**Related works**

The amount of works concerning video classification remains small compared to image classification. Classifying videos presents unique challenges for machine learning models.  Deep learning tools such as the convolutional neural networks are extensively used for image analysis and classification tasks, but they become relatively expensive to use for a corresponding analysis in videos by requiring memory provision for the additional temporal information. Simonyan et al. [1] achieved very competitive performance by training two CNNs on spatial (static frames) and temporal (optical flows) streams separately and then fusing the two networks. Ji et al. and Karparthy et al. extended the image-trained CNN into temporal domain by stacking static frames, upon which convolution can be performed with space-temporal filters [2, 3]. Shengxin et.al conducted an in-depth exploration of different strategies for doing event detection in videos using convolutional neural networks (CNNs) trained for image classification. They studied different ways of performing spatial and temporal pooling, feature normalization, choice of CNN layers as well as choice of classifiers [4]. While it is now clear that CNN-based approaches outperform most state-of-the-art handcrafted features for image classification [5], until very recent [4] it was not sure for video classification.

Our work is closely related to other research challenges towards the efficient use of CNN for video classification. The aim of our effort is to compare architectures of different complexity. We propose an efficient approach to exploit off-the-shelf image-trained CNN architecture for video classification. We advance to a more complex model based on CNN architecture tailored for capturing the video characteristics proposed by [3]. On top of the single-stream CNN architectures, we build a two-stream neural network for temporal and spatial analysis. This work aims at filling a gap in the existing works, where the focus is placed on designing new classification pipelines or deeper network structures [4] without trying to evaluating and adjust implementation details of common successful image-trained architectures.

[1] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Fisher networks for large-scale image classification. In NIPS, 2013.

[2] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. In ICML, 2010.

[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR, 2014.

[4] Z. Shengxin, Florian Luisier, Walter Andrews, Nitish Srivastava. Exploiting Image-trained CNN Architectures for Unconstrained Video Classification, 2015.

[5] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN

features off-the-shelf: an astounding baseline for recognition. In IEEE Conference on

Computer Vision and Pattern Recognition (CVPR). IEEE, 2014.