Action Recognition based on Pose Estimation

Presented by:
Anshu Kumari
(39/CSE/17019/217)
4th Year, 8th Semester

Mentored by:
Dr. Oishila Bandyopadhyay
Assistant Professor
Department of Computer Science
IIIT Kalyani

Overview

- Objective
- Motivation
- Challenges
- Related Works
- Proposed Method
- Pipeline
- Data Description
- Technical Dependencies
- Training Methodology
- CNN Model Architecture
- Model Efficiency
- Inferences
- Conclusion
- Future Work
- Reference

Objective

Suspicious human activity recognition from surveillance video is an active research area of image processing and computer vision.

I proposed a deep learning approach to classify human activities by tracking multiple humans in real-time scenarios or through videos.

It consists of mainly 4 different types of human activities such as sitting, standing, running, & falling-down.

The objective is to recognize the suspicious human activities such as running and falling-down in surveillance videos, mostly applicable on shopping mall.

Motivation

- Inspection of illegal activities
- In Sports to track players
- For autonomous training
- Tracking any specific person
- Surveillance and security
- Pedestrian Tracking
- Traffic control
- In shopping mall to track individuals

Challenges

- **Accuracy:** Dependency of tracking accuracy on detection accuracy
- **Occlusion:** Loses track of humans in case of Occlusion
- **Visual Similarity:** Overlapping bounding boxes with low Euclidean distances are associated without taking into consideration the visual similarity between them.
- **Annotating Training Dataset:** Difficult to collect data about any emergent and unpredictable events
- **GPU is required:** For better results in real-world applications

Related Work

- **Traditional Features Based**
- Extracted features: aspect ratio, fall angle, centre speed, and head speed
- Classifier: SVM, CNN

- **Depth Features Based**
- Extracted features: 18 keypoints of human body skeleton
- Feature extractor: VGG-16 network
- Classifier: RNN with LSTM & other CNN classifier

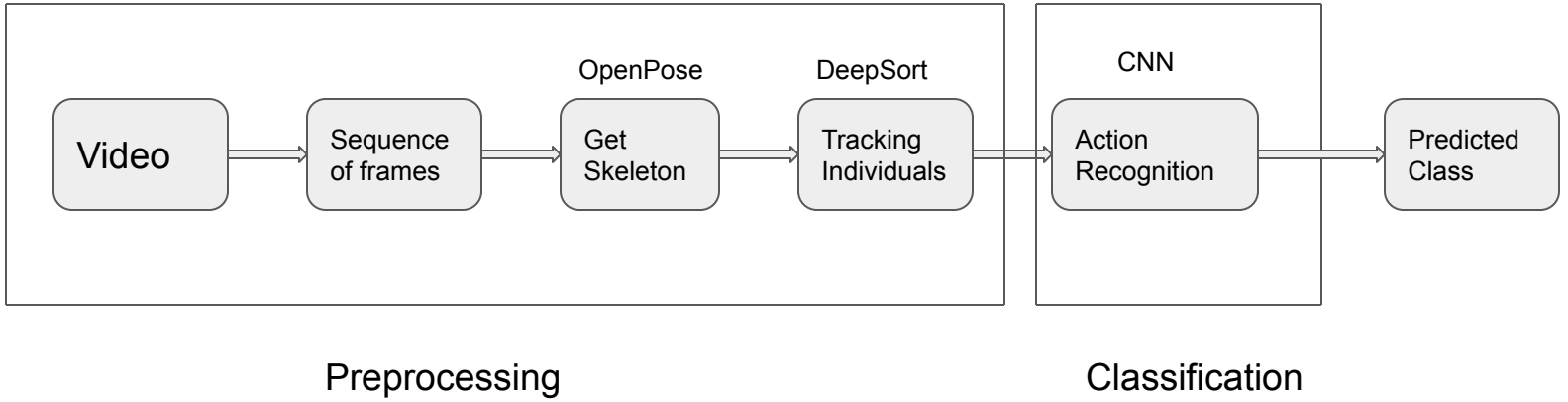
Proposed Method

The **Skeleton-based Convolutional Networks for Action Recognition approach** aims to classifying and recognizing individuals based on framewise joints, which can be used for safety monitoring.

The implementation steps of this approach are:

- Realtime pose estimation by **OpenPose.**
- Online human tracking for multi-people scenario by **DeepSort** algorithm.
- Action recognition with DNN for each person based on single framewise joints detected from OpenPose.

