Short-Term Temporal Convolutional Networks for Dynamic Hand Gesture Recognition

Yi Zhang*, Chong Wang†, Ye Zheng, Jieyu Zhao, Yuqi Li and Xijiong Xie ‡

January 17, 2020

Abstract

The purpose of gesture recognition is to recognize meaningful movements of human bodies, and gesture recognition is an important issue in computer vision. In this paper, we present a multimodal gesture recognition method based on 3D densely convolutional networks (3D-DenseNets) and improved temporal convolutional networks (TCNs). The key idea of our approach is to find a compact and effective representation of spatial and temporal features, which orderly and separately divide task of gesture video analysis into two parts: spatial analysis and temporal analysis. In spatial analysis, we adopt 3D-DenseNets to learn short-term spatio-temporal features effectively. Subsequently, in temporal analysis, we use TCNs to extract temporal features and employ improved Squeeze-and-Excitation Networks (SENets) to strengthen the representational power of temporal features from each TCNs' layers. The method has been evaluated on the VIVA and the NVIDIA Gesture Dynamic Hand Gesture Datasets. Our approach obtains very competitive performance on VIVA benchmarks with the classification accuracies of 91.54%, and achieve state-of-the art performance with 86.37% accuracy on NVIDIA benchmark.

Index terms— Gesture Recognition, 3D-DenseNets, TCNs, multimodal.

1 Introduction

Gesture recognition is a fast expanding field with applications in human-computer interaction[1], sign language recognition[2] and etc.. Due to subtle differences among similar gestures, complex scene background, different observation conditions, and noises in acquisition, robust gesture recognition is very challenging.

The main task of gesture recognition is to extract features from an image or a video and then classify or determine each sample to a certain label. Gesture

1

recognition aims to recognize and understand meaningful movement of human bodies in which arms and hands play crucial roles. Only few gestures can be identified from their spatial or structure information in an image or a single frame. In fact, motion cues and structure information simultaneously characterize a unique gesture. How to learn spatiotemporal features effectively is always the key in gesture recognition. Although in the past decades, many methods have been proposed for this issue, ranging from static to dynamic gestures, and from motion silhouettes-based to the convolutional neural network-based, there are still many challenges associated with the recognition accuracy.

At present, although most existing models have reached a high performance for isolated gesture recognition, most methods have been developed based on Convolutional Neural Networks (CNNs)[3][4] or Recurrent Neural Networks (RNNs)[5]. With the development of deep learning, more and more new architectures of CNNs have been proposed, especially DenseNets [6] what have powerful feature extraction ability. Meanwhile, a new architecture to solve sequence problem named TCNs [7] have been proposed. Compared to RNNs and their canonical recurrent architectures such as LSTMs and GRUs, TCNs have comparable clarity and simplicty. In our approach, we adopt 3D-DenseNet to extract short-term stapio-temporal features, then these features are input into the TCNs to finish the task of Classification.

However, recently, for extracting more complete temporal features, a few methods have been proposed based on attention mechanism. The research prove that there are various relationships between features' interior in neural networks. SENets [8] are new architectural unit with the goal of improving the quality of representations produced by a network by explicitly modelling the interdependencies between the channels of its convolutional features. And in our approach, we reform SENets and combine them into TCNs to strengthen capacity of TCNs in temporal features extracting.

The pipline of our method is depicted in Figure 1, and the main contribution can be summarized as following:

- **Spatial analysis.** We design a multi-stream truncated 3D-DenseNet, which extracts spatio-temporal features from a video, and through local temporal pooling, obtain the decomposed short-term spatio-temporal features, to solve problem that single frame image can not carry enough spatial or structure information of gesture and reduce repetitive training for video clips.

- **Temporal analysis.** We employ TCN to replace RNN as the main model of sequence information feature analysis. In addition, we improve SENets and apply them in temporal domain to rescale the weights between temporal features and extract more effective temporal features to achieve higher classification accuracy.
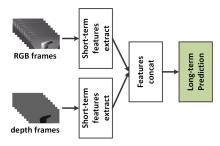
Figure 1: An overview of the proposed method. The proposed deep architecture is composed of two main steps: (a) Multimodal short-term spatio-temporal feature sequence extracting by truncated 3D-Densenet (T3D-Dense), local temporal average pooling (LTAP) and multimodal features concatenation. (b)Long-term feature sequence recognizing via TCN and TSE.

