Università degli Studi di Milano – Bicocca
Corso di Laurea Magistrale in Data Science
Anno Accademico 2021/2022

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

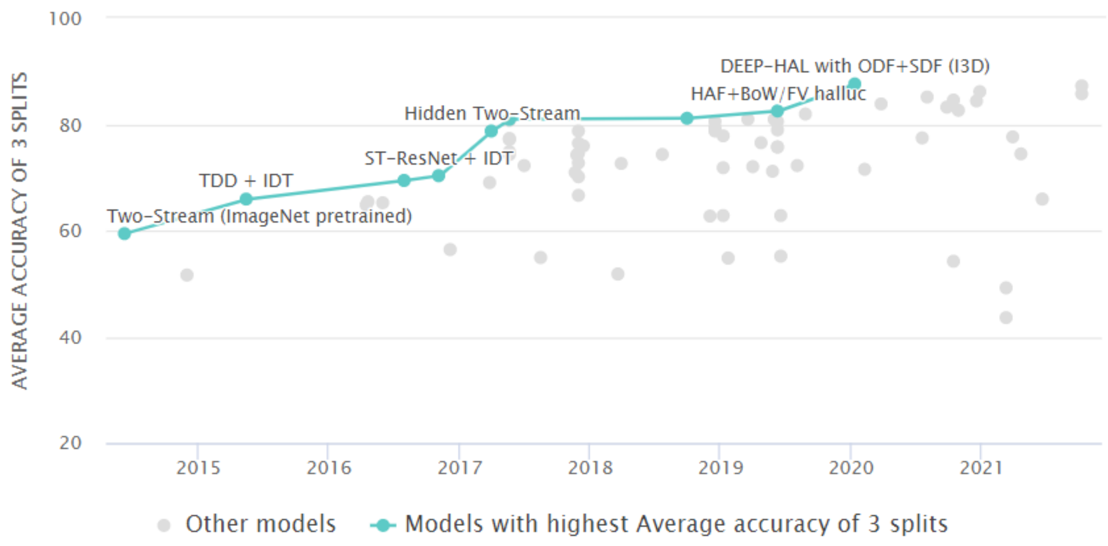
Task: Vision-based activity recognition

- **Human Activity Recognition:** Type of time series classification problem that involves classifying an action performed by someone
- **Video Classification And Human Activity Recognition:** Identification of different actions performed in video clips (a sequence of 2D frames)
- **Differences with image** classifications:
  - **Temporal** information
  - Higher **computational cost**
  - Capturing **long context** (many actions and camera movements)
  - No standard **benchmark datasets**

Literature and State-of-the-Art

- Approaches (spatial and temporal):
  - Single frame (2D) or stack of frames (3D) analysis for **spatial analysis**
  - Optical Flow or RNN for **temporal analysis**
  - **Combinations** of these (Two-stream networks)
- State-of-the-Art:
  - DEEP-HAL with ODF+SDF (I3D) 2020 -> **87.56%**

