electronics

MDPI

Article

Two Stage Continuous Gesture Recognition Based on Deep Learning

Huogen Wang [1,2]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; hw823@uowmail.edu.au
[2] Hithink RoyalFlush Information Network Co., Ltd., Hangzhou 310012, China

Abstract: The paper proposes an effective continuous gesture recognition method, which includes two modules: segmentation and recognition. In the segmentation module, the video frames are divided into gesture frames and transitional frames by using the information of hand motion and appearance, and continuous gesture sequences are segmented into isolated sequences. In the recognition module, our method exploits the spatiotemporal information embedded in RGB and depth sequences. For the RGB modality, our method adopts Convolutional Long Short-Term Memory Networks to learn long-term spatiotemporal features from short-term spatiotemporal features obtained from a 3D convolutional neural network. For the depth modality, our method converts a sequence into Dynamic Images and Motion Dynamic Images through weighted rank pooling and feed them into Convolutional Neural Networks, respectively. Our method has been evaluated on both ChaLearn LAP Large-scale Continuous Gesture Dataset and Montalbano Gesture Dataset and achieved state-of-the-art performance.

Keywords: gesture segmentation; gesture recognition; weighted rank pooling; dynamic image; 3D Convolutional LSTM Network

Citation: Wang, H. Two Stage Continuous Gesture Recognition Based on Deep Learning. *Electronics* **2021**, *10*, 534. https://doi.org/10.3390/electronics10050534

Academic Editor: Hugo Proença

Received: 2 December 2020
Accepted: 20 February 2021
Published: 25 February 2021

1. Introduction and Related Works

Gesture recognition is an attractive research direction because of its wide application in virtual reality, human–computer interaction, and sign recognition. However, it is also a big challenge for the research of continuous gesture recognition, because the number, order, and boundaries of gestures were unclear in a continuous gesture sequence [1]. Both the temporal segmentation and the recognition problems need to be solved in continuous gesture recognition. In fact, temporal segmentation and gesture recognition can be solved separately.

One typical challenge in continuous gesture recognition is temporal segmentation. The position and motion of hands were often employed for temporal segmentation [2,3]. However, these methods were sensitive to the complex background and built upon accurate hand detection. Sliding window is also a promising skill to obtain gesture instances with 3D convolutional neural networks (3DCNN) [4]. Therefore, the computation of 3DCNN is expensive and the length of the sliding volume is fixed. To overcome the drawbacks of these works, a binary classification is proposed for temporal segmentation. As shown in Figure 1, video frames can be classified into gesture frames that cover useful hand movement and transitional frames between adjacent gestures. We believe that appearance information and hand motion information are complementary in temporal segmentation. Therefore, a novel temporal segmentation method was proposed to distinguish between gesture frames and transitional frames by combining both appearance information and hand motion information.

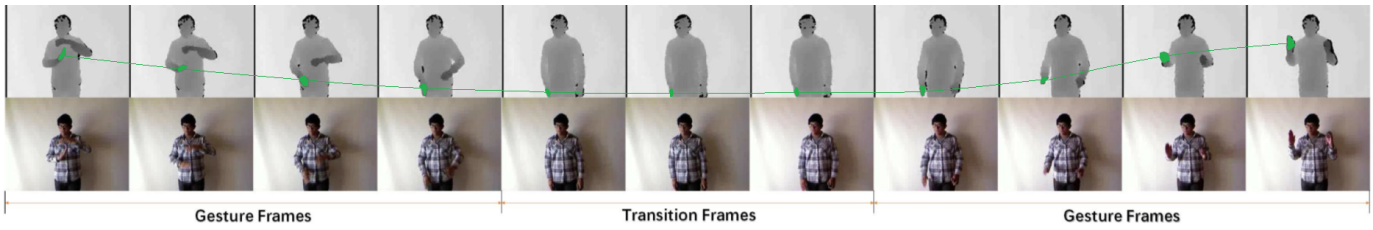

Figure 1. The sample gesture sequence. A continuous gesture sequence is composed of gesture frames and transitional frames between two gestures. We found that both the appearance information and hand motion information are useful for temporal segmentation.

After temporal segmentation, a continuous gesture sequence can be divided into several isolated gesture sequences. Therefore, isolated gesture recognition methods can be employed for the final recognition. Several attempts have been made to recognize gestures from RGB-D sequences with deep learning, including ConvNets combined with an RNN [5–10], 3D CNN [11–19], Two-stream CNNs [20–26], and Dynamic Image (DI)-based methods [27–32]. However, we argue that appropriate gesture recognition methods need to be selected according to the difference characteristics of RGB modality and depth modality. Therefore, we propose a novel gesture recognition network, which deals with RGB and deep modality in different ways, respectively. For the RGB modality, the proposed method adopts 3D ConvLSTM [9] to learn spatiotemporal features from video frames of a RGB sequence and its saliency sequence. An example of a RGB sequence and its saliency sequence was shown in Figure 2. For depth modality, inspired by the outstanding performance of rank pooling [27,28,30,31,33–35], this paper employs weighted rank pooling [36] to encode depth sequences into Depth Dynamic Images (DDIs). To overcome temporal information loss, DMDI is also extracted from the absolute differences (motion energy) between consecutive frames of a depth sequence with weighted rank pooling. Then, both DDIs and DMDIs are fed into ConvNets for final recognition. Finally, multiple 3D ConvLSTMs and ConvNet are fused together by late fusion.

