

VIDEO CLASSIFICATION: HUMAN ACTION RECOGNITION ON HMDB51 DATASET

Elaborato di:

Alessio PASINATO matricola n. 887000

Gianluca SCURI matricola n. 886725

Giorgio CARBONE matricola n. 811974

Task: Human action recognition in videos

- ❑ **Human Action Recognition (HAR): time series classification** problem that involves classifying an action performed by someone
- ❑ **Vision-based Human Action Recognition (V-HAR):** classification of actions performed in **video clips**
- ❑ **Video classification vs image classification:**
 - Higher **computational cost**
 - **Spatial** and **temporal** information
 - Capturing **long spatiotemporal context**
- ❑ **Simple Action Recognition:** model that classifies singular global actions in short video clips



Action: **waving**

Simple Action Recognition: Video Classification Methods

❑ Hand-crafted Features

❑ Single Stream Networks

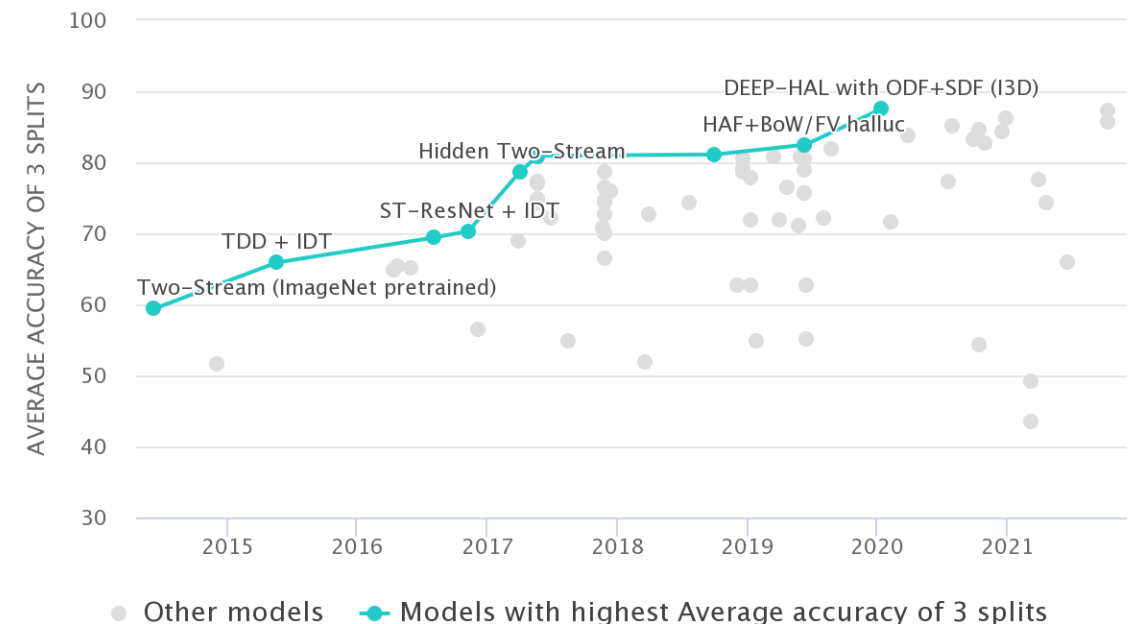
- Single-Frame CNN (2D CNN's)
- Late Fusion / Early Fusion / Slow Fusion (2D CNN's)
- 3D ConvNet (C3D)
- CNN with LSTM's
- Pose Detection and LSTM