Action Recognition on HMDB-51



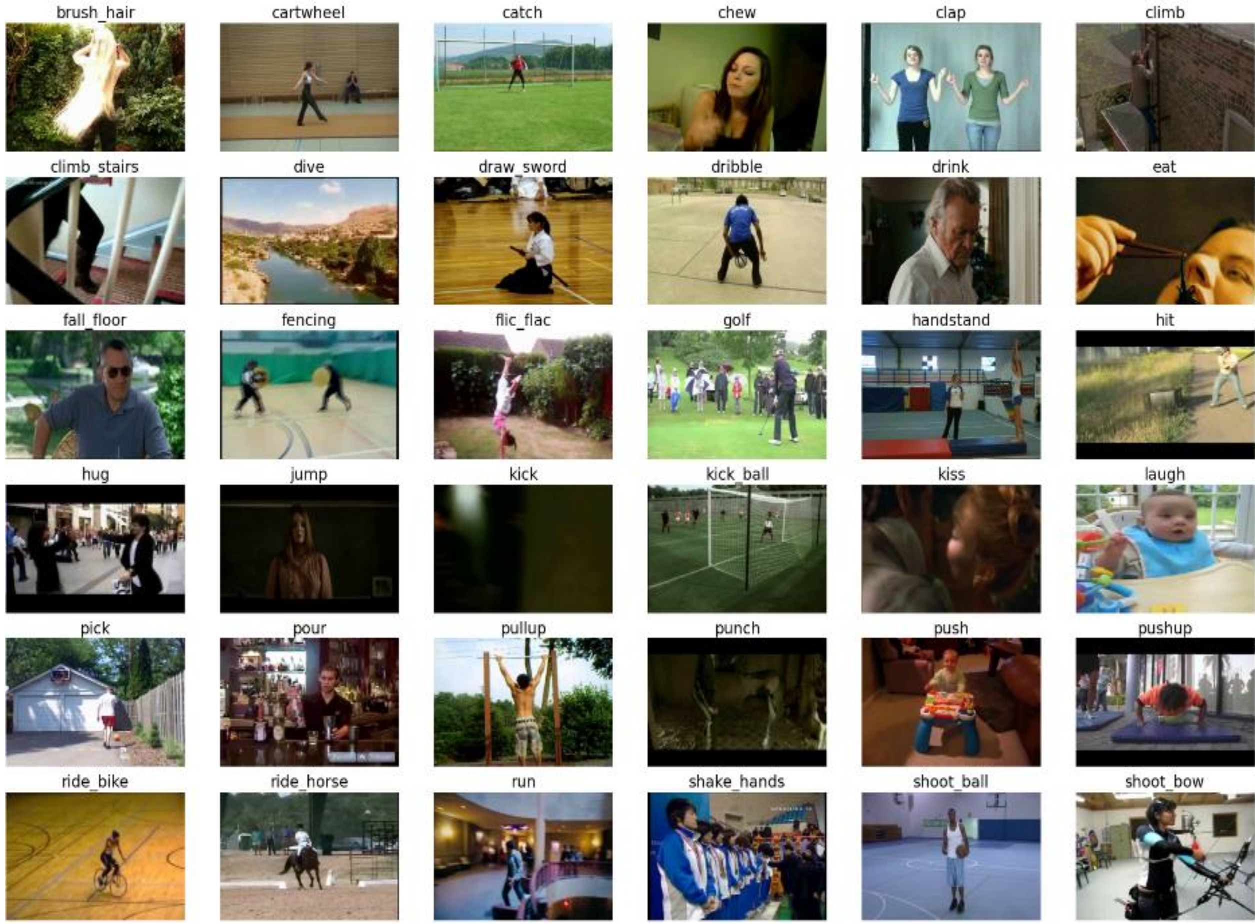
HMDB: A Large Video Database for Human Motion Recognition

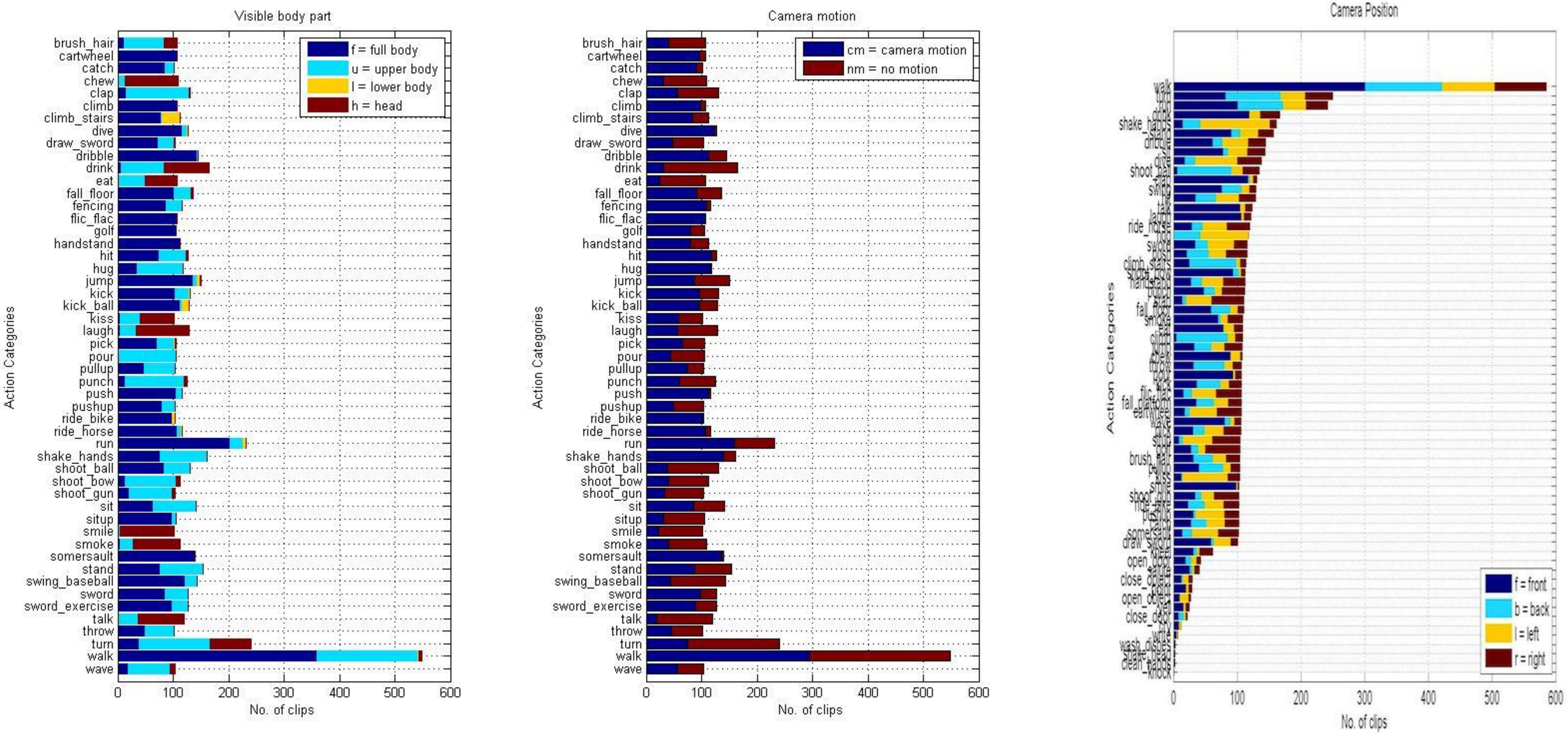
- 6849 clips
- Different **sources**: Youtube, Google videos, movies.
- **51 categories** with a minimum of 101 clips per action
- Those **categories** can be grouped in **five types**:
    - General **facial actions**
    - **Facial actions with object**
    - General **body movements**
    - **Body movements with object** interaction
    - **Body movements for human interaction**

