Design Engineering

ISSN: 0011-9342
Issue: 11 | Pages: 714 - 722

# First-person View Gesture Recognition Based on Region Convolutional 3D Network

**Shentao Wang, Shang Zhang***
College of Computer and Information Technology, China Three Gorges University, Yichang, China
*Corresponding Author: Shang Zhang

***Abstract:***

Gesture is a visual and effective method of interaction with VR/AR devices. However, there are still many challenges for video-based first-person view gesture recognition. This paper proposes a method based on Region Convolutional 3D Network to detect and recognize gestures in undivided RGB-D video data. RGB data and Depth data were extracted with a model and feature fusion was conducted. Then generate candidate regions containing gestures in time. After that, classifies selected generates candidate temporal regions into specific gestures. By characteristic design of first-person view gestures, numbers and size of anchor boxes were optimized to accelerate the process of generating gesture candidate regions, where the quality of gesture candidate regions was enhanced and computation is saved. To verify the method and test effects under different backgrounds and illuminations, we tested with EgoGesture dataset, indicating validity of the method. Our method showed superiority in respect of interaction with wearable devices of VR/AR.

***Keywords:*** *Video analysis, Neural networks, Gesture recognition.*

## I. INTRODUCTION

A gesture is a common form for human exchange. Mankind also uses such a communication form in man-machine interaction. In interaction with wearable device (VR/AR helmets and glasses), a gesture is the most natural way with the wearable equipment interaction. Moreover, the image-oriented gesture data can be gained by a camera on the wearable equipment, dispensing with the extra hardware equipment. Therefore, accurate positioning gesture and recognition of the gesture in continuous video streaming data are considered to be one of the important parts in human-computer interaction research. But due to the non-determinacy of gesture activities (from large arm movements to tiny finger movements and even various gestures), it still has difficulties.

The early research on gesture recognition by computer vision was identified by using hand-crafted features [1]. These hand-made functions can express appearance and movement changes corresponding to gesture performance, but they only focus on single frames in continuous video.

ISSN: 0011-9342
Issue: 11 | Pages: 714 - 722

Recently, deep learning is used to many aspects of computer vision. End-to-end learning framework based on CNN and RNN is adopted to the gesture recognition [2-4]. Some research objects in studies on gesture recognition are video data of isolated gestures. Each part of video data only contains a gesture motion [5,6]. Gesture recognition is similar to temporal action detection. Both of them need to identify all motions in continuous video recognition. Gesture location is one of the main issues of continuous hand signal recognition task, because unlike continuous hand signal, continuous hand signal recognition is more difficult than isolated gestures at ambiguous boundaries of unsegmented sequences. In the Region Convolutional 3D Network method, proposal subnet is used to realize time division for time proposal with motions [7]. This paper improves the proposal subnet to accelerate the proposal process and to gain the high quality of activity segments for proposals.

Due to complicated illumination and background, depth images are introduced to obtain information with more body distance information. Depth image information is good for gesture localization of gestures from the first-person perspective. Relevant work of gesture recognition has proved that introduction of depth images can effectively improve recognition effect [8-10]. This paper targets at RGB-D video data to do end-to-end training and integrates with features of RGB image data and depth image data in the feature extraction stage. The main steps of the proposed method is given in Fig 1.

Work contributions are involved in the following aspects. Firstly, feature fusion is used to effectively introduce depth image information to the end-to-end training. Secondly by aiming duration features of gesture motion, proposal process of time quantum is optimized. Finally, EgoGesture dataset is verified to prove the frame effectiveness.

RGB image features are integrated with depth image features in the network. The active gesture segment is proposed. Finally, boundary and motion types of the active gesture segment are refined.

