Video-based Action Recognition

Contents

Dataset Description.

For this study, the UCF 101 dataset, which is an extension of the UCF50 dataset, is being used. This dataset contains more than 13,000 videos which constitutes to 101 action categories.

The action categories can be divided into five types:

1) Human-Object Interaction.

2) Body-Motion Only.

3) Human-Human Interaction.

4) Playing Musical Instruments.

5) Sports.

The original action categories are as follows: The action categories for UCF101 data set are: *Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Baseball Pitch, Basketball Shooting, Basketball Dunk, Bench Press, Biking, Billiards Shot, Blow Dry Hair, Blowing Candles, Body Weight Squats, Bowling, Boxing Punching Bag, Boxing Speed Bag, Breaststroke, Brushing Teeth, Clean and Jerk, Cliff Diving, Cricket Bowling, Cricket Shot, Cutting In Kitchen, Diving, Drumming, Fencing, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, Front Crawl, Golf Swing, Haircut, Hammer Throw, Hammering, Handstand Pushups, Handstand Walking, Head Massage, High Jump, Horse Race, Horse Riding, Hula Hoop, Ice Dancing, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Knitting, Long Jump, Lunges, Military Parade, Mixing Batter, Mopping Floor, Nun chucks, Parallel Bars, Pizza Tossing, Playing Guitar, Playing Piano, Playing Tabla, Playing Violin, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Sitar, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rafting, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Shaving Beard, Shotput, Skate Boarding, Skiing, Skijet, Sky Diving, Soccer Juggling, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Swing, Table Tennis Shot, Tai Chi, Tennis Swing, Throw Discus, Trampoline Jumping, Typing, Uneven Bars, Volleyball Spiking, Walking with a dog, Wall Pushups, Writing On Board, Yo Yo.*

Video conversion.

For better training and testing purposes, the videos are clipped and fps reduced with the help of the `imageio.mimsave` function. The videos will be interpreted as easy to learn, resized gifs.

Kinetics 400 Labeling.

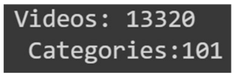
The dataset contains human action classes, with at least 400 video clips for each action. Each clip lasts around 10s and is taken from a different YouTube video. The actions are human focused and cover a broad range of classes including human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands.

For our current use case, we will use these labels in accordance with our UFC 101 dataset. There are about 400 labels present in this dataset.

```
Total:400 labels
```

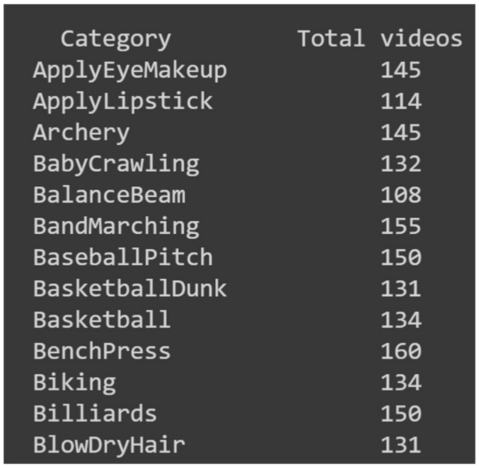
UCF Dataset summary.

After an initial scan of the dataset, we find that there are total 13,320 videos and 101 categories.



```
Videos: 13320
 Categories:101
```

A few of the categories and number of videos are as follows:

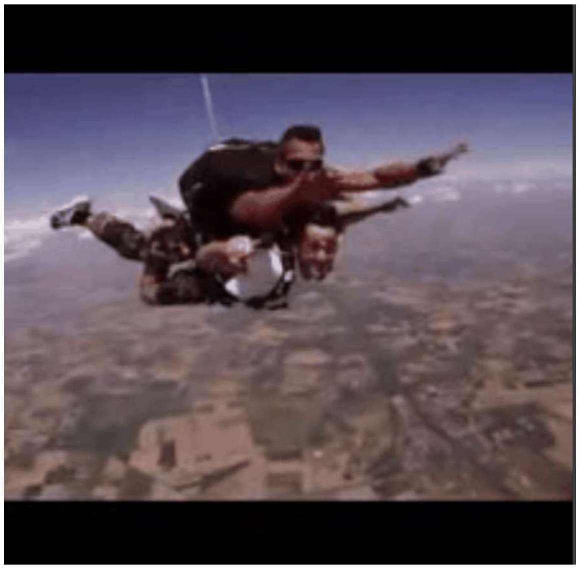


```
   Category         Total videos
 ApplyEyeMakeup           145
 ApplyLipstick            114
 Archery                  145
 BabyCrawling             132
 BalanceBeam              108
 BandMarching             155
 BaseballPitch            150
 BasketballDunk           131
 Basketball               134
 BenchPress               160
 Biking                   134
 Billiards                150
 BlowDryHair              131
```

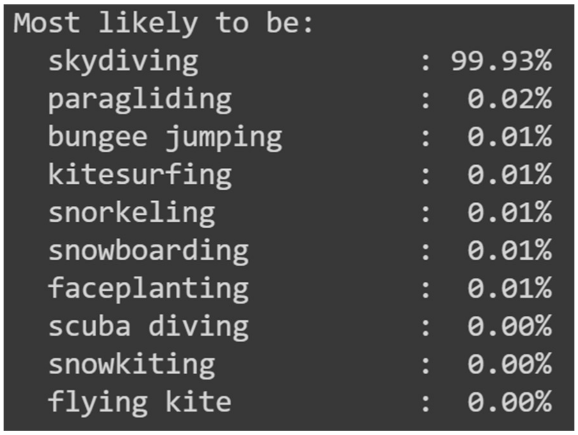
Highest probabilities results with Tensorflow.

This study will leverage Tensorflow for training and testing the input videos.

Example input video 1 (seen better in notebook):

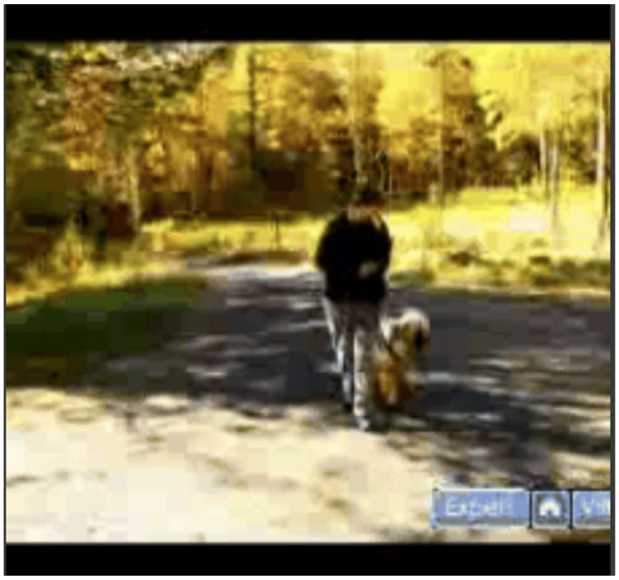


Results:

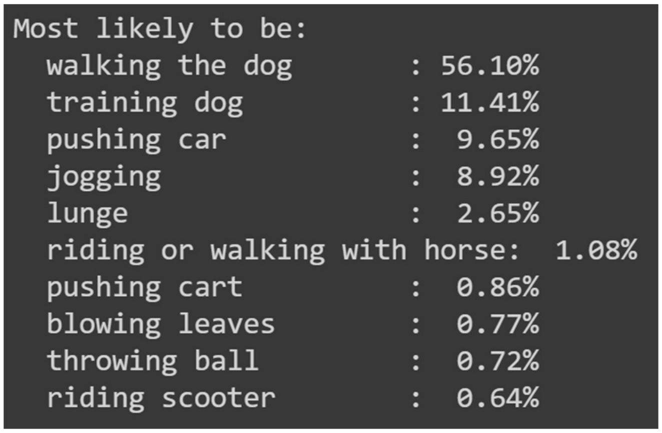


```
Most likely to be:
  skydiving             : 99.93%
  paragliding           :  0.02%
  bungee jumping        :  0.01%
  kitesurfing           :  0.01%
  snorkeling            :  0.01%
  snowboarding          :  0.01%
  faceplanting          :  0.01%
  scuba diving          :  0.00%
  snowkiting            :  0.00%
  flying kite           :  0.00%
```

