Deep Learning for Computer Vision:

Video Game Genre Prediction Using Youtube as a Data Source

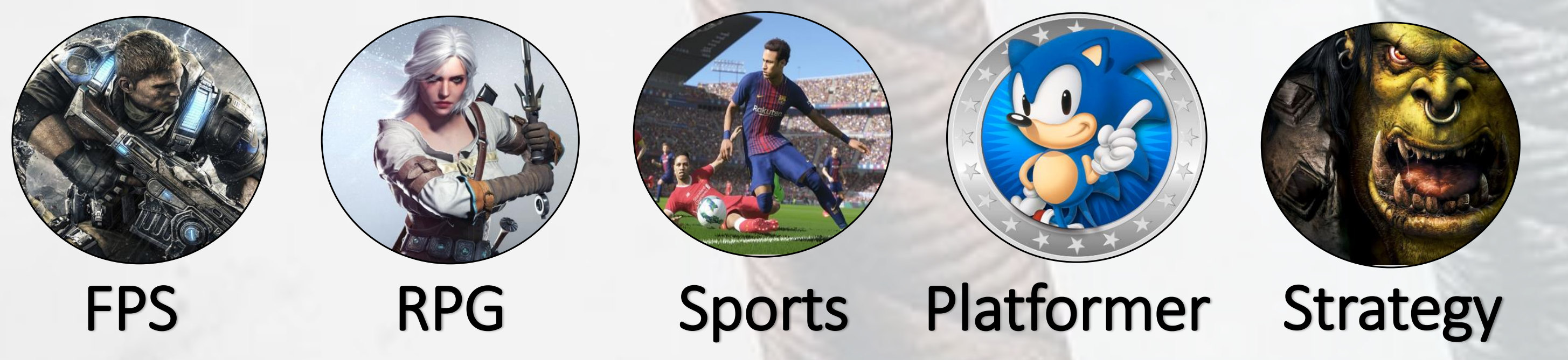
Piotr Tatarczyk *piotr.tatarczyk@tum.de*
Janis Postels *janis.postels@tum.de*
Damian Bogunowicz *damian.bogunowicz@tum.de*
Peter Nagy *peter.nagy@tum.de*

INTRODUCTION

Within the recent years Youtube was able to accumulate a vast amount of labelled videos making it a very resourceful source of data. Especially gaming videos - e.g. gameplay reviews or commentated matches - make up a considerate amount of this data. In this project we want to utilize this data and build game-genre-classifier. The input of our classifier is thereby a video of the game which is subject to classification and the output responds to the category of the game. The goal is to enable our classifier to accurately predict the genre of a so far unknown game.

