The paper presents the method to learn global description of video by feature pooling and recurrent neural networks. The purpose is to obtain global video features with numbers of frames to increase the classification performance. AlexNet and GoogleLeNet are used to process still images and optical flow frames.

The paper introduces multiple feature representations to aggregate video information by using pooling and LSTM. Conv pooling performs on the last convolutional layer, which enables to backtrack the spatial information of features. Late pooling performs before the soft max layer so that high level features can be combined together. Slow pooling performs on both convolutional layer and fully connected convolutional layer with small size of window to capture local information. Local pooling only performs on the last convolutional layer separately such that local information is guaranteed to be preserved. Time-domain Conv adds pooling on time domain with frame stride of 5. LSTM cell keeps a value to preserve the information of neighborhood cells and update basing on the incoming information. Optical flow is trained to represent the motion pattern of objects. Combing optical flow and image frames with convolutional pooling and LSTM achieve the state-of-art performances.