applied sciences

MDPI

Article

Gesture Recognition Based on 3D Human Pose Estimation and Body Part Segmentation for RGB Data Input

Ngoc-Hoang Nguyen, Tran-Dac-Thinh Phan, Guee-Sang Lee *, Soo-Hyung Kim and Hyung-Jeong Yang

Department of Electronics and Computer Engineering, Chonnam National University, 77 Yongbong-ro, Gwangju 500-757, Korea; hoangnguyenkcv@gmail.com (N.-H.N.); phantrandacthinh2382@gmail.com (T.-D.-T.P.); shkim@jnu.ac.kr (S.-H.K.); hjyang@jnu.ac.kr (H.-J.Y.)
* Correspondence: gslee@jnu.ac.kr

Received: 31 July 2020; Accepted: 4 September 2020; Published: 6 September 2020

Abstract: This paper presents a novel approach for dynamic gesture recognition using multi-features extracted from RGB data input. Most of the challenges in gesture recognition revolve around the axis of the presence of multiple actors in the scene, occlusions, and viewpoint variations. In this paper, we develop a gesture recognition approach by hybrid deep learning where RGB frames, 3D skeleton joint information, and body part segmentation are used to overcome such problems. Extracted from the RGB images are the multimodal input observations, which are combined by multi-modal stream networks suited to different input modalities: residual 3D convolutional neural networks based on ResNet architecture (3DCNN_ResNet) for RGB images and color body part segmentation modalities; long short-term memory network (LSTM) for 3D skeleton joint modality. We evaluated the proposed model on four public datasets: UTD multimodal human action dataset, gaming 3D dataset, NTU RGB+D dataset, and MSRDailyActivity3D dataset and the experimental results on these datasets proves the effectiveness of our approach.

Keywords: dynamic gesture recognition; human action recognition; multi-modalities network

1. Introduction

Gesture recognition has recently attracted much attention because of its wide applications such as the human–computer interaction, telecommunications, and robotics, but it still remains as one of the major challenges because of the inherent complexity of human motions. In early times, gesture recognition based on conventional techniques of classification with handcrafted features, such as support vector machine (SVM), bag-of-features and multiclass SVM, and hidden Markov model (HMM), have been proposed [1–3]. Recently, deep learning-based methods are increasingly employed due to their advantages of end-to-end learning by automatic extraction of spatiotemporal features from raw data. The development of deep learning methods based on a convolution neural network (CNN) and recurrent neural network (RNN) or long short-term memory network (LSTM) have achieved positive results in handling gesture recognition tasks [4–8]. However, there are limitations in the performance of gesture classification due to the complexity of the scene, e.g., the presence of multiple actors in the background, occlusions, illumination changes, or viewpoint variations.

In existing methods, to overcome the challenges caused by the issue of background or viewpoint variations, gesture recognition is usually developed by combining multiple modalities of data inputs (such as skeleton joints information, human body shape, RGB, optical flow, and depth frames) with newly developed deep learning models [9–11]. By utilizing skeleton joints information or depth

information, gesture recognition performance has been significantly improved because they are helpful in the representation of gestures and played an important role in gesture identification. Although human skeleton joints and depth data can be collected directly from time-of-flight (ToF) cameras, gesture recognition on RGB video input is a substantial challenge because the human pose should be estimated with high accuracy. Skeleton joints convey vital information to represent gesture from the human pose, but it is not enough to identify complicated motions when it does not match to the shape of body parts correctly. For instance, if there are other actors present in the scene together with the main target, the action of other objects from the background can cause confusion to the correct extraction of the target person's skeleton. Figure 1b shows possible errors in skeleton joints extraction of the main person due to the presence of other moving objects. However, when the body parts are segmented beforehand, the skeleton of a target can be obtained with much higher accuracy because the layout of the skeleton is restricted by the body part. Moreover, skeleton joints of the frame sequence of a video can be temporally incoherent due to independent errors in each frame as we illustrate in Figure 2; and this can cause incorrect classification.

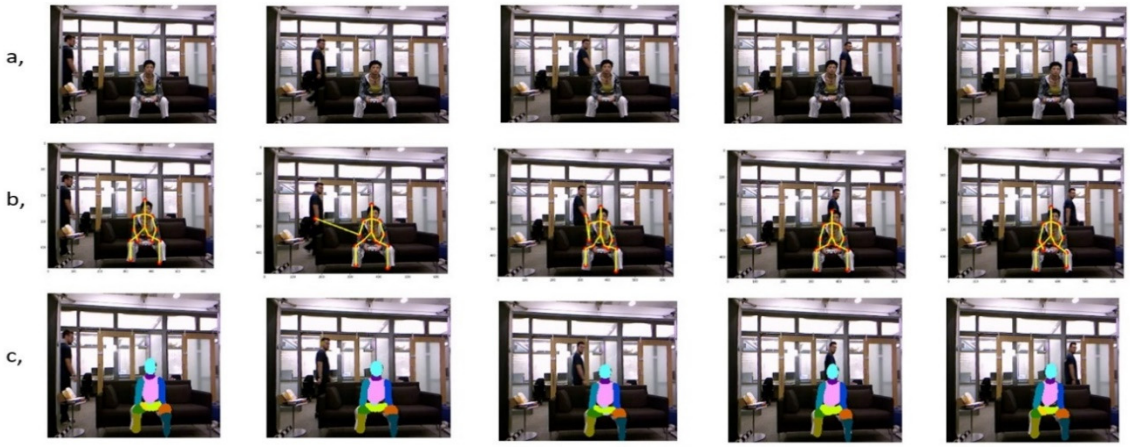

Figure 1. Illustration of the case in which many actors are present in the scene. The first row (**a**) RGB input images, the second row (**b**) extracted skeleton joints images, and the last row (**c**) color-encoded body part segmentation images. In this case, the skeleton joints extraction can be distracted by other objects in the background, but the color-encoded body part segmentation can help avoiding such a case.