DATASET

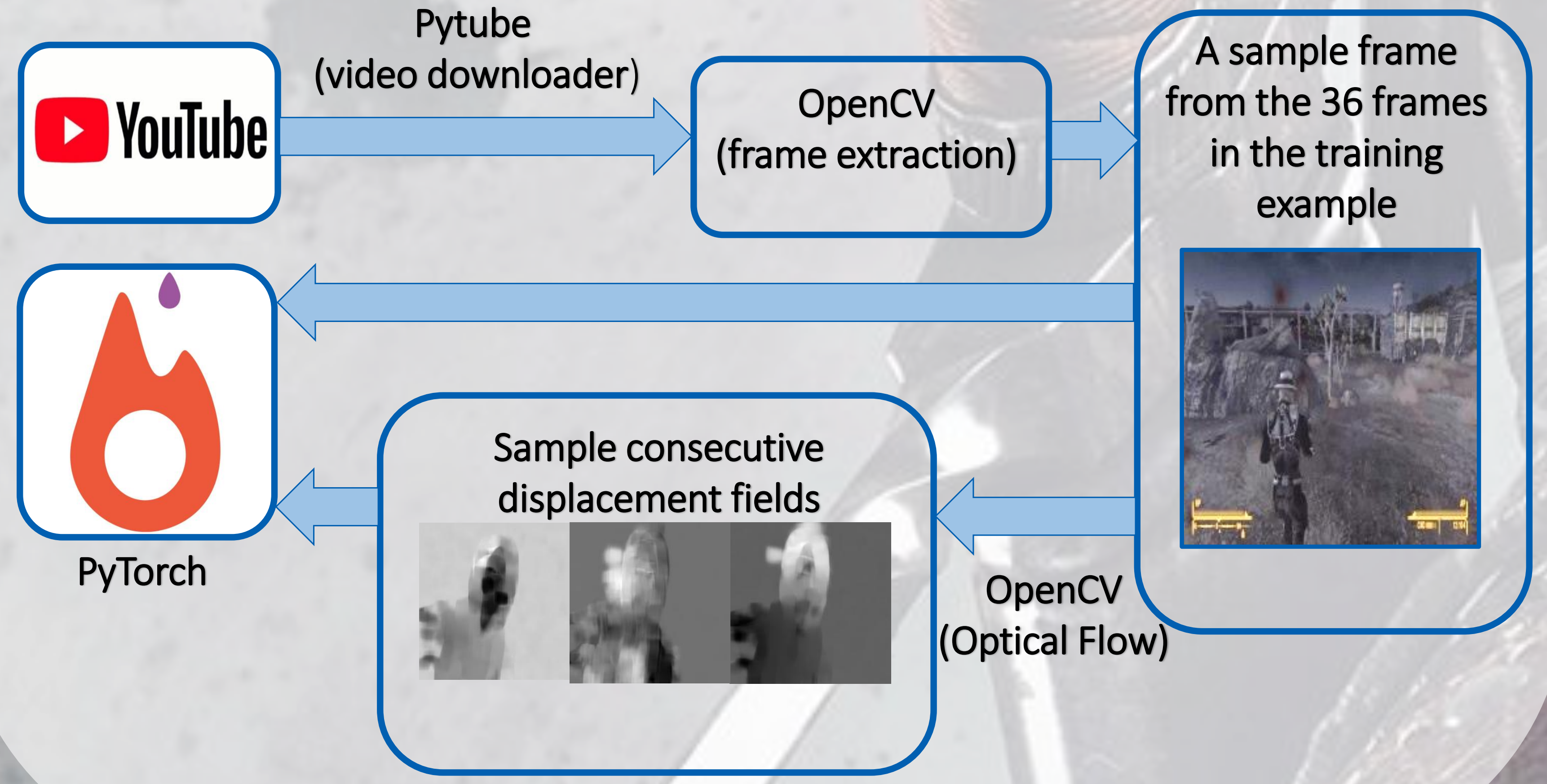
The training set used for the network is obtained through downloading various gameplays from YouTube. For each of classifier's five classes:



we gather around 180 hours of footage in total. From that, we sample uniformly about 20 hours, so every network is effectively trained with 4 hours of gameplay per class. The footage was checked by our team for its quality. The important traits of the videos were:

- the variety of dissimilar game levels, characters, menus and cut-scenes
- least amount of advertisements, running commentaries (eSport broadcasts) and other noise

For pre-processing, we take advantage of the techniques and parameters described in related works. For the spatial stream, we split every gameplay into 6 second videos. Given the frame rate of 6 fps, each data point is made up of 36 (6 times 6) frames. Each frame is resized to 224x224. To maintain the high variance and computational efficiency, for the Single Stream Image Classification, we sample 5 random frames out of 36. In case of Two-Stream Video Classification approach we treat data in the following way. For the temporal stream, for each data point we compute the optical flow between frames. This means that one data point for the temporal stream consists of 35 displacement fields in the horizontal direction and 35 fields in the vertical direction. The test set is made up of 30 minutes of gameplay containing new game titles, which are not included in the original training set.