ISSN: 0011-9342
Issue: 11 | Pages: 714 - 722

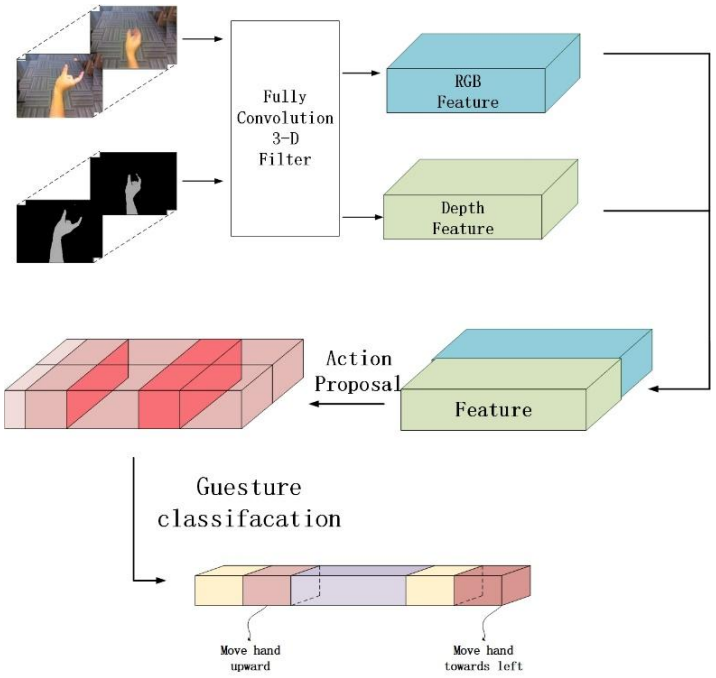


Fig 1: The framework is detected by aiming at the continuous RGB-D video stream.

## II. RELATED WORK

There are many accumulations of research results on the gesture recognition. In the initial phase, handcrafted functions for gesture recognition were used, such as hidden Markov model (HMM) [10], and so on. Afterwards, traditional features were expanded to the time-space domain by virtue of 3D HOG or other achievements, enhancing effectiveness of features [11]. These handmade features can represent changes in look, shape and movement that match gesture performance, but these methods rely on dense sampling of the video. With the rapid development of deep learning, neutral network is used to each field of computer vision, without exception for the gesture recognition. For example, Gesture recognition is performed by way of modeling the context information of the gesture sequence by RNN and C3D [11,12]. The above method uses hand positions to achieve temporal segmentation, which is not adaptable to complex environments. Recurrent 3D CNN is used to solve a problem that model long-term dependencies are always lost in the gesture time-domain.

Bandback et al. [13] Segmenting a hand from an egocentric video using CNN and using the segmented handmask image as input to another CNN to identify one's hand movement. Then use these heatmaps and hand covers with various classifiers to classify your hand activity. Use to detect self-hand activity. This method iteratively extracts the current region associated with the hand based on previous hand positions and computes features on the extracted regions. Then enter these functions into another network to recognize gestures.

In the past several years, video analysis task of classifies from the isolated action video to temporal action detection in continuous video. In earlier research, detection of action in

ISSN: 0011-9342
Issue: 11 | Pages: 714 - 722

continuous video is to segment the video through the time sliding window, and then use the classifier algorithm to classify the actions in each window to solve the task, the active boundary of such method positioning not flexible. Later, the method of judging the integrity of the action by dividing. The actions in the video into three stages has improved the flexibility of the proposal boundary to some extent. Region Convolutional 3D Network (R-C3D) uses the multi-task study to combine proposal candidate section with classification to realize end-to - improving learning. Such a procedure avoids iterative calculations and creates flexible supply limits. This article also suggests an R-C3D gesture-oriented model for effective gesture recognition because of its excellent performance.

## III. CONVOLUTIONAL 3D NETWORK OF FEATURE REGIONS

In this section, we describe improvements to Region Convolutional 3D Network (R-C3D) so that it can effectively detect gestures in RGB-D video.

R-C3D is an effective temporal action detection framework as shown below Fig 1. Inspired by Faster-RCNN, it divides the temporal action detection task into two subnets: Proposal Subnet and the Classification Subnet. The network consists of three components. Frist, R-C3D method extracts the feature of the input video stream data through the feature extraction network. Import features into the Proposal Subnet and the Classification Subnet, respectively, and then the Proposal Subnet generates the proposal segment that may contain the actions. Finally, the Classification Subnet classifies the proposals into specific action types or backgrounds and further refine the proposal segment.

The improvement on R-C3D is involved in two aspects. The RGB and depth image features are extracted from the C3D model, then new features are formed by merging and entered into subsequent networks for assessment. The first is the amalgamation of two features. Adding depth features by merging features on the time axis can effectively improve the recognition effect. Second, adding the priori knowledge of the gesture duration to the proposal self-network and designing anchor segments with pertinence.