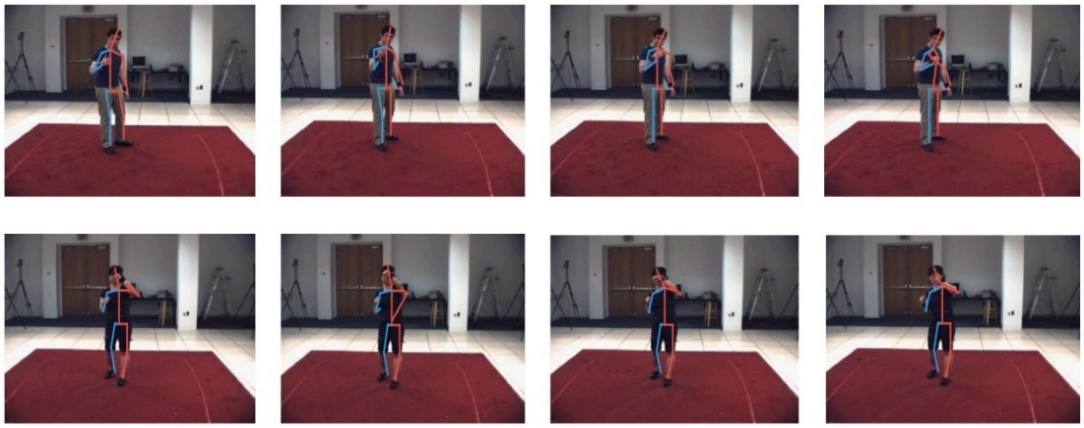

Figure 2. Illustration of skeleton joints of sequence sequential frames can be incoherent due to independent error in each frame.

In this paper, we propose a multi-modal gesture recognition method for RGB data input with a multi-modal algorithm. The algorithm consists of three submodels: two residual 3D convolutional neural networks based on ResNet architecture (3DCNN_ResNet) [12] and a long short-term memory

network (LSTM) to perform on three data modalities: RGB images, color body part segmentations, and 3D human skeleton joints, respectively. We extract color body part segmentations and 3D human skeleton joints from RGB input.

The color body part segmentations with 12 semantic parts are obtained by the segmentation CNN network RefineNet [13]. By utilizing body part segmentation, the gesture recognition network can focus on only the target person in the presence of multiple objects in the scene. The 3D human skeleton joints information of the target person is extracted by the 3D human pose estimation network called Pose3D [14]. Pose3D is a deep learning network that obtains the sequence of 3D poses based on the temporal information across the sequence of 2D joint locations in order to prevent temporally incoherent estimation. The gesture classes are predicted by a combination of these three submodels that are effectively fused by an integrated stacking module as the late fusion layers.

The contributions of the paper are:

- The reason for the distraction of skeleton joints extraction has been addressed, which hinders from the proper functioning of gesture recognition methods. In other words, the existence of noise or extra persons in the background can cause such distractions.
- The solution to the problem of multiple actors in the scene, which caused the distraction of skeleton joints extraction, has been presented with target person extraction. The target person is segmented, and used for eliminating distraction in skeleton joints extraction.

This idea of using target person extraction has never been addressed before in the literature as far as we know.

The remainder of this paper is organized as follows: related works are given in Section 2. In Section 3, the proposed algorithm for gesture recognition is presented. The experiments on four public datasets to evaluate the effectiveness of our approach are given in Section 4. Finally, a conclusion is given in Section 5.

2. Related Works

In this section, we briefly described the previous methods relevant to our work for gesture recognition. Although significant improvement on gesture recognition has been reported, but new challenges appear with different input modalities and restrictions.

In early years, gesture or action recognition problem has been dealt with classical machine learning methods such as support vector machine (SVM) [1], bag-of-features [2], and hidden Markov model (HMM) [3]. In these methods, gestures are classified based on the features extracted by a hand-engineered extractor. Hussein et al. [1] presented a discriminative descriptor for action classification based on the covariance matrix for 3D skeleton joint locations. Dardas [2] similarly used the SVM classifier to identify gesture classes, but via a bag-of-words vector mapped from key points extracted by scale-invariant feature transform (SIFT). Lee [3] presented a method using HMM based on the likelihood threshold estimation of the input pattern. The gesture recognition task has been tackled by conventional machine learning methods, but there are significant limitations in these approaches. For instance, the parameters of the model depend on experience, and the system is sensitive to noise.