2 Related Work

Gesture taxonomies and representations have been studied for decades.The vision based gesture recognition techniques include the static gesture oriented and the dynamic gesture oriented methods [1].

Recently, convolution neural networks (CNNs) [9] have made a great breakthrough on computer vision related tasks by their powerful feature extraction ability, thus the features extracted by CNNs are widely used in many action classification tasks instead of hand-crafted features for better performance. Features are extracted by 2D-CNN from the starts. bi-directional rank pooling [10][11] was used to encode the spatial and temporal information of videos. Temporal convolutions for gesture recognition in videos Beyond temporal pooling [12] was proposed to solve gesture recognition problem in videos by a new temporal pooling method. On the other hand, C3D[13] model is developed and provides a better performance and main contribution in this research is proposed an architecture to extract spatio-temporal features from a video clip. Concurrently, a multi-stream 3D-CNN[14] was designed for hand gesture recognition and the classifier consisted of two subnetworks: a high-resolution network (HRN) and a low-resolution network (LRN) in this model.

Meanwhile, with the development of convolutional neural networks, more and more architectures of CNNs were proposed, like AlexNet [9], VGGNet [15], GoogleNet [16] [17] [18] [19], ResNet [20] and DenseNet [6]. All of these models have one target that is building a higher architectures of CNNs to dig deeper and more complete statial features from low-level image frames, and then classify. In the area of isolated gesture recognition, Res-C3D model[21] was used and won the first place twice in ChaLearn LAP Multi-modal Isolated Gesture Recognition Challenges 2016 [22] and 2017 [23]. Whatmore, DenseNets as one of the latest convolutional architectures, was adopted in action recognitions especially face recognitions and gesture recognitions gradually. A face recognition model
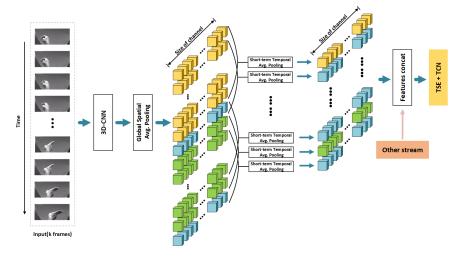
Figure 2: The architecture of 3D-DenseNet.

named Dense Face[24] was proposed to explore the performance of densely connected network in face recognition. DenseNets[25] also was used to classifier the different actions in recent researches.

Regarding the temporal information of the video sequences, Long Short Term Memory(LSTM) networks is a common choice to gesture recognition. For instance, convolutional LSTM[26] was introduced for spatio-temporal feature maps. 2S-RNN(RGB and Depth)[27] was used for continuous gesture recognition. However, RNNs including LSTMs and GRUs have some weaknesses on temporal domain like short-range information learning, oversized memory capacity. To make these weaknesses up, TCNs is proposed and applied in the gesture reconition. Res-TCN[28] was proposed for skeleton-based dynamic hand gesture recognition. Whatmore, a model based on TCN[29] was proposed for gesture recognition.

Other important works based on attention mechanism. Attention mechanism or attention model firstly was applied to neural networks by Vaswani et al[30]. After that, more and more researches are proposed based on attention mechanism, so as SENets [8] that improve ResNets to win first place of ILSVRC 2017 classification.

3 Our Approach

In the video recognition, both of the spatial and temporal information are important. Although there have been impressive progress in spatial feature extraction using 2D-CNNs based networks[14][3], how to effectively learn the temporal features is still a very challenging problem. Unlike the 2D-CNNs focusing on the single image, various 3D-CNN based networks[5][31][32][33] have been proposed

4

to process the successive frames simultaneously. For the video of dynamic hand gestures, adjacent frames are usually similar and containing the same static gesture, while the static gestures change several times during the whole video. Thus, in this paper we decompose the video to two different parts. One is the short-term spatio-temporal information in the adjacent frames, and the other is the long-term temporal information analysed by a sequential model. Based on this consideration, we raised two major questions,

- how to learn short-term spatio-temporal features effectively from video clips in the same video.