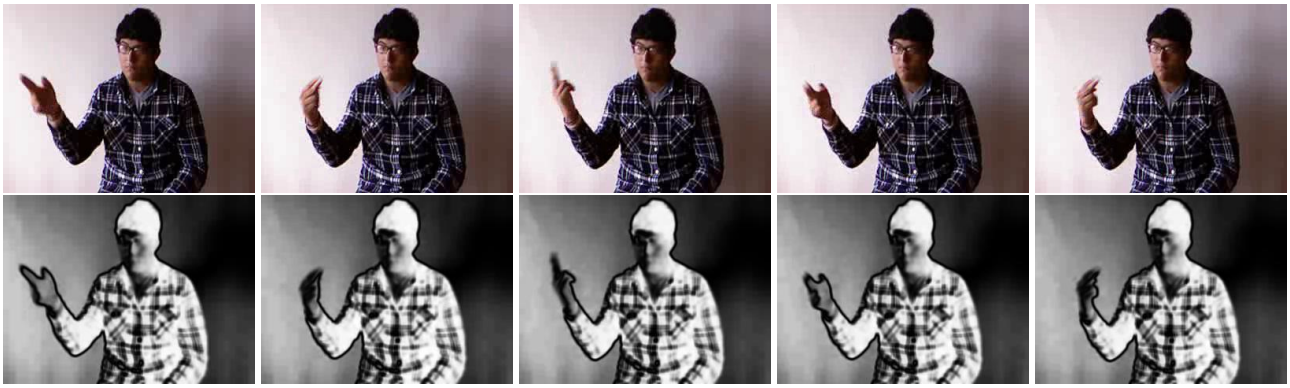

Figure 2. Illustration of a RGB sequence (**top**) and its saliency sequence (**bottom**).

The proposed method achieved state-of-the-art performance on the ChaLearn LAP ConGD Datasets [37] and Montalbano Gesture Recognition Dataset [38]. Part of the work [39] was reported in Chalearn Challenges on Action, Gesture, and Emotion Recognition: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions @ICCV17 [40]. The key contribution of this paper is to segment the continuous gesture with both the appearance information and the hand motion information, and to encode the geometric, motion and structural information based on the different characteristics of the RGB modality and depth modality. Compared with the conference paper [39], the extension includes:

1. Temporal segmentation with both the appearance information and the hand motion information;

2. The bidirectional rank pooling in [39] is replaced with the weighted rank pooling [36] to capture sequence-wide temporal evolution;
3. The method is also evaluated on Montalbano Gesture Dataset in addition to the ChaLearn LAP ConGD Datasets and state-of-the-art results are achieved;
4. More analysis and discussion are presented in this paper.

The remainder of this paper is organised as follows. Section 2 gives the details of the proposed temporal segmentation and gesture recognition method. Section 3 presents the experiments to verify the effectiveness of the proposed method and the discussions. The paper is concluded in Section 4.

2. Proposed Method

As shown in Figure 3, our proposed method consists of two steps: temporal segmentation and gesture recognition. Given a continuous gesture sequence, we must determine beginning and ending frames of gestures, this problem refers to temporal segmentation. Then, each segmented gesture must be assigned a label.

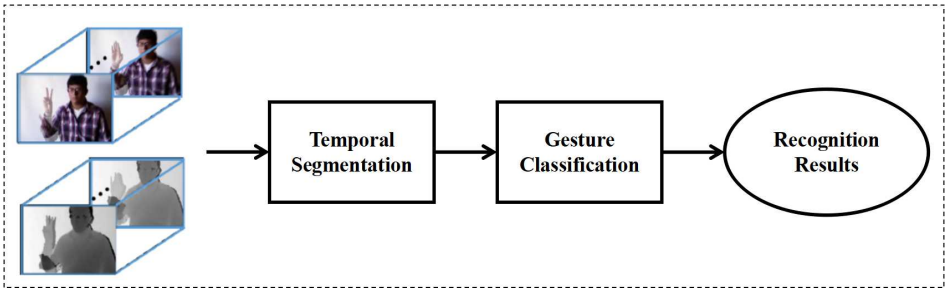

Figure 3. The overview of our proposed method for continuous gesture recognition. The proposed method consists of two phases: temporal segmentation and gesture recognition.

2.1. Temporal Segmentation

As shown in Figure 1, video frames can be divided into gesture frames and transition frames with a binary classification problem. To address this problem, both the appearance information and the hand motion information are employed to classify video frames in RGB and depth sequences. Generally, one will put their hands down after performing a gesture. Therefore, hand positions can be a wise way to realize temporal segmentation. Faster R-CNN [41] was adopted to detect the hand regions, due to the excellent performance of Faster R-CNN in object detection. Then, the height of hands in each frame was obtained and the average height of the initial several frames was treated as the height threshold. As shown in Figure 4, if one hand was first higher than the height threshold, it could be considered as the beginning of a new gesture. If both hands were lower than the height threshold, it could be considered as the ending of a gesture.

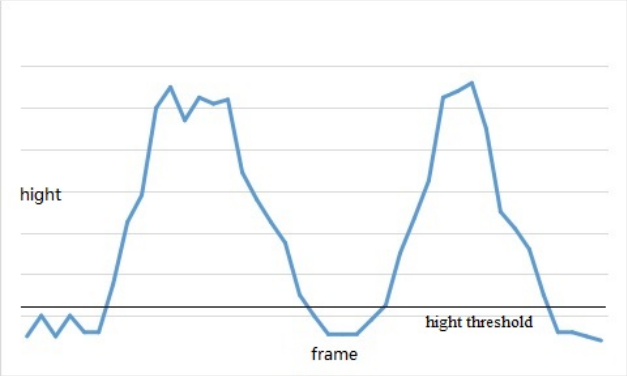

Figure 4. An example of the temporal segmentation result with hand positions for a continuous gesture sequence.

To take full advantage of the appearance information, two stream ConvNets were used for temporal segmentation. As shown in Figure 5, two stream ConvNets are combined by late fusion. The details of the training are presented in Section 3.1.1. We can use this method to assign "transition frames" or "gesture frames" to each frame.

