A Review on Multimodal Fusion Method for Gesture Recognition

Dong Jae Lee
Department of Electronics Engineering
Kookmin University
Seoul, South Korea
dj777@kookmin.ac.kr

Sunwoong Choi
Department of Electronics Engineering
Kookmin University
Seoul, South Korea
schoi@kookmin.ac.kr

***Abstract*— Recent research using deep learning has been actively conducted in various fields, including computer vision, reinforcement learning, classifiers, and more. AlphaGo, which learned to play Go and beat professional players, was developed based on reinforcement learning research. This paper focuses on computer vision in particular, which also has multiple subfields such as image restoration and image compression. This paper examines the use of deep learning with video data in computer vision. Video data can be divided into RGB and Depth, and the fusion of these two types of data will be used, referred to as multimodal fusion. By reviewing several papers, this method will be applied to gesture recognition research for potential improvements.**

Keywords—Multimodal fusion, CNN, Deep learning, Gesture recognition

I. INTRODUCTION

Deep learning has become a popular trend in artificial intelligence recently, and various models such as basic Convolutional Neural Network(CNN), Recurrent Neural Network(RNN), and Residual 3D CNN(R3D) with residual blocks are being developed. Deep learning is widely applied in different fields, including natural language processing, computer vision, and reinforcement learning. Machine learning and deep learning can be used to conduct various research in fields such as sleep, dementia, defect detection, and more. Another example is AlphaGo, which is the name of a device that was trained with machine learning to play the game of Go. AlphaGo gained popularity for its record of defeating world-class Go players.

In this research, the focus is on computer vision, specifically on motion recognition using video data. Different fusion methods are explored to create various situations and improve the accuracy of the model.

The technology of gesture recognition can be applied in various fields, such as drone operation, as long as a camera can be used for control. For instance, if the issue is too much light, a method can be employed to decrease the overall brightness while elevating the illumination of the object to be identified. Furthermore, video data is interwoven with chronological data, commonly known as Optical-Flow-Data. Two of the most representative methods for this are Gunner [1] and Lucas-Kanade [2]. By using the previous and current frames, the technique can determine the distance and direction of the object's movement. Learning can be achieved by combining RGB data, depth data, and optical flow data using their respective techniques. There are several learning approaches available for this purpose. There are various fusion methods, including data fusion, feature fusion, and decision fusion. Data fusion refers to the method of performing fusion before learning, followed by training. Feature fusion means performing fusion before the fully connected layer. Decision fusion means performing fusion after the fully connected layer. With the existence of various fusion methods, we will review several papers related to the requirements for multimodal fusion.

II. MULTIMODAL FUSION METHOD

A. RC3D(Residual Convolutional Neural Network)

First, as it is shown in Fig. 1, to create the RC3D[3] model, each of the three data types - RGB, Depth, and Optical Flow - is trained using the ResC3D model[4] as a base structure. After that, a technique called feature fusion is used to combine the learned features from each type of data. Finally, an algorithm called Support Vector Machine(SVM) is used to complete the learning process.

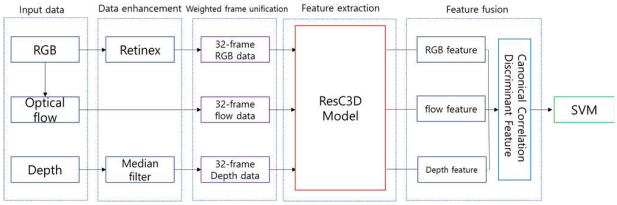

Fig. 1. Residual Convolutional Neural Network (RC3D) Model

Firstly, The RGB data used in the RC3D Model will be described. It is a type of data that contains the basic RGB colors - red, green, and blue. In this paper, Retinex Theory [5] is used to process the RGB data. Retinex is a method that adjusts the lighting conditions of an image. It averages the light reflected from the object's surface and the surrounding environment, which helps to make the object recognition more accurate. Bright areas are made darker, while dark areas are made brighter to enhance the image quality.

Optical flow[2] is a technique that converts temporal data into spatial data by extracting both position and direction information. By utilizing this information, the precision of subsequent video data can be enhanced. It involves deriving motion information between frames from RGB data and transforming it into optical flow data.

A median filter was utilized for pre-processing the depth data. The pre-processing of the data was performed using

three different methods. In the video, 32 frames need to be extracted as training data, and the entire data can be divided into three segments: beginning, climax, and ending. Since there are many important frames in the climax segment, more frames are extracted from the climax, and slightly fewer are extracted from the beginning and ending segments to form 32 frames. This method is applied to RGB, Depth, and optical flow. The training is carried out, and fusion is performed at the final decision fusion stage, followed by SVM for finalization.

The large-scale RGB-D gesture dataset - the Chalearn LAP IsoGD database[6] was used as the training data, consisting of 47,933 gestures and 249 labels. The conclusion is that fusion outperforms standalone RGB, Depth, optical flow, and C3D+LSTM models in terms of performance.

