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 extendable 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.

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.

## Experimantal 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.

### 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.