❑ Two Stream Networks

- Optical Flow and CNN's
- SlowFast Networks

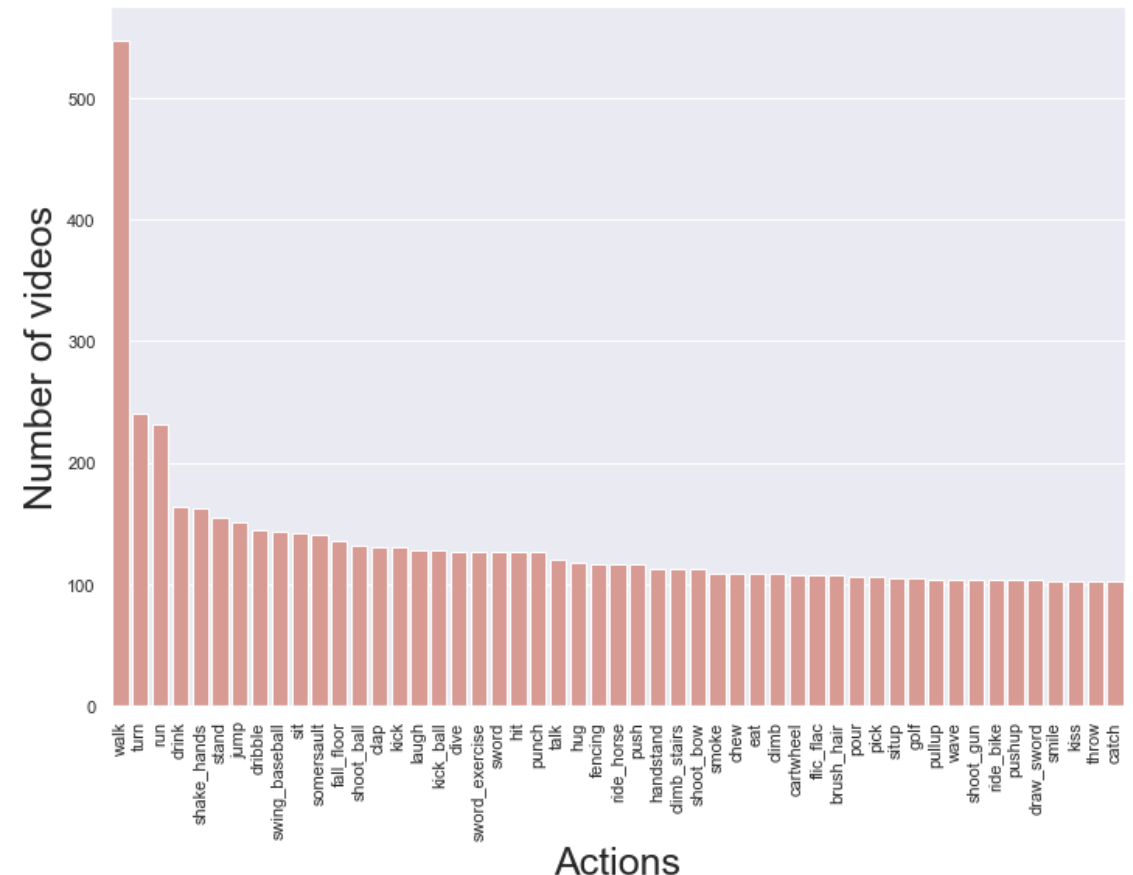
❑ Action Recognition on HMDB-51: State-of-the-Art:

- DEEP-HAL with ODF+SDF (I3D) 2020 -> **87.56%**



HMDB51: A Large Video Database for Human Motion Recognition

- ❑ 6766 annotated clips
- ❑ Different **sources**: Youtube, Google videos, movies.
- ❑ 51 **categories** with a minimum of 101 clips per action
- ❑ **Categories** grouped in **five types**:
 - General **facial actions**
 - **Facial actions** with **object manipulation**
 - General **body movements**
 - **Body movements** with **object interaction**
 - **Body movements** for **human interaction**



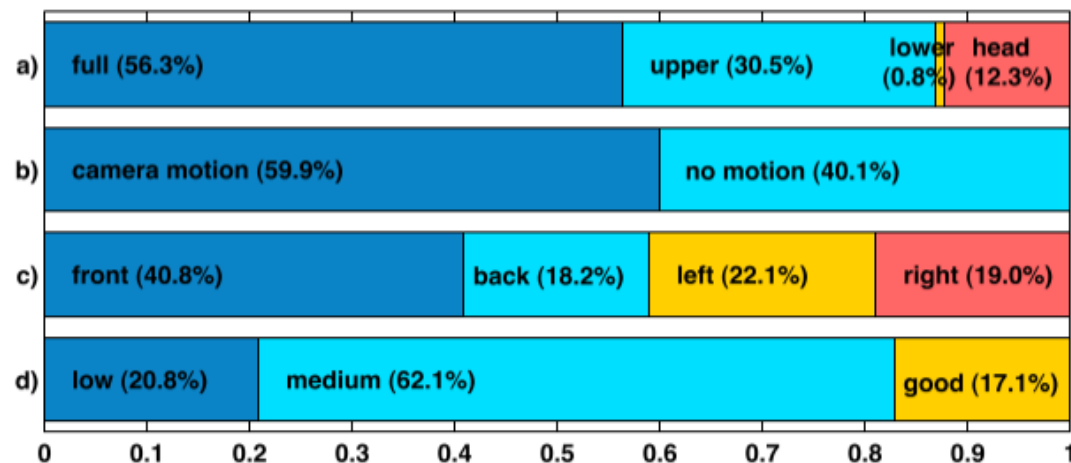
HMDB51: Data Exploration

❑ Additional meta information:

- a) **Visible body parts / occlusions:** full, upper, lower, head
- b) **Camera motion:** motion or static
- c) **Camera viewpoint:** front, back, left, right
- d) **Video quality:** good, medium or bad
 - **Number of people** involved