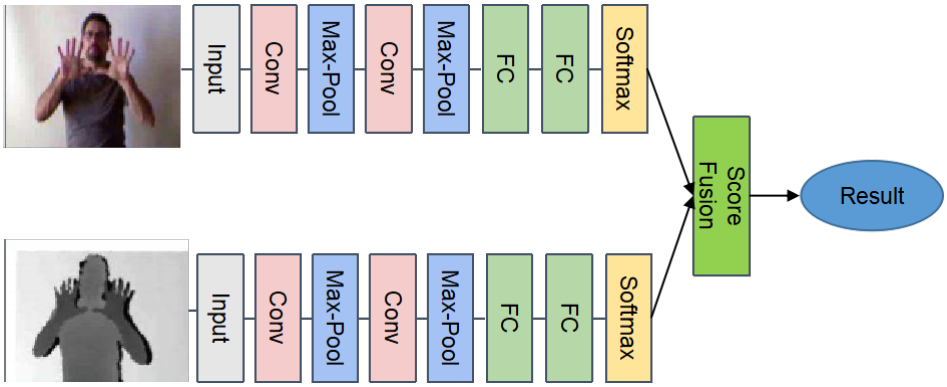

Figure 5. Two stream Convolutional Neural Networks (CNNs) for temporal segmentation.

Finally, the segmentation result was obtained by the fusion of both the above results. As shown in Figure 6, the beginning and the end of each gesture are typically transitional frames. The middle frame of transitional frames is treated as the final boundary between two gestures.

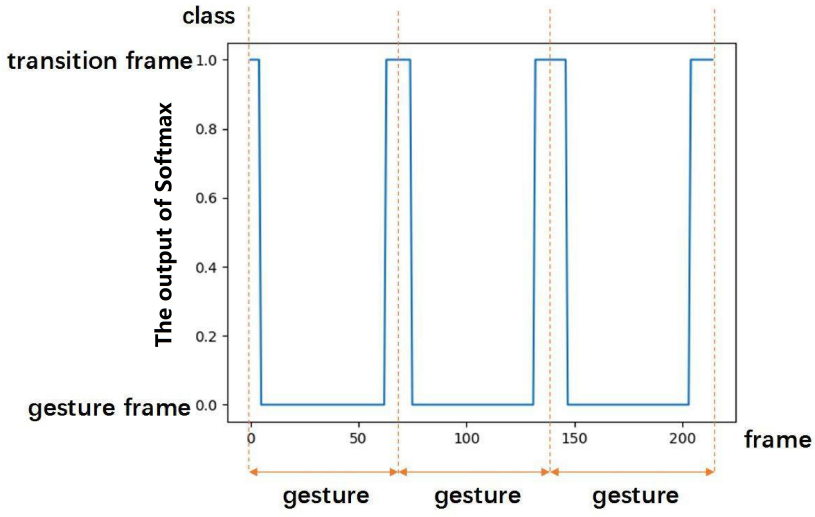

Figure 6. An example of the temporal segmentation results. The transitional frames between adjacent gesture are the boundaries. The sequence is segmented into three isolated gesture sequences, the middle point of transitional frames is defined as the boundary of two gestures.

2.2. Proposed Gesture Recognition Network

Taking into account the different characteristics of the RGB and depth modality, a novel gesture recognition framework is proposed. The overview of the proposed gesture recognition framework is shown in Figure 7.

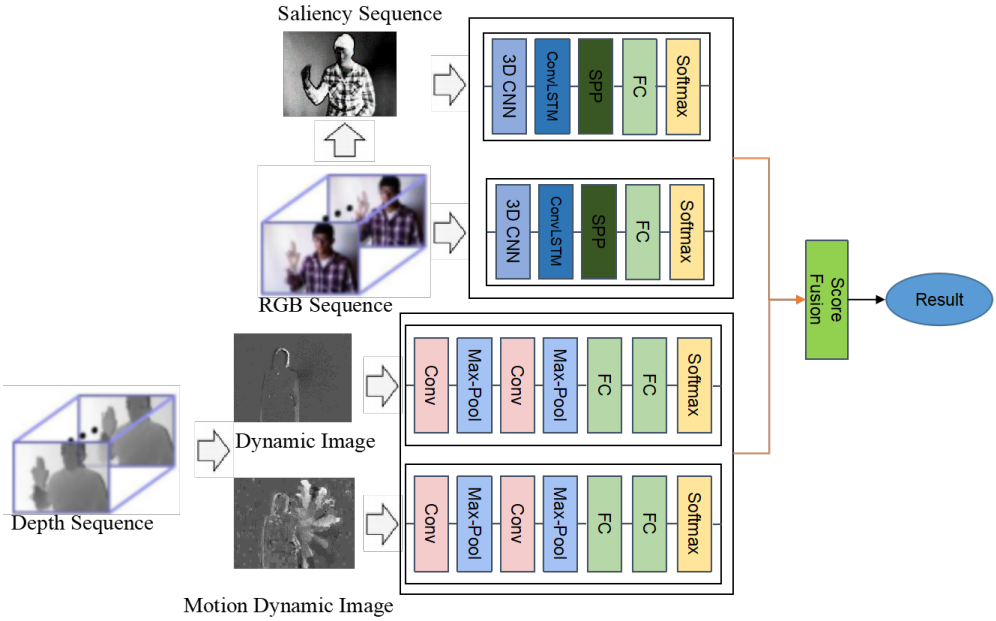

Figure 7. The overview of the proposed gesture recognition framework.