B. MMTM(Multimodal Transfer Module for CNN Fusion)

This paper discusses the use of late fusion in multimodal applications, where each modality is processed separately in a unimodal CNN stream, and their scores are combined at the end. Despite its simplicity, late fusion is still widely used in state-of-the-art multimodal applications. However, the paper proposes a new approach called the Multimodal Transfer Module (MMTM)[7], shown in Fig. 2, to overcome the limitations of late fusion. The MMTM is a simple neural network module that leverages knowledge from multiple modalities in CNN. It can be added at different levels of the feature hierarchy, allowing for slower modality fusion. The MMTM uses squeeze and excitation operations to recalibrate channel-wise features in each CNN stream, utilizing knowledge from multiple modalities. Overall, the paper presents an interesting approach to improve upon the limitations of late fusion in multimodal applications.

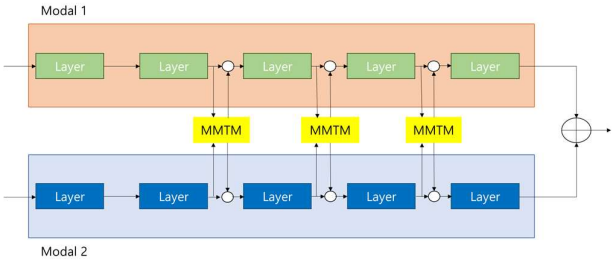

Fig. 2. Network Multimodal Transfer Module (MMTM)

This paper will provide a detailed explanation of the proposed MMTM module and how it can be added among unimodal branches with minimal changes to their network architectures, allowing each branch to be initialized with existing pretrained weights. Experimental results show that our framework improves the recognition accuracy of well-known multimodal networks. The authors demonstrate state-of-the-art or competitive performance on four datasets that span the task domains of dynamic hand gesture recognition, speech enhancement, and action recognition with RGB and body joints. For multimodal gesture recognition in video, video and skeleton modalities are widely used. These methods have disadvantages. Video recognition methods have difficulty processing background noise and non-action movements because there is no explicit human body model, Most of the context and global signals may be lost if only the body posture is relied on. Recently, an architecture is being developed to further improve action recognition by fusing the characteristics of these methods. Pose Map [8] is a space-time pose heat map and A two-stream network is used to process the skeleton separately, and late fusion is used for final prediction. An end-to-end-trainable multitask network for joint pose estimation and action recognition is also proposed.

In convolutional neural networks, the output features of convolutional layers are limited in their receptive field size, resulting in a lack of global context. To address this issue, a technique proposed in [9] is to squeeze the spatial information into the channel descriptors by performing global average pooling over the spatial dimensions of the input features. This process helps to capture the most salient features while reducing the dimensionality of the input. This approach has been shown to improve the performance of convolutional neural networks in various image recognition tasks.

The authors of the paper investigated the impact of various model choices on the performance of their multitask deep architecture for gesture recognition. They compared different architectures in the transfer layer and explored using different numbers of transfer layers. Their findings indicated that the convolutional MMTM variations did not significantly improve over the late fusion method, highlighting the importance of extracting information with global receptive field information in the squeeze unit. Moreover, the results of the self-excitation approach with no intermediate fusion indicated that most of the performance gain in MMTM was due to the slow fusion of the modalities rather than the pure squeeze and excitation method. The authors also discovered that the best performance was obtained when the output of half of the last inception modules was fused by MMTM, suggesting that mid-level and high-level features were more beneficial than low-level features from this approach.

C. Multi-Scale Attention 3D CNN for Multimodal Gesture Recognition

Gesture recognition is an important field within this area, where information from the hand is a key factor. However, the current method of using estimated key points has limitations, as it requires continuous attention to the hand and can lead to incorrect key point estimations, resulting in increased time and complexity, and potentially incorrect results.

To address these issues, the paper proposes a multi-scale attention C3D [10], as shown in Fig. 3, that utilizes a combination of local and global attention mechanisms to achieve overall attentional processing through multimodal data fusion. The local attention focuses on the hand region by utilizing a hand detector, reducing the interference of unrelated factors. Global attention is achieved through a dual spatio-temporal attention module that considers both the human posture context and channel context.

To take advantage of the differences between RGB and depth data, the proposed network uses a multimodal fusion method to combine their features. Overall, this approach improves upon existing methods by utilizing attention mechanisms and multimodal fusion techniques to achieve more accurate and efficient gesture recognition.

The network used in the study is called I3D. The authors start with RGB data, the network extracts information from both hands using a hand detection algorithm. The data from both hands is then input into the I3D network, and then immediately transferred to the multimodal fusion network.

The combined data of RGB and both hands then goes through I3D and two attention models, namely the Spatiotemporal Vision Attention Module (SVAM) and the Spatiotemporal Channel Attention Module (SCAM), which perform element-wise addition. After this, average pooling is performed, and the result is transferred to the multimodal fusion network.

