ICCV 2021 Workshop - SRVU

Spatio-Temporal Video Representation Learning for AI Based Video Playback Style Prediction

Rishubh Parihar, Gaurav Ramola, Ranajit Saha, Raviprasad kini,
Aniket Rege, Sudha Velusamy, Samsung Bangalore

Agenda

1) The Relevance of Video understanding for Mobile Devices
2) Current State of Video understanding approaches
3) Motion patterns in human action videos - mHMDB51 dataset
    a) Motion Type Classifier Architecture
    b) Quantitative Results
    c) Video playback style recommendation
4) Conclusion

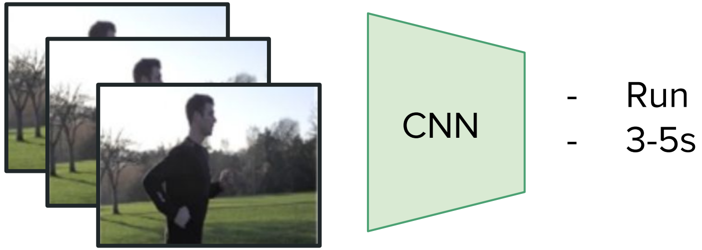
Video Analysis on Mobile Devices

- A large number of videos are captured on mobile phones each day that are shared various short video platforms like tik-tok, snapchat, reels.
- In current scenario there are a range of tools available where the user has to manually select and try of the filters
- Their is a necessity of automated tools to edit the videos on mobile devices to make them more shareable
- Intelligence capability for mobile devices

[1] Best Short Video Apps for Socializing 2021
https://www.apowersoft.com/best-short-video-apps.html

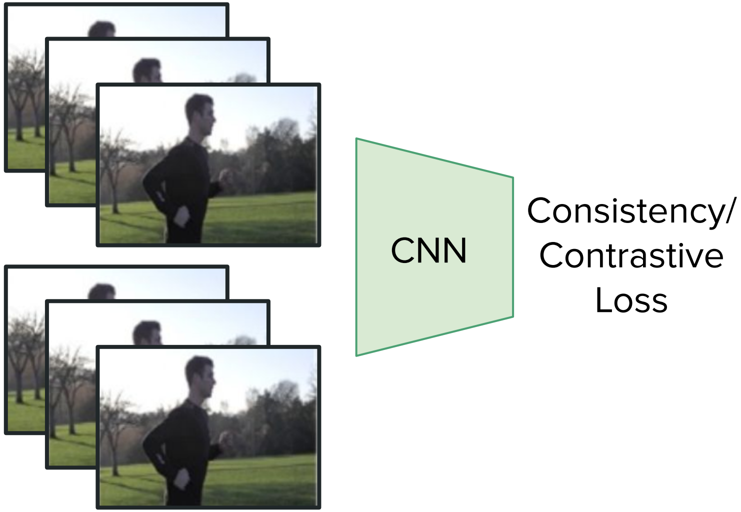
Current State of Video Understanding

Action recognition and localization



- Training with large scale labeled datasets
- Supervised Training with 3D CNNs

Unsupervised representation learning



- Training with unlabelled data to learn spatio-temporal representations

Motion Classification

Every common world human actions can be categorized into one of the following five primitive motion type classes: linear, projectile, oscillatory, local and random - mHMDB51

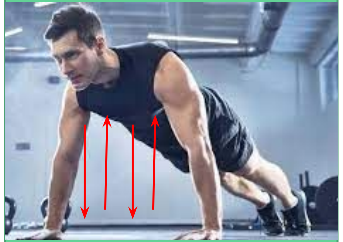
Linear

Ex. Run, Walk, Brush-hair, Climb, Push, Pull

Projectile

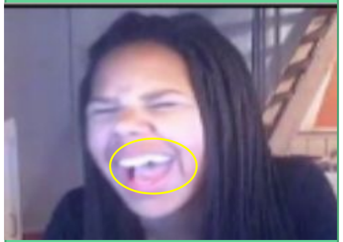
Ex. Shoot ball, Cartwheel, Dive, Jump, Golf

Oscillatory



Ex. Pushups, Dribble, Situps, Clap

Local



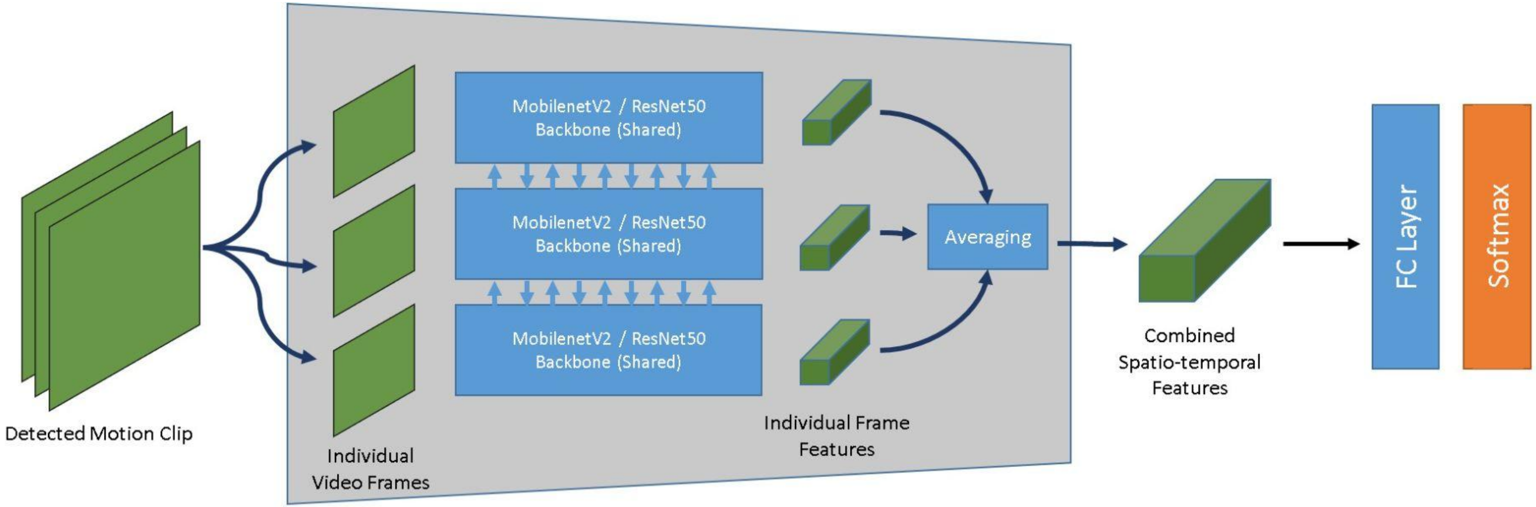
Ex. Smile, Chew, Talk, Smoke, Shake-hands

