Journal of Energy Research and Environmental Technology (JERET)
p-ISSN: 2394-1561; e-ISSN: 2394-157X; Volume 7, Issue 2; October-December, 2020 pp. 48-52

http://www.krishisanskriti.org/Publication.html

Realtime Human Pose Estimation for Search and Rescue Operations

Rishabh Jain

Undergraduate Student, Krishna Institute of Engineering and Technology
rishujain1405@gmail.com

Abstract: ***Disaster management requires evacuation of people from the affected regions to safety. However, large masses of people get displaced as the aftermath, which in turn, poses a challenge for the evacuation team. In such a scenario, search and rescue team requires accurate coverage and information of all the regions where the victims are present. To provide a faster and accurate identification of the number of displaced and their locations, this paper proposes a novel approach to search and rescue using human pose identification in any visual input. In this paper, a complex human pose classification problem is solved using the proposed OpenPose model. This model uses spatial as well as temporal features of the video for the classification of different human poses, to recognize what the rescuee is trying to of people with varying body shapes, size, and locations in the frame. The proposed dataset is MPII Human Pose Dataset which is one of the best state-of-the-art datasets comprising of images and video burst-frames of people showing different body postures; useful in accurate human pose estimation. Also, the testing of the model is done by using self-filmed drone videos and several internet sources. The major contribution of this paper is (1) a novel 2DCNN powered model for keypoint classification, (2) identification of the pose of the human body after joining all the keypoints and determining the correct pose. The model accurately predicts all the keypoints present on all the human bodies present in the frame and then, predicts the best pose identified, among the following poses: jump, kick, wave, punch, run, etc.***

Keypoints: OpenPose, HRnet, Human Pose Estimation, Search and Rescue, Deep Learning, Convolutional Neural networks.

1. INTRODUCTION

1.1 Background of the Research paper

More than seventy percent of a person's communication is based on his body postures. Under intense life-threatening circumstances in disaster-struck locations, it is natural for a person to shout for help, but more than that, a "wave" sounds more persuasive when help seems to be at a distance. Such a pose can be used to identify the needy. A CCTV camera can record visual footage of a violent encounters between two persons like punches, kicks, headbutts, etc. in real-time that can be recognized to enhance violence detection in real-time. Henceforth, if a program can identify the pose of not just a single person, but every person in a frame of video, the problem of identification of the intention of the frame can be identified using poses of different people in it. A recon drone can be created that feeds its real-time captured video to the program, and thereafter, can help in identifying the needy and also, locating them in the videoframe.

1.2 Common Challenges in the Research

The human body has many attributes such as height, width, arm length, neck size, amount of fat in different regions of the body and many more factors, which makes it difficult for the model to identify the accurate position of keypoints in a human body. The modifications in the open source code of the model OpenPose and HRnet posed a challenge because of frequent changes in its source codes and increased complexity of the entire files database. Also, the drone footage data required for the fruitful training of the model was not readily available. Many errors occurred during the implementation of all the different approaches locally as well as on Google Collaboratory.

1.3 Most applied Approaches and their analysis

The motive of the project was met with the idea of using an object detection machine learning model for human body detection and then, using keypoints in the body to recognize its instantaneous posture. The OpenPose and HRnet models were considered as the favorable models to serve the purpose. OpenPose was applied using TensorFlow base and the model was trained on MPII dataset. HRnet model was also trained with the same dataset regardless of the base frame of the model. However, as per the use case of the project, the HRnet model did not provide good performance in comparison to the OpenPose model.

1.4 Objective of the paper

The paper describes the most accurate way of recognizing multiple human poses in real-time from a video input which can be the fundamental root for solving automated search and rescue operations using drone surveillance. The proposed

model takes in video input and then, identifies the keypoints of each person in the frame of the video and returns the output in the form of an output video running in 10 fps (Frames per second). The paper's novelty lies in the architecture of the OpenPose model and its accuracy. The training of the model took around 57 hours to complete, and after that, the model gives an astounding 97% training accuracy along with 96% validation accuracy. The proposed model has also been deployed to help a general user, use the model to estimate the pose of the people present in the video input of a webcam/ drone cam. Such a solution can be helpful in recognizing different activities done by humans in real-time videos.