Pytube
(video downloader)
OpenCV
(frame extraction)
A sample frame from the 36 frames in the training example
OpenCV
(Optical Flow)
Sample consecutive displacement fields
PyTorch

METHODOLOGY

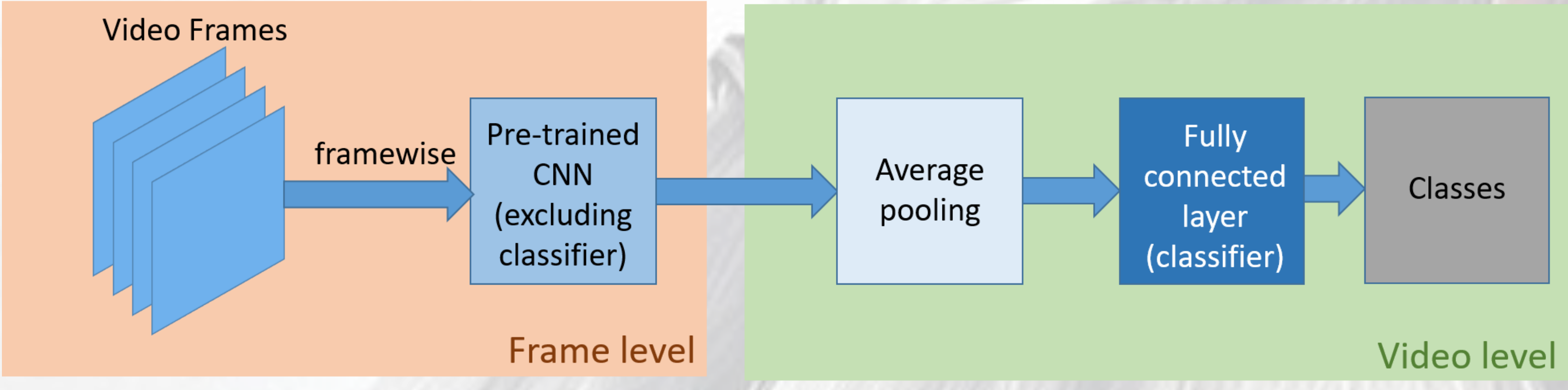
In order to tackle the research endeavour effectively, we start with the simple CNN and progressively develop the complexity of the architecture.

1. Single Stream Image Classification Approach

Preliminary architecture used in the approach is the Dense Convolutional Network (DenseNet). It is chosen among several other architectures (VGG11, Resnet16), because it gives best performance in the least amount of epochs and allows the use of big batch sizes.