2.2.1. Gesture Recognition for Depth Modality

Firstly, four sets of dynamic images, including Depth Dynamic Images (DDIs) and Depth Motion Dynamic Images (DMDIs), are generated from a depth sequence through bidirectional weighted rank pooling [36]. Weighted rank pooling takes into account the fact that frames in a sequence and regions in frames have varying importance.

Construction of Dynamic Images

Dynamic images are formed by applying weighted ranking pool in a bidirectional way directly to the pixels of the video sequence. DDIs are generated from depth sequence, whereas DMDIs are constructed from the absolute differences between consecutive frames through an entire depth sequence. In this paper, the temporal weight of the frame is calculated with the average flow magnitude and the spatial weight of each pixel is the flow magnitude of that pixel. Then each dynamic image is fed into a ConvNet for classification. Figure 8 gives an example of the dynamic images, showing that DMDIs can be used as a complement to DDIs to preserve both structural information and motion cues.

2.2.2. Gesture Recognition for RGB Modality

The 3D ConvLSTM network proposed in Zhu et al. [9] includes input preprocessing, 3D Convolutional Network (3D CNN), Convolutional LSTM (ConvLSTM), Spatial Pyramid Pooling (SPP), Fully Connected Layer (FC), and Softmax. Input preprocessing adopts uniform sampling with temporal jitter based on pyramid input to sample each sequence into a fixed length. Then, the video sequence is fed into the 3D CNN [12] to learn short-term spatiotemporal features. Two-level ConvLSTM is adopted to learn long-term spatiotemporal features from short-term spatiotemporal features. The output of the top ConvLSTM layer is fed into SPP [42]. The full-connected layer is added on the top of SPP and connected to Softmax. Different from [9], both RGB sequence and its saliency sequence extracted using the algorithm described in [43] are fed into the 3D ConvLSTM networks.

2.2.3. Score Fusion for Classification

Given a pair of RGB and depth video sequences, the RGB sequence and its saliency sequence are fed into trained 3D ConvLSTM networks, and DDIs and DMDIs are fed into trained ConvNets separately. The outputs of all networks are nomalized using L_1 norm

and fused by average-score fusion in an element-wise way. The index of the max score in the resultant vector is assigned as the label.

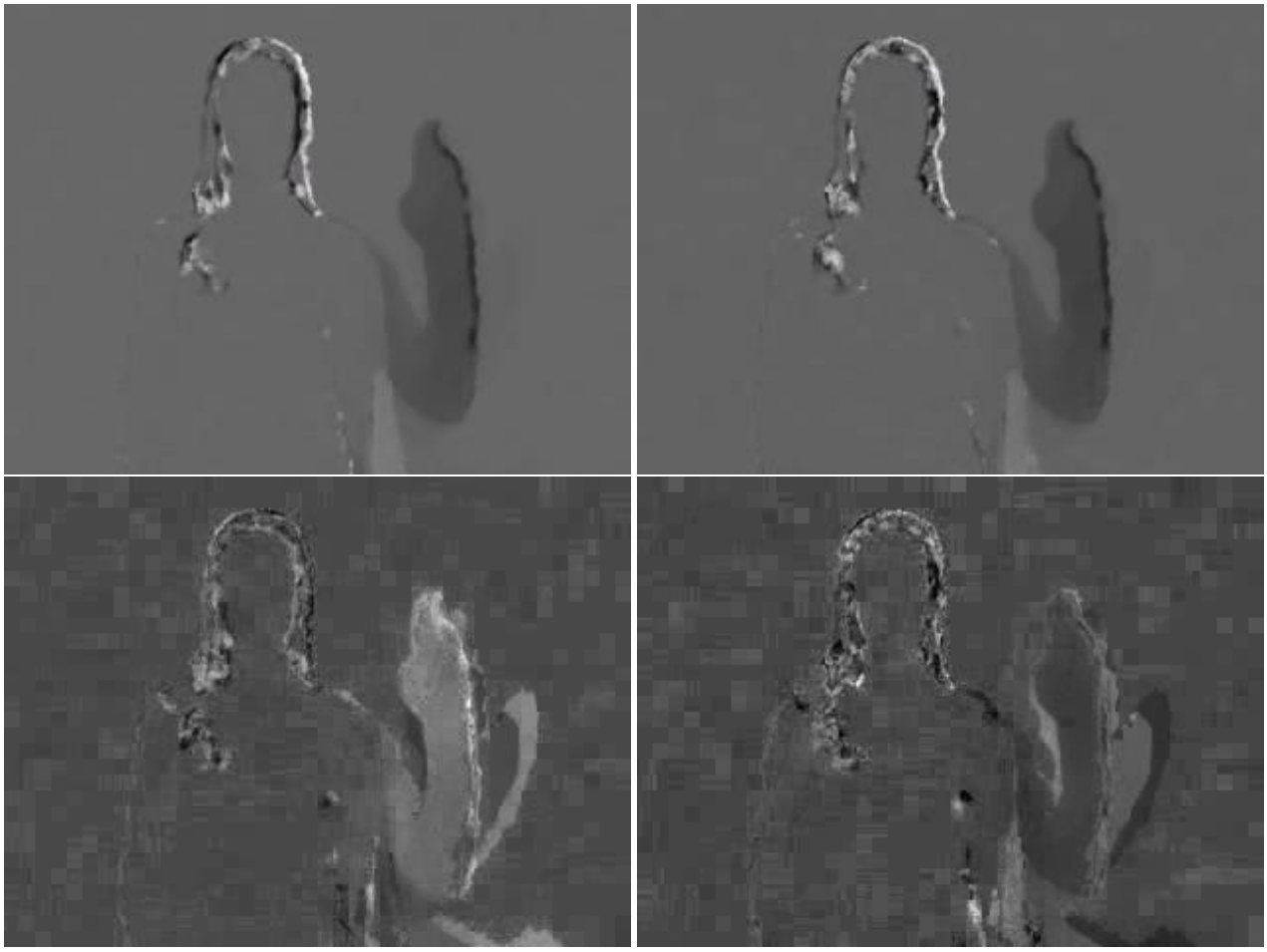

Figure 8. Samples of generated forward and backward Depth Dynamic Images (DDIs) and Depth Motion Dynamic Images (DMDIs) for gesture Mudra1/Ardhapataka, the left images are dynamic images for forward, the right images are dynamic images. From top to bottom: DDIs and DMDIs.

