Video Event Categorization

Computer Vision (CS6350)
TPA - 2

1 Problem Statement

The aim of the project is to build a system that can categorize large number of videos according to a set of complex human actions being performed. The videos feature large intra-class variations as well as several challenges in the form of camera motion, jitter, multiple persons and low quality.

2 Input

- A set of videos of arbitrary time length and action class (from dataset)

3 Output

- Table of classification accuracy on the dataset splits

Table 1: Sample Table

Method	Accuracy (Split1)	Mean Accuracy over 3 Splits

- Demo to run on a given video clip

4 Dataset

- **UCF-101**: This dataset contains 13320 annotated videos belonging to 101 classes having 180 frames/video on average. The evaluation will be done on the three standard train/test splits following the scheme of THUMOS challenge.
- **HMDB-51**: A large collection of real-world videos gathered from various sources such as web and movies. This dataset comprises of 6766 videos across 51 classes, where each class consists of at least 100 variable length video clips.

5 References

1. K. Simonyan & A. Zisserman. (2014) Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (pp. 568-576).

2. Mahasseni, Behrooz, & Sinisa Todorovic. (2016) Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (pp 3054-3062)

3. Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1933-1941).

4. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 4724-4733).

5. Cosmin Duta, I., Ionescu, B., Aizawa, K. & Sebe, N. (2017), Spatio-Temporal Vector of Locally Max Pooled Features for Action Recognition in Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3097-3106).

6. Sun, S., Kuang, Z., Sheng, L., Ouyang, W. & Zhang, W., (2018). Optical flow guided feature: a fast and robust motion representation for video action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1390-1399), 2018.

7. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A Closer Look at Spatiotemporal Convolutions for Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6450-6459).