Learning Spatiotemporal Features using 3DCNN and Convolutional LSTM for Gesture Recognition

Liang Zhang, Guangming Zhu, Peiyi Shen, Juan Song, Syed A. Shah, Mohammed Bennamoun
{liangzhang, gmzhu, pyshen, songjuan}@xidian.edu.cn {afaq.shah, mohammed.bennamoun}@uwa.edu.au

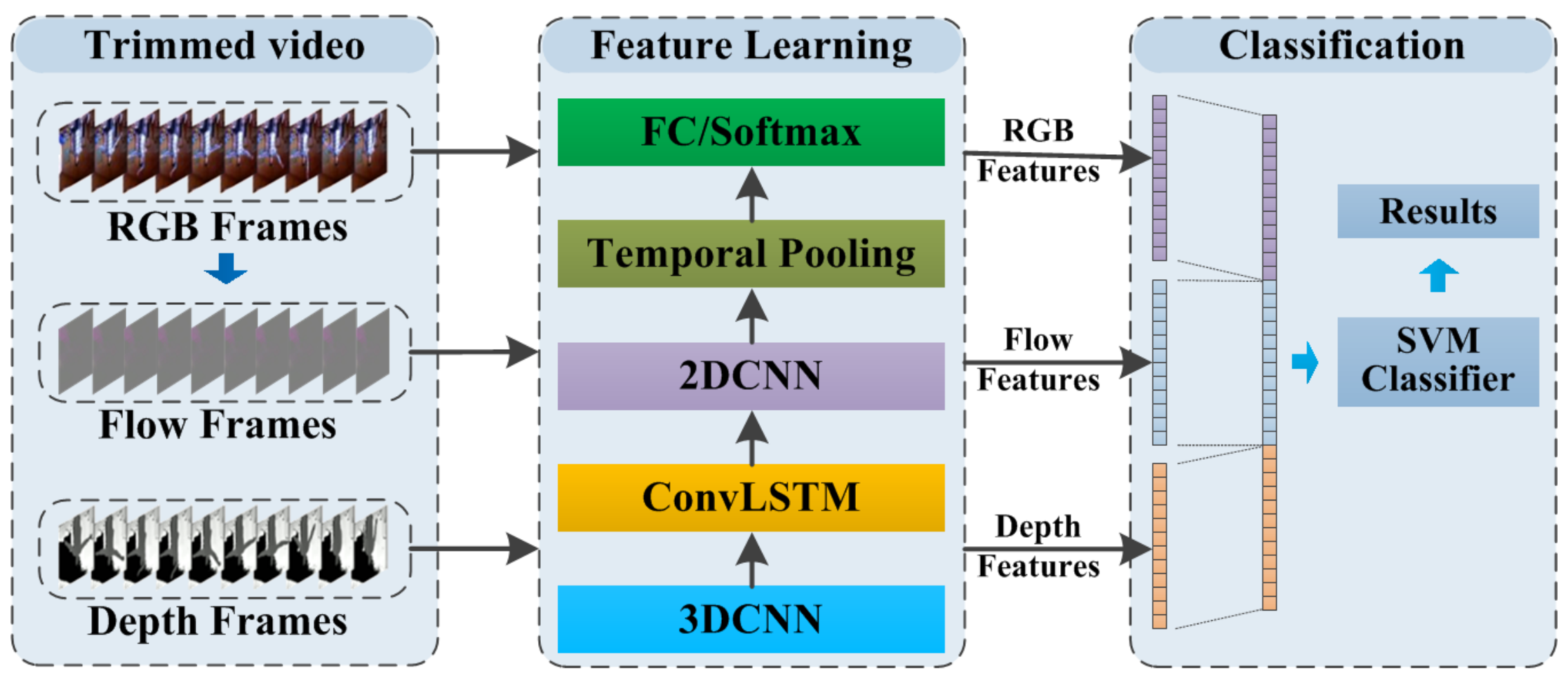
Contributions

We propose to first learn short-term spatiotemporal features using a **shallow 3DCNN**, and then learn long-term spatiotemporal features further using **bidirectional convolutional LSTM** (ConvLSTM), lastly recognize gestures using **2DCNN** based on the learnt 2D spatiotemporal feature maps.

- 2D spatiotemporal feature maps are learnt using 3DCNN and bidirectional convolutional LSTM. The 2D feature maps can encode the global temporal information and local spatial information. Spatiotemporal correlation information is kept through the whole feature map learning process.
- The proposed deep architecture can transform video files into 2D spatiotemporal feature maps. This transformation makes the deep architecture more extensible to utilize the state-of-the-art 2DCNN to learn the higher-level spatiotemporal features for gesture recognition.
- To the best of our knowledge, this is the first time to learn 2D spatiotemporal feature maps using 3DCNN and bidirectional ConvLSTM, and then to learn higher-level spatiotemporal features using 2DCNN for the final gesture recognition.