3. Experiments

The proposed method was evaluated on ChaLearn LAP ConGD Dataset [37] and Montalbano Dataset [44]. The evaluation protocols of continuous gesture recognition is Jaccard index (the higher the better). The network training and experimental results of the proposed methods on the dataset were reported.

3.1. Network Training

3.1.1. Network Training for Temporal Segmentation

To train the ConvNets for temporal segmentation, a dataset was collected for the binary classification. In the dataset, training samples of the class "transitional frames" were collected from eight frames around the bounary of two gestures, and training samples of the class "gesture frames" were picked from the rest frames. VGG-16 [45] was fine-tuned for temporal segmentation from the pre-trained models on ImageNet [29]. Both networks were trained using mini-batch SGD with the momentum and weight decay being set to 0.9 and 0.0001, respectively. The batch-size was 64. The activation functions in all hidden layers were RELU. To fit the input size of VGG-16, the input images were resized into 224×224. The initial learning rate was 0.01 and decreased to $\frac{1}{10}$ its every 40K iterations.

The training underwent 90K iterations. The VGG-16 was implemented with Tensorflow and trained on one TITAN X Pascal GPU.

3.1.2. Network Training for Depth Modality

Four ConvNets were trained on the DDIs and DMDIs individually. In this paper, the ResNet-50 [46] was adopted as the ConvNet model. For ChaLearn LAP ConGD Dataset, We fine-tuned the ConvNets for DDIs and DMDIs with pre-training models on ImageNet [29]. The networks were fine-tuned for Montalbano Gesture Dataset based on the models trained on ChaLearn LAP ConGD Dataset. The data augmentation such as horizontal flip and standard color augmentation was used. We adopted batch normalization right after each convolution and before activation function. All hidden weight layers used the RELU. The network weights were learned using mini-batch SGD with the momentum and weight decay being set to 0.9 and 0.0001, respectively. The batch-size was set to 16. To fit the input size of ResNet-50, all inputs were resized to 224×224. The learning rate was initially set to 10^{-4} and then dropped to its $\frac{1}{10}$ every 40K iterations. The total training iterations was 90K and early stopping was also used to reduce the overfitting. The optical flow was extracted by the TVL1 optical flow algorithm implemented in OpenCV with CUDA. The ResNet-50 was implemented with Tensorflow and trained on one TITAN X Pascal GPU.

3.1.3. Network Training for RGB Modality

The 3D ConvLSTM was trained separately on RGB sequences and saliency sequences. For ChaLearn LAP ConGD Dataset, the network was fine-tuned on RGB modality from the pre-training model on SKIG [47] provided by Zhu et al. [9] and then this model was fine-tuned on saliency sequences. The network was fine-tuned for the Montalbano Gesture Dataset based on the models trained on ChaLearn LAP ConGD Dataset. Batch normalization was introduced to accelerate the training processes. The learning rate was set to 0.1 and then dropped to its $\frac{1}{10}$ every 15K iterations. The weight decay was initially 0.004. At most 60K iterations are needed for training. The batch-size was set to 13, the number of frames in each clip was 32, and each image was cropped into 112×112. The 3D ConvLSTM was implemented based on the Tensorflow and Tensorlayer platforms and trained on one TITAN X Pascal GPU.

3.2. Evaluation of Different Settings and Comparision

3.2.1. Temporal Segmentation Evaluation

To evaluate the effectiveness of the proposed temporal segmentation method, we compared the performance of the proposed temporal segmentation method with the one of only using the hand motion information and the one of only using the appearance information on ChaLearn LAP ConGD Dataset. The continuous gesture sequence was divided into isolated gesture sequences with different segmentation methods, and then the isolated gesture sequences were recognized with our proposed gesture recognition network. The comparison on the validation set of Chalearn LAP ConGD Dataset is shown in Table 1. Our proposed temporal segmentation method outperforms the method with only the hand motion information used and only the appearance information used. These results also demonstrated that the hand motion information and the appearance information were complementary in temporal segmentation.

Table 1. Comparison of the performance of the proposed temporal segmentation method with the one of only using the hand motion information and the one of only using the appearance information on on the validation set of ChaLearn LAP ConGD Dataset.

Segmentation Methods	Mean Jaccard Index
Hand motion information	0.5103
Appearance information (two stream CNNs) [39]	0.5214
Proposed segmentation method	0.6453

3.2.2. Rank Pooling vs. Weighted Rank Pooling

Table 2 compares the performance using rank pooling and weighted rank pooling on the validation set of ChaLearn LAP ConGD Dataset. The results of three groups rank pooling are listed, including a convenient rank pooling, different spatial weight estimation methods, and different temporal weight estimation methods. In the second group, flow-guided aggregation is better than background-foreground segmentation and salient region detection. The foreground area was segmented by the most reliable background model (MRBM) [48], and the salient region was extracted by global contrast-based salient region detection [49]. The spatial weight of the pixel in the foreground area/the salient region is assigned to 1. Otherwise, the spatial weight is assigned to 0. In the third group, the flow-guided frame weight is better than the selection key frames. The key frames were selected by an unsupervised learning method [50]. The temporal weight of key frames is assigned to 1, and the temporal weight of other frames is assigned to 0. These results show that flow-guided aggregation method outperforms rank pooling 0.0401 and flow-guided frame weight method outperforms rank pooling 0.0378. This verifies that weighted rank pooling are more robust and more discriminative in gesture recognition.