2. RELATED WORK

2.1 Overview of the previous approaches

Previously, CMU's implementation of OpenPose in C++, and HRnet were used as the base models to serve the purpose. The main dataset chosen for training and validation is MPII human pose dataset which proved to be a better choice in serving the model. The testing of the model is done on several drone shots filmed manually as well as scavenged from different internet sources. CMU's **OpenPose** is a great real-time multi-person system to simultaneously identify different keypoints on the human body including hand, face and foot, viz. total 135 keypoints, on bursts of images. The model takes an image, local video or a webcam feed and an IP camera as its input and displays basic image/video along with keypoint display/saving (PNG, JPG, and AVI) as output. Using the CMU's OpenPose GitHub repository, the model got implemented on Google Collaboratory. However, the model gave many distinctive errors concerning model version incompatibility due to which it was held astray.

2.2 Different Approaches, Different solutions

The base idea was to propose an accurate video classification approach that would use a dual channel classifying model, i.e., one for valid keypoint recognition and the other for estimation of the human posture by connecting all those keypoints together with the use of two models, OpenPose and HRnet.

The OpenPose separates the spatial and the temporal features of the videoframe and then, uses them to declare keypoints in the frame at different locations in the frame. The input frame is passed through a recurring series of two 3x3 Convolutional layers and then, a single 2x2 Pooling layer until the different features of the frame are obtained (VGG-19 in the figure).

The extracted features are then introduced into two parallel channels of convolutional layers. The first channel is responsible for the prediction of a set of 18 confidence maps, wherein each map corresponds to a certain body part of the human pose skeleton; while the second channel works in predicting a set of 38 Part Affinity Fields (PAFs) which is the probability that the parts can be associated. Such can be inferred by observing the architecture of OpenPose in figure 1.

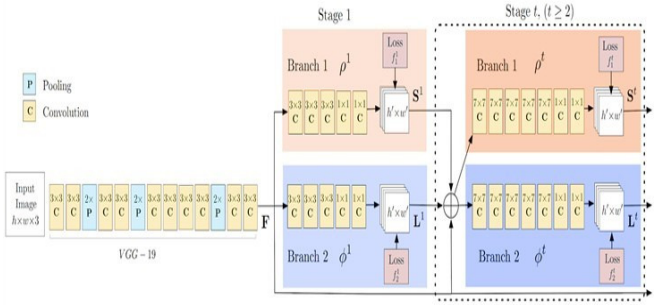

Fig. 1. Architecture of OpenPose model

The HRnet (high-resolution representation for ImageNet classification) is a convolutional neural network-based model that works impressively on high resolution video feeds, but doesn't correspond with the required dataset, due to the person being far away from the capturing lens. Firstly, the image vector of each videoframe is segregated into four resolution feature maps impartially. These maps are fed into a convolutional neural network, which analyzes the features and returns the output in four channels with 128, 256, 512 and 1024 streams respectively. Secondly, the downsampling of the features with high resolution is done with the use of alternating pooling and two 3x3 convolutional layers returning an output of 256 channels, just like OpenPose. Downsampled features are now merged with all the extracted features. This process recursively executes four times to store position of human body(s) in 1024 channels (still in small resolution). Henceforth, a 1x1 convolution, succeeded by a comprehensive average pooling operation, is used to convert 1024 channels to 2048 output channels. Furthermore, these channels are classified into certain classes corresponding to specific human poses using softmax classifier [2]. This is how a human pose can be accurately estimated using HRnet.