Framework

The proposed deep architecture is mainly composed of two components: *2D spatiotemporal feature map learning* and *classification based on the 2D feature maps*. The former learns 2D spatiotemporal feature maps from the normalized inputs using a shallow 3DCNN and bidirectional convolutional LSTM. The latter learns higher-level spatiotemporal features further using 2DCNN, and then uses a linear Support Vector Machine (SVM) classifier for the final gesture recognition. Multimodal fusion is employed to improve the recognition accuracy.



References

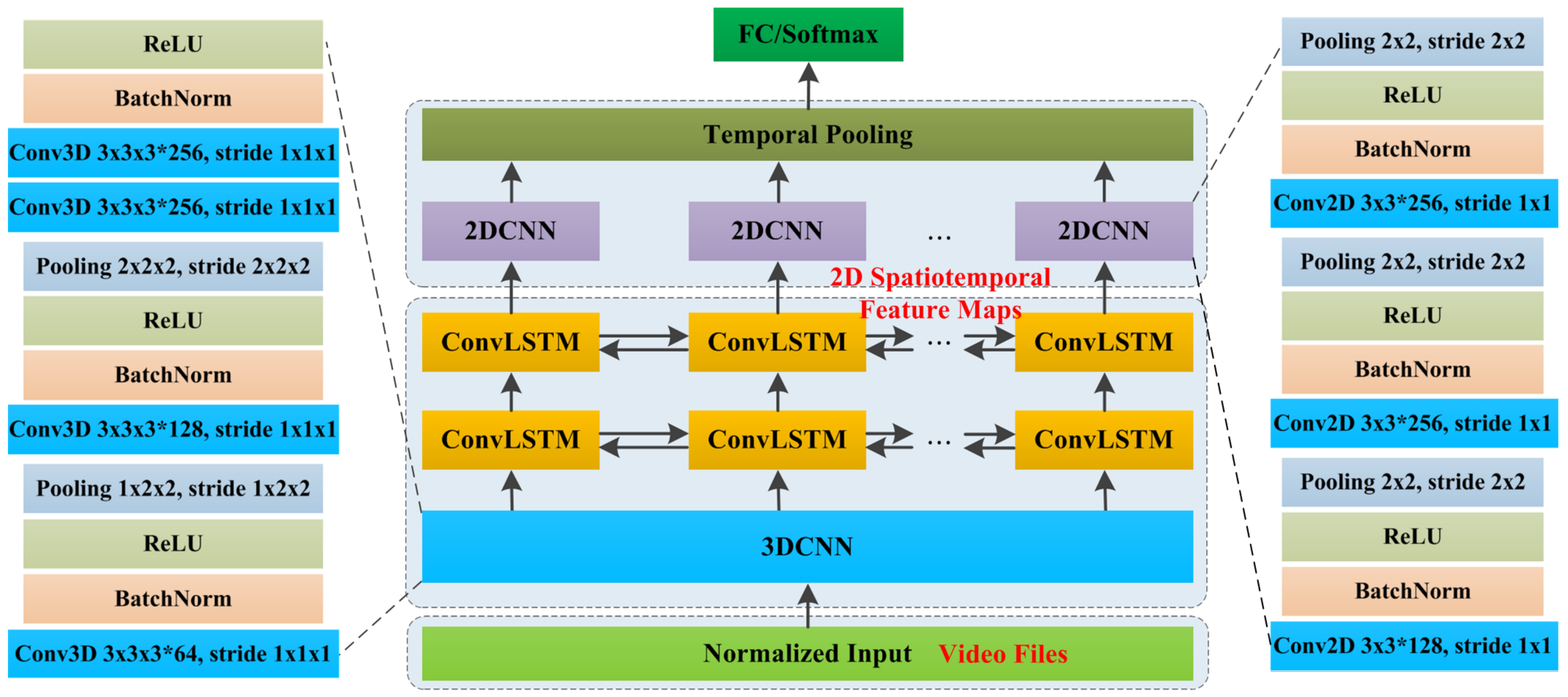
[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in ICCV, 2015, pp.4489-4497.

[2] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. WOO, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in NIPS, 2015, pp. 802–810.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China under Grant No.61702390, No.61401324, and No.61305109, and the China Postdoctoral Science Foundation under Grant No.2016M592763.

Proposed Deep Architecture


ReLU
BatchNorm
Conv3D 3x3x3*256, stride 1x1x1
Conv3D 3x3x3*256, stride 1x1x1
Pooling 2x2x2, stride 2x2x2
ReLU
BatchNorm
Conv3D 3x3x3*128, stride 1x1x1
Pooling 1x2x2, stride 1x2x2
ReLU
BatchNorm
Conv3D 3x3x3*64, stride 1x1x1
FC/Softmax
Temporal Pooling
2DCNN
2DCNN
...
2DCNN
2D Spatiotemporal Feature Maps
ConvLSTM
ConvLSTM
...
ConvLSTM
ConvLSTM
ConvLSTM
...
ConvLSTM
3DCNN
Normalized Input
Video Files
Pooling 2x2, stride 2x2
ReLU
BatchNorm
Conv2D 3x3*256, stride 1x1
Pooling 2x2, stride 2x2
ReLU
BatchNorm
Conv2D 3x3*256, stride 1x1
Pooling 2x2, stride 2x2
ReLU
BatchNorm
Conv2D 3x3*128, stride 1x1