Table 2. Comparison of recognition accuracy using rank pooling and weighted rank pooling on the validation set of ChaLearn LAP ConGD Dataset.

Methods	Mean Jaccard Index
Rank Pooling	0.6453
Background-foreground segmentation	0.6503
Salient region detection	0.6645
Flow-guided aggregation	0.6854
Selection key frames	0.6736
Flow-guided frame weight	0.6831

3.2.3. Different Features Evaluation

In this section, the features extracted from the RGB component and depth component were evaluated. The performance using features extracted by the DDIs + ConvNet, DMDIs + ConvNet, RGB + 3D ConvLSTM, Saliency + 3D ConvLSTM, and their combination was evaluated respectively. Average score fusion is used for the combination in this experiment. The evaluation result was listed in Table 3, the symbol • denotes that the corresponding feature is selected for gesture recognition, and the symbol × denotes that the corresponding feature is not included for gesture recognition.

The ConvNet features from DDIs and DMDIs were compared on the validation set of ChaLearn LAP ConGD Dataset in Table 3. Although the performance of DMDI was slightly lower than the one of DDI, the fusion of the ConvNet features extracted from DDIs and DMDIs achieved 0.1196 improvement (i.e., 0.6414 vs. 0.5218). The result demonstrated that variations in the background, shadows, or sudden changed variations in lighting conditions can have substantial impact on the performance and the ConvNet features extracted from DDIs and DMDIs are complementary.

Then the 3D ConvLSTM features extracted from RGB and Saliency were compared on the validation set of ChaLearn LAP ConGD Dataset. From Table 3, we can see the performance of Saliency outperformed the one of RGB, which proved that the background can reduce the performance. The fusion of the 3D ConvLSTM features extracted from RGB and Saliency achieve 0.1302 improvement (i.e., 0.6127 VS. 0.4825). The results have also demonstrated that the 3D ConvLSTM features extracted from RGB and Saliency are complementary.

The Mean Jaccard Index achieved 0.6127 based on RGB modality, and the Mean Jaccard Index was 0.6414 based on depth modality. In addition, the fusion of all features offered 0.1686 improvement (i.e., 0.6904 vs. 0.5218) on the validation set of ChaLearn LAP ConGD Dataset. These results have demonstrated that all features from ConvNet and 3D

ConvLSTM are complementary and different discriminative. The result also verified the effectiveness of our proposed network.

Table 3. The evaluation of different features on the validation set of ChaLearn LAP ConGD Dataset.The symbol • in the Table 3 denotes that the corresponding feature is selected for gesture recognition, and the symbol × denotes that the corresponding feature is not included for gesture recognition.

DDIs + ConvNet	DMDIs + ConvNet	RGB + 3D ConvLSTM	Saliency + 3D ConvLSTM	Mean Jaccard Index
•	×	×	×	0.5218
×	•	×	×	0.5132
×	×	•	×	0.4617
×	×	×	•	0.4825
•	•	×	×	0.6414
×	×	•	•	0.6127
•	×	•	×	0.6351
•	•	•	×	0.6562
•	×	•	•	0.6479
•	•	•	•	0.6904

3.2.4. Score Fusion Evalution

In this paper, score fusion was employed to fuse the classification obtained from the ConvNets and 3D ConvLSTMs. The common score fusion methods are average, maximum, and multiply score function. The comparisons among the three score fusion methods were shown in Table 4. These results showed that the average score fusion method achieved the best result.

Table 4. Comparison of three different score fusion methods on the validation set of ChaLearn LAP ConGD Dataset.

Score Fusion Methods	Mean Jaccard Index
Maximum	0.6851
Multiply	0.6874
Average	0.6904

3.3. Evaluation on ChaLearn LAP ConGD Dataset

3.3.1. Description

The ChaLearn Gesture Dataset (CGD) includes color and depth video sequences recorded by Microsoft Kinect [51]. There are 22,535 RGB-D gesture videos and 47,933 RGB-D gesture instances in the ChaLearn LAP ConGD Dataset. A total of 249 gestures are included and performed by 21 different individuals. Detailed information is shown in Table 5.

Table 5. Statistics of the ChaLearn LAP ConGD Dataset

Sets	# of Gestures	# of RGB Videos	# of Depth Videos	# of Subjects
Training	30,442	14,134	14,134	17
Validation	8889	4179	4179	2
Testing	8602	4042	4042	2
All	47,933	22,535	22,535	21

3.3.2. Experimental Results

Table 6 compared the performance of the proposed method and that of exiting methods on validation set. MFSK [37] and MFSK + DeepID [37] segmented the continuous gesture sequence to isolated gesture firstly and recognized the isolated gesture with the hand-craft features. Wang et al. [52] employed the QOM method to segment the continuous gesture

sequence and then extracted an improved depth motion map using color coding method over the segmented sequence, and CNN was adopted to train and classify the segmented gesture. Chai et al. [3] first adopted Faster R-CNN to extract the hand for the temporal segmentation, and then two-stream RNNs were adopted to fuse multi-modality features for the recognition. Camgoz et al. [4] applied 3D convolutional networks to RGB video and jointly learned the features and classifier. It can be seen that our proposed method achieved state-of-the-art results compared with existing methods.

Table 6. Comparison of the proposed method and other methods on the validation set of ChaLearn LAP ConGD Dataset.