- how to reasonably classify a sequence which is combined from these consecutive features.

In order to address these issues, we designed a novel architecture to extract a sequence of short spatio-temporal features in order to recognize dynamic gestures.

As depicted in Figure 1, the overall process can be divided into two parts: 1) multi-modal short-term spatio-temporal feature extraction based on 3D-DenseNets and 2) spatio-temporal sequence classify with and temporal SENets embedded TCNs. To be specific, the details of the proposed network structure is presented in Figure **??** and Figure **??**.

3.1 temporal local pooling to extract short-term features

Due to the availability of various data types and the nature of signing videos, a more robust feature representation can acquired from the incorporation of multi-modal hand gesture information. To effectively present the the location, shape and sequential information in the adjacent gesture frames, we design a multi-stream DenseNet based on the C3D[13] to extracts short-term spatio-temporal features. Assume a given video V with n frames, it is firstly re-sampled to k frames. Thus, the input video V_S is denoted as,

$$V_S = [v_1, v_2, ..., v_k] \tag{1}$$

where v_k is the k-th frame image of video sequence in the input.

As aforementioned, we consider multiple modalities of gesture video data as the input. Each type of the data is set as one data stream and fed to the same network structure. The outputs of them will be fused together later as shown in Figure 1. The proposed model contains 4 dense blocks, containing 6, 12, 24, 16 layers respectively. Following the basic design in DenseNet[6] and C3D[13], the detailed network configurations are shown in Table 1. It is worth noting that most of the convolution layers are with $3 \times 3 \times 3$ filters, which limits the process only on the local spatial and temporal domain. Moreover, the temporal pooling size and stride in all the transition layers are set as 1 to avoid the fusion of the short-term temporal information, which is one major difference from the other conventional 3D-CNNs[13].

Table 1: **3D-DenseNet architectures. The growth rate of network is k = 12. Note that each "conv" layer shown in the corresponds the sequence BN-ReLU-Conv.**

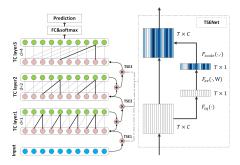| Layers	Filter Size
Convolution	$5 \times 5 \times 5$ conv, stride $2 \times 2 \times 1$
Pooling	$3 \times 3 \times 1$ max pool, stride $2 \times 2 \times 1$
Dense Block 1	$\begin{bmatrix} 1 \times 1 \times 1 \ conv \\ 3 \times 3 \times 3 \ conv \end{bmatrix} \times 6$
Transition Layer 1	$1 \times 1 \times 1$ conv $2 \times 2 \times 1$ average pool, stride $2 \times 2 \times 1$
Dense Block 2	$\begin{bmatrix} 1 \times 1 \times 1 \ conv \\ 3 \times 3 \times 3 \ conv \end{bmatrix} \times 12$
Transition Layer 2	$1 \times 1 \times 1$ conv $2 \times 2 \times 1$ average pool, stride $2 \times 2 \times 1$
Dense Block 3	$\begin{bmatrix} 1 \times 1 \times 1 \ conv \\ 3 \times 3 \times 3 \ conv \end{bmatrix} \times 24$
Transition Layer 3	$1 \times 1 \times 1$ conv $2 \times 2 \times 1$ average pool, stride $2 \times 2 \times 1$
Dense Block final	$\begin{bmatrix} 1 \times 1 \times 1 \ conv \\ 3 \times 3 \times 3 \ conv \end{bmatrix} \times 16$
Classification Layer	global spatial average pool global temporal average pool fully-connected, softmax

Figure 3: Architecture and architectural elements in a TCN. There is an example that dilated causal convolution with dilation factors d = 1,2,4 and filter size k = 2 in figure. The receptive field is able to cover all values from the input sequence. And adjacent layers are connected by residual block.Before temporal convolution layer, the inputs need to go through the corresponding Temporal Squeeze-and-Excitation(TSE) layer to adjust weight of input in temporal domain.

Since the 3D-Densenet is served as a short-term spatio-temporal features extractor, we truncate it to obtain the features only. To be specific, the global temporal average pooling layer, last softmax and fully-connected layers are discarded, after the model is first pre-trained with isolated gesture data.