Pipeline


OpenPose
DeepSort
CNN
Video
Sequence of frames
Get Skeleton
Tracking Individuals
Action Recognition
Predicted Class
Preprocessing
Classification

Data Description

- **The Le2i Fall Detection Dataset**
  - It consists of 191 human activity videos in four scenes: Office, Home, Coffee Room, and Lecture Room.
  - The format of the video is 320*240, 25frame/s.

- **The UR Fall Detection Dataset**
  - It consists of 70 depth video and corresponding RGB video, which contains 30 fall videos and 40 daily activity videos from different angles and different lighting conditions.
  - The image format is 1920*1080.

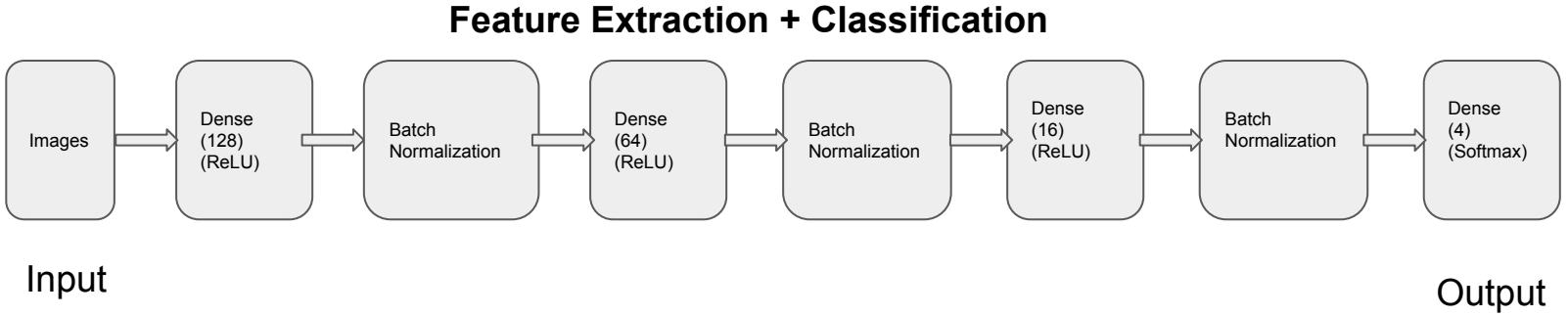
Technical Dependencies

- Python
- TensorFlow & Keras
- OpenCV
- NumPy & SciPy
- Scikit-Learn Library

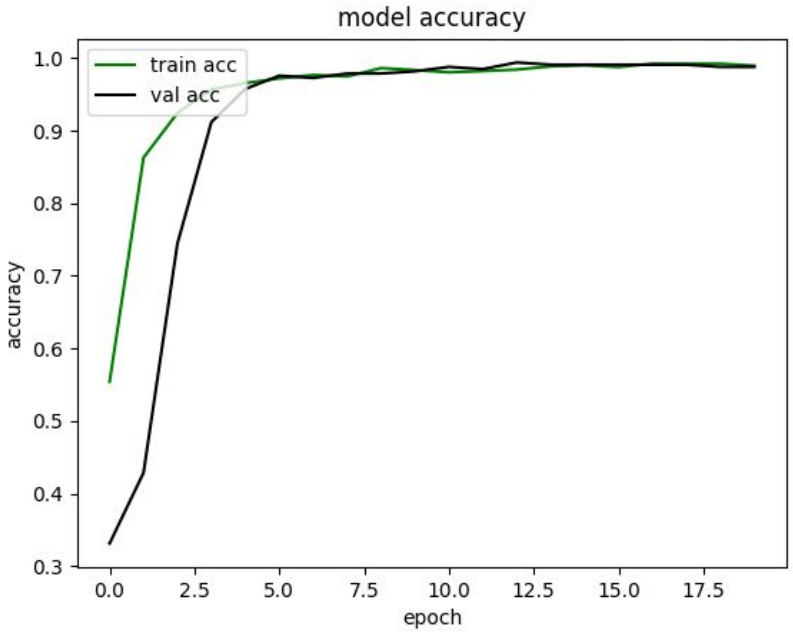
Training Methodology

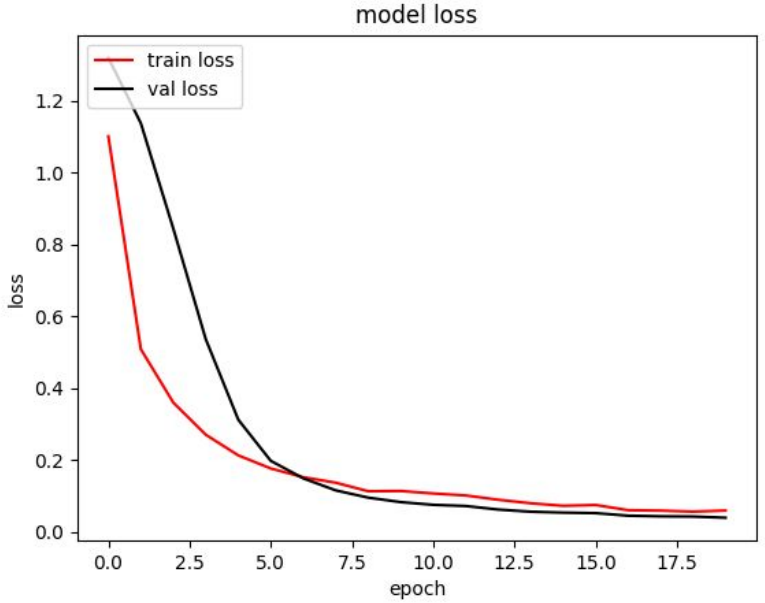
1. First, I have used OpenPOse pre-trained VGG_origin, training with the VGG net or mobilenet_thin model, training with mobilenet on COCO dataset to get human skeleton bodies on each frame and their corresponding 18 keypoints with their (x,y) coordinates on individuals with respect to each frame.
2. The sequence of frames is tracked using DeepSort, to track the multiple persons with longer period of occlusions. It uses deep association metric to reduce the occlusion visual challenge.
3. After preprocessing the sequence of frames, it passages through sequential CNN model classifier to first train the model on the chosen dataset and then the model is validate again test dataset.
4. Then, a performance metrics is calculated to measure the reliability of the model.