Recently, due to the revolution of deep learning, gesture recognition approaches have been presented with impressive performances compared to the traditional methods. The deep learning-based methods became popular due to its capability to extract spatiotemporal features from the raw video input automatically. A convolutional neural network (CNN) was initially used for extracting spatial features for static images, but it has been extended to deal with different input types or to extract different types of features. Various approaches [4,5,15,16] have utilized CNN to treat sequential frames of a video as multi-channel inputs for the purpose of video classification. Feichtenhofer [4] incorporated a two-stream network with separate ConvNets for RGB images and optical flow images to extract motion information for gesture classification. Kopuklu [15] proposed data level fusion to combine RGB and optical flow modalities with static images to extract action features. CNN can be incorporated

with RNN or LSTM to learn both spatial and temporal features of a video for action classification. Donahue [8] deployed long-term recurrent convolutional network (LRCN), in which CNN is used to extract spatial features of images and LSTM is applied to capture temporal dependencies in the sequence of such features.

For dynamic gesture recognition, [7,17,18] applied a 3D convolutional neural network to capture discriminative features along both spatial and temporal dimensions due to 3D convolutions and 3D pooling. Tran [17] used spatiotemporal features extracted by 3DCNN to classify gesture with SVM classifier. Molchanov [16] proposed recurrent 3D convolutional neural networks (R3DCNN) to recognize gestures online, where 3DCNN is used as a feature extractor. The gesture recognition based on a human pose has also achieved impressive results. Utilizing RNN or LSTM to capture temporal features from human skeleton joints for gesture classification is gaining popularity [19–21], in which skeleton joints are extracted by depth information with a ToF camera, but it becomes more challenging when only RGB data input is used. In other works, multiple data modalities or multiple deep learning models are combined to achieve better performance in gesture classification. Duan [9] presented a convolutional two-stream consensus voting network based on 3DCNN for RGB and depth channels to identify the gesture classes in a video input. Chai [10] also proposed a multi-stream model based on RNN with hand location information extracted from RGB-D input. The summary of the related works is given in Table 1.

Table 1. The summary of related works.

Author	Approach/Features	Details	Comments
Hussen et al. [1]	Support Vector Machine (SVM)	Discriminative classification from 3D skeleton joints	Classical machine learning methods.
Dardas et al. [2]	Bag of features	Classification via bag-of-words vector	
Lee et al. [3]	Hidden Markov Model	HMM based on the likelihood threshold estimation	
Feichtenhofer et al. [4]	Two-stream network with separate ConvNets	Extract motion from RGB images and optical flow images	Early deep learning-based methods
Kopuklu et al. [15]	Data level fusion	Combine RGB and optical flow modalities with static images	
Donahue et al. [8]	Long-term recurrent convolutional network	CNN and LSTM for spatio-temporal features	
Tran et al. [17]	3DCNN, SVM classifier	Spatio-temporal features extracted by 3DCNN	3DCNN to capture spatial and temporal features
Molchanov et al. [16]	Recurrent 3D Convolutional Neural Networks	Online gesture recognition by 3DCNN	
Yan et al. [19], Li et al. [20], and Omran et al. [21]	RNN or LSTM, human skeleton joints	Capture temporal features from skeleton joints and depth	Deep learning + human pose
Duan et al. [9]	Convolutional two-stream consensus voting network	Combine the results from RGB and depth in a video input	Multi-modal or multiple deep learning models
Chai et al. [10]	Multi-stream model based on RNN	Extract hand location information from RGB-D input	

Most of the gesture recognition methods exploit the human pose estimation, however often the pose estimation in the video can often be inconsistent because of independent errors in the sequence of frames. Additionally, the extraction of skeleton joints may not be successful when the

background contains multiple objects or human beings. In this paper, we try to solve these problems through the use of segmentation of the target person. We propose a novel method for gesture recognition with multi-modalities: RGB images, color body part segmentation images, and 3D skeleton joints information.

3. Proposed Method

In this section, we present in detail our proposed approach for gesture recognition. The proposed method consisted of three submodels: two 3D_ ResNet networks and an LSTM network to deal with RGB frames, color body part segmentation images, and 3D skeleton joints information. These three submodels were effectively fused by integrated stacking module as a late fusion layer in order to decide the gesture class. We show an overview of the proposed algorithm in Figure 3.