Therefore, we can get the global spatio-temporal feature F_k after the global spatial average pool layer,

$$F_k = [f_1, f_2, , f_k] \tag{2}$$

where the temporal length is k and represent respective spatial feature of k frames.

Then T short-term spatio-temporal features are cut and pooled from the global feature F_k. The t-th short-term spatio-temporal feature x_t is constructed as,

$$x_t = ltap[f_{t-\frac{k}{T}}, f_{t-\frac{k}{T}+1}, , f_{t+\frac{k}{T}-1}] \tag{3}$$

where $ltap$ is local temporal average pool layer in truncated 3D-Densenet, $\frac{k}{T}$ is half of temporal feature interval. In this way, the adjacent $ltap$ windows also overlapping that assure the relevance and completeness of the front and back frame information.

After local temporal average pooling, we can get a sequence of short-term features in single modality. Multimodal feature sequences are fused into one sequence before input into TCN. In this paper, all feature sequences of different modality are concated in channel dimension.

3.2   TSENet + TCN for long-term prediction

Based on the short-term spatio-temporal features extracted from all kinds of data modalities (RGB, optic flow, depth, etc.), the long-term temporal features
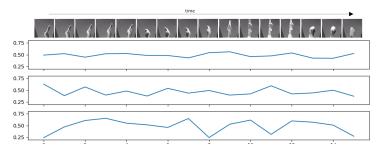
Figure 4: An example sequence from VIVA gesture and its corresponding temporal weghts from TSE-Nets.

of the whole video is considered to classify the category of the given hand gesture. In this work, a sequence recognition model named TCNs is employed and modified to process the long-term temporal information. The main characteristics of TCNs are the use of causal convolutions and the mapping of an input sequence to an output sequence of the same length. In addition, accounting for sequences with long history, this model uses dilated convolutions that enable a large receptive field as well as residual connections that allow training deeper networks. Considering that our task is to classify the category of hand gesture videos, the output layer of TCN is further processed by one fully connection layer to obtain a single class label for each gesture sequence. The structure of the proposed modified version of the TCN model is depicted in Figure **??**.

The short-term temporal features $X = [x_1, , x_T]$ are utilized as the input sequence of the proposed modified TCN with the outputs $Y = [y_1, , y_T]$, while the calculation of y_t, $t < T$ depends only on $X = [x_1, , x_T]$. The reason is that the dilated convolutions are calculated as,

$$y_t = (x *_d h)_t = \sum x_{t-d_m} h_m \tag{4}$$

where $*_d$ is the operator for dilated convolutions, d is the dilation factor and h is the filter's impulse response. For a TCN with L layers, the output of the last layer y^L is used for the sequence classification. The class label $\widehat{o}$ attributed to the sequence is found through a fully connected layer with a softmax activation function,

$$\widehat{o} = softmax(W_o \cdot y^L + b_o) \tag{5}$$

where W_o, b_o are trainable parameters.

It is noting that the short-term spatio-temporal features $x_1, , x_T$ actually have different contributions to the recognition in the long-term temporal information processing. For instance, the gesture "swipe +" in Figure 4(b) contains three paths. The first path is extremely similar to the gesture "swipe left" (Figure 4(a)) when $t < 9$. The same phenomenon occurs between the third path of the gestures "swipe +" and "swipe down" (Figure 4(c)) when $t > 23$. In order to assign different temporal weight to $X = [x_1, , x_T]$, a temporal

8

Squeeze-and-Excitation network (TSENet) block is inserted between each temporal convolution layers.

