Yavar Write Up

- Ashwanth.L 21pc05

Problem Statement : People Fall detection in areas like staircases, escalators,steps etc. from 2d video

Main tasks:

- People detection from given video
- Posture estimation
- Classification of the posture as either fall or no fall
- Evaluation

Requirements: high accuracy and minimal false positives

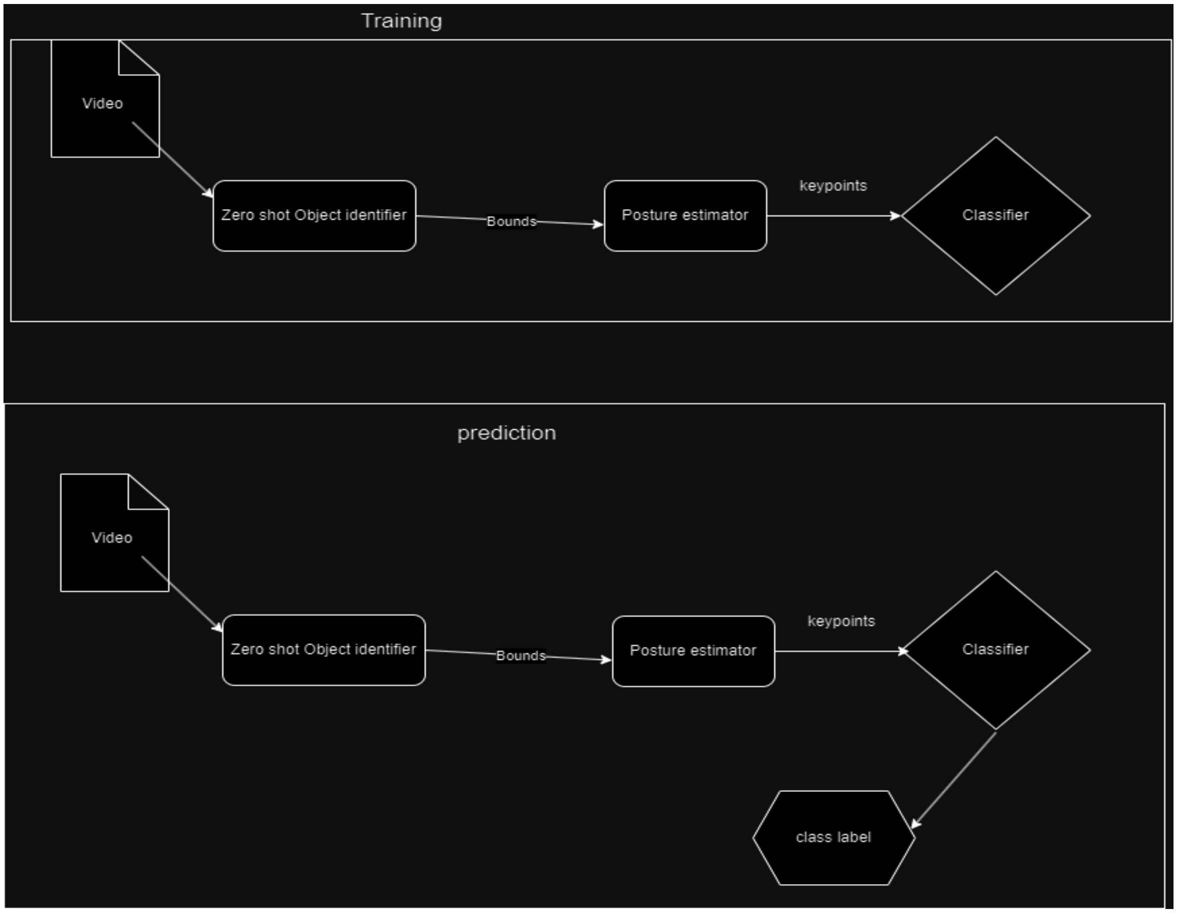
General Approach: Consider windows of time in the video(1 second) and split the one second video into its frames and perform fall detection on each frame and if any of the frames are predicted to have a person falling mark the window and intimate the user

Training Dataset :

- For people detection : COCO
- Falling people : UR fall dataset **http://fenix.ur.edu.pl/~mkepski/ds/uf.html**

Intuition: Transform data and use the class label(since labeled data in the context of if the person is falling or not), use a model to find the association between the the transformed data and class labels,here the transformed data is key points in the 2d plane and based on it identify the associated class label for test data,also since falling is not a single uniform action(fall forward,backward,to the sides etc) make decisions based on the closer(or more similar) points in the feature space rather than the whole dataset

Approach 1 : Using Zero shot Object Detection



Step 1 Object identification :
Use a pre trained model with Zero shot object detection capabilities such as OWL-ViT or OWL-V2,But OWL-V2 performs similar to OWL-Vit but has the advantage of objectness(likelihood of object being present in the box rather than the background) classifier which can be used to rank people detection in multiple people scenario in cases such as stairwell where there are often multiple people,text prompts such as human or person can be used to identify people.