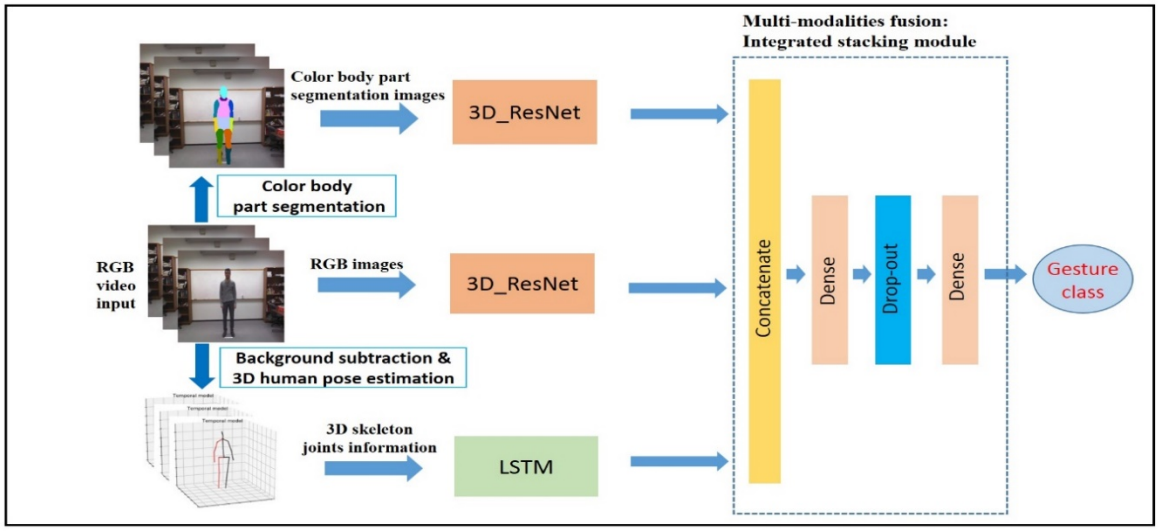

Figure 3. The proposed approach. This algorithm consists of three subnetworks and the fusion module.

In our algorithm, given the RGB image sequence of a video, the color body part segmentation images were generated by the segmentation network—RefineNet [22]. Some of the datasets, e.g., MSRDailyActivity3D, contained sometimes more than one person on the scene. The target person was a person whose position was nearest to the center of the screen and closest to the camera in most of the frames of the video. Therefore, a person closer to the center area was selected first. If more than one person were around the center of the screen, the person closer to the camera was selected from the depth information. From the color body part segmentation, the background of the input image was excluded from the scene, leaving only the target person, which was used for extracting 3D human skeleton joints by temporal 3D human pose estimation network—Pose3D. Two-stream 3D_ResNet networks were used to learn the features from the RGB images and color body part segmentation images for gesture classification. In the other subnetwork, the extracted 3D human skeleton joints were utilized by the LSTM network.

- Color body part segmentation: The RefineNet model, given the RGB input frames, produced the color-coded 12 body parts segmentation. The RefineNet is a multi-path refinement network for semantic segmentation via multi-level features and potentially long-range connections. The RefineNet model typically consists of four blocks: adaptive convolution, multi-resolution fusion, chained residual pooling, and output convolution. The batch-normalization layers were simplified from the convolution block but it still contained the remaining convolution units of the original ResNet. Multi-resolution fusion performed feature map fusion by convolutions and a summation. Multiple resolution feature maps extracted from varied input paths were fused into a high-resolution feature map. Additionally, the output feature map was fused through

multiple pooling blocks of chained residual pooling blocks. The final prediction was given by the output convolution block, which had another remaining convolution unit and soft-max layer. The RefineNet network was based on ResNeXt-101 [23] with trained weights by the UP-3D dataset [24] for color-coded 12 body parts segmentation. Figure 4 shows the example of color body parts segmentation on sampled frames on a video.

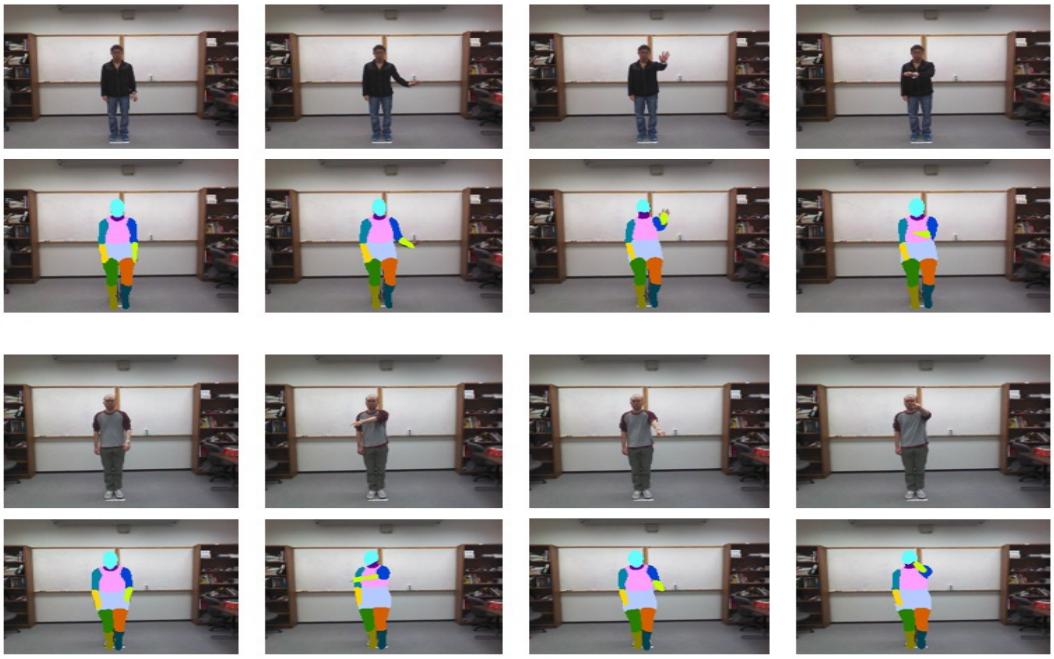

Figure 4. Example of color-encoded 12 body parts segmentation from RGB images sequence by the RefineNet model.