3.1 Feature Extraction and Feature Integration

When the input data is a video rather than a static picture, the features in the time dimension also need to be learned. C3D model can automatically learn about these features simultaneously from time and space [14]. The whole architecture of the C3D consists of 8 convolutional layers, 5 pooling layers, 2 fully connected layers of size 4096 and final softmax layer to output predicted label. Our model inputs a series of RGB video frames and depth video frames with some dimensions (R3*L*H*W 和 D3*L*H*W). The convolutional layer of C3D model (conv1a to conv5b) respectively extracts features of C3D model (conv1a to conv5b) respectively extracts features for RGB data and depth data to obtain the feature (Rxxx and Dxxx). Considering that C3D model is used as the pre-trained model, the height (H) and width (W) of the frames are 112. The frame (L) can be arbitrary, but it is only limited by the equipment. The main steps of the proposed method is given in Fig 2:

ISSN: 0011-9342
Issue: 11 | Pages: 714 - 722

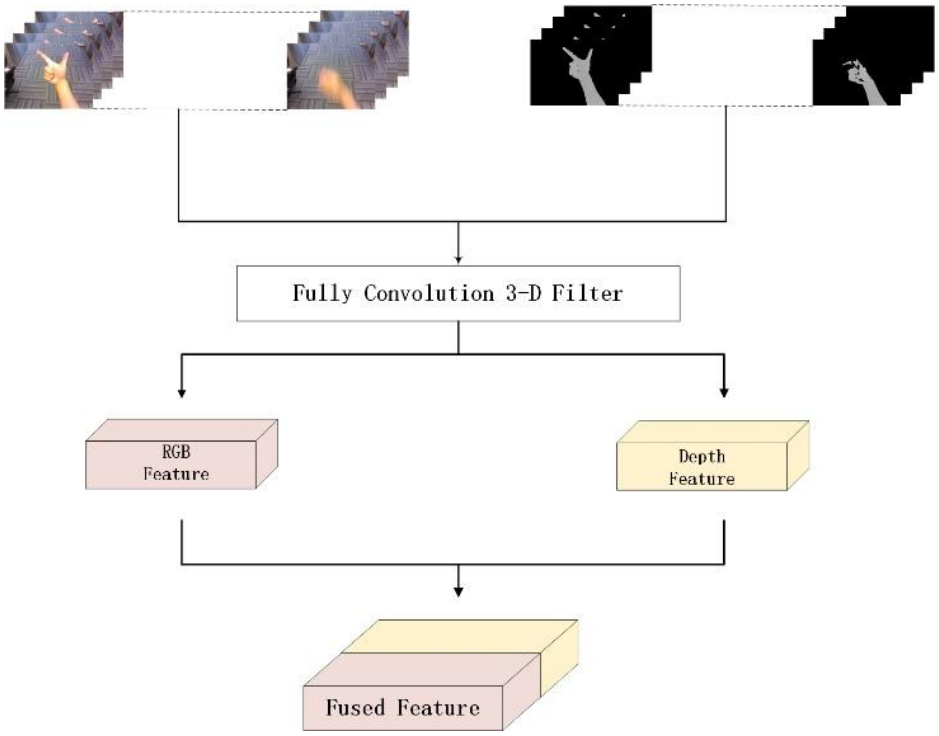


Fig 2: The process of feature fusion

Hand distance change information introduced by depth features causes a great influence on gesture localization. Performs a functional integration of RGB and depth functions to effectively use the visual information provided by the different channels. In order to utilize the pre-training C3D model, we adopt the fusion of the feature level and connect the feature vector of RGB and deep channels. The fusion process is as follows:

$$CONV5b_{rgb} = C3D(v_{prgb})$$
$$CONV5b_{depth} = C3D(v_{pdepth})$$
$$F = CONV5b_{rgb} \oplus CONV5b_{depth} \quad (1)$$

3.2 Anchor Segments Design

The Proposal Subnet phase of the R-C3D model introduces anchor segments into the time proposal subnet so that the model can predict variable length proposals. Predict if there are potential suggestions sections and if the forecast suggestions contain activity. When generating suggestions, we generate a set of sparse, class-independent suggestions by classifying a set of rescaling anchors at each location on the feature map.

The EgoGesture dataset consists of 83 gestures designed for human-computer interaction, each of which is meaningful, natural and rememberable by the users. They cover most operations and communication operations for wearables devices. According to the features of man-machine interaction, the duration of gestures used to operate the device is not too long, and

ISSN: 0011-9342
Issue: 11 | Pages: 714 - 722

the difference in the duration of gestures is not too large. On the grounds of the statistics of the continuous frames of EgoGesture dataset, the duration of most gesture ranges from m to n frames. The anchor segments are designed according to this feature. Because the frame dimension L in the original video is compressed to L/8 after feature extraction, the pre-defined anchor segment is k segment evenly distributed between m/8 and n/8. The anchor segments designed based on the priori knowledge of the gesture duration are more targeted that it can reduce the number of k and further reduce the total number of segments (L/8)*k. Ultimately, the time required for temporal proposal process is reduced and the quality of generated proposal segments is improved. The major steps of the proposed method is given in Fig 3.