In contrast to RGB data, depth data first goes through the I3D network and then passes through SVAM and SCAM attention models. After that, average pooling is performed, and the result enters the multimodal fusion network.

The three types of data, RGB, both hands, and depth, undergo element-wise multiplication in the multimodal fusion network section. Finally, the network extracts the label for the input data.

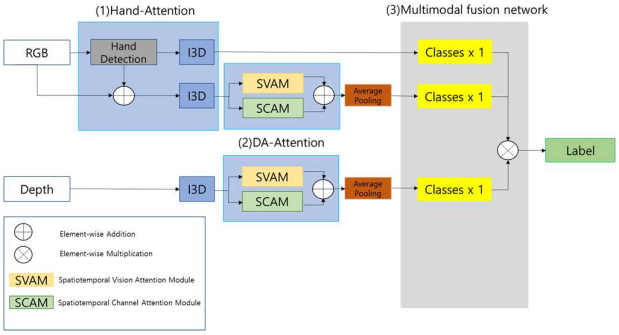

Fig. 3. Multi-scale Attention 3D Convolutional Neural Network. (1) A local attention module to enhance the network's attention to the hand region. (2) A dual spatiotemporal attention module to extract global spatiotemporal posture context information. (3) A multimodal fusion network for fusion different modality features.

SVAM and SCAM are attention modules used in the study. SVAM takes input information of size C x L x H x W and inserts a convolutional layer with a filter size of 1x1x1. It then proceeds with different reshapes and is divided into three types: Q, K, and V. Q is of the form [(H x W x L) x C], K is of the form [C x (H x W x L)], and V is of the form [(C x (H x W x L)], which has the same form as K. Q and K are multiplied element-wise, resulting in a matrix of size [(H x W x L) x (H x W x L)]. The result is then multiplied element-wise with V again. The output is obtained after performing a reshape and addition at the beginning.

This attention mechanism allows the network to focus on specific spatiotemporal regions in the input data and adjust the weights accordingly, enabling the network to learn relevant features for classification tasks.

SCAM is another attention module used in the study, which is similar to SVAM but with a different structure. The input data for SCAM has the structure [C x T x H x W]. The data is divided into three sections using reshape. The first section is of size [C x (H x W x T)], the second section is of size [(H x W x T) x C], and the third section is also of size [(H x W x T) x C].

The first and second sections undergo element-wise multiplication, resulting in a matrix of size [(H x W x T) x (H x W x T)]. The matrix is then normalized in the form of [C x C]. The third section is then multiplied element-wise with the result of the normalization, and the final output is obtained after performing a reshape and addition with the initial data.

This attention mechanism allows the network to focus on specific spatiotemporal regions in the input data, adjust the weights accordingly, and capture the inter-channel relationships, which can improve the network's classification performance.

The dataset used in the study is the ChaLearn LAP IsoGD [6] dataset, which is the same dataset used in the first paper referenced in the study. The I3D network used in the study was pre-trained using the Kinetics 400 dataset.

III. Result and Conculsion

We have reviewed three papers on multimodal gesture recognition, focusing on the field of computer vision gesture recognition, which is their current research area. Through this paper review, we have learned about the novel use of attention modules and various fusion methods. Additionally, we have discovered data preprocessing methods such as the median filter and Retinex, which can increase accuracy, as well as optical flow, which can provide information on the direction and location of movement.

We believe that this review provides an opportunity to develop further in their research field and apply these techniques to improve the accuracy of gesture recognition systems.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) (No. 2021R1F1A1062285).

References

[1] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29-July 2, 2003 Proceedings, vol. 13. Springer-Verlag, 2003.

[2] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," International Journal of Computer Vision, vol. 61, pp. 211-231, 2005.

[3] Q. Miao, X. Ding, Y. Zhu, and X. Ma, "Multimodal gesture recognition based on the ResC3D network," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017.

[4] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "ConvNet architecture search for spatiotemporal feature learning," arXiv preprint arXiv:1708.05038, 2017.

[5] E. H. Land and J. J. McCann, "Lightness and Retinex theory," JOSA, vol. 61, no. 1, pp. 1-11, 1971.

[6] I. Guyon, V. Athitsos, P. Jangyodsuk, and H. J. Escalante, "The ChaLearn Gesture Dataset (CGD 2011)," Machine Vision and Applications, vol. 25, no. 8, pp. 1929-1951, 2014.

[7] H. R. Vaezi Joze, M. S. Hosseini, H. R. Rabiee, and S. Shirazi, "MMTM: Multimodal Transfer Module for CNN Fusion," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[8] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[10] H. Chen, L. Wu, Z. Liu, and L. Zhang, "Multi-Scale Attention 3D Convolutional Network for Multimodal Gesture Recognition," Sensors, vol. 22, no. 6, pp. 2405, 2022.

[11] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.