- Temporal video 3D human pose estimation: We extracted 3D joints skeleton information for gesture recognition from temporal video 3D human pose estimation called Pose_3D [14] from background subtracted RGB images. The Pose_3D network is a sequence-to-sequence network that predicts a sequence of temporally consistent 3D human pose from the sequence of 2D human poses. The 2D human pose was obtained by the state-of-art 2D human pose estimation framework—the stacked-hourglass network [25] trained on the Human3.6M dataset [26]. The decoder of the Pose_3D consists of LSTM units and residual connection to predict temporally consistent 3D poses of the current frame using the 3D poses of previous frames and 2D joints information of all frames, which were taken from the final state of the encoder. The temporal smoothness constraint was imposed on the 3D pose extraction of a video. Since the stacked-hourglass network was used for 2D pose estimation on individual frames, this constraint made the predicted 3D poses more stable and reliable even with 2D pose estimation failure in a few frames within the temporal window. Figure 5 shows the example of temporal 3D skeleton joints of video frames extracted by the Pose_3D network.

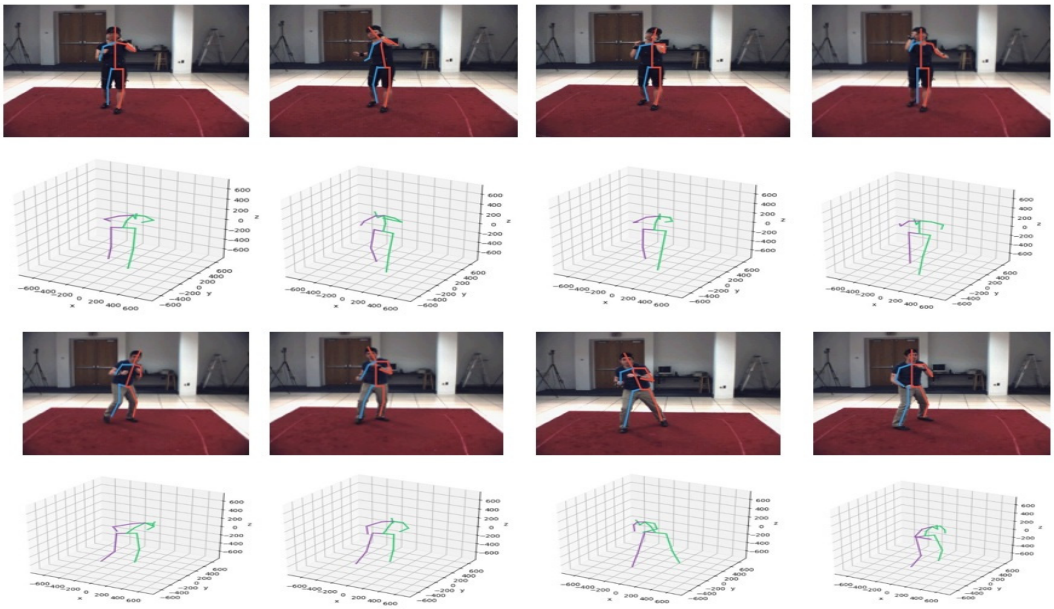

Figure 5. Example of skeleton joints extraction from RGB images sequence by temporal video 3D human pose estimation-Pose_3D.

- Two-stream 3D_ResNet networks for RGB and color body part segmentation modalities: The residual 3D_CNN network based on ResNet architecture [23] was applied to benefit from two data modalities: the background subtracted RGB images and color-coded 12 body parts segmentation images for gesture classification. For the input of 3D_CNN of the RGB image branch, we subtracted the background of the RGB image by color body part segmentation and fed it into the network. Due to spatiotemporal feature learning by 3D convolution and 3D pooling, the 3D_CNN network is known as one of the essential frameworks for video classification. Residual 3D_CNN could significantly improve the classification performance of basic 3D_CNN framework. The 3D_ResNet is one of the current residual 3D_CNN versions. Various ResNet-based architectures with 3D convolutions were studied, but the 3D_ResNet network based on ResNeXt-101 was employed because of the quality performance for the proposed method.

Different from other original bottleneck blocks with a standard convolutional layer, the ResNeXt block employs a group convolution layer with its capacity to divide the feature maps into small groups. The single 3D_ResNet stream modality network in our proposed method included five ResNeXt blocks. The structure of each ResNeXt block consisted of convolution layers (group convolution layer), a batch normalization (BatchNorm) layer, and a rectified-linear unit (ReLu) layer, as shown in Figure 6. The input of each modality stream network was a fixed number of T sampled frames of a video: $V_c = \{x_{c1}, x_{c2}, \ldots, x_{cT}\}$ for RGB modality and $V_{ps} = \{x_{ps1}, x_{ps2}, \ldots, x_{psT}\}$ for color body part segmentation modality. The operation function of these stream-networks was $\Theta_c(.)$ and $\Theta_{ps}(.)$ respectively. Then, the prediction probabilities of the stream networks for RGB input and color body part segmentation for i classes can be described as:

$$P_c\{p_1|V_c, p_2|V_c, \ldots, p_i|V_c\} = \Theta_c(V_c) \quad (1)$$

$$P_{ps}\{p_1|V_{ps}, p_2|V_{ps}, \ldots, p_i|V_{ps}\} = \Theta_{ps}(V_{ps}) \quad (2)$$

where p_i is the prediction probability of the video belonging to the class ith, and P_c and P_{ps} denote the network outputs, which are the vectors or class prediction probabilities.