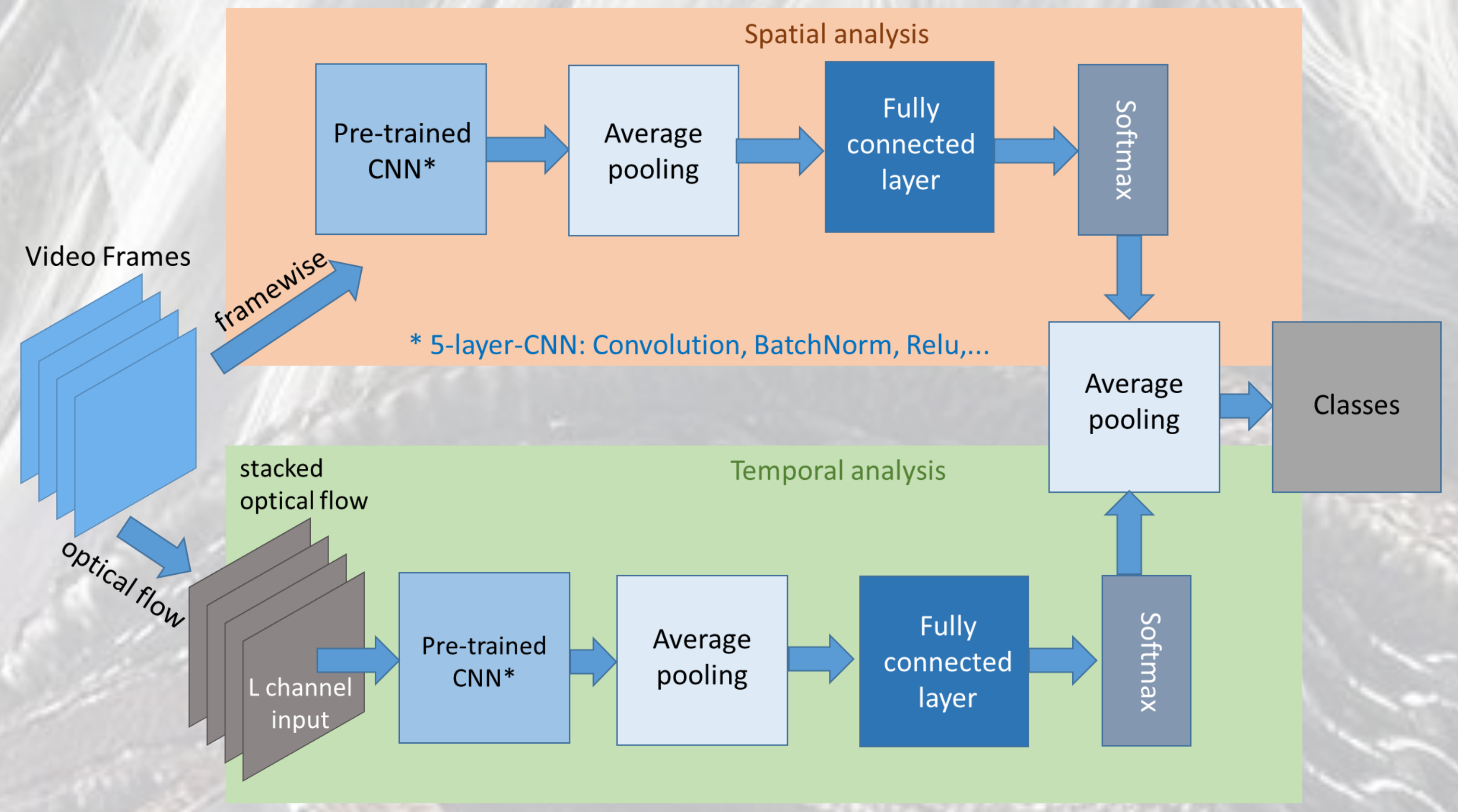
2. Single Stream Video Classification Approach

The frames are processed separately by a pretrained CNN. Afterwards we use average pooling to go over from frame level representation to video level representation followed by a three layer FCN for the final classification. The parameters of the CNN stay untouched throughout the training process. We experiment with two different architectures (VGG11 and DenseNet).


Video Frames
framewise
Pre-trained CNN (excluding classifier)
Average pooling
Fully connected layer (classifier)
Classes
Frame level
Video level