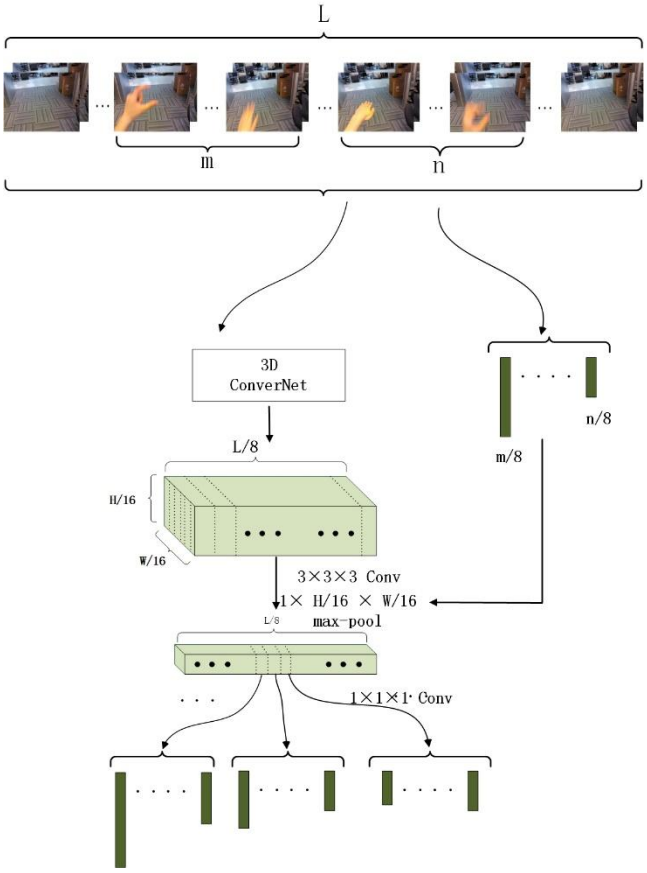


Fig 3: The process of proposes candidate activities of variable length

## IV. EXPERIMENT AND DISCUSSION

4.1 Dataset

We tested our model on the latest and largest self-action center gesture dataset EgoGesture [14]. The data set consists of 2081 RGB-D videos recorded by 50 different Subjects in 6 Scenes, containing 24161 gesture samples from 83 gesture categories. The EgoGesture dataset has a certain complexity. It contains background changes, lighting changes, changes in the

ISSN: 0011-9342
Issue: 11 | Pages: 714 - 722

subject's motion state, etc. in different Scenes. The dataset simulates the use of wearable devices as much as possible. We conducted the experiment by dividing the EgoGesture dataset into a training set (80%) test set (20%). The dataset schematic is given in Fig 4.

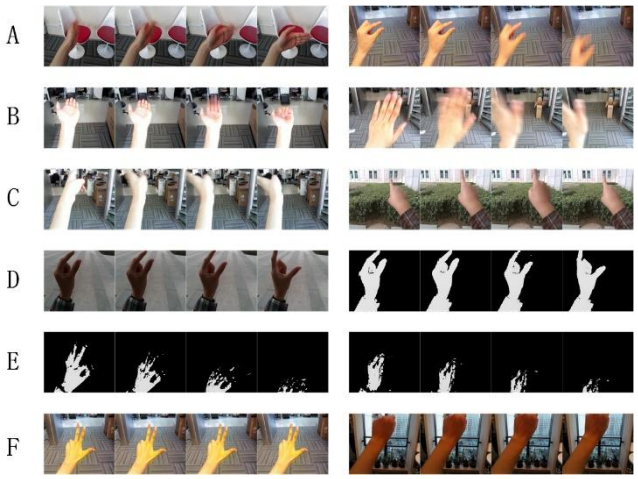


Fig 4: Some examples to demonstrate the complexity of EgoGesture dataset. (A)Different background; (B) Motion blur; (C) Indoor & outdoor environment; (D) RGB & Depth; (E) Motion blur; (F) Illumination change

4.2 Performance Evaluation Indicators

Our goal is to locate and recognize gestures in successive videos. We use the Jaccard index to evaluate the performance of the model.