CNN Model Architecture

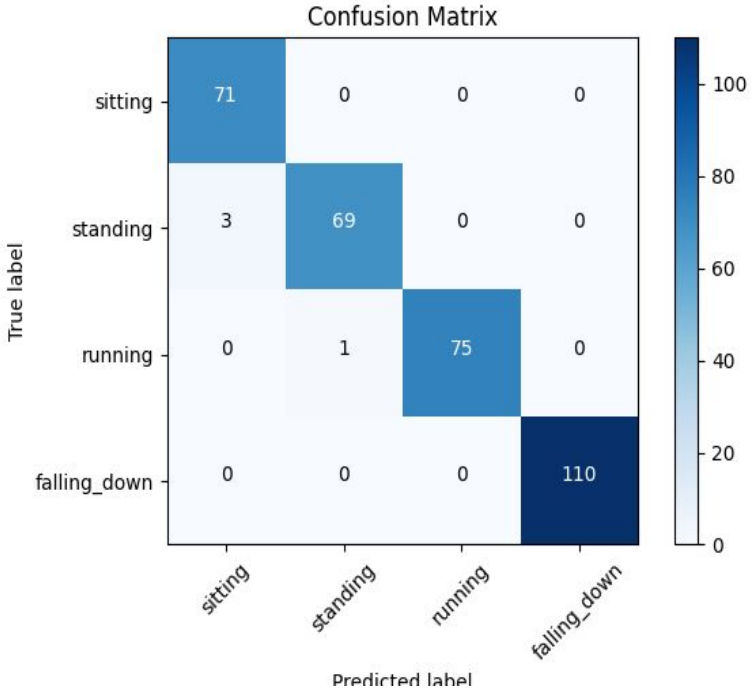


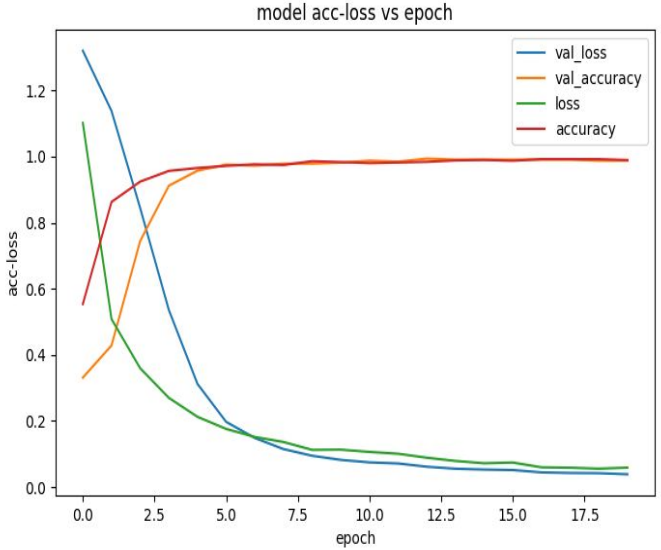
Model Efficiency



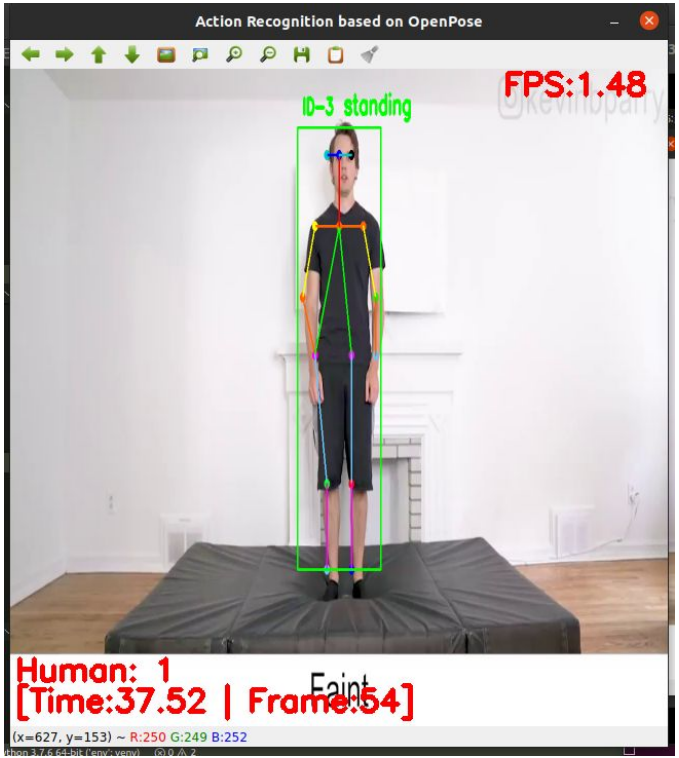


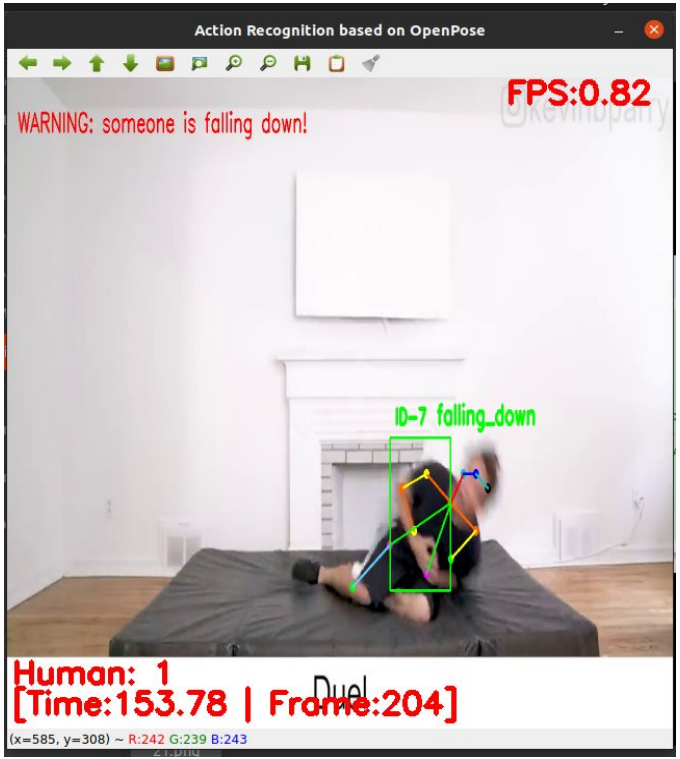
Model Efficiency





Inferences

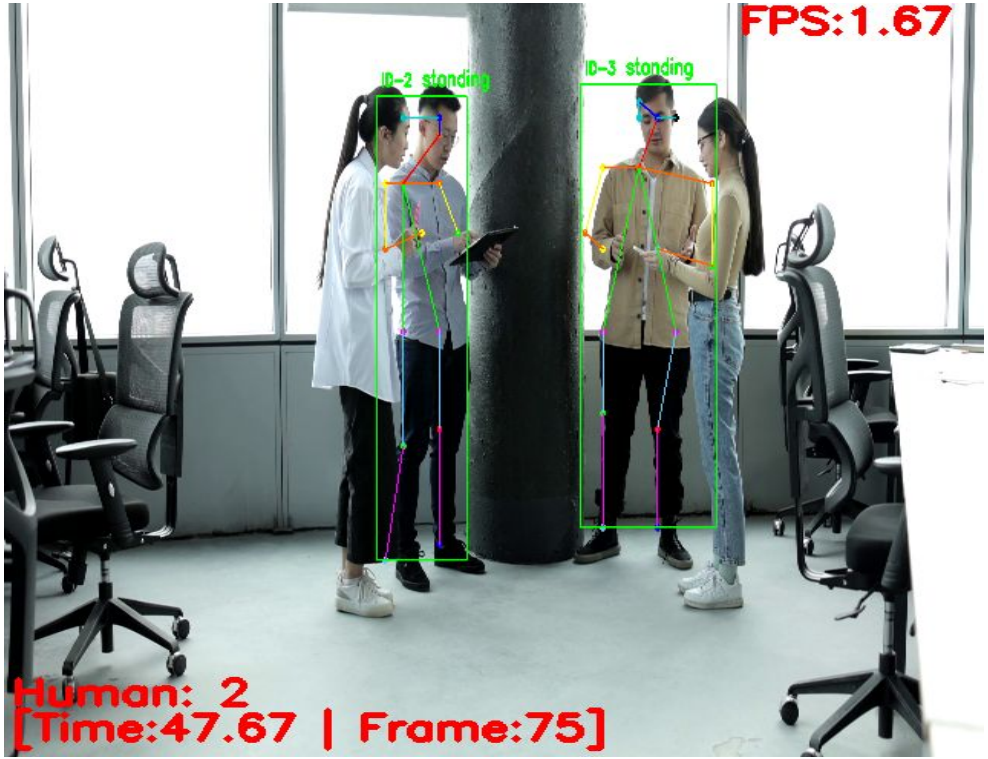

Action Recognition based on OpenPose
FPS:1.48
ID-3 standing
Human: 1
[Time:37.52 | Frame:54]
Faint
(x=627, y=153) ~ R:250 G:249 B:252


Action Recognition based on OpenPose
FPS:0.82
WARNING: someone is falling down!
ID-7 falling_down
Human: 1
[Time:153.78 | Frame:204]
Duel
(x=585, y=308) ~ R:242 G:239 B:243

Inferences

Wrong Prediction

- In terms of occlusion
- Four persons in the frame
