# Multilayer-Pooling Selection Fusion with 3D Convolutional Neural Networks for Classification of Videos

Abhishek Kapoor[1, a)], Abhishek Kumar [2)], Ayush Mahajan [3)] and Dashiv Sharma [4)]

[a)]Corresponding author: abhishek3764.beai23@chitkara.edu.in

**Abstract.** 3D Convolutional Neural Networks are employed for video classification and understanding. However, their global spatio-temporal context basis can limit local learning. To address this, a compact multi-layer feature fusion network has been developed that incorporates both local and global viewpoints. This innovative multi-layer feature synthesis network collaborates with local and global perspectives to mitigate this issue. The current network utilizes the global view branch to comprehend underlying video semantics and the local view branch to grasp contextual local semantics. Unlike conventional C3D modeling, the large view branch is only fed video features that are most activated from the broader 3D receptive field. The addition of such shallow view contexts enables the local view branch to learn more robust and discriminative spatiotemporal representations for video classification. We suggest 3D convolutional networks with multi-layer pool selection fusion for video classification, where the embedded deep global feature is amalgamated with information from a shallow layer of the feature extraction network. This local functionality is achieved through spatial-temporal pyramid synthesis and adaptive synthesis. By assembling three different aggregation units, diverse spatio-temporal feature information is acquired, which is subsequently cascaded and utilized for classification. Upon testing on the UCF-101 and HMDB 51 datasets, classification accuracy rates of 95.0% and 72.2% were achieved, respectively. These results indicate that the proposed 3D convolutional networks with multi-layer pooled selection fusion exhibit superior classification performance.