Let Gs, Ps be the binary indicator vector of the sequence s, where 1 value corresponds to the frame in which the i-th gesture is being performed, and the following is the definition of the Jaccard index value of the i-th class:

$$J_{s,i} = \frac{G_{s,i} \cap P_{s,i}}{G_{s,i} \cup P_{s,i}} \tag{2}$$

Here Gs is the ground truth of the i-th gesture label of sequence s, and Ps is the prediction. When both Gs and Ps are empty, Js is 0 at this time. For a sequence s with a unique real label, its Jaccard index is calculated as follows:

$$J_s = \frac{1}{l_s} \sum_{i=1}^{L} J_{s,i} \tag{3}$$

Finally, the final evaluation metric is the average of the Jaccard index for all testing sequences.

$$\overline{J}_S = \frac{1}{n} \sum_{j=1}^{L} J_{sj} \tag{4}$$

4.3 Experimental Results

We tested the model on a single GTX1080ti GPU and Intel i5-8600K CPU to get the speed of our model and the Jaccard index. We used PyTorch for software, NVIDIA RTX 2080ti is

ISSN: 0011-9342
Issue: 11 | Pages: 714 - 722

used to train the network model. The experimental are shown in the following TABLE I. Compared with existing methods, our model performs better in terms of runtime and Jaccard.

In the real-time gesture detection task, the method in this paper has a good result compared with the reference algorithm on the basis of comprehensively considering the two indicators of running speed and accuracy.

It should be noted that the experimental of other comparison models on the dataset are based on the results of Cao [4]. The experimental platform is a single GTX Titan. X GPU platform [13].

**TABLE I. Experimental results**

| Method | Jaccard | Runtime |
|---|---|---|
| C3D-l16s16 | 0.618 | 312fps |
| C3D-l16s8 | 0.698 | 156fps |
| C3D+STTM-l16s8 | 0.709 | 111fps |
| Ours (on GTX1080ti) | 0.684 | 286fps |

## V. CONCLUSIONS AND PROSPECTS

In this work, we improved the R-C3D for self-centered gesture recognition tasks and demonstrated the validity of the model on the EgoGesture dataset. We merged the depth features with the RGB features and proposed. The stage introduces a priori knowledge of the duration of the gesture to achieve faster detection speeds.

We will further explore the following work: 1) Study the effects of different feature extraction models on RGB-D feature fusion. 2) Explore the impact of different anchor segments on the proposal stage and study the changes in the perception.

## REFERENCES

[1] Liu L., Shao L., (2013) Learning discriminative representations from rgb-d video data, IJCAI, 1.

[2] Molchanov P., Gupta S., Kim K., Kautz J., (2015) Hand gesture recognition with 3d convolutional neural networks, CVPRW.

[3] Molchanov P., Yang X., Mello S. D., Kim K., Tyree S., Kautz J., (2016) Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks, CVPR.

[4] Cao C., Zhang Y., Wu Y., Lu H., Cheng J., Egocentric Gesture Recognition Using Recurrent 3D Convolutional Neural Networks with Spatiotemporal Transformer Modules.

[5] Wang H., et al. (2017) Large-scale Multimodal Gesture Recognition Using Heterogeneous Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[6] Wang P., et al. (2016) Large-scale isolated gesture recognition using convolutional neural networks. Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE.

[7] Xu H., Das A., Saenko K. (2017) R-C3D: region convolutional 3d network for temporal activity detection, CoRR, abs/1703.07814. 8.

ISSN: 0011-9342
Issue: 11 | Pages: 714 - 722

[8] Duan J., Wan J., Zhou S., Guo X., Li S., (2017) A unified framework for multi-modal isolated gesture recognition, ACM Transactions on Multimedia Computing Communications and Applications (TOMM)(Accept).

[9] Wan J., Zhao Y., Zhou S., Guyon I., Escalera S. Li S. Z, (2016) Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition, CVPR Workshops,.

[10] Chai X., et al. (2016) Two streams recurrent neural networks for large-scale continuous gesture recognition. Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE.

[11] Klaser A., Marszałek M., Schmid C., (2008) A spatio-temporal descriptor based on 3d-gradients, Proc. BMVC 19th Brit. Mach. Vis. Conf., 1-10.

[12] Liu Z., Chai X., Liu Z., Chen X. (2017) Continuous gesture recognition with hand-oriented spatiotemporal feature. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. ages 3056-3064. 3, 5.

[13] Bambach Sven S., Lee Stefan, Crandall David J., Chen Yu, (2015) Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions, The IEEE International Conference on Computer Vision (ICCV), December.

[14] Cao C, Zhang Y., Wu Y., Lu H., Cheng J., (2017) Egocentric Gesture Recognition Using Recurrent 3D Convolutional Neural Networks with Spatiotemporal Transformer Modules, In Proceedings of IEEE International Conference On Computer Vision (ICCV), Venice, Italy.