- The LSTM network for 3D skeleton joints modality: The LSTM network was proposed as a submodel for gesture recognition to benefit from the extracted 3D skeleton joints data. The 3D skeleton data provides useful information about the temporal features such as the localization of the relevant body joints over a time series to recognize the performed action. The LSTM networks are utilized to capture the contextual information of a temporal sequence for long periods by the gates and memory cells. In an LSTM network, the operation of gates and memory cells by time can be described as follows:

$$\begin{cases} i_t = \sigma(W_i[x_t, h_{t-1}] + b_i), \\ f_t = \sigma\big(W_f[x_t, h_{t-1}] + b_f\big), \\ o_t = \sigma(W_o[x_t, h_{t-1}] + b_o), \\ \widetilde{c_t} = \tan h(W_c[x_t, h_{t-1}] + b_c), \\ c_t = f_t * c_{t-1} + i_t * \widetilde{c_t}, \\ h_t = \tan h(c_t) * o_t \end{cases} \tag{3}$$

where i, f, and o are the vectors of the input gate, the forget gate and the output gate, respectively. $\widetilde{c_t}$ and c_t denote the "candidate" hidden state and internal memory of the unit respectively. ht is the output hidden state and σ(.) is a sigmoid function while W and b represent connected weights matrix and bias vectors, respectively.

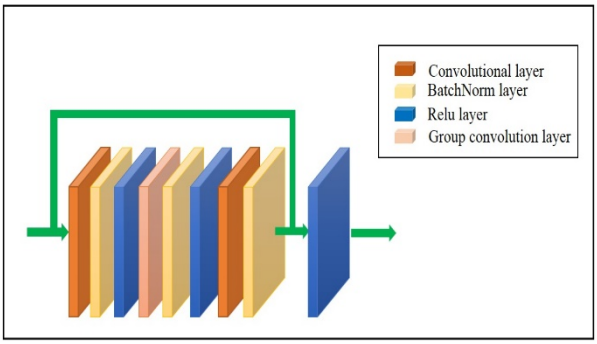

Figure 6. The overview of ResNet blocks. It consists of two convolution layers (1 × 1 × 1 kernel size), one group convolution layer (3 × 3 × 3 kernel size), and three batch normalization (BatchNorm) layers and rectified-linear unit (ReLu) layers.

In our approach, we used the relevant 17 human skeleton joints extracted by the Pose_3D network to perform the gesture classification. The input of the LSTM submodel was $T \times 51$ vector, which contains 3D location of 17 skeleton joints in a sequence of T frames of a video. This input vector is defined as $V_s = \{v_{s1}, v_{s2}, \ldots, v_{sT}\}$, where v_{sk} is a 51 × 1 vector of 3D position information of all skeleton joints at the k^{th} frame. The prediction probability of the LSTM modality network for i gesture classes by a given input V_j is:

$$P_j\{p_1|V_j, p_2|V_j, \ldots, p_i|V_j\} = \Theta_j(V_j) \tag{4}$$

where $\Theta_j(.)$ represents the operation function of the LSTM network.

- Multi-modality fusion by an integrated stacking module: The final gesture class was predicted by the fusion score of three submodels. The multi-modalities fusion block combined the prediction score of submodels in the integrated stacking module, as shown in Figure 3. This module is a more extensive multi-headed neural network, which consists of a concatenation layer, two fully connected layers and a dropout layer to avoid overfitting. We used the outputs of each subnetworks as separate input-heads to the module. The fusion score computed by the fusion block with non-linear operation function $\Theta_{\text{fusion}}(.)$ to predict gesture classes as follows:

$$P\{p_1|V_c, p_2|V_c, \ldots, p_i|V_c\} = \Theta_{fusion}\left\{ \begin{array}{l} P_c\{p_1|V_c, p_2|V_c, \ldots, p_i|V_c\}, P_{\text{ps}}\{p_1|V_{\text{ps}}, p_2|V_{\text{ps}}, \ldots, p_i|V_{\text{ps}}\}, \\ P_j\{p_1|V_j, p_2|V_j, \ldots, p_i|V_j\} \end{array} \right\} \quad (5)$$

4. Experimental Results

To evaluate the performance of the proposed method, we did experiments on four public datasets: UTD multimodal human action dataset [27], gaming 3D dataset [28], NTU RGB+D dataset [29] and MSRDailyActivity3D dataset [30]. In these datasets, we used only RGB modality for our work.

4.1. Datasets

- UTD Multimodal Human Action dataset [27]: is a multimodal human action dataset that was collected by using a Microsoft Kinect sensor and a wearable inertial sensor. This data includes 27 different types of actions performed by eight gestures, where each actor repeats each gesture four times. This dataset contains a total of 861 gesture videos after removing corrupted sequences.
- Gaming 3D dataset [28]: is a collection of 600 videos for action recognition in the gaming scenario. This dataset consisted of 20 action classes performed by ten subjects, and each subject repeated each action three times. Besides providing the action class label for each video, this dataset also had the ground truth for the peak frame of each action.
- NTU RGB+D dataset [29]: is a large dataset that contains 56,880 RGB-D videos captured by three Kinect V2 sensors concurrently. This dataset consisted of 60 human activities related to daily actions, mutual actions, and medical conditions. In our work, we only focused on the daily actions category with the RGB video data.
- MSRDailyActivity3D dataset [30]: is a dataset captured for daily human activities in a living room. This collection contained 16 regular activity classes performed by ten different individuals. There are 320 activity videos. This dataset was made with a noisy background with other activities from untargeted people.