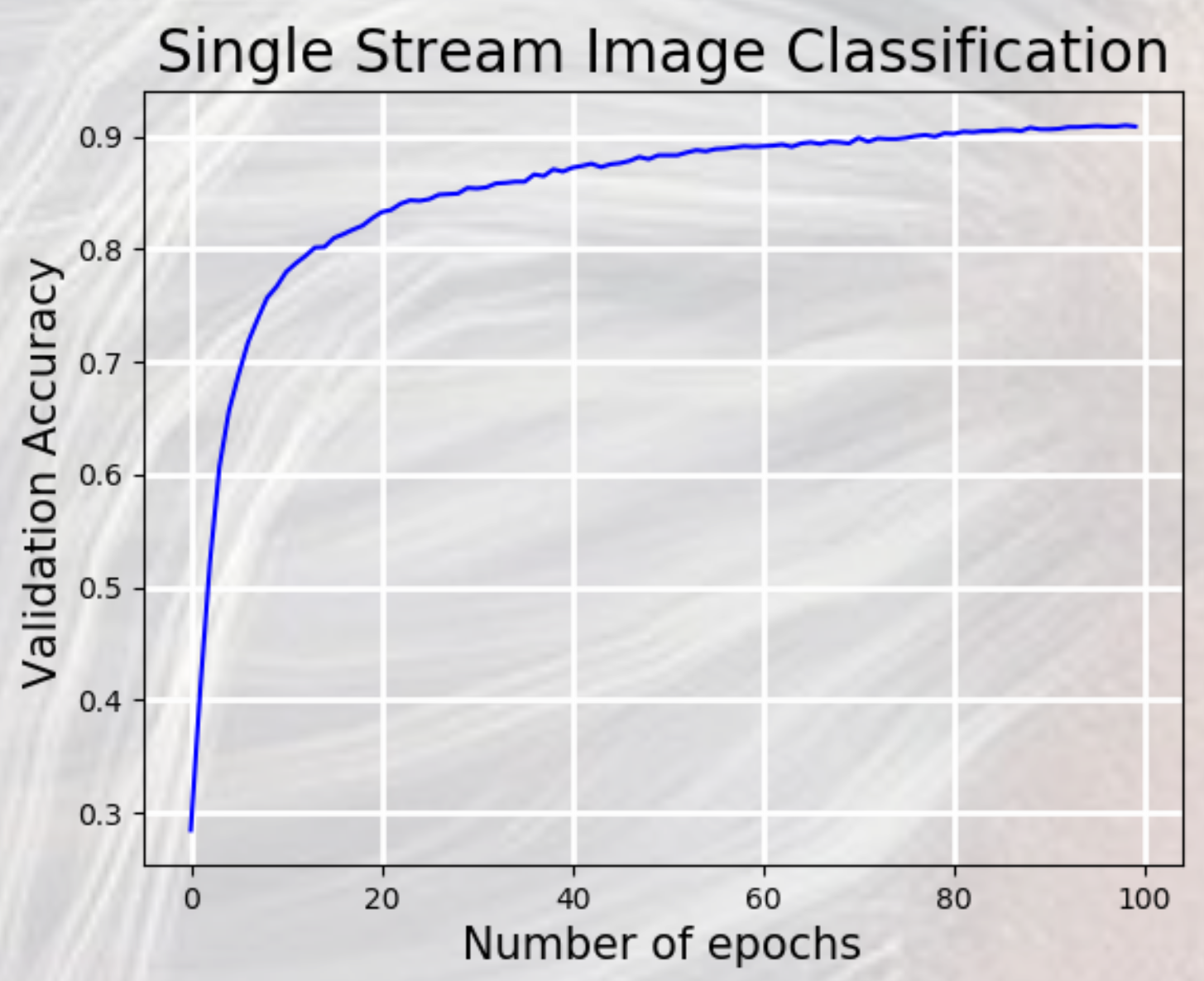
3. Two-Stream Video Classification Approach