Data Exploration

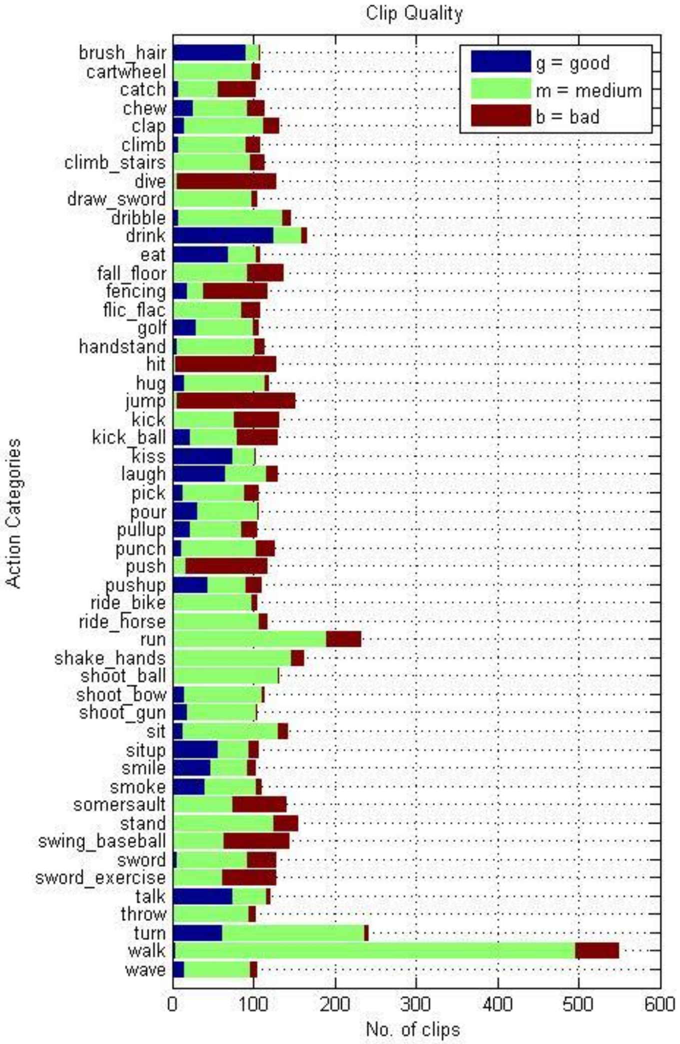
- Video **normalization**:
  - **Height** of all the frames is scaled to **240** pixels
  - **Width** is scaled to maintain aspect ratio
  - Frame rate of **30 fps**