Both the models were trained on the MPII Dataset which contains approximately 25K images available for human pose estimation. For the use case scenario, the OpenPose model gave quick and more accurate prediction on an exercise drone video while validation, whereas, the HRnet model was too slow to give a less accurate prediction on the same video due to the varying sizes of the human bodies in the frame of the video being too small or too big in certain time instants.

3. METHODOLOGY

This paper mainly uses OpenPose which is a library written in C++ using OpenCV and Caffe which can be used efficiently for real-time multi-person keypoint detection and multithreading. It gathers three sets of trained models on giving an input feed: one for body pose estimation that is

concerned with classification theories, another one for hands and a last one for faces, just because maximum keypoints are present on a person's face. And each set has several models as the model has been trained on MPII [2-3].

3.1 OpenPose Pipeline

The OpenPose pipeline involves different fields like deep learning, calculus, graph theory and set theory.

1. A numerical matrix that holds the maximum chance that a specific pixel in the input image holds a specific part is known as a 'heatmap'. Each body part of the human body is associated with 18 such heatmaps, along with an additional matrix for storing the extra features. The location of each part of the body from its corresponding heatmaps, is extracted. [5]

2. The matrices that give information about the orientation and the location of the part in the image, is known as Part Affinity Fields. These fields are always stored in a duo as x, y directional pairs for each part in respective totals of 38 in an indexed fashion. Now, the association of such couples of parts into pairs is done, thanks to these 38 matrices [5].

3. Next step is the detection of the parts in the image (or the videoframe). The extraction of parts' locations out of a heatmap is needed now (just like extraction of points in done from a function). This can be done by analyzing the local maximums, which can be implemented using an algorithm that rejects all the non-maximum points (NMS algorithm) [5].

4. Now that the best recognized body parts have been extracted, pairs are formed using these parts. Moreover, a complete bipartite graph, where the edges of the graph are the linked parts and the vertices are the pairs is required to satisfy the solution[5].

5. Furthermore, a line integral is calculated along the edge linking each pair of selected body part, over the respective part affinity fields. This allots each pair a score that is held in a weighted bipartite graph and will help in solving the assignment issue [3].

6. The weighted bipartite graph, after compiling all scores, will now, show each and every possible link between any two keypoints and stores a total for every link. The objective now is to extract the links that maximize the grand total [5].

7. Finally, the extracted links need to be merged and converted into the final keypoint skeleton. Thereafter, the output is a collection of human sets, where each human is a set of parts, over keypoints, where each part contains its index, its relative coordinates and its score [5]. The best human set with the maximum total score is selected as the prediction.