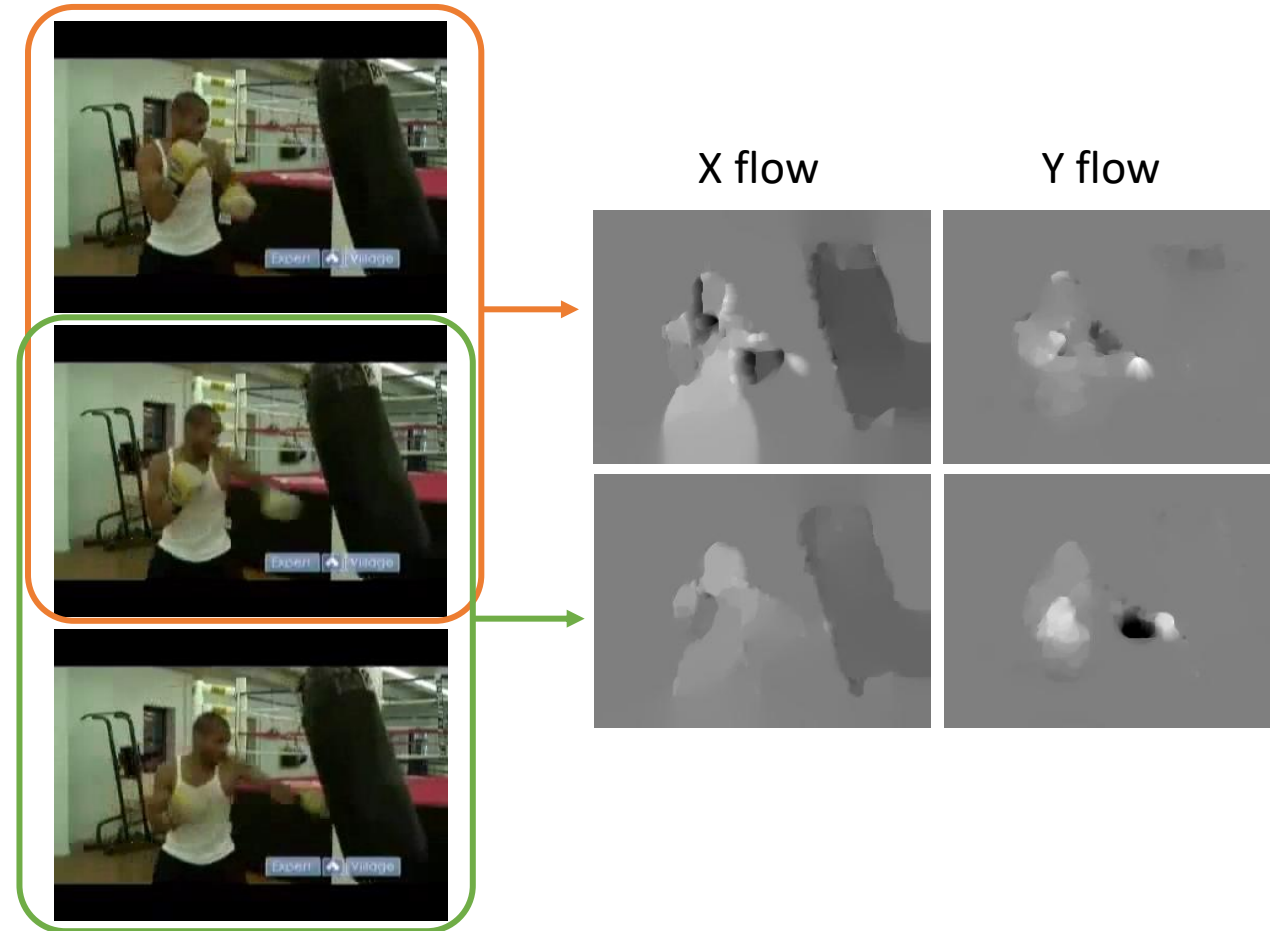
❑ Video normalization:

- **Height** of all the frames is scaled to **240** pixels
- **Width** is scaled to maintain aspect ratio
- Frame rate of **30 fps**



Data Preparation

1. **Training** (70 clip/class) and **test** (30 clip/class) split
2. **Frame extraction** (625775 total frames)
3. **Dense Optical Flow extraction**
 - 2 consecutive frames
 - Dual TV-L1 Optical Flow
4. **Sampling**
5. **Data augmentation** on frames & optical flows:
 - Random **Horizontal flip** (50% probability)
 - Random **Crop** (224x224)
 - Random **Rotation** (0.15)
6. **Centering, scaling** and **Resizing** (224x224)