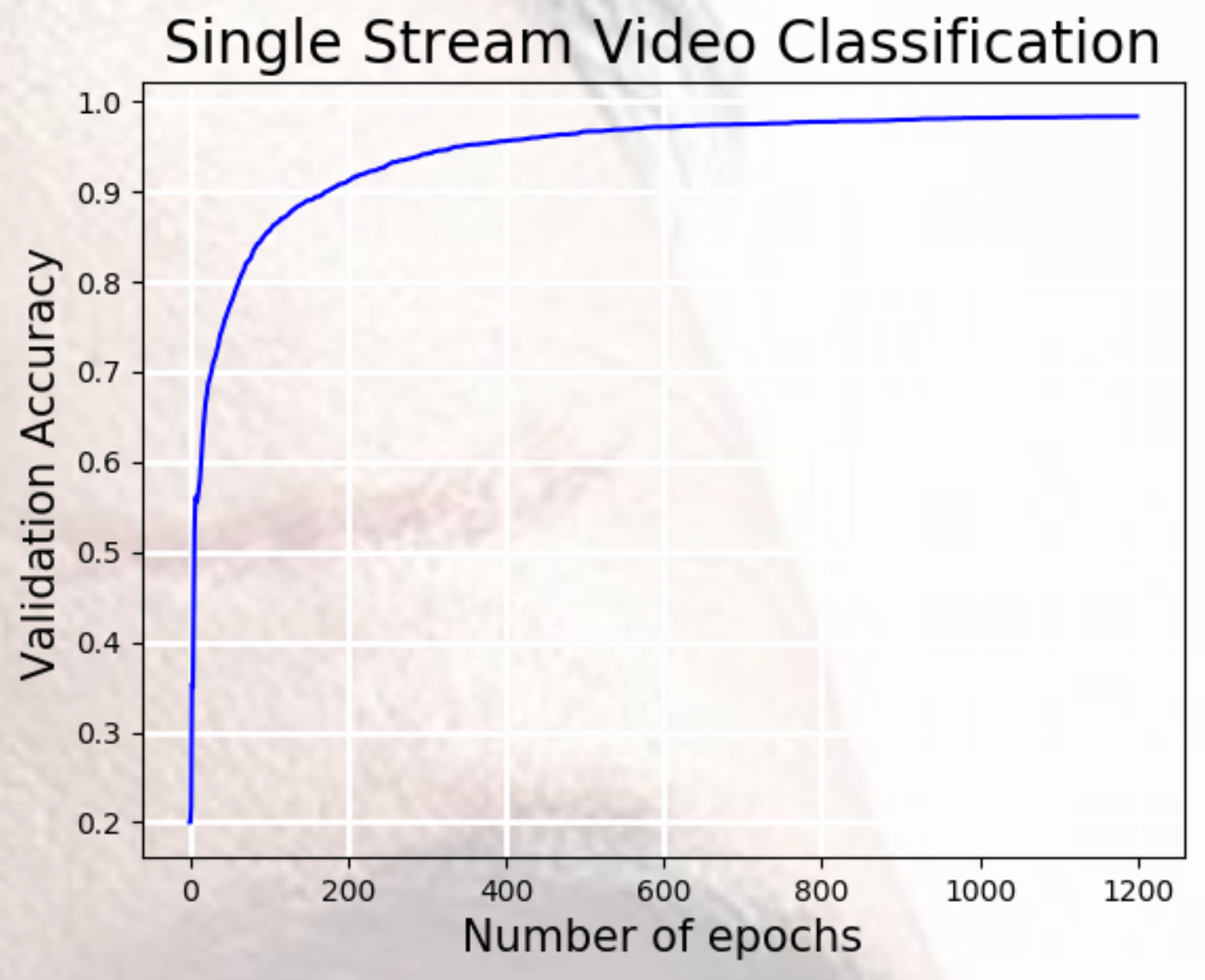
The spatial stream is similar to the Single Stream Video Classification approach. In temporal steam we compute the optical flow between the consecutive frames. We stack the resulting greyscale images on top of each other in groups of ten and put them into a CNN. Images flow through an average pooling layer to go over into video level and are classified by a FCN. At the end the output of both FCNs is processed by a softmax layer and finally averaged to yield the overall result.