Step 2 Posture estimation :
Use a top down Posture estimation approach in this case use Alphapose along with PoseFlow, an open source model which is know to have good accuracy while having inaccurate bounding boxes(helps in mitigating the effects of prediction errors in the previous steps) using alpha pose the key points can be obtained as well as be stored in different formats such as json which can be used for the next step

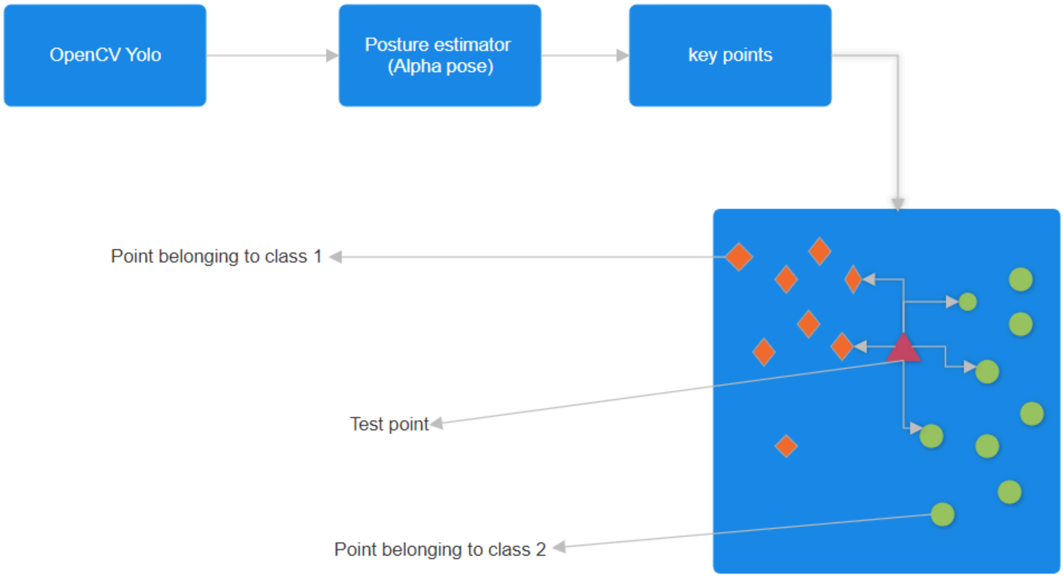
Step 3 Classification of posture:
Consider the key points from the previous steps as features,and since the dataset we use here is a labeled data set we can use the relationship between the

coordinates of the the key points in the 2d plane and the class variable to make decisions, also in order to account for the variations we might face with regards to the scale and resolution of the video , normalize the coordinates by dividing with scale/resolution.Prioritize key points such as the ones that denote the head, shoulders , waist and legs .Using the data which we have obtained till now train a strong classifier such as an SVM or an ensemble of weak classifiers.This concludes the training phase

Prediction Phase: Similar to the training phase obtain the coordinates of the key points (by object identification followed by posture estimation) normalize it and use it as the input for for the classifier the output can be used to predict if a fall has occurred or not

Considerations:
The dataset used contains only data for flat surfaces
Overfitting while using classifiers must be prevented

Approach 2


OpenCV Yolo
Posture estimator (Alpha pose)
key points
Point belonging to class 1
Test point
Point belonging to class 2

Similar to approach 1 but instead of using OWL-Vit use Opencv Yolo the next step of key points identification is the same as the previous approach but here the decision can be made using KNN , ie find N closest neighbors in the n-dimensional space and assign the majority class label to the test data Here the idea is to classify the data set based on the most similar or closest data available, the primary advantage using KNN is that it reduces the training complexity significantly and has high accuracy

Issues and challenges:

- Available dataset contains only 70 clips in total out of which only 30 clips are that of people falling
- The data set contains people falling in flat surfaces only
- A possible solution could be to generate large amounts of video clips of people falling using models such as GAN
- The above approach relies only on a single frame to make decisions, which increases the likelihood of false positives i.e it only considers if people are lying down or not based on a single frame, to solve this previous frames must also be taken into account
- Rather than normalizing the coordinates as suggested other values such as ratio between height and width and angles can also be used