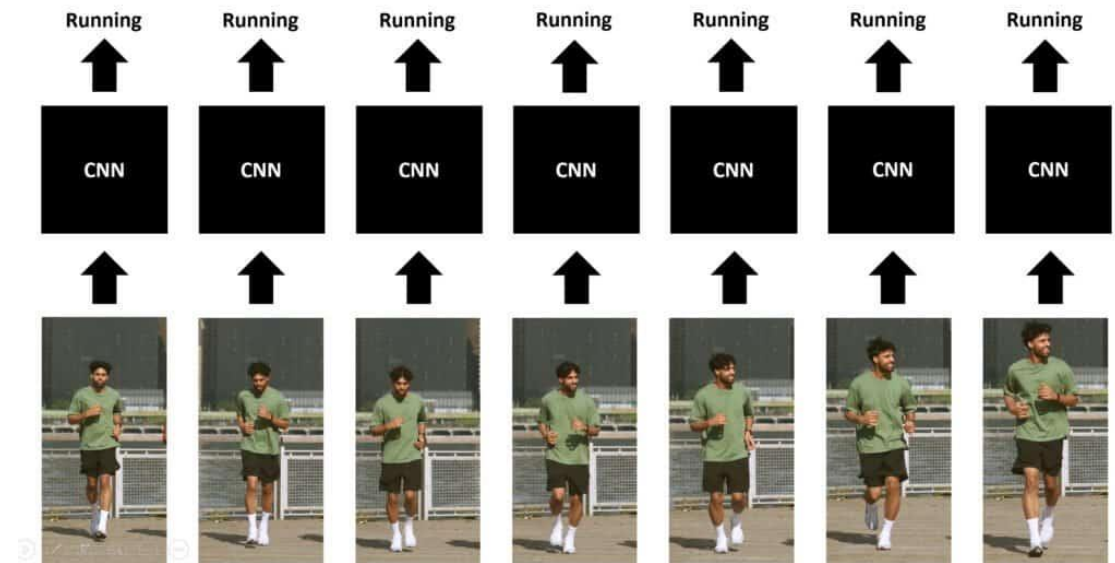
First approach: CNN single frame classification

❑ Architecture

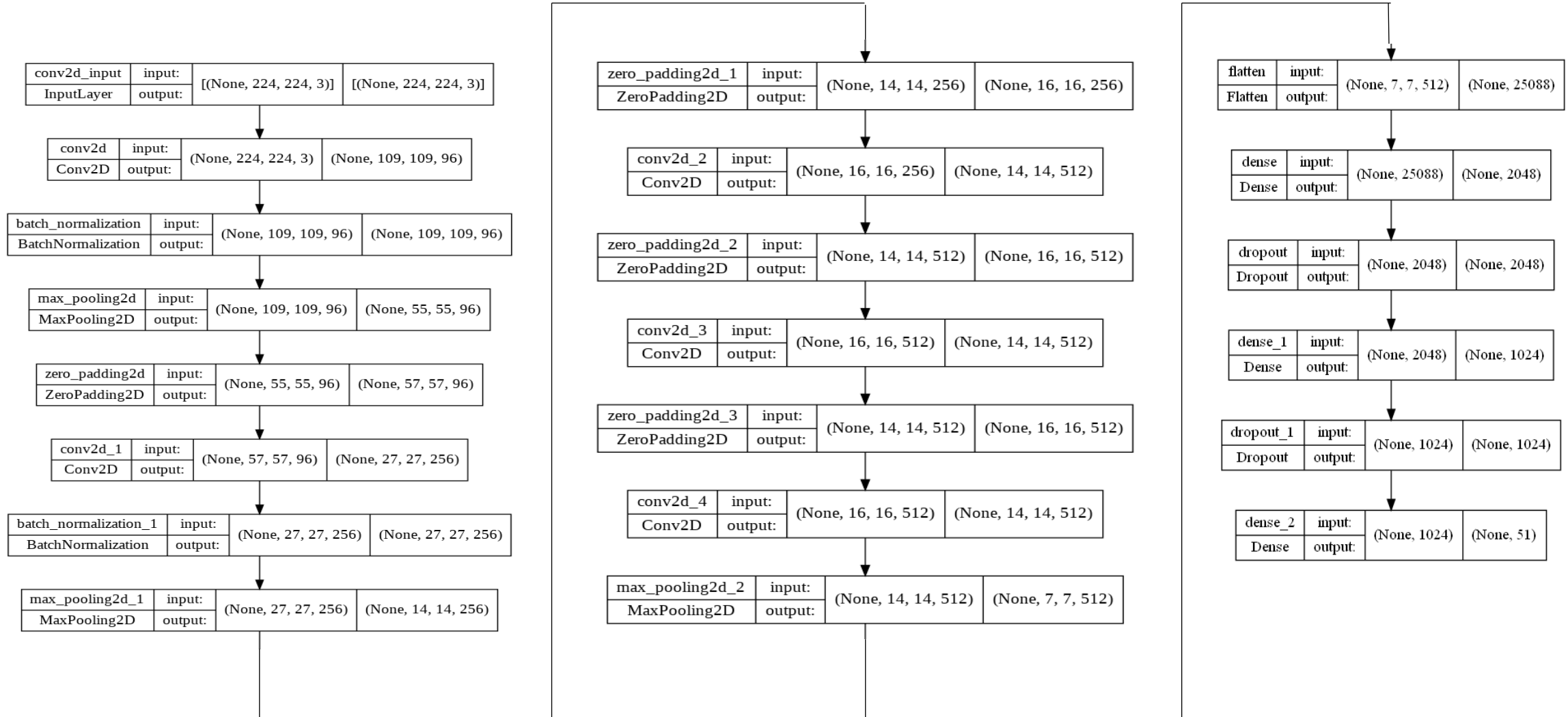
- Convolutional Neural Network
- 17 frames per video

❑ Layers:

- Preprocessing (224 x 224 x 3)
- Convolutional2D (activation: Relu)
- Zero padding
- Batch normalization
- MaxPooling (pool size: 3x3, stride: 2)
- Dense (dropout: 0.5, activation: ReLu, Softmax)



First approach: CNN single frame classification



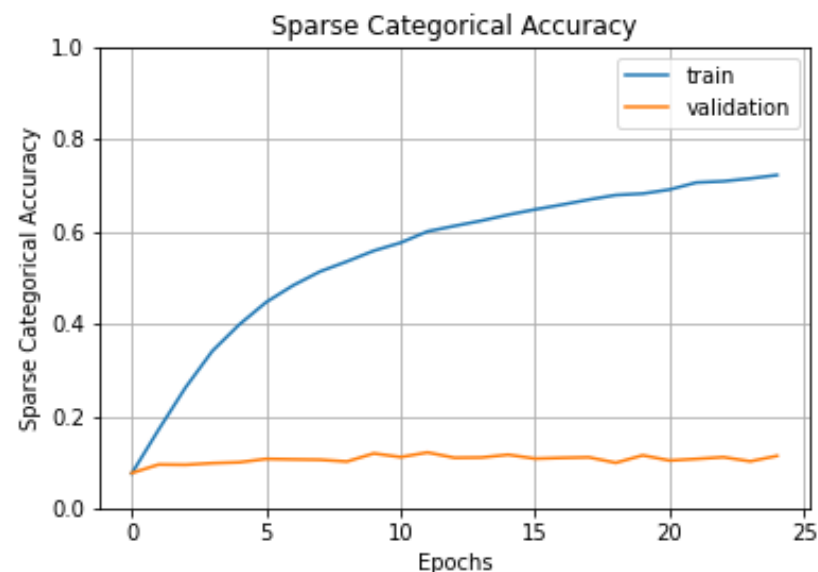
First approach: CNN single frame classification

❑ First train:

- Validation loss: **13.73**
- Top 1 accuracy: **6.6%**

❑ Best train:

- Validation loss: **4.74**
- Top 1 accuracy: **12.1%**
- Top 5 accuracy: **34.4%**



	Optimizer	Epochs	L.R.	Batch	Train. P.	Data Aug.	Norm.
First	Adam	100	0.001	64	117,789,043	Resize	[0, 255]
Best	Adam	25	0.001	128	60,142,035	Flip, Rot, Crop	[-1, 1]

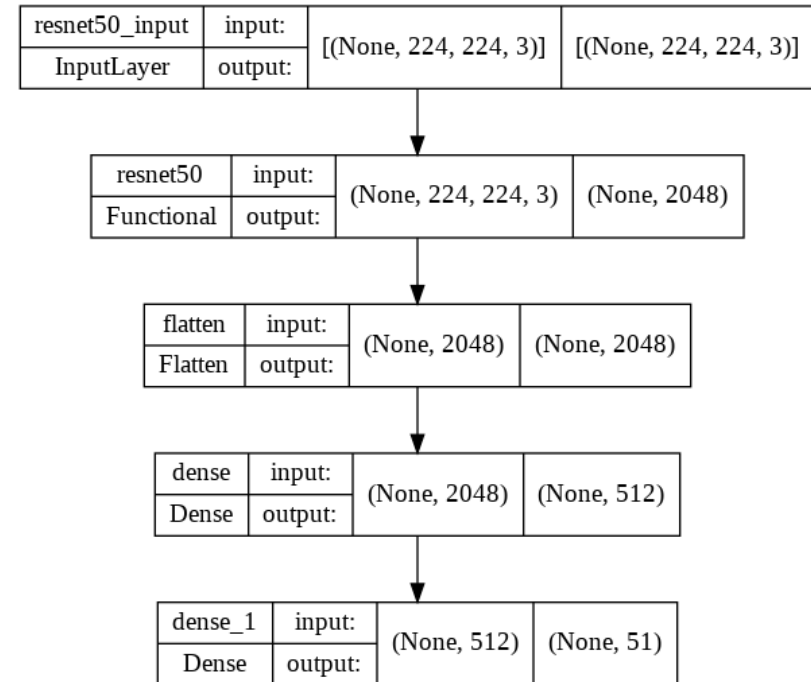
Second approach: Transfer Learning

❑ Architecture

- Finetuned CNN based of ResNet50
- 17 frames per video

❑ Layers:

- ResNet50 (not trainable, weights from ImageNet)
- Flatten
- Dense (activation: ReLu, SoftMax)