Random

Ex. Fencing, Fall, Sit, Stand, Hug

Motion Classifier Architecture

- Our model architecture is inspired by Temporal Segment Networks with TSM blocks
- We sample T frames from the video and process them through a MobileNet based TSN backbone



- Linear
- Projectile
- Oscillatory
- Local
- Random

[1] Temporal Segment Networks: Towards Good Practices for Deep Action Recognition
https://arxiv.org/abs/1608.00859

Motion Classification Results

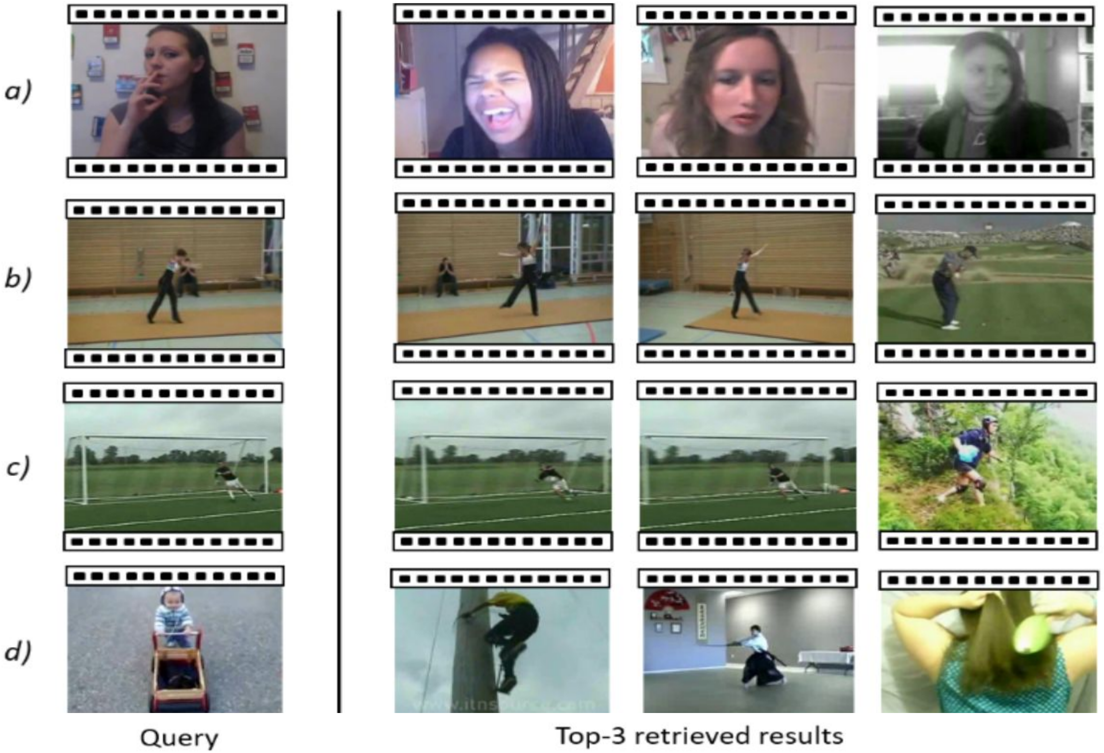
Table1. Model Performance Comparison

Method	Accuracy
Flow Baseline Classifier	25.64
$\text{Ours}_{\text{Scratch}}$	38.56
$\text{Ours}_{\text{ImageNet}}$	57.58
$\text{Ours}_{\text{Kinetics}}$	72.68

Table2. Ablation on number of input frames

Segments	Accuracy	MACs
1	61.76	0.41G
2	71.05	0.82G
3	72.68	1.23G
8	68.17	3.28G

Results on the Downstream Task of Video Retrieval



Video Playback Style Recommendation

Input Video Clip	Motion Type Predicted	Playback Style Assigned
Jogging 	***Linear***	*Reverse*
Diving 	***Projectile***	*Boomerang*
Drinking 	***Local***	*Loop*
Fencing 	***Random***	*Forward*

Conclusions

- A novel direction for video understanding by motion type classification
- Inference time of 200ms for a 10s video clip on a Samsung S20 phone
- Learned rich motion representations that generalize well to downstream task of video retrieval
- An application of Video Playback style recommendation system based on predicted motion type classification