The proposed deep architecture is illustrated as in the above figure. Three facts are taken into consideration when constructing the deep architecture: **a)** 3DCNN is a representative and outstanding deep architecture for spatiotemporal feature learning; **b)** RNN/LSTM networks are more suitable for long-term temporal information learning; **c)** Spatiotemporal correlation information plays an important role for gesture recognition. Therefore, we propose to use 3DCNN [1] and ConvLSTM [2] for spatiotemporal feature learning. ***3DCNN is designed to learn local or short-term spatiotemporal features***, so it does not need to be deep. ***Bidirectional ConvLSTM is designed to learn global or long-term spatiotemporal features***. The spatiotemporal correlation information is encoded during the recurrent process. The spatial size and the temporal length are only shrunk by a ratio of 4 and 2 respectively in the 3DCNN component, and ConvLSTM does not change the spatial size. Thus, large 2D spatiotemporal feature maps are learnt.

Generally, video files need to be decoded into separate image files or encoded into special feature images when the state-of-the-art 2DCNN networks are employed in video-based applications. In this paper, the proposed deep architecture can encode video files into 2D spatiotemporal feature maps. This enables 2DCNN to be used in video-based applications in an alternative way.

Exerimental Results

Experimental Setup. A variant of the proposed architecture, in which the softmax classifier is executed on the output of 2DCNN at each recurrent step first and then the recognition results are fused to report the final recognition accuracy, is evaluated to compare with the proposed architecture. The superscripts ^{a} and ^{b} in Table 1 indicate the variant architecture and the proposed deep architecture respectively. MaxFusion and AvgFusion in Table 1 denote the final score fusion strategies in the variant deep architecture. MaxPooling and AvgPooling in Table 1 denote the temporal pooling methods in the proposed deep architecture.

Fusion Methods	Modality	Accuracy(%)
^{a}MaxFusion	RGB	50.48
^{a}MaxFusion	Depth	47.93
^{a}MaxFusion	RGBD	54.55
^{a}AvgFusion	RGB	50.97
^{a}AvgFusion	Depth	48.89
^{a}AvgFusion	Flow	45.28
^{a}AvgFusion	RGBD	55.29
^{a}AvgFusion	RGBD+Flow	**57.09**
^{b}MaxPooling	RGB	50.38
^{b}MaxPooling	Depth	49.65
^{b}AvgPooling	RGB	51.31
^{b}AvgPooling	Depth	49.81
^{b}AvgPooling	Flow	45.30
^{b}AvgPooling	RGBD+Flow	**57.50**
^{b}AvgPooling+SVM	RGBD+Flow	**58.65**

Table 1. Recognition results on the validation subset of IsoGD.

Method	Valid Set(%)	Testing Set(%)
Action Map	36.27	-
Pyramidal C3D	45.02	50.93
Wang et al.	39.23	55.57
Li et al.	49.20	56.90
3DDSN-Fusion	-	56.37
2SCVN-3DDSN	-	**67.26**
Proposed + SVM	58.65	62.14

Table 2. Recognition results on the IsoGD dataset. (2SCVN-3DDSN employs ensemble learning which integrates Two Stream Consensus Voting Network (2SCVN) and 3D Depth-Saliency Network (3DDSN). Three kinds of neural networks are trained on the data of four modalities to get the final optimal recognition accuracy. However, only the proposed deep architecture is used to report our recognition accuracy. The proposed deep architecture still obtains remarkable accuracy, only with a simple new neural network.)

Method	Accuracy(%)
RGGP+RGB-D	88.70
Choi et al.	91.90
4DCOV	93.80
Depth Context	95.37
Tung et al.	96.70
MRNN	97.80
DLEH2(DLE+HOG2)	98.43
3DCNN+RNN+CTC	98.60
Zhu et al.	98.89
Proposed	**99.52**
Proposed + SVM	**99.53**

Table 3. Recognition Results on the SKIG dataset.

A. How to Fuse? Max pooling is more frequently used to reduce the dimensionality of homogeneous convolutional feature maps in the "Conv-Pooling" blocks in the state-of-the-art neural networks. But, averaging is superior to fuse high-level features learnt from different perspectives.
B. What to Fuse? Early feature fusion is superior to late score fusion.
C. Spatiotemporal Feature Learner. The deep architecture (3DCNN + ConvLSTM + 2DCNN) is an effective spatiotemporal feature learner. It is robust to various scene backgrounds and illumination conditions theoretically and actually, and it can also process gestures with various time durations effectively.

Conclusion. ***The proposed deep architecture provides an alternative method to transform video files into 2D feature maps***. The paper only presents the preliminary version of the deep architecture. The state-of-the-art skills of 2DCNN, 3DCNN and LSTM networks can be further utilized to construct an improved version in order to obtain higher recognition accuracy.