As shown in Figure **??**, the average pooling is applied on the channel dimensions C of $X = [x_1, , x_T]$ to squeeze channel-wise information. Such obtained temporal descriptor $z = [z_1, , z_T]$ is a $T \times 1$ vector, while the t-th element of z is calculated as,

$$z_t = F_{sq}(x_t) = \frac{1}{C} \sum_{i=1}^{C} x_t(i) \tag{6}$$

Then another excitation operation is followed to capture the temporal dependencies, i.e. the temporal weights. To fulfil this objective, we opt to employ a simple gating mechanism with the activations:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z))) \tag{7}$$

where σ refers to the sigmoid fuction, δ refers to the ReLU fuction, $W_1 \in \mathbb{R}^{\frac{T}{r} \times T}$ and $W_2 \in \mathbb{R}^{T \times \frac{T}{r}}$, and r is the size of squeeze channel. The final output of the block is obtained by rescaling the transformation output U with the activations:

$$\widetilde{x} = F_{scale}(u_t, s_t) = s_t \cdot u_t$$

where $\widetilde{X} = [\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_T]$ and $F_{scale}(u_t, s_t)$ refers to temporal-wise multiplication between the scalar s_t and the feature map $u_t \in \mathbb{R}^T$.

An example of the weights on different TSENet layers is illustrated in Figure 5. It can be seen that the values of the weights changes corresponding to the input gesture sequence as desired.

4   Experiment

The proposed network architecture is implemented by tensorflow, and trained using one NVIDIA Quadro GP100 GPU. Multimodal 3D-DenseNet models have same structures and are pre-trained using RGB and optic flow(if optic flow existed or can be calculated) data respectively. Adam optimizer is used for training 3D-DenseNet and the learning rate is initialized to $6.4e-4$ and decayed by 10 every 25 epochs. The weight decay is set to $1e-4$. And the dropout rate is set to 0.2. The compression rate and the growth k in the DenseNet block are set as 0.5 and 12, respectively. For the TCN model, we use Adam optimizer for training, and the learning rate is initialized to $1e-4$, epsilon is $1e-8$.

4.1   Dataset

In this section, we compare our method with the other state-of-the-art dynamic hand gesture methods. Two publicly available multi-modal dynamic hand gesture datasets (VIVA[14] and NVGesture[5]) are used to evaluate our proposed model in the experiments.

VIVA[14] The VIVA challenges dataset is a multimodal dynamic hand gesture dataset specifically designed with difficult settings of cluttered background, volatile illumination, and frequent occlusion for studying natural human activities in real-world driving settings. This dataset was captured using a Microsoft Kinect device, and contains 885 intensity and depth video sequences of 19 different dynamic hand gestures performed by 8 subjects inside a vehicle. Figure 4 shows some gesture sequences.

NVGesture[5] The NVGesture dataset has been captured with multiple sensors and from multiple viewpoints for studying human-computer interfaces. It contains 1532 dynamic hand gestures recorded from 20 subjects inside a car simulator with artificial lighting conditions. This dataset includes 25 classes of hand gestures. The gestures were recorded with SoftKinetic DS325 device as the RGB-D sensor and DUO-3D for the infrared streams. In the experiments, we use RGB, depth and optical flow modalities, while the optical flow is calculated from the RGB stream using the method presented in [34].

4.2   Data Preprocessing

In VIVA dataset, data augmentation is comprised of three other operations: reverse ordering of frames, horizontal mirroring, and applying both operations together. With these operations we generated additional samples for training. For example, applying both operations transforms the original gesture "Swipe Left" with the right hand to a new gesture "Swipe Left" with the left hand . In NVGesture dataset, for special augmentation, videos are resized to have the smaller video size of 256 pixels, and then randomly cropped with a 224x224 patch.

Data normalization is also applied on both datasets, since a fixed dimension of input data is required in the C3D model and TCN model. For the videos with different temporal lengths, uniform normalization with temporal upsampling and downsampling is used. To compress or extend a given video V with n frames to k frames, 1) If $n > k$, we split the video V into a k section video set V_S averagely, where $V_S = [V_1, V_2, ..., V_k]$. For each piece in the video set V_S, we randomly choose one frame as the representation of the sub-video fragment. Finally we concatenate all the represent frames and make them as the result of the normalization. 2) If $n < k$, we randomly choose $k - n$ frame in the video, then repeat them follow by themselves.

In our experiments, the average number of frames k is set as 32 for VIVA dataset and 64 for NVGesture dataset. Due to the high complexity of 3D convolutional calculating, the spatial size of the inputs is restricted to 112×112.