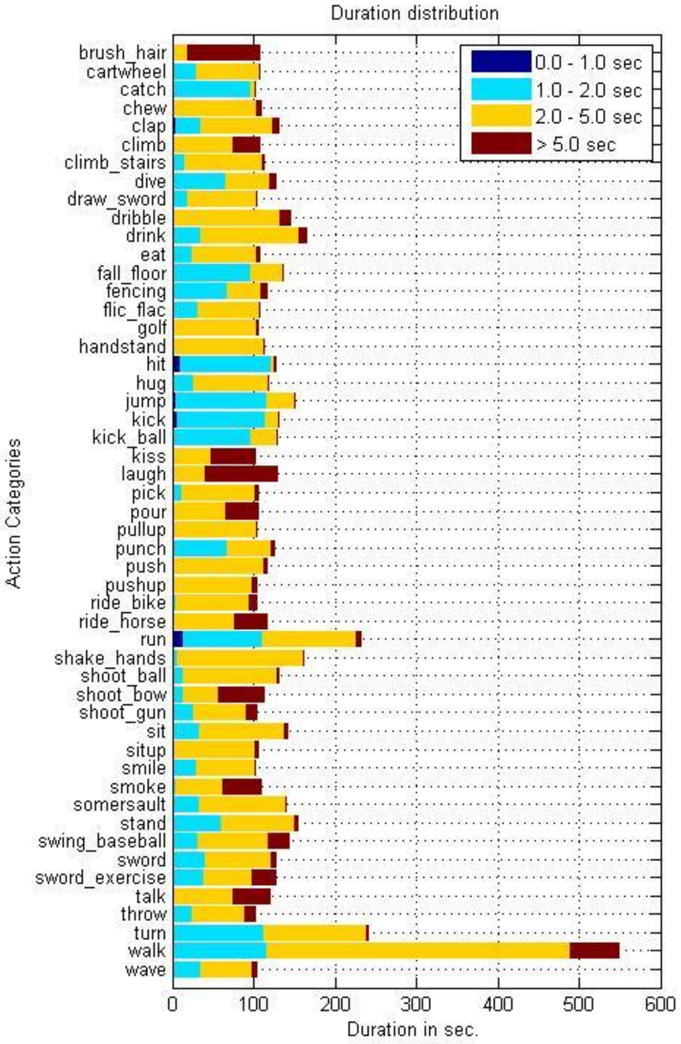
- Videos are **classified** as following:
  - Visible **body parts**: full, upper or lower body
  - **Camera motion**: motion or static
  - **Camera viewpoint**: front, back, left, right
  - **Number of people** involved in the action
  - **Video quality**: good, medium or bad





