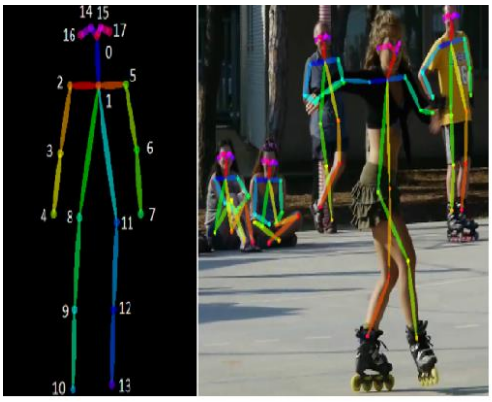
Electronics & ICT Academy
National Institute of Technology, Warangal

Post Graduate Program in

Machine Learning and Artificial Intelligence

Action recognition using pose estimation

Capstone Project II AI ML PG Program by NITW E&ICT Academy



Project Synopsis

Submitted by

Jiss Peter

Technical Architect – Tata America International Ltd (TCS Ltd.)

Jiss07@gmail.com

+1(754)-248-3941

Aim of the Project

Aim of this project is to automatically recognize human actions based on analysis of the body landmarks using pose estimation.
Learning Outcome

1. Implementation of Convolutional Neural Network based pose estimation for body landmark detection
2. Implementation of pose features-based action recognition and its improvement using graphical feature representation and data augmentation of body landmarks
3. Preparation and preprocessing of image datasets
4. Fine tuning and improvement of the action recognition model with better feature representation and data augmentation
5. Development, error analysis and deep learning model improvement
6.

Problem Statement

Analysis of people's actions and activities in public and private environments are highly necessary for security. This cannot be done manually as the number of cameras for surveillance produce lengthy hours of video feed every day. Real-time detection and alerting of suspicious activities or actions are also challenging in these scenarios. This issue can be solved by applying deep learning-based algorithms for action recognition.

Solution Implemented

1.Human Pose Estimation
Like most computer vision problems today, the state-of-the-art approach towards the pose estimation problem is to use a deep learning network called Convolutional Neural Network (CNN). A CNN model is the backbone of any AI-enabled video analytics solution and it needs to be trained using hundreds of thousands of annotated datasets before it can be of any use. There are many open source annotated datasets that support the development of CNN models.
2.Human Action Recognition
Every human action, no matter how trivial, is done for some purpose. For example, in order to complete a physical exercise, a patient is interacting with and responding to the environment using his/her hands, arms, legs, torsos, bodies, etc. An action like this denotes everything that can be observed, either with bare eyes or measured by visual sensors. Through human vision system, we can understand the action and the purpose of the actor. One of the ultimate goals of artificial intelligence research is to build a machine that can accurately understand humans' actions and intentions, so that it can better serve us.

Architecture Used

Convolution operation is one of the fundamental components in deep networks for action recognition, which aggregates pixel values in a small spatial (or spatiotemporal) neighborhood using a kernel matrix. 2D vs 3D Convolution: 2D convolution over images is one of the basic operation in deep networks, and thus it is straightforward to use 2D convolution on video frames. The work in presented a single-frame architecture based on 2D CNN model, and extracted a feature vector for each frame. As multiple frames are presenting in videos, 3D convolution is more intuitive to capture temporal dynamics in a short period of time. Using 3D convolution, 3D convolutional networks (3D

ConvNets) directly create hierarchical representations of spatio-temporal data. The CNN network based architecture is used to resolve the problem of Human pose estimation and an Nural Network (NN) based architecture is used to provide solution for Human action recognition in this project.

Real-World Applications

Pose & Action recognition and prediction algorithms empower many realworld application

- Assisted living
  Personal care robots may be deployed in future assisted living homes.
- Video games
  Commercially, pose estimation has been used in the context of video games, popularized with the Microsoft Kinect sensor (a depth camera). These systems track the user to render their avatar in-game.
- Visual Surveillance
  Security issue is becoming more important in our daily life, and it is one of the most frequently discussed topics nowadays. Places under surveillance typically allow certain human actions, and other actions are not allowed
- Video Retrieval
  Nowadays, due to fast growth of technology, people can easily upload and share videos on the Internet. However, managing and retrieving videos according to video content is becoming a tremendous challenge as most search engines use the associated text data to manage video data
- Human-Robot Interaction
  Human-robot interaction is popularly applied in home and industry environment
- Autonomous Driving Vehicle
  Action prediction algorithms could be one of the potential and may be most important building components in an autonomous driving vehicle.

Conclusion

This project was all about building a convolutional neural network (CNN) classifier to solve the problem of estimating 3D human poses using frames captured from movies as well as to predict the action from photos and videos. Our hypothetical use case was to enable visual effects specialists to easily estimate the pose of actors (from their shoulders, necks, and heads from the frames in a video. Our task was to build the intelligence for this application
The modified VGG16 architecture we built using transfer learning has a test mean squared error loss of ~ 450 squared units over 200 test images for each of the 14 coordinates (that is, the 7(x, y) pairs). We can also say that the test root mean squared error over 200 test images for each of the 14 coordinates is 21.326 units.The root mean squared error (RMSE), in this case, is a measure of how far off the predicted joint coordinates/joint pixel location are from the actual joint coordinate/joint pixel location.

Summary

In this chapter, we successfully built a deep convolution neural network/VGG16 model in Keras on FLIC images. We got hands-on experience in preparing these images for modeling. We successfully implemented transfer learning, and understood that doing so will save us a lot of time. We defined some key hyperparameters as well in some places, and reasoned about why we used what we used. Finally, we tested the modified VGG16 model performance on unseen data and determined that we succeeded in achieving our goals. Also defined a new NN based model for action recognition as well as video action recognition

References

[1] Matthew Lamons, Rahul Kumar, Abhishek Nagaraja - Python Deep Learning Projects , July 2018
[2] TASWEER AHMAD, HUIYUN MAO2, LUOJUN LIN1, GUOZHI TANG - Action Recognition using Attention - Joints Graph Convolutional Neural Networks
[3] Md Matiqul Islam1 , Antony Lam, Hisato Fukuda, Yoshinori Kobayashi and Yoshinori Kuno - An intelligent shopping support robot
[4]Wanqing Li ,Zhengyou Zhang, Zicheng Liu - Action Recognition Based on A Bag of 3DPoints
[5] Online medias such as kaggle, medium, analytics vidya etc.