For the MSRDailyActivity3D dataset and Gaming 3D dataset, we stratified a random split of each dataset into 5 folds with a train/valid/test set by ratio 8:1:1, respectively. For the UTD-MHAD dataset, we applied the 8-fold cross-validation method. Seven sets were used for training and the remaining set was used for testing. For the NTU-RGB+D dataset, there are two kinds of benchmarks recommended by the creator of the NTU-RGB+D dataset, which are cross-subject and cross-view benchmarks. We chose the cross-view benchmark, which included 37,462 clips for training and 18,817 clips for testing without validation.

4.2. Implementation

Three subnetworks were trained with the corresponding data modalities extracted from the datasets described above. The color body parts segmentation images and 3D skeleton joints were extracted first from the dataset to separately train corresponding subnetworks. The color body parts segmentation images were obtained by the RefineNet network, pretrained with UP-3D dataset as described in Section 3. We used the Pytorch framework of the RefineNet network, which was provided publicly. The 3D skeleton joints information was estimated by the Pose_3D network with the pretrained weight on the large dataset Human3.6M.

For two-stream networks for RGB and color-encoded body parts segmentation modalities, they were pretrained with UCF101—action recognition dataset [31]. The optimal settings for the LSTM network comprise of three memory blocks, and 256 LSTM Cells per memory blocks as in [32].

We combined the output scores of three subnetworks and fed it into the integrated stacking module to generate the final gesture class prediction. This multi-modality fusion is trained after subnetworks generate their own outputs.

The adaptive moment estimation [33] was used to optimize the parameters during the training process due to its effectiveness with a large number of parameters. We utilized the adaptive learning rate method to get the optimal parameters of the model.

All of the above codes were run with Nvidia GTX 1080 Ti GPU. The experiment was conducted with PC of CPU i7-7700 and 32GB of memory. It took 2–4 days per dataset for the experiment but the primary concern was the performance, which we mainly focused on.

4.3. Results

The performance of the proposed approach for gesture recognition was evaluated by the experiments conducted on four presented datasets: UTD multimodal human action dataset, gaming 3D dataset, NTU RGB+D dataset, and MSRDailyActivity3D dataset. Comparisons of the results with previous approaches are reported in Table 1.

Table 2 shows the comparison of the proposed network with existing networks on different datasets [7,8,19,34–40]. The comparison shows that the proposed network outperformed the existing works. Note that the combination of the three subnetworks was better than each single data modality network as shown in Table 3. The color body part segmentation modality produced better performance than the RGB modality due to the removal of background noise. Classification of gestures by important change detection with the spatiotemporal interest point (STIP) [36] also exhibited the effectiveness of the proposed method. The proposed method was outperformed by only [37] in which additional sensor devices like an accelerometer or gyroscope were utilized in addition to the RGB input. These sensors were available only on special circumstances.

Table 2. Gesture classification performance with different datasets ('-': not available).

Datasets/Methods	UTD Multimodal Human Action	Gaming 3D	NTU RGB+D	MSRDaily Activity3D
C3D [7]	85.3	89.1	83.3	87.5
LRCN [8]	83.0	-	-	-
ST-GCN [19]	-	-	88.3	-
I3D [34]	90.7	93.8	85.8	88.4
T_C3D [35]	89.5	90.3	85.7	88.9
STIP [36]	70.3	-	-	-
[37] (RGB, Accelerometer, Gyroscope)	97.6	-	-	-
TSSI + GLAN + SSAN [38]	-	-	89.1	-
Structure Preserving Projection (RGB+ Depth) [39]	-	-	-	89.8
ScTPM + CS-Mltp(RGB+ Depth) [40]	-	-	-	90.6
Proposed method	96.7	95.3	90.4	90.3

Table 3. Gesture classification performance in each modality and the fusion result on the UTD multimodal human action dataset.

Method	Accuracy (%)
3D_ResNet (RGB)	92.1
3D_ResNet (color body part segmentation)	94.6
LSTM (3D skeleton joints)	95.4
Fusion Result	96.7

For the NTU RGB+D dataset, to compare with the related model using only the RGB data input, our proposed method achieved outstanding results. While the ST-GCN network [19], which used a spatial–temporal graph convolutional network for a sequence of skeleton graphs, and the TSSI + GLAN + SSAN network [38], which utilized a two-branch attention architecture on skeleton images, obtained the highest accuracy in the literature, our method outperformed as shown in Table 1.

In the MSRDailyActivity3D dataset, our method outperformed the quality networks for video classification with only RGB video data input such as C3D, T_C3D, or I3D. Comparing with two

state-of-the-art models, the structure preserving projection method [39] and ScTPM + CS-Mltp [40], our method had the same or even a close performance.

5. Conclusions

In this paper, we presented a novel approach for dynamic gesture recognition by using RGB images. This approach is a combination of two 3D_ResNet networks and one LSTM network to benefit from multi-modalities of RGB frames, 3D skeleton joint information, and color body part segmentation. The output scores of the submodels were fused at the integrated stacking module to obtain the final gesture class prediction. The effectiveness of the proposed method was shown by the experimental results on four public datasets.