Methods	Mean Jaccard Index $\overline{J_S}$
MFSK [37]	0.0918
MFSK+DeepID [37]	0.0902
Wang et al. [52]	0.2403
Chai et al. [3]	0.2655
Camgoz et al. [4]	0.2809
Wang et al. [39]	0.5214
Proposed method	**0.6904**

The proposed method was also compared with the methods in ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge [40] in Table 7. The mean Jaccard Index of our proposed method achieved 0.6976 in the test set. Our proposed method achieved state-of-the-art results.

Table 7. Performance comparison with other teams in ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge.

Team	ICT_NHCI	AMRL	PaFiFA	Deepgesture	Wang et al. [39]	Proposed Method
Mean Jaccard Index $\overline{J_S}$ (valid set)	0.5163	0.5957	0.3646	0.3190	0.5214	**0.6904**
Mean Jaccard Index $\overline{J_S}$ (test set)	0.6103	0.5950	0.3744	0.3164	0.5307	**0.6976**

3.4. Evaluation on Montalbano Gesture Dataset

3.4.1. Description

Montalbano gesture dataset [38] was also recorded by Microsoft Kinect Sensor. It contains 20 Italian cultural/anthropological. Four modalities, including RGB, depth, mask, and skeleton, can be found in this dataset. It is labeled frame-by-frame. The characteristics of Montalbano Gesture Dataset were:

- the duration of each gesture varied greatly and there was no self-occlusion;
- there was no information on the number or order of gestures;
- the intra-class variability of gesture samples was high, while the inter-class variability of some gesture categories was low.

These characteristics brought lots of challenges. The detail of Montalbano Gesture Dataset was shown in Table 8.

Table 8. Information of Montalbano Gesture Dataset.

Training Sequences	Validation Sequences	Test Sequences
393 (7754 gestures)	287 (3362 gestures)	276 (2742 gestures)
Sequence Duration	Number of Users	Labeled Frame
1–2 min	27	1,720,800

3.4.2. Experimental Results

Table 9 showed the result on Montalbano Gesture Dataset. Our proposed method achieved state-of-the-art performance. Left and right hand regions are treated as independent streams to improve the performance [53] and skeleton information is used in [54] to crop the specific area in videos. However, only RGB and depth modalities were used in our proposed method. The promising performance demonstrated the effectiveness of our proposed method.

Table 9. Comparison of the proposed method and other methods on Montalbano Gesture Dataset.

Methods	Mean Jaccard Index $\overline{J_S}$
MRF, KK, PCA, HoG [55]	0.827
AdaBoost, HoG [56]	0.834
Multi-scale DNN [53]	0.870
Temp Conv + LSTM [54]	0.906
Proposed Method	**0.923**

3.5. Discussion

Temporal segmentation is crucial for continuous gesture recognition. The temporal segmentation and gesture recognition in continuous gesture recognition were performed separately in this paper. We assume that there are some transition frames between two consecutive gestures and one will puts hands down after performing a gesture. Although our proposed method has achieved good performance on both ChaLearn LAP ConGD Dataset and Montalbano Gesture Dataset, these assumptions limited the wider application of the proposed method. In our future work, we will explore a more general approach to address the problem of continuous gesture recognition.

In addition, current continuous gesture recognition methods can not address the problem of online gesture recognition. In actuality, important real-time applications including sign language interpreter and driver assistance systems require identifying gestures as soon as each video frame comes. How to improve the proposed method for online gesture recognition will be a good research direction.

4. Conclusions

The paper presents an effective method for large-scale multimodal gesture segmentation and recognition. The video sequences are first segmented into isolated gesture sequences by classifying the frames into gesture frames and transition frames. For each segmented gesture sequence, our proposed method explores the effective spatiotemporal information based ConvNets for depth modality and 3D ConvLSTMs for RGB modality. Experimental results on the ChaLearn LAP ConGD Dataset and Montalbano Gesture Dataset verified the effectiveness of our proposed method. In our future work, we will explore a more general approach to address the problem of continuous gesture recognition and improve the proposed method for online gesture recognition.

Funding: Huogen Wang gratefully acknowledges the financial support from the Chinese Scholarship Council.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of Tianjin University.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are open access and available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jiang, F.; Zhang, S.; Wu, S.; Gao, Y.; Zhao, D. Multi-Layered Gesture Recognition with Kinect. *J. Mach. Learn. Res.* **2015**, *16*, 227–254.
2. Peng, X.; Wang, L.; Cai, Z.; Qiao, Y. Action and Gesture Temporal Spotting with Super Vector Representation. In *Workshops at European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 518–527.
3. Chai, X.; Liu, Z.; Yin, F.; Liu, Z.; Chen, X. Two Streams Recurrent Neural Networks for Large-scale Continuous Gesture Recognition. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 31–36.
4. Camgoz, N.C.; Hadfield, S.; Koller, O.; Bowden, R. Using Convolutional 3d Neural Networks for User-independent Continuous Gesture Recognition. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 49–54.
5. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
6. Sharma, S.; Kiros, R.; Salakhutdinov, R. Action Recognition Using Visual Attention. *arXiv* **2015**, arXiv:1511.04119.
7. Jain, A.; Zamir, A.R.; Savarese, S.; Saxena, A. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016, pp. 5308–5317.
8. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.c. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; pp. 802–810. pp. 802–810.
9. Zhu, G.; Zhang, L.; Shen, P.; Song, J. Multimodal Gesture Recognition Using 3D Convolution and Convolutional LSTM. *IEEE Access* **2017**, *5*, 4517–4524.
10. Sun, L.; Jia, K.; Chen, K.; Yeung, D.Y.; Shi, B.E.; Savarese, S. Lattice Long Short-term Memory for Human Action Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
11. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231.
12. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3d Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
13. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
14. Sun, L.; Jia, K.; Yeung, D.Y.; Shi, B.E. Human Action Recognition using Factorized Spatio-temporal Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4597–4605.
15. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 305–321.
16. Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-temporal Representation with Pseudo-3d Residual Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5533–5541.
17. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
18. Zhou, Y.; Sun, X.; Zha, Z.J.; Zeng, W. Mict: Mixed 3d/2d convolutional tube for human action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 449–458.
19. Wang, L.; Li, W.; Li, W.; Van Gool, L. Appearance-and-relation networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1430–1439.
20. Simonyan, K.; Zisserman, A. Two-stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems*, Montréal, Canada, 8–13 December, 2014; pp. 568–576.
21. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-stream Network Fusion for Video Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.
22. Feichtenhofer, C.; Pinz, A.; Wildes, R.P. Spatiotemporal Multiplier Networks for Video Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA. 21–26 July 2017; pp. 7445–7454.
23. Zhu, W.; Hu, J.; Sun, G.; Cao, X.; Qiao, Y. A Key Volume Mining Deep Framework for Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1991–1999.
24. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European Conference on Computer Vision*. Springer: Cham, Switzerland, 2016; pp. 20–36.
25. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-Time Action Recognition with Enhanced Motion Vector CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2718–2726.

26. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A. Hidden Two-stream Convolutional Networks for Action Recognition. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2018; pp. 363–378.
27. Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A.; Gould, S. Dynamic Image Networks for Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3034–3042.
28. Fernando, B.; Gavves, E.; Oramas, J.M.; Ghodrati, A.; Tuytelaars, T. Modeling Video Evolution for Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5378–5387.
29. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252.
30. Fernando, B.; Gould, S. Learning End-to-End Video Classification with Rank-pooling. In *International Conference on Machine Learning*, New York, USA, 19–24 June 2016; pp. 1187–1196.
31. Fernando, B.; Anderson, P.; Hutter, M.; Gould, S. Discriminative Hierarchical Rank Pooling for Activity Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1924–1932.
32. Cherian, A.; Fernando, B.; Harandi, M.; Gould, S. Generalized Rank Pooling for Activity Recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1581–1590.
33. Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Rank Pooling for Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 773–787.
34. Wang, P.; Li, W.; Liu, S.; Gao, Z.; Tang, C.; Ogunbona, P. Large-scale Isolated Gesture Recognition using Convolutional Neural Networks. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 7–12.
35. Wang, P.; Li, W.; Gao, Z.; Zhang, Y.; Tang, C.; Ogunbona, P. Scene Flow to Action Map: A New Representation for RGB-D based Action Recognition with Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 595–604.
36. Wang, H.; Song, Z.; Li, W.; Wang, P. A Hybrid Network for Large-Scale Action Recognition from RGB and Depth Modalities. *Sensors* **2020**, *20*, 3305.
37. Wan, J.; Zhao, Y.; Zhou, S.; Guyon, I.; Escalera, S.; Li, S.Z. Chalearn Looking at People Rgb-d Isolated and Continuous Datasets for Gesture Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 56–64.
38. Escalera, S.; Athitsos, V.; Guyon, I. Challenges in multi-modal gesture recognition. In *Gesture Recognition*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 1–60.
39. Wang, H.; Wang, P.; Song, Z.; Li, W. Large-scale multimodal gesture segmentation and recognition based on convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 3138–3146.
40. Wan, J.; Escalera, S.; Anbarjafari, G.; Jair Escalante, H.; Baro, X.; Guyon, I.; Madadi, M.; Allik, J.; Gorbova, J.; Lin, C.; Xie, Y. Results and Analysis of ChaLearn LAP Multi-Modal Isolated and Continuous Gesture Recognition, and Real Versus Fake Expressed Emotions Challenges. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Venice, Italy, 22–29 October 2017.
41. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, Canada, 7–12 December 2015; pp. 91–99.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916.
43. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned Salient Region Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1597–1604.
44. Escalera, S.; Baró, X.; Gonzalez, J.; Bautista, M.A.; Madadi, M.; Reyes, M.; Ponce-López, V.; Escalante, H.J.; Shotton, J.; Guyon, I. Chalearn looking at people challenge 2014: Dataset and results. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014, pp. 459–473.
45. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Liu, L.; Shao, L. Learning Discriminative Representations from RGB-D Video Data. *Int. Jt. Conf. Artif. Intell.* **2013**, *4*, 8.
48. Liu, Y.; Yao, H.; Gao, W.; Chen, X.; Zhao, D. Nonparametric background generation. *J. Vis. Commun. Image Represent.* **2007**, *18*, 253–263.
49. Cheng, M.M.; Mitra, N.J.; Huang, X.; Torr, P.H.; Hu, S.M. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 569–582.
50. Sheng, L.; Xu, D.; Ouyang, W.; Wang, X. Unsupervised collaborative learning of keyframe detection and visual odometry towards monocular deep slam. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 4302–4311.

51. Zhang, Z. Microsoft Kinect Sensor and Its Effect. *IEEE Multimed.* **2012**, *19*, 4–10.
52. Wang, P.; Li, W.; Liu, S.; Zhang, Y.; Gao, Z.; Ogunbona, P. Large-scale Continuous Gesture Recognition using Convolutional Neural Networks. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 13–18.
53. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1692–1706.
54. Pigou, L.; Van Den Oord, A.; Dieleman, S.; Van Herreweghe, M.; Dambre, J. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *Int. J. Comput. Vis.* **2018**, *126*, 430–439.
55. Chang, J.Y. Nonparametric Gesture Labeling from Multi-modal Data. In *Workshop at the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 503–517.
56. Monnier, C.; German, S.; Ost, A. A multi-scale boosted detector for efficient and robust gesture recognition. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 491–502.