production of holoscopic 3D image.* Ph.D. Dissertation. Brunel University London.

[3] C. Wang M. Peng and T. Chen. 2018. Attention Based Residual Network for Micro-Gesture Recognition. *13th IEEE International Conference on Automatic Face & Gesture Recognition* (2018), 790–794. https://doi.org/10.1109/FG.2018.00127

[4] V. Ojansivu, E. Rahtu, and J. Heikkila. 2008. Rotation invariant local phase quantization for blur insensitive texture analysis. In *19th International Conference on Pattern Recognition.* 1–4. https://doi.org/10.1109/ICPR.2008.4761377

[5] R. Qin, Y. Liu, M.R. Swash, H Meng, T. Lei, and T. Chen. 2019. A Fast Automatic Holoscopic 3D Micro-gesture Recognition System for Immersive Applications. In *The 15th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery.*

[6] G. Sharma, S. Jyoti, and A. Dhall. 2018. Hybrid Neural Networks Based Approach for Holoscopic Micro-Gesture Recognition in Images and Videos. *13th IEEE International Conference on Automatic Face & Gesture Recognition* (2018), 808–814. https://doi.org/10.1109/FG.2018.00130

[7] M. R. Swash, A. Aggoun, O. Abdulfatah, B. Li, J. C. Fernández, E. Alazawi, and E. Tsekleves. 2013. Pre-processing of holoscopic 3D image for autostereoscopic 3D displays. In *2013 International Conference on 3D Imaging*. 1–5. https://doi.org/10.1109/IC3D.2013.6732100

[8] Y. Zhang Y. Zhang X. Su and S. Liu T. Lei, X. Jia. 2018. Holoscopic 3D Micro-Gesture Recognition Based on Fast Preprocessing and Deep Learning Techniques. *13th IEEE International Conference on Automatic Face & Gesture Recognition* (2018), 795–801. https://doi.org/10.1109/FG.2018.00128

[9] C. Wu, M. McCormick, A. Aggoun, and S. Y. Kung. 2008. Depth Mapping of Integral Images Through Viewpoint Image Extraction With a Hybrid Disparity Analysis Algorithm. *Journal of Display Technology* 4, 1 (March 2008), 101–108. https://doi.org/10.1109/JDT.2007.904360

[10] M. R. Swash Y. F. Gaus Y. Liu, H. Meng and R. Qin. 2018. Holoscopic 3D Micro-Gesture Database for Wearable Device Interaction. *13th IEEE International Conference on Automatic Face & Gesture Recognition* (2018), 802–807. https://doi.org/10.1109/FG.2018.00129

[11] S. Yang, Y. Zhang, D. Feng, and C. Wang M. Yang, J. Xiao, K. Long, S. Shan, and X. Chen. 2019. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8.

[12] W. Zhang, W. Zhang, and J. Shao. 2018. Classification of Holoscopic 3D Micro-Gesture Images and Videos. *13th IEEE International Conference on Automatic Face & Gesture Recognition* (2018), 815–818. https://doi.org/10.1109/FG.2018.00131