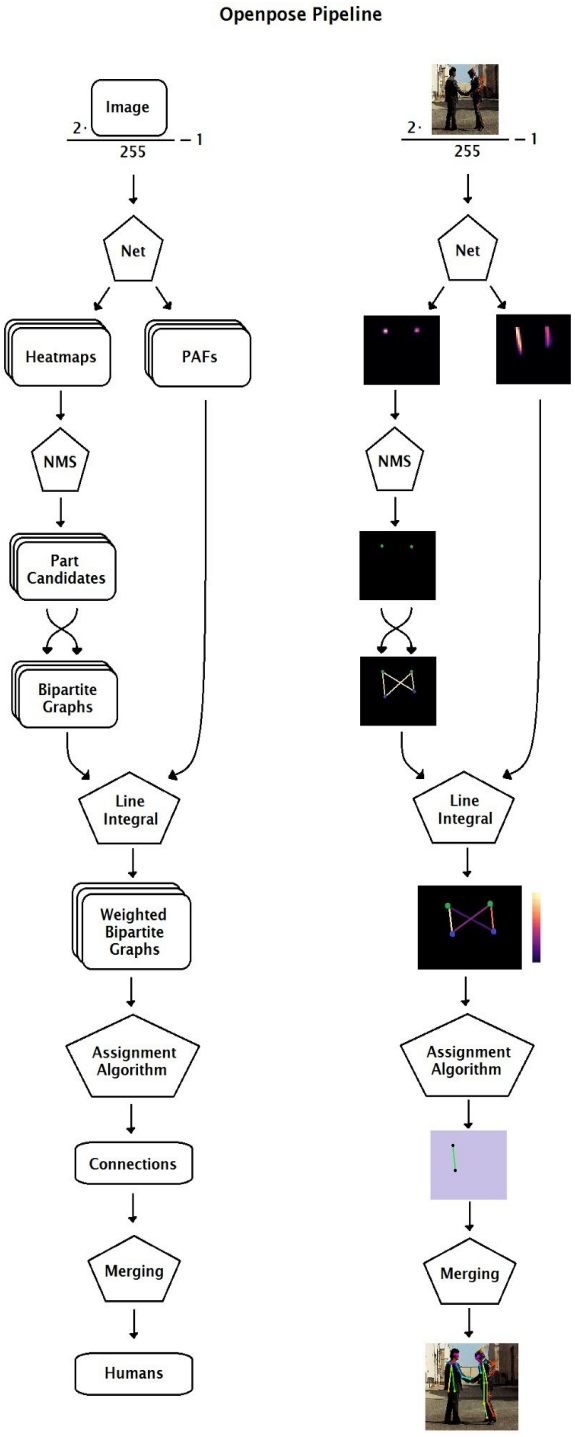

Fig. 2. The OpenPose pipeline explained through a flowchart

4. ANALYSIS

4.1 Choice of Dataset

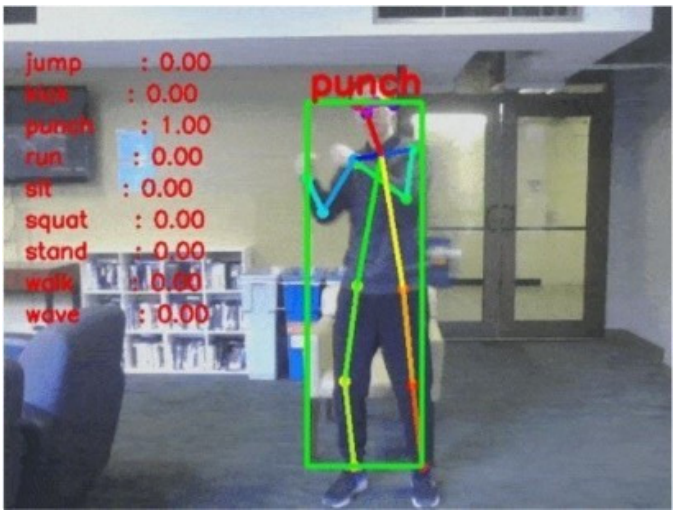

Fig. 3: Model gives output as a 10-fps video with all the keypoints and the recognized pose.