Example input video 2:

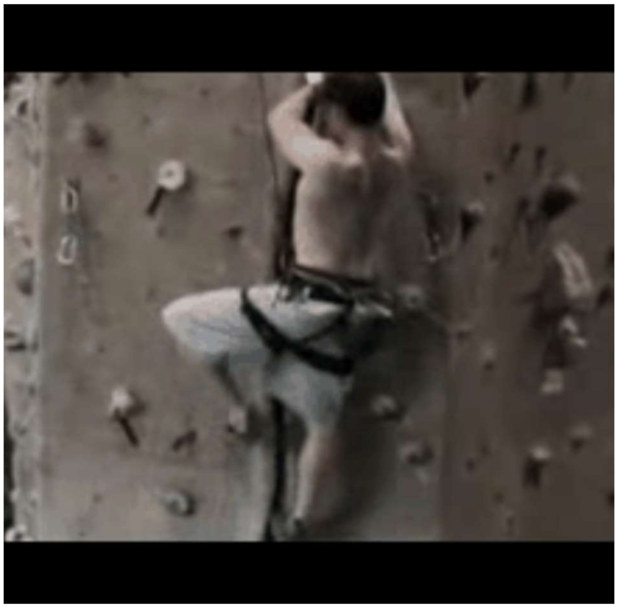


Results:

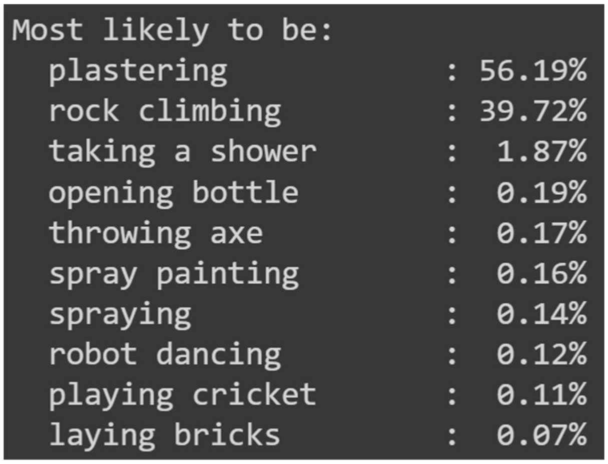


```
Most likely to be:
  walking the dog       : 56.10%
  training dog          : 11.41%
  pushing car           :  9.65%
  jogging               :  8.92%
  lunge                 :  2.65%
  riding or walking with horse:  1.08%
  pushing cart          :  0.86%
  blowing leaves        :  0.77%
  throwing ball         :  0.72%
  riding scooter        :  0.64%
```

Example input video 3:



Result:



```
Most likely to be:
  plastering             : 56.19%
  rock climbing          : 39.72%
  taking a shower        :  1.87%
  opening bottle         :  0.19%
  throwing axe           :  0.17%
  spray painting         :  0.16%
  spraying               :  0.14%
  robot dancing          :  0.12%
  playing cricket        :  0.11%
  laying bricks          :  0.07%
```

This is a misclassified instance as the action in the input video is most recognized to be ‘plastering’ and ‘rock climbing’. This input video is classified as ‘rock climbing’ in the dataset.

References

1) CRCV | Center for Research in Computer Vision at the University of Central Florida (ucf.edu)

2) [1705.06950] The Kinetics Human Action Video Dataset (arxiv.org)

3) TensorFlow Hub (tfhub.dev)