Detecting Key Actors In a Multiperson Video

Gauri Chandgude[1], Viraj Hajare[2], Nahush Dange[3], Shubham Keni[4], Gauri Salunkhe[5]

[1](Electronics and telecommunication, Atharva College Of Engineering/ Mumbai University, India
[2](Electronics and telecommunication, Atharva College Of Engineering/ Mumbai University, India)
[3](Electronics and telecommunication, Atharva College Of Engineering/ Mumbai University, India)
[4](Electronics and telecommunication, Atharva College Of Engineering/ Mumbai University, India)
[5](Electronics and telecommunication, Atharva College Of Engineering/ Mumbai University, India)

Abstract: *Recognision of a key person in multiperson video is a challenging task, with many people active in the scene but only detecting the desired person. Detection of a desired person in videos has hugely benefited from the introduction of recent large-scale datasets and models. This is mainly limited to the domain where the videos contain only one actor performing only one action. In our project, we are presenting a model and dataset for this particular setting. Videos captured in public places typically contain many people interacting with each other. In our work, we use annotation to highlight the detected key person. We show how to use a neural network to represent information of a desired person; the annotation model is trained with selecting the most relevant figure of that person in each frame. We are introducing an annotation based model for detection in multi-person videos. Our method can generalize to any multi-person setting.* ***Keywords:****API, dataset bias, deep learning, Labelling, Tensorflow, Open CV.*

I. Introduction

Person detection and object detection is one of the active research topics in computer vision. Wide range of applications exists for tracking and behaviour recognition were studied by many researchers. Automated system for tracking moving objects have received a lot of attention and importance from industries and academia for its potential applications in the fields of surveillance and engineering such as video police investigation, content based image retrieval, and gait recognition. Current research in this section focuses on three main steps: 1.Low level (Detection) 2.Intermediate level (Tracking) 3.High level (Behavioural Analysis) .Deep Learning never fails to amaze us. It has a profound impact on several domains. Image classification using convolutional neural networks (CNNs) is fairly easy today, especially with the advent of powerful front-end wrappers such as Keras with a TensorFlow backend. But what if you would like to spot quite one object in Associate in nursing image? This drawback is named “object localization and detection.” It is much more difficult than simple classification. Until 2015, image localization victimisation CNNs was terribly slow and inefficient. But no need to Worry, TensorFlow’s Object Detection API comes to the rescue! They have completed most of the work for you. All you have to do is to prepare the dataset properly and set some configurations.

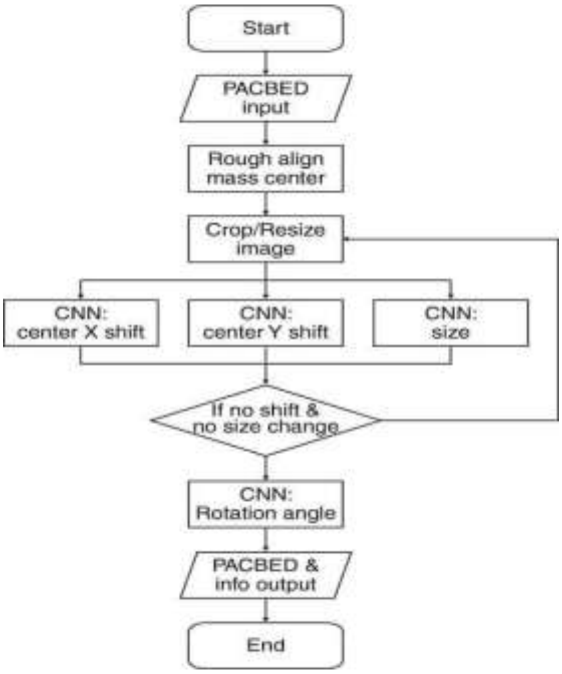
You can train your model and use then it for illation. TensorFlow also provides pre-trained models, trained on the MS COCO, Kitti, or the Open Images datasets. We could use them as such, if you just want to use it for standard object detection. In this age of cheap drones and affordable satellite launches, there has never been that much data of our world from above. There already exists companies using satellite imagery from companieslike Planet and Descartes Labs, applying object detection to count cars, trees and ships. This has resulted in high quality data, which was impossible to get before, now reaching a broader audience. Some companies are making use of drone footage for automatic inspections to reach places or using object detection for general purpose analysis. Also some companies add automatic detection and location of problems without the need for human intervention. Deep learning has been a true game changer in machine learning, especially in computer vision. In a similar manner that deep www.iosrjournals.org learning models have crushed alternative classical models on the task of image classification, they're currently state of the art in object detection as well.

II. Literature review

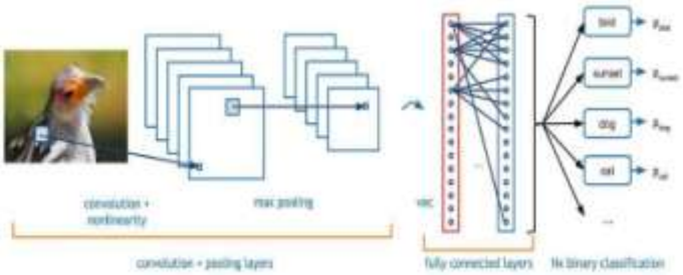
In case paper of A Krenker, J Bešter and A Kos,” Introduction to the Artificial Neural Networks” (2011) neural network-based human detection system was presented. Unlike similar systems which are limited to detecting upright, frontal faces, this system detects humans at any degree of rotation in the video plane.