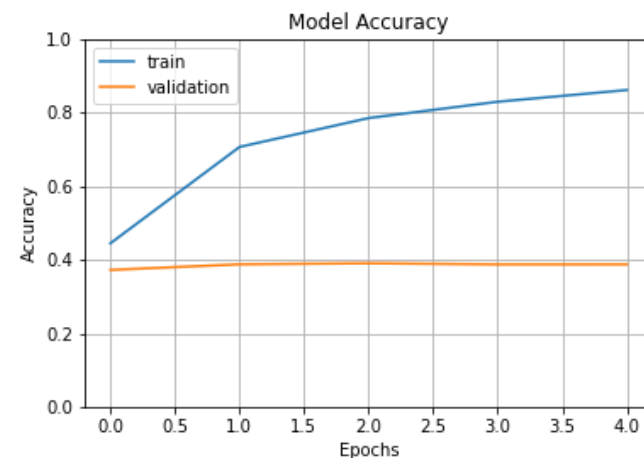
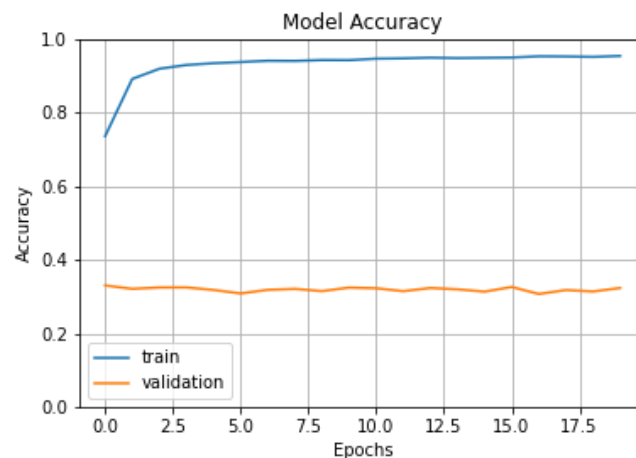
Second approach: Transfer Learning

❑ First train:

- Validation loss: **2.91**
- Top 1 accuracy: **31.9%**
- Top 5 accuracy: **59.1%**

❑ Best train:

- Validation loss: **2.30**
- Top 1 accuracy: **39.0% -> 45.0%** (integral video)
- Top 5 accuracy: **71.6% -> 78.5%** (integral video)



	Optimizer	Epochs	L.R.	Batch	Train. P.	Data Aug.	Norm
First	Adam	20	0.001	64	1,075,251	Resize, Flip	[0,255]
Best	Adam	5	0.0001	64	1,075,251	Resize, Flip	Centered

Third approach: two-stream CNN

❑ Architecture:

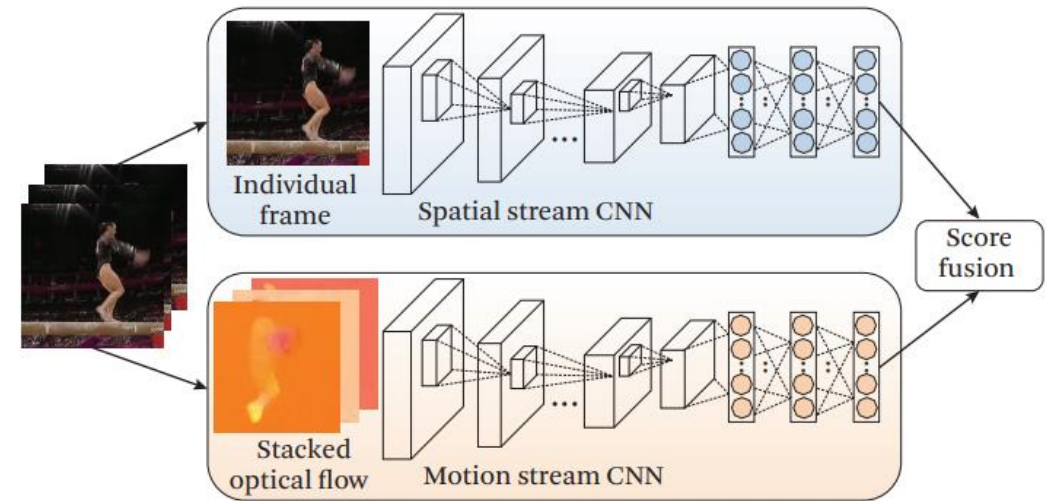
- Spatial CNN + Temporal CNN
- Averaging SoftMax results

❑ Spatial stream:

- Finetuned ResNet

❑ Motion stream:

- Same architecture as “first approach CNN” with input size (224, 224, 20)
- Semi-randomly selected batch of N stacked optical flow from N randomly selected videos
- Each flow stack is composed of 10 x-channels and 10 y-channels consecutive optical flows



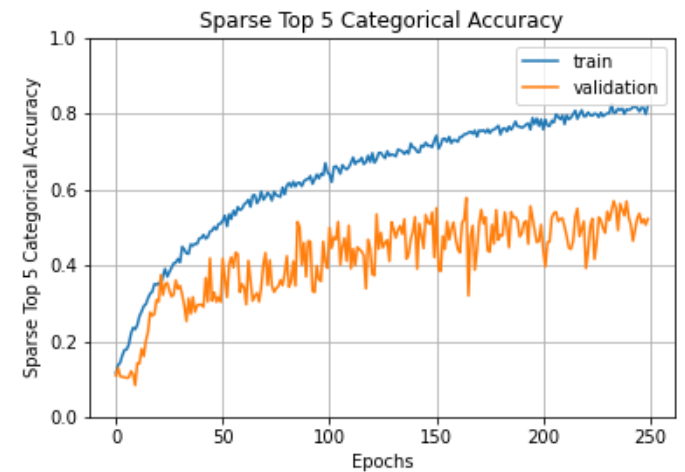
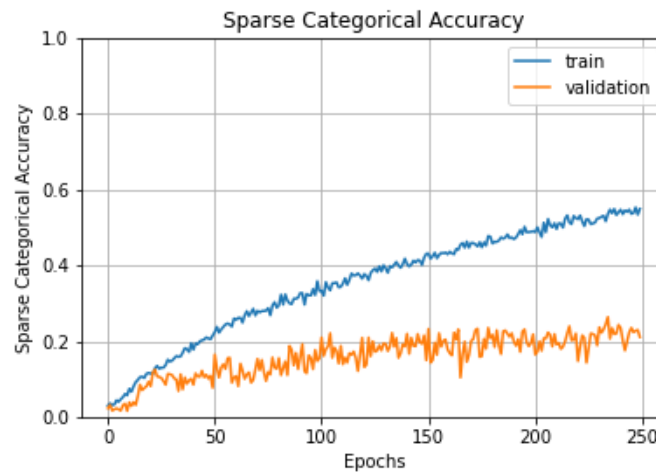
Third approach: two-stream CNN (motion stream)

❑ First train:

- Validation loss: **3.46**
- Top 1 accuracy: **15.0%**
- Top 5 accuracy: **42.4%**

❑ Best train:

- Validation loss: **3.27**
- Top 1 accuracy: **25.8%**
- Top 5 accuracy: **54.3%**

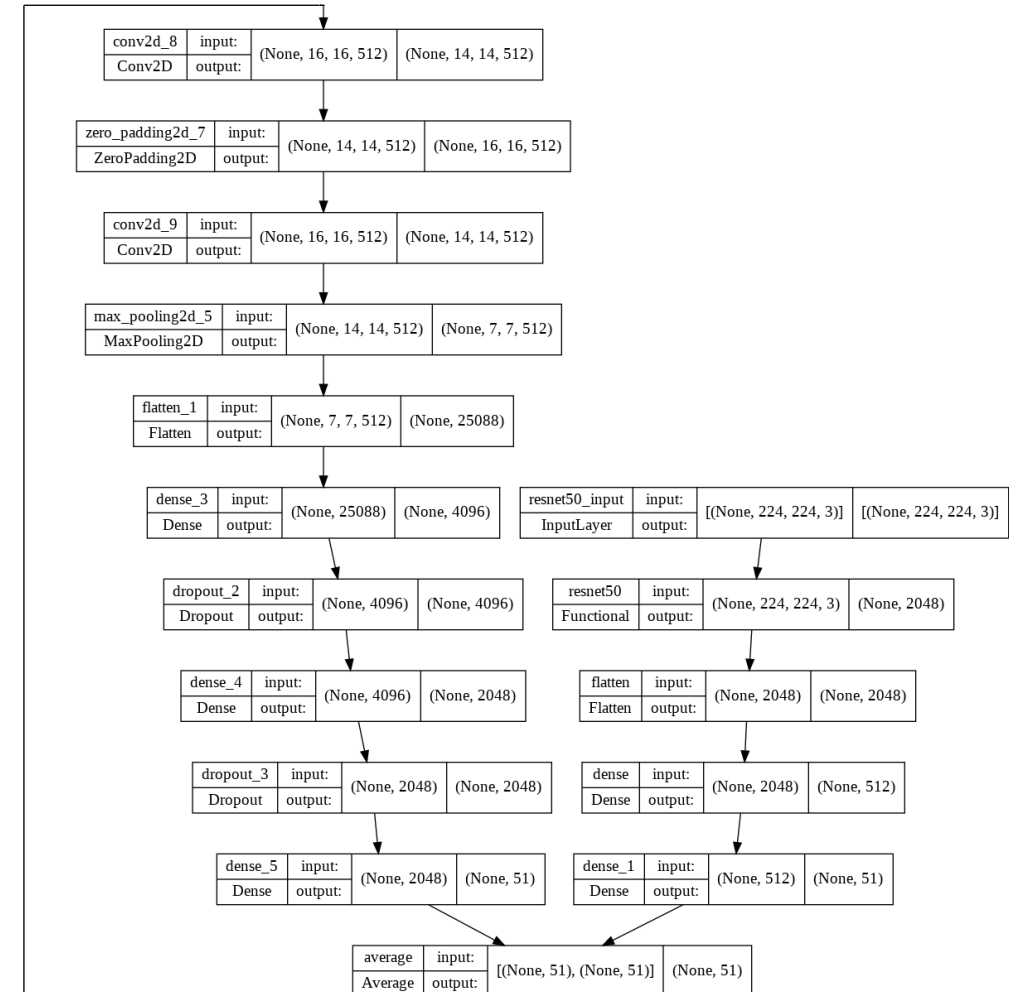
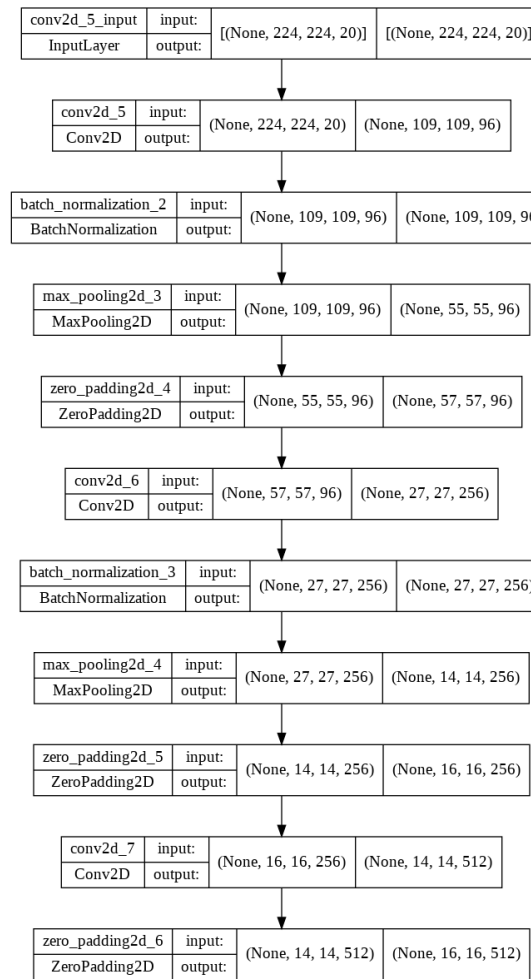


	Optimizer	Epochs	L.R.	Batch	Train. P.	Data Aug.	Norm
First	SGD (mom. 0.9)	100	0.01	64	117,789,043	Resize	[0, 255]
Best	SGD (mom. 0.9)	250	0.01	128	60,142,035	Flip, Crop, Resize	Cent + Scal

Third approach: two-stream CNN

Two-stream CNN:

- Validation loss: **2,33**
- Top 1 accuracy: **38%**
- Top 5 accuracy: **70%**



Models prediction on integral video (example)

❑ Spatial net:

- golf: 98,38 %
- shoot_bow: 0,81%
- catch: 0,40 %
- kick_ball: 0,25 %
- handstand: 0,03 %

❑ Temporal net:

- golf: 66,3 %
- climb: 6,1 %
- handstand: 5,0 %
- walk: 2,0 %
- pick: 2,0 %



❑ Spatial net:

- ride_bike: 75,6 %
- push: 17,1 %
- draw_sword: 1,1 %
- ride_horse: 0,9 %
- dive: 0,7 %

❑ Temporal net:

- catch: 19,2 %
- cartwheel: 6,8 %
- sommersault: 6,6 %
- dide_horse: 5,3 %
- climb: 5,3 %



Final evaluation

❑ Best method:

- ResNet (accuracy: **45,0%** , top 5 accuracy: **78,5%**)
- Possibly the two-stream CNN in case of steady or stabilized videos

❑ Possible causes of **low accuracy**:

▪ Dataset:

- Nuisances (camera motion, scenes cuts, low quality videos)
- Limited number of individual videos (6849 clips extracted from 1407 videos)
- Short action duration compared to video length (can be missed during the sampling phase)

▪ Models:

- Not optimal training parameters
- Too many weights compared to data (first model)

Bibliografia

- ❑ Kuehne, Hildegard, et al. "HMDB: a large video database for human motion recognition." 2011 International conference on computer vision. IEEE, 2011.
- ❑ Wang, Lei, and Piotr Koniusz. "Self-supervising action recognition by statistical moment and subspace descriptors." Proceedings of the 29th ACM International Conference on Multimedia (2021.)
- ❑ Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems 27 (2014).
- ❑ Wang, Limin, et al. "Temporal segment networks: Towards good practices for deep action recognition." European conference on computer vision. Springer, Cham, (2016).
- ❑ Sargano, Allah Bux, Plamen Angelov, and Zulfiqar Habib. "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition." applied sciences 7.1 (2017).
- ❑ Laptev, Ivan, et al. "Learning realistic human actions from movies." 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, (2008).

Bibliografia

- ❑ Zach, Christopher, Thomas Pock, and Horst Bischof. "A duality based approach for realtime tv-L1 optical flow." Joint pattern recognition symposium. Springer, Berlin, Heidelberg, (2007).
- ❑ Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2014).

Sitografia

- ❑ *Serre Lab*. URL: <https://serre-lab.clips.brown.edu/resource/hmdb-a-large-human-motion-database/>
- ❑ Introduction to Video Classification and Human Activity Recognition URL: <https://learnopencv.com/introduction-to-video-classification-and-human-activity-recognition/>
- ❑ Deep Learning for Videos: A 2018 Guide to Action Recognition: <https://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review>
- ❑ HMDB-51 Benchmark (Action Recognition) <https://paperswithcode.com/sota/action-recognition-in-videos-on-hmdb-51>