4.3   Evaluation on VIVA Dataset.

Table 2 shows the performance of the dynamic hand gestures tested on the RGB and depth modalities of the VIVA dataset. The compared methods include the hand-crafted approach HOG+HOG2[35], the recurrent CNN-based
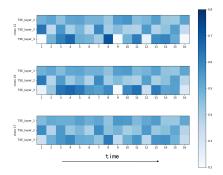
Figure 5: Examples of temporal weights.

method(CNN:LRN)[14], the C3D model which were pretrained on Sport-1M dataset, the I3D method[32] that performs very well in action recognition, and the Multimodal Training / Unimodal Testing (MTUT) model[33] which shows promising performance in dynamic hand gesture recognition. All the results are reported by averaging the classification accuracies. It can be seen that the proposed model achieves the highest accuracy, which is 5.46% higher than the state-of-the-art method MTUT. This experiment shows that our model is effective to extract both short-term and long-term spatio-temporal information for dynamic hand gesture recognition.

To validate the effect of the proposed TSENet layers, the accuracy obtained by vanilla TCN is also shown in Table 2. It can be seen that the presence of the TSENet layers in the TCN can improve the recognition rate by around 0.8%. Three examples of the temporal weights produced by TSENet layers are shown in Figure 6. It is interesting to see that the weights in the third layer contain obvious large and small values, which means it does select the important ones from the short-term features. Moreover, if we change the 3D-Dense networks to Res3D which is used for extracting the short-term features. The accuracy will further drop about 4.8%. It proves the effectiveness of the structure of the proposed model.

Figure 6a shows the confusion matrix as well for the experiment. It can be seen that the proposed model confused between the Swipe and Scroll gestures performed along the same direction. Many gestures were mis-classified as the Swipe down gesture, the Rotate CW/CCW gestures were difficult for the proposed model. In some case, the propose model may have difficulties with distinguishing between the Swipe + and the Swipe X gestures.

4.4   EVALUATION ON NVgesture.

The NVGesture dataset, containing RGB, depth and optical flow modalities, is also used to test the proposed model. Table 3 tabulates the results of our method in comparison with the recent state-of-the-art methods: HOG+HOG2[35], im-

(a) confusion matrix on the VIVA dataset   (b) confusion matrix on the NVGesture dataset
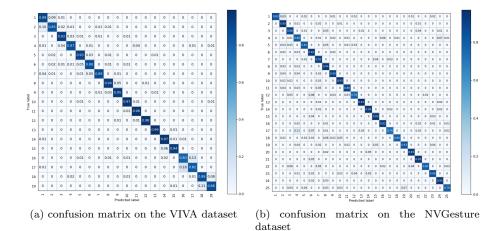
Figure 6: The confusion matrices obtained by comparing the grand-truth labels and the predicted labels from the RGB+depth modalities on the VIVA dataset and the RGB+opt.flow+depth modalities on the NVGesture dataset by our model. Best seen on the computer, in color and zoomed in.

proved dense trajectories(iDT)[31], R3DCNN[5], two-stream CNNs[3], and C3D as well as human labeling accuracy. The iDT method is often recognized as the best performing hand-crafted method. However, we observe that similar to the pervious experiments the 3D-CNN-based methods outperform the other hand gesture recognition methods, and among them, our method provides the better performance in all the modalities. Nonetheless, compare to the latest method MTUT, our method accuacies are close to the MTUT. Our method has the better performance in both of RGB and optical flow modalities, it improve accuracy by 0.73%. But in RGB+Depth modalities and in RGB+Depth+Opt.flow modalities, our method is not performing good enough. This is in part due to the knowledge that gestures in NVGesture are more complex and have more invalid information. Although through TCN and TSE, our method can key information in the frames and weaken the influence of irrelevant information, the redundant non gesture information, especially in temporal, always affects the final results of the experiment.

5 Conclusion

We developed an effective method for multi-modal (RGB, depth and optic flow data) dynamic hand gesture recognition with 3D-DenseNets and TCNs. And in TCNs, we improved and applied an attention model named SENets to learn and extract deeper temporal features. The experiments show that the proposed model achieved the highest accuracy in VIVA dataset, as well as competitive

results in NVGesture dataset.

However, our model is still not an end-to-end model and has to be trained step by step. Meanwhile, NVGesture still have a large room for improvement, we still have a lot of work to enhance the accuracy of the model.