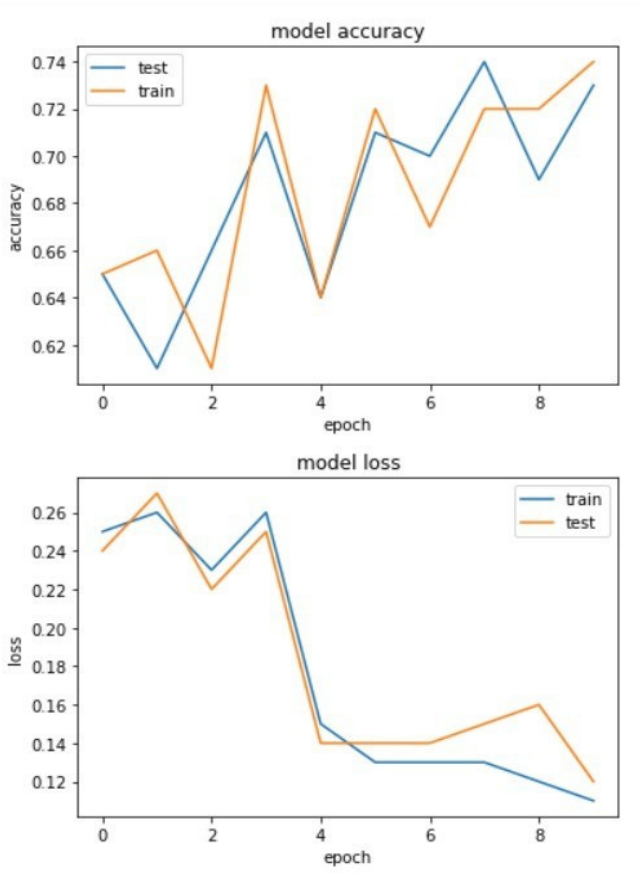

Fig. 4. Graphs showing accuracy of models

MPII human pose dataset serves its purpose at the best when a large amount of pictorial data of one or more human body is required because it comprises of nearly 25,000 images having over 40,000 people with labeled body joints. Therefore, it was used to train the OpenPose model. In addition to the MPII dataset, some real-world data related to the problem statement was also compiled from YouTube and the internet. Moreover, few videos were also recorded manually for testing of the model.

4.2 Experimental Process and Results

The model was tested on videos which contained different actions being performed by people to check if the model detects the actions. The results were highly successful, and the model was able to predict the gestures in these videos. This model gave 76.3 accuracy on the dataset after running for 10 epochs. Due to lack of enough computational power, the model was not able to train with more data and consequently, took too much time in its training.

5. CONCLUSION

5.1 Inference and Future Scope

During this project, we were able to identify the need for pose estimation and its application in different fields.

We were able to recognize postures of the human body from drone surveillance videos. This can prove effective in various Search and Rescue (SAR) Operations by minimizing the risk taken for a manual search operation. This project can further be expanded into various other domains such as using in armed forces to keep an eye on the enemy activities. This can also be used to detect any violent activity taking place. This can also be used in sports to analyze the position of a player while playing, e.g. analyzing the movements of the batsmen while hitting a ball. However, application in each field would require certain changes in the basic model in order to fit the changes required by that particular field. In short, Pose Estimation can be expanded into a number of fields having a number of applications.

Several approaches were tried to predict hand gestures by tuning hyperparameters to a certain range. This instilled in-depth knowledge about neural networks, working on pre-trained models, various architectures of different models and learning the importance and implementation of the project. I also learnt the technique to build and deal with datasets required for this purpose i.e. COCO, MPII and building the dataset according to my own requirement.

5.2 Limitations

Like any other project there are some possible limitations that can faced during deployment. Some of these can be-

- Too much lighting/brightness can affect the video quality being captured.
- Gesture recognition can be difficult during night.
- Camouflaging with the background can also affect the recognition.

Camera quality and height form which the video is captured can also affect the recognition.

REFERENCES

[1] Gines H. Martínez; "OpenPose: Whole Body Estimation", *Carnegie Mellon University*, April 2019

[2] Jingdong W., Ke S., Tianheng C., Borui J., Chaorui D. and Yang Z., "Deep High-Resolution Representation Learning for Visual Recognition", arXiv:1908.07919v2 [cs.CV], 13, March 2020.

[3] Zhe C, Tomas S, Shih-En W, Yaser S; OpenPose: Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields, 30 May 2019.

[4] Balmukund M, Deepak G, Pratik N, Vipul M; "A hybrid approach for search and rescue using 3DCNN and PSO"; 2 June 2020.

[5] Ale Solano, Medium Blog: "Human pose estimation using OpenPose with TensorFlow (Part 1 and 2)", 2 October 2017

[6] MPII Dataset Implementation and Citation; URL: http://human-pose.mpi-inf.mpg.de/

[7] Bharath Raj; Medium Blog: "An Overview of Human Pose Estimation with Deep Learning"; 28 April 2019

[8] "High-Resolution Network: A universal neural architecture for visual recognition"; Microsoft; 17 June 2020