# Human Action Recognition

This assignment will work with the UCF-101 human action recognition dataset. The dataset consists of 13,320 videos between ~2-10 seconds long of humans performing one of 101 possible actions. The dimensions of each frame are 320 by 240.

This homework will compare a single frame model (spatial information only) with a 3D convolution-based model (2 spatial dimensions + 1 temporal dimension).

- Part one: Fine-tune a 50-layer ResNet model (pretrained on ImageNet) on single UCF-101 video frames
- Part two: Fine-tune a 50-layer 3D ResNet model (pretrained on Kinetics) on UCF-101 video sequences

There are three sets of results for comparison: 1.) single-frame model, 2.) 3D model, 3.) combined output of the two models. For each of these three, report the following:

- `(top1_accuracy,top5_accuracy,top10_accuracy)`: Did the results improve after combining the outputs?
- Use the confusion matrices to get the 10 classes with the highest performance and the 10 classes with the lowest performance: Are there differences/similarities? Can anything be said about whether particular action classes are discriminated more by spatial information versus temporal information?
- Use the confusion matrices to get the 10 most confused classes. That is, which off-diagonal elements of the confusion matrix are the largest: Are there any notable examples?

Put all of the above into a report and submit as a pdf. Also zip all of the code (not the models, predictions or dataset) and submit.