In a paper of P.kumar (2011) ,proposed FNN combined with RNN approach to reduce the computation time for locating human faces.The experimental results of comparison with conventional neural networks showed that the high speed is achieved once applying FNN.In a case paper of Tanvi Patel (2015), described a rule-based algorithm for robust human action recognition combined with event detection using a convolutional neural network (CNN).Different architecture, approach, programming language, processor and memory requirements, database for training/testing images and performance measure of human detection system were used in each study of different ANNs algorithmIn the paper of Vignesh Ramanathan1, Jonathan Huang2, Sami Abu-El-Haija2, Alexander Gorban2,Kevin Murphy2, and Li FeiFei, IEEE 2016,they proposed a model that combine the RNN and CNN (RCNN) which will Give full play to their respective advantages: RNN can learn temporal and context features, especially long-term dependency between two entities, while CNN is capable of catching more potential. Recognition were studied by many researchers.In case paper of A Krenker, J Bešter and A Kos," Introduction to the Artificial Neural Networks" (2011) neural network-based human detection system was presented. Unlike similar systems which are limited to detecting upright, frontal faces, this system detects humans at any degree of rotation in the video plane.The system employs multiple networks; the primary could be a "router" network that processes every input window to see its orientation then uses this data to arrange the window for one or more detector networks.In a paper of P.kumar (2011) ,proposed FNN combined with RNN approach to reduce the computation time for locating human faces. The experimental results of comparison with conventional neural networks showed that the high speed is achieved when we apply FNN. In a case paper of Tanvi Patel (2015), described a rule-based algorithm for robust human action recognition combined with event detection using a convolutional neural network (CNN).

III. Flowchart



CNNs have wide applications in image and video recognition, recommender systems and linguistic communication process. In this article, the instance that i'll take is expounded to digital Vision. However, the essential thought remains an equivalent and might be applied to the other use-case! CNNs have been made up of neurons with learnable weights and biases.Each somatic cell receives many inputs, takes a weighted sum over them, pass it through an activation function and responds with an output.The whole network includes a loss operate and every one the information and tricks that we tend to developed for neural networks still apply on CNNs.In deep learning, CNN is a class of deep neural networks and most commonly used to analyse visual images. Convolutional neural network uses a variety of multilayer perceptrons designed to want stripped preprocessing. They are additionally called shift invariant or house invariant artificial neural networks (SIANN), supported their shared-weights design and translation invariance characteristics.Little pre-processing compared to other image classification algorithms. This means that the network learns the filters that in traditional algorithms were hand-engineered. This prior knowledge and human effort in featured design is a huge advantage. A convolutional neural network consists of associate input associated an output layer, similarly as multiple hidden layers.The hidden layers of a CNN usually encompass convolutional layers, RELU layer i.e.activation operate, pooling layers, absolutely connected layers and social control layers.Description of the method as a convolution in neural networks is by convention. Mathematically it's a cross-correlation instead of a convolution (although cross-correlation may be a connected operation).This solely has significance for the indices within the matrix, and so that weights ar placed at that index.



IV. Annotations

Every image localization task requires ground truth annotations. The annotations used here are XML files with 4 coordinates representing the location of the bounding box surrounding a person, and its label. We use the Pascal VOC format. Here we've taken an example of Ronaldo. We've labeled around 250 images to create database.Separation of our dataset into two folders, namely images and annotations was must. We placed the label_map.pbtxt and trainval.txt inside your annotations folder. We Created a folder named as xmls inside the annotations folder and place all your XMLs inside that. While performing a standard image classification, given an input image, we presented it to our neural network, and we obtained a single class label and perhaps a probability associated with the class label as well. This class label is meant to characterize the contents of the entire image,visible contents of the image.

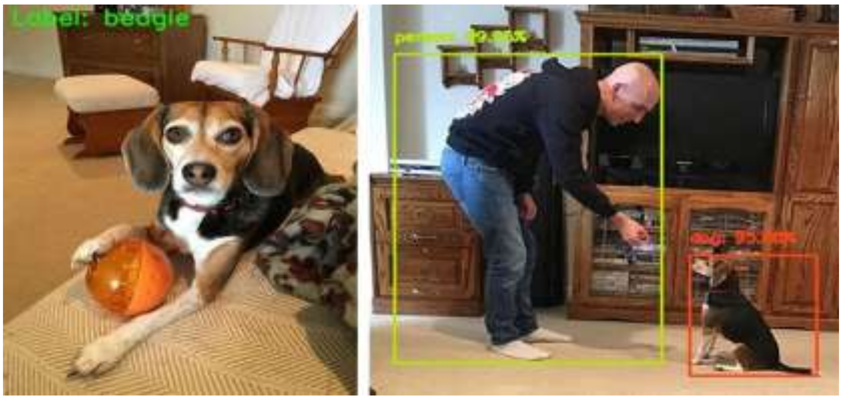


Figure 1 (right) explains an example of implementation of deep learning object detection. We can easily notice how both the person and the dog are localized with their bounding box

When an input image, we wish to obtain: 1) A list of bounding boxes, or the (x, y)-coordinates for each