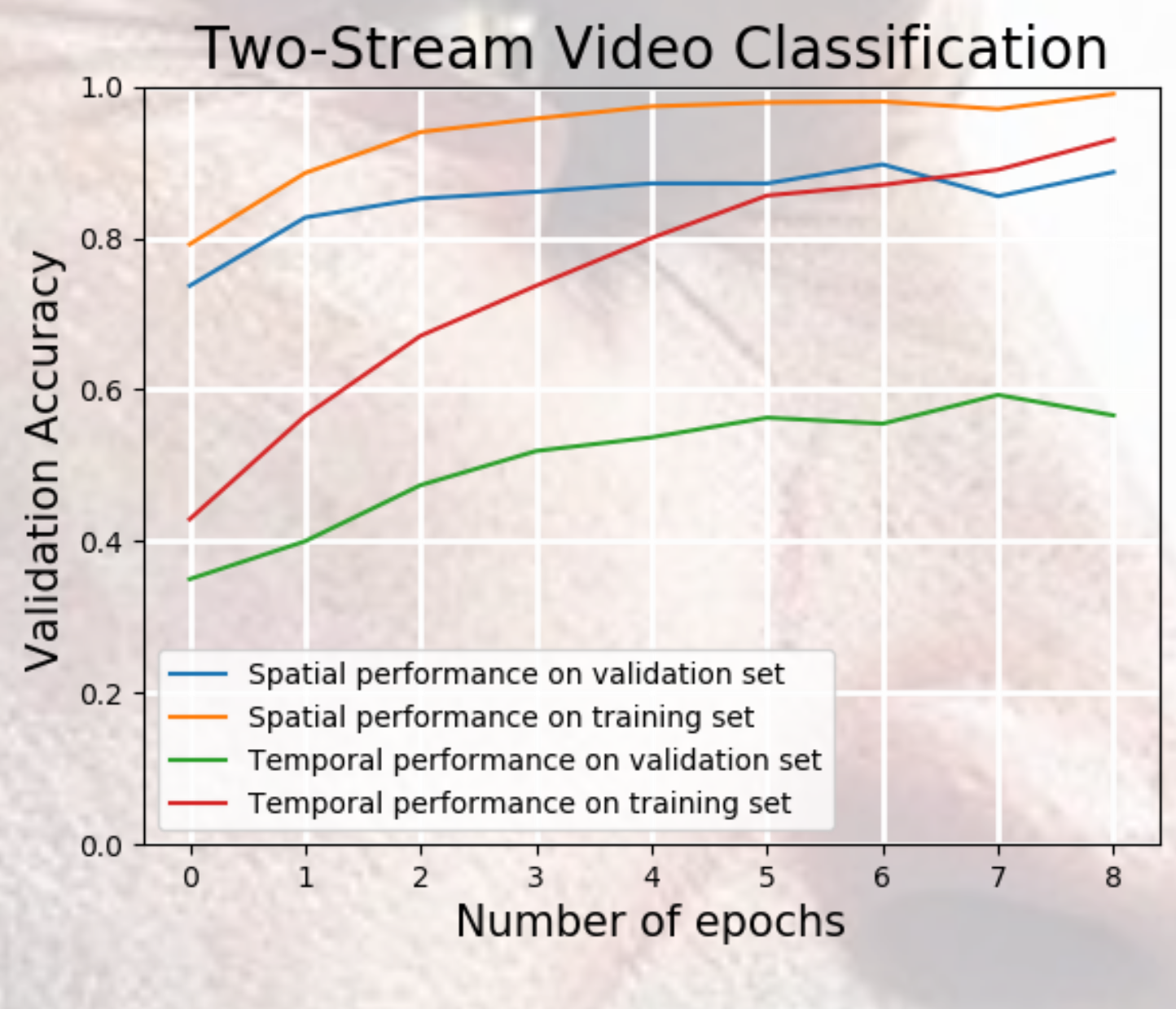
[1] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Fisher networks for large-scale image classification. In NIPS, 2013.
[2] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. In ICML, 2010.
[3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, 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.

RESULTS


Single Stream Image Classification


Single Stream Video Classification

Approach	Optimizer	Accuracy on internal test set	Accuracy on external test set
SSIC	Adam	91.7%	78.7%
SSVC	SGD+Nesterov Momentum	83%	37.%
TSVC	Adam	88,6%	44,2%


Two-Stream Video Classification

RELATED WORKS

The amount of works concerning video classification remains small compared to image classification. Deep learning tools such as the convolutional neural networks (CNNs) are extensively used for the 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 CNNs trained for image classification. 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 CNNs 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.