The paper introduces a two-stream convolutional architecture for action recognition in video. Still image frame provides important information for action recognition as the strong relation between activities and corresponding object. Action pattern can be represented by multiple frames by using temporal representation. A spatial ConvNet takes a still image as input with similar architecture of AlexNet. Regularization and max pooling are used for each hidden layer. Input of temporal ConvNet is computed by different motion representation on 10 consecutive frames. Dense optical flow is computed at the every pixel of current frame and next frame. Trajectory tracking consists by couple points. The displacement of a point at a position is computed by the current point and next point. Bi-directional optical flow starts from the intermediate frame and computes the optical flow in both forward and backward direction. Mean flow subtraction subtracts a mean vector for each displacement vector. It cooperates with nonlinear rectification as it thresholds at 0 and mean flow subtraction produces positive and negative values. Training is done by mini-bash SGD of 256 samples across the classes. Each video sample produces one image for spatial ConvNet and optical flow volume for temporal ConvNet.

The architecture takes the advantages of using pre-trained AlexNet on large image dataset. Pre-trained model outperforms the model trained from the scratch. And two-stream architecture does show the competitive result s comparing to state-of-art methods.