References

[1] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 1–54, 2015.

[2] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 3075–3084.

[3] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[4] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *European Conference on Computer Vision*. Springer, 2014, pp. 474–490.

[5] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4207–4215.

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[7] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona, "Large-scale isolated gesture recognition using convolutional neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 7–12.

[11] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 773–787, 2016.

[12] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 430–439, 2018.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[14] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3d convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 1–7.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[19] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Q. Miao, Y. Li, W. Ouyang, Z. Ma, X. Xu, W. Shi, and X. Cao, "Multimodal gesture recognition based on the resc3d network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3047–3055.

[22] H. J. Escalante, V. Ponce-López, J. Wan, M. A. Riegler, B. Chen, A. Clapés, S. Escalera, I. Guyon, X. Baró, P. Halvorsen *et al.*, "Chalearn joint contest on multimedia challenges beyond visual analysis: An overview," in *2016 23rd international conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 67–73.

[23] J. Wan, S. Escalera, G. Anbarjafari, H. Jair Escalante, X. Baró, I. Guyon, M. Madadi, J. Allik, J. Gorbova, C. Lin *et al.*, "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*, 2017, pp. 3189–3197.

[24] T. Zhang, R. Wang, J. Ding, X. Li, and B. Li, "Face recognition based on densely connected convolutional networks," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*. IEEE, 2018, pp. 1–6.

[25] W. Hao and Z. Zhang, "Spatiotemporal distilled dense-connectivity network for video action recognition," *Pattern Recognition*, vol. 92, pp. 13–24, 2019.

[26] L. Zhang, G. Zhu, P. Shen, J. Song, S. Afaq Shah, and M. Bennamoun, "Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3120–3128.

[27] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen, "Two streams recurrent neural networks for large-scale continuous gesture recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 31–36.

[28] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, "Spatial-temporal attention res-tcn for skeleton-based dynamic hand gesture recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[29] P. Tsinganos, B. Cornelis, J. Cornelis, B. Jansen, and A. Skodras, "Improved gesture recognition based on semg signals and tcn," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1169–1173.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[31] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 219–238, 2016.

[32] 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, pp. 6299–6308.

[33] M. Abavisani, H. R. V. Joze, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1165–1174.

[34] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[35] E. Ohn-Bar and M. M. Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE transactions on intelligent transportation systems*, vol. 15, no. 6, pp. 2368–2377, 2014.

Table 2: **Accuracies of different multimodal fusion-based hand gesture methods on the VIVA dataset. The top performer is denoted by boldface.**

Method	Fused modalities	Accuracy
HOG+HOG2[35]	RGB+Depth	64.5
CNN:LRN[14]	RGB+Depth	74.4
CNN:LRN:HRN[14]	RGB+Depth	77.5
C3D[13]	RGB+Depth	77.4
I3D[32]	RGB+Depth	83.10
MTUT[33]	RGB+Depth	86.08
3D-Dense	RGB+Depth	88.21
Res3D+TCN	RGB+Depth	85.97
3D-Dense+TCN	RGB+Depth	90.73
3D-Dense+TCN$_{tse}$(*proposed*)	RGB+Depth	**91.54**

Table 3: **Accuracies of different multimodal fusion-based hand gesture methods on the NVGesture dataset. The top performer is denoted by boldface.**

Method	Fused modalities	Accuracy
HOG+HOG2[35]	RGB+Depth	36.9
I3D[32]	RGB+Depth	83.82
MTUT[33]	RGB+Depth	**86.10**
Ours	RGB+Depth	84.87
Two Stream CNNs[3]	RGB+Opt. flow	65.6
iDT[31]	RGB+Opt. flow	73.4
I3D[32]	RGB+Opt. flow	84.43
MTUT[33]	RGB+Opt. flow	85.48
Ours	RGB+Opt. flow	**86.21**
R3DCNN[5]	RGB+Depth+Opt. flow	83.8
I3D[32]	RGB+Depth+Opt. flow	85.68
MTUT[33]	RGB+Depth+Opt. flow	**86.93**
Ours	RGB+Depth+Opt. flow	86.37
Human labeling accuracy:		88.4