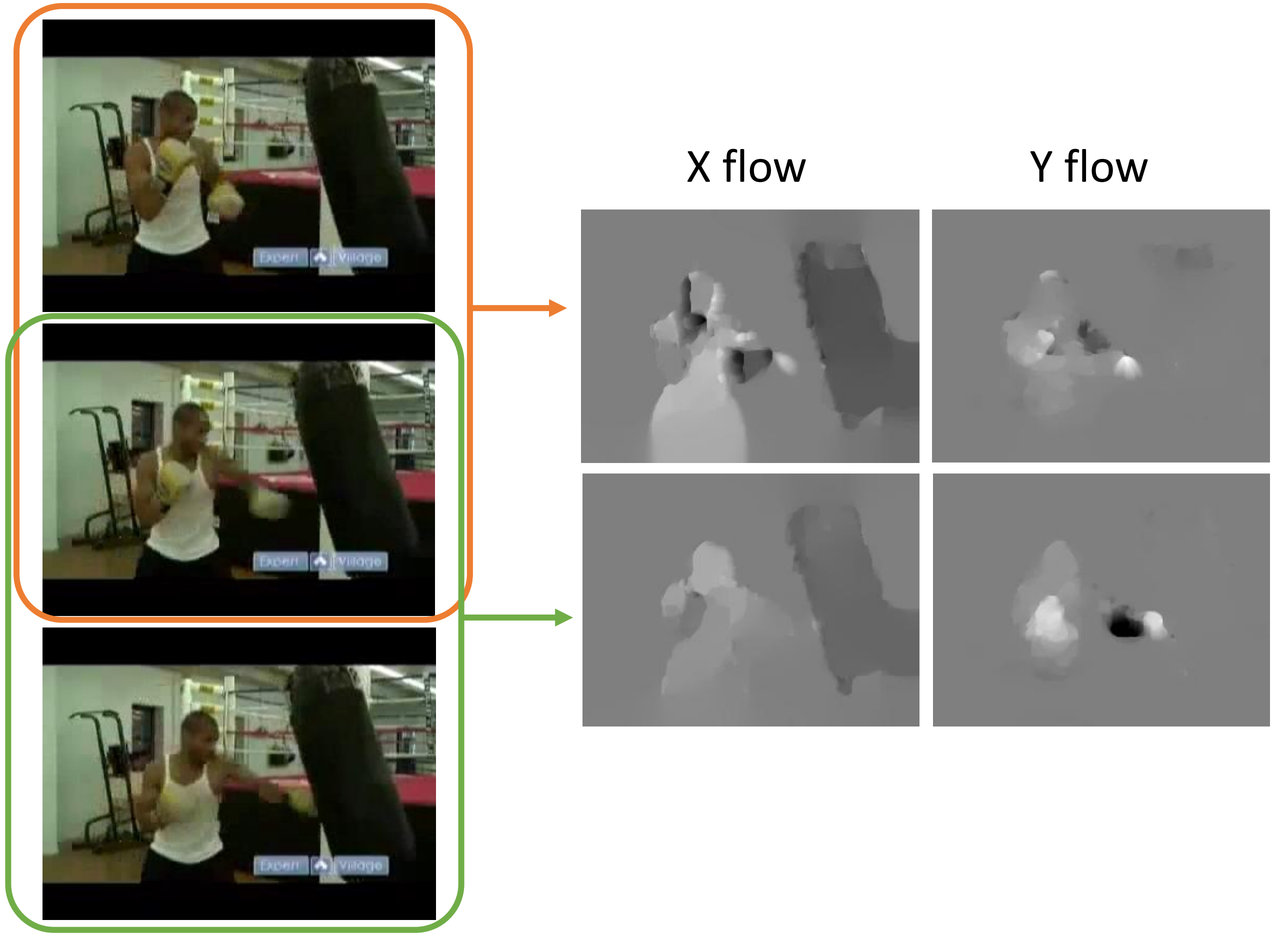
Visible body part
f = full body
u = upper body
l = lower body
h = head
Action Categories
brush_hair
cartwheel
catch
chew
clap
climb
climb_stairs
dive
draw_sword
dribble
drink
eat
fall_floor
fencing
flic_flac
golf
handstand
hit
hug
jump
kick
kick_ball
kiss
laugh
pick
pour
pullup
punch
push
pushup
ride_bike
ride_horse
run
shake_hands
shoot_ball
shoot_bow
shoot_gun
sit
situp
smile
smoke
somersault
stand
swing_baseball
sword
sword_exercise
talk
throw
turn
walk
wave
0
100
200
300
400
500
600
No. of clips
Camera motion
cm = camera motion
nm = no motion
Action Categories
brush_hair
cartwheel
catch
chew
clap
climb
climb_stairs
dive
draw_sword
dribble
drink
eat
fall_floor
fencing
flic_flac
golf
handstand
hit
hug
jump
kick
kick_ball
kiss
laugh
pick
pour
pullup
punch
push
pushup
ride_bike
ride_horse
run
shake_hands
shoot_ball
shoot_bow
shoot_gun
sit
situp
smile
smoke
somersault
stand
swing_baseball
sword
sword_exercise
talk
throw
turn
walk
wave
0
100
200
300
400
500
600
No. of clips
Camera Position
f = front
b = back
l = left
r = right
Action Categories
walk
shake_hands
shoot_ball
ride_horse
climb_stairs
handstand
fall_floor
brush_hair
0
100
200
300
400
500
600
No. of clips


Clip Quality
g = good
m = medium
b = bad
Action Categories
brush_hair
cartwheel
catch
chew
clap
climb
climb_stairs
dive
draw_sword
dribble
drink
eat
fall_floor
fencing
flic_flac
golf
handstand
hit
hug
jump
kick
kick_ball
kiss
laugh
pick
pour
pullup
punch
push
pushup
ride_bike
ride_horse
run
shake_hands
shoot_ball
shoot_bow
shoot_gun
sit
situp
smile
smoke
somersault
stand
swing_baseball
sword
sword_exercise
talk
throw
turn
walk
wave
0
100
200
300
400
500
600
No. of clips


Duration distribution
0.0 - 1.0 sec
1.0 - 2.0 sec
2.0 - 5.0 sec
> 5.0 sec
Action Categories
brush_hair
cartwheel
catch
chew
clap
climb
climb_stairs
dive
draw_sword
dribble
drink
eat
fall_floor
fencing
flic_flac
golf
handstand
hit
hug
jump
kick
kick_ball
kiss
laugh
pick
pour
pullup
punch
push
pushup
ride_bike
ride_horse
run
shake_hands
shoot_ball
shoot_bow
shoot_gun
sit
situp
smile
smoke
somersault
stand
swing_baseball
sword
sword_exercise
talk
throw
turn
walk
wave
0
100
200
300
400
500
600
Duration in sec.