- Detecting only two persons



Conclusion

- The approach achieves 97.91 accuracy on the Le2i Fall Detection Dataset and achieves 98.33% accuracy on the UR Fall Detection Dataset.
- This approach can be used to identify sudden changes in the shape of the human body.
- It partly handles problems caused due to occlusion by introducing the DeepSort framework.

Future Work

- In the future, I will improve this method to focus on more complex environments, such as crowded public areas, malls, etc.
- Additionally, the model will be extended to identify other dangerous behaviours that may occur in our daily life.
- Furthermore, I hope to develop a practical application of the activity detection system to integrate with the Internet of things.

Reference

[1] Wu, Zuxuan and Yao, Ting and Fu, Yanwei and Jiang, Yu-Gang, “Deep Learning for Video Classification and Captioning”, Association for Computing Machinery and Morgan & Claypool, 2017, https://doi.org/10.1145/3122865.3122867

[2] Jeffin Gracewell, J., Pavalarajan, S. “Fall detection based on posture classification for smart home environment”. *J Ambient Intell Human Comput* 12, 3581–3588 (2021). https://doi.org/10.1007/s12652-019-01600-y

[3] Nicolai Wojke and Alex Bewley and Dietrich Paulus, “Simple Online and Realtime Tracking with a Deep Association Metric”, CoRR, 2017, https://dblp.org/rec/journals/corr/WojkeBP17.bib


Thank
you