Author Contributions: Conceptualization, G.-S.L. and N.-H.N.; methodology, N.-H.N.; writing—review and editing, N.-H.N., T.-D.-T.P. and G.-S.L.; supervision, G.-S.L., S.-H.K. and H.-J.Y.; project administration, G.-S.L., S.-H.K. and H.-J.Y.; funding acquisition, G.-S.L., S.-H.K. and H.-J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058) and NRF funded by MSIT(NRF-2020R1A4A1019191).

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

References

1. Hussein, M.E.; Torki, M.; Gowayyed, M.A.; El-Saban, M. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, Beijing, China, 3–19 August 2013; Volume 13, pp. 2466–2472.
2. Dardas, N.H.; Georganas, N.D. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Trans. Instrum. Meas.* **2011**, *60*, 3592–3607. [CrossRef]
3. Lee, H.K.; Kim, J.H. An HMM-based threshold model approach for gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 961–973.
4. 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 (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1933–1941.
5. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
6. Ohn-Bar, E.; Trivedi, M.M. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2368–2377. [CrossRef]
7. 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.
8. 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.
9. Duan, J.; Zhou, S.; Wan, J.; Guo, X.; Li, S.Z. Multi-Modality Fusion based on consensus-voting and 3D convolution for isolated gesture recognition. *arXiv* **2016**, arXiv:1611.06689.
10. 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 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 31–36.
11. Wu, D.; Pigou, L.; Kindermans, P.J.; Le, N.D.H.; Shao, L.; Dambre, J.; Odobez, J.M. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1583–1597. [CrossRef] [PubMed]

12. Hara, K.; Kataoka, H.; Satoh, Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6546–6555.
13. Lin, G.; Milan, A.; Shen, C.; Reid, I. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
14. Rayat Imtiaz Hossain, M.; Little, J.J. Exploiting temporal information for 3D human pose estimation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 68–84.
15. Kopuklu, O.; Kose, N.; Rigoll, G. Motion fused frames: Data level fusion strategy for hand gesture recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2103–2111.
16. Molchanov, P.; Yang, X.; Gupta, S.; Kim, K.; Tyree, S.; Kautz, J. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4207–4215.
17. Tran, D.; Ray, J.; Shou, Z.; Chang, S.F.; Paluri, M. ConvNet architecture search for spatiotemporal feature learning. *arXiv* **2017**, arXiv:1708.05038, preprint.
18. Luvizon, D.C.; Picard, D.; Tabia, H. 2D/3D Pose estimation and action recognition using multitask deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5137–5146.
19. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
20. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3595–3603.
21. Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; Schiele, B. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In Proceedings of the International Conference on 3D Vision, Verona, Italy, 5–8 September 2018; pp. 484–494.
22. Fu, K.; Zhao, Q.; Gu, I.Y.H. Refinet: A deep segmentation assisted refinement network for salient object detection. *IEEE Trans. Multimed.* **2018**, *21*, 457–469. [CrossRef]
23. 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.
24. Lassner, C.; Romero, J.; Kiefel, M.; Bogo, F.; Black, M.J.; Gehler, P.V. Unite the people: Closing the loop between 3d and 2d human representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6050–6059.
25. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
26. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
27. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Proceedings of the IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015; pp. 168–172.
28. Bloom, V.; Makris, D.; Argyriou, V. G3D: A gaming action dataset and real time action recognition evaluation framework. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 7–12.
29. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A large scale dataset for 3D human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
30. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1290–1297.

31. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human action classes from video in the wild. *arXiv* **2012**, arXiv:1212.0402, preprint.
32. Sarkar, A.; Gepperth, A.; Handmann, U.; Kopinski, T. Dynamic hand gesture recognition for mobile systems using deep LSTM. In Proceedings of the International Conference on Intelligent Human Computer Interaction, Evry, France, 11–13 December 2017; pp. 19–31.
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.
34. Carreira, J.; Zisserman, A. Quo vadis, action recognition? In 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.
35. Liu, K.; Liu, W.; Gan, C.; Tan, M.; Ma, H. T-C3D: Temporal convolutional 3D network for real-time action recognition. In Proceedings of the Thirty-second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
36. Mahjoub, A.B.; Atri, M. Human action recognition using RGB data. In Proceedings of the 11th International Design & Test Symposium, Hammamet, Tunisia, 18–20 December 2016; pp. 83–87.
37. Ehatisham-Ul-Haq, M.; Javed, A.; Azam, M.A.; Malik, H.M.; Irtaza, A.; Lee, I.H.; Mahmood, M.T. Robust human activity recognition using multimodal feature-level fusion. *IEEE Access* **2019**, *7*, 60736–60751. [CrossRef]
38. Yang, Z.; Li, Y.; Yang, J.; Luo, J. Action recognition with spatio-temporal visual attention on skeleton image sequences. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 2405–2415. [CrossRef]
39. Yu, M.; Liu, L.; Shao, L. Structure-preserving binary representations for RGB-D action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1651–1664. [CrossRef] [PubMed]
40. Luo, J.; Wang, W.; Qi, H. Spatio-temporal feature extraction and representation for RGB-D human action recognition. *Pattern Recognit. Lett.* **2014**, *50*, 139–148. [CrossRef]