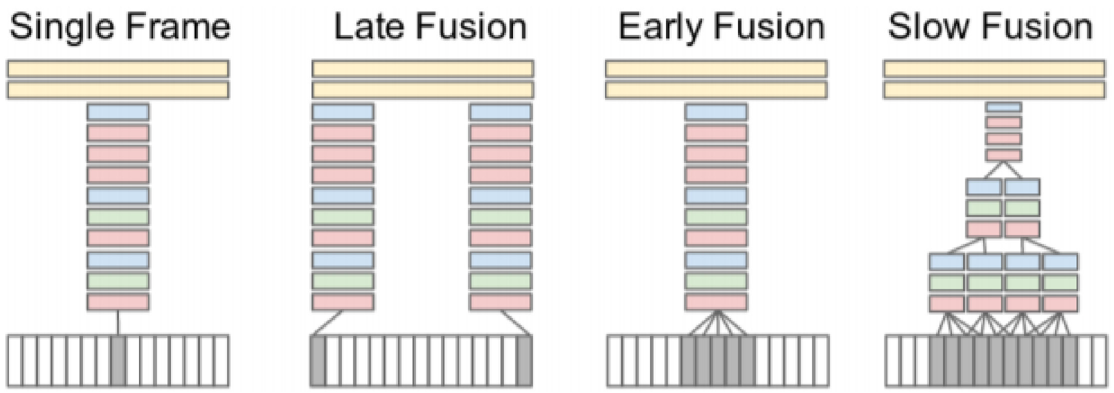
Data Preparation

1. Split **training and test** (respecting class balance)
2. **Frame extraction** from videos (625775 total frames)
3. **Dense optical flow**
   - Between 2 consecutive frames
   - TV-L1 Optical Flow Estimation
4. **Sampling**
   - Frames: 17 frames per video equally spaced
   - Stacked Optical Flows (224, 224, 20)
5. **Data augmentation** on frames & optical flows:
   - Random **Horizontal flip** (50% probability)
   - Random **Crop** (224x224)
   - Random **Rotation** (0.15)



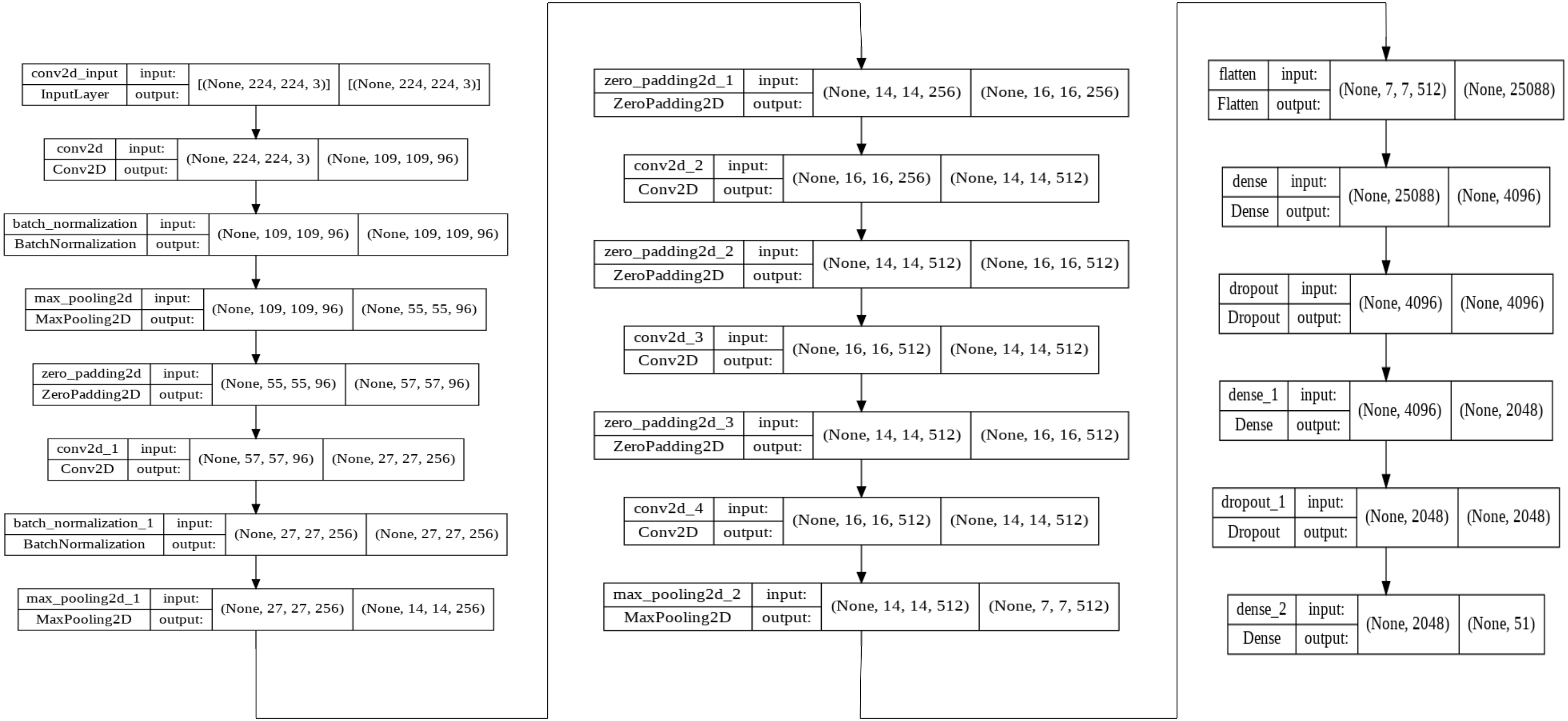
First approach: CNN image classification

- Layers:
  - Convolutional2D (activation: Relu)
  - Batch normalization
  - MaxPooling (pool size: 3x3, stride: 2)
  - Dense (dropout: 0.5, activation: ReLu, Softmax)
- Training parameters:
  - Optimizer: Adam (Learning rate: 0.001)
  - Batch size: 64
  - Epochs: 100
- Results:
  - Validation loss: **13.73**
  - Top 1 accuracy: **6.6%**



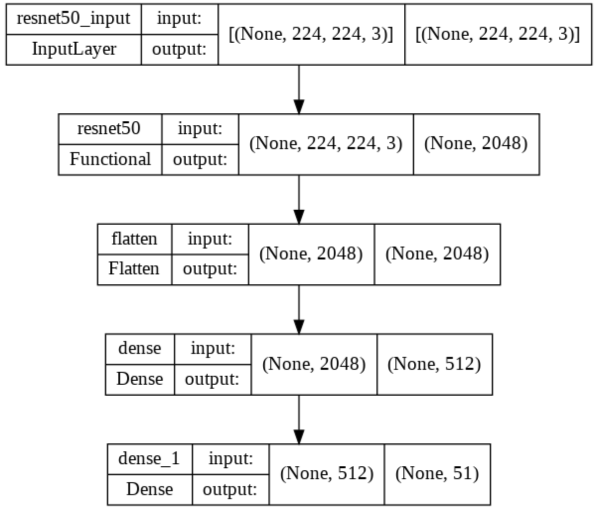
Parameters	Number
Total	117,789,747
Trainable	117,789,043
Non-trainable	704

First approach: CNN image classification



Second approach: finetuned ResNet50

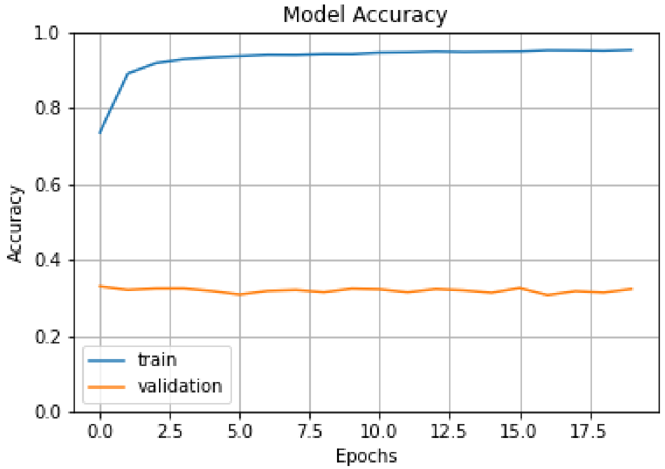
- Layers:
  - ResNet50 pre-trained on ImageNet
  - Flatten
  - Dense (activation: ReLu, SoftMax)
- Training parameters:
  - Optimizer: Adam
  - Learning rate: 0.001 -> 0.0001
  - Loss: sparse categorical crossentropy
  - Batch size: 64
  - Epochs: 20 -> 5

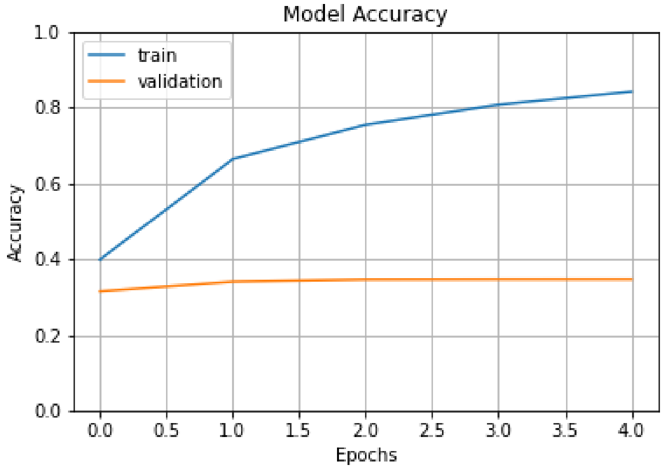


Parameters	Number
Total	24,662,963
Trainable	1,075,251
Non-trainable	23,587,712

Second approach: finetuned ResNet50

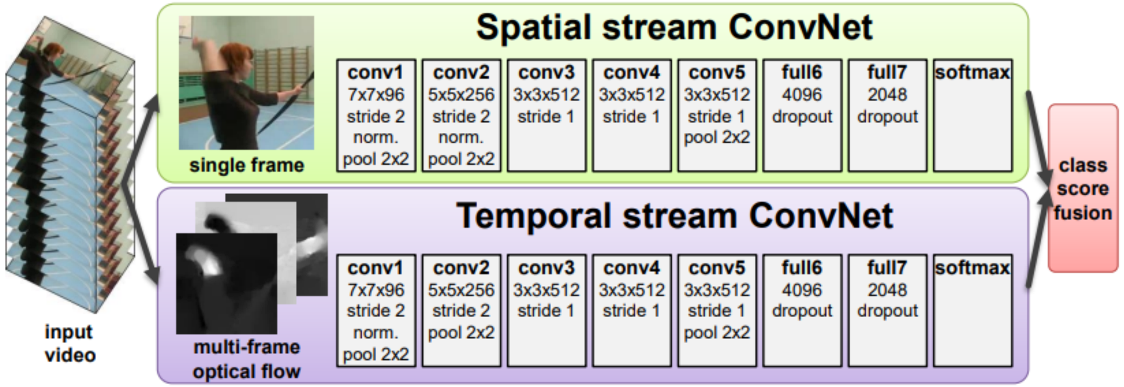
- First train:
  - Validation loss (sparse categorical crossentropy): 2.91
  - Top 1 accuracy: 31.9%
  - Top 5 accuracy: 59.1%
- Second train:
  - Validation loss (sparse categorical crossentropy): 2.72
  - Top 1 accuracy: 34.6%
  - Top 5 accuracy: 68.1%





Third approach: two-stream CNN

- ❑ Architecture:
  - Spatial CNN + Temporal CNN
  - Averaging SoftMax results
- ❑ Spatial stream:
  - Finetuned ResNet
  - 17 frames (224, 224, 3)
- ❑ Temporal stream:
  - Same architecture as “first approach CNN” (optimizer: SGD, learning rate: 0.01, momentum: 0.9)
  - Randomly selected batch of 32 stacked optical flow from 32 randomly selected videos
  - Each flow stack is composed of 10 x-channels and 10 y-channels consecutive optical flows (224, 224, 20)



Third approach: two-stream CNN

- Spacial stream:
  - Validation loss: 2.72
  - Top 1 accuracy: 34.6%
  - Top 5 accuracy: 68.1%
- Temporal stream:
  - Validation loss: 3,46
  - Top 1 accuracy: 15%
  - Top 5 accuracy: 42.4%
- Class score fusion:
  - Validation loss: NA
  - Top 1 accuracy: NA
  - Top 5 accuracy: NA

Final evaluation

- **Best method**:
  - NA
- Possible causes of **low accuracy:**
  - **Dataset:**
    - Image degradation (Camera motion, scenes cuts, low quality videos, noise)
    - Short action duration compared to video length (can be missed during the sampling phase)
  - **Models:**
    - Not optimal train 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

- 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): 110.
- Zach, Christopher, Thomas Pock, and Horst Bischof. "A duality based approach for realtime tv-l 1 optical flow." Joint pattern recognition symposium. Springer, Berlin, Heidelberg, 2007.

Sitografia

- *Serre Lab.* URL: https://serre-lab.clps.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/
- Browse State-of-the-art: https://paperswithcode.com/dataset/hmdb51
- Deep Learning for Videos: A 2018 Guide to Action Recognition: https://blog.qure.ai/notes/deep-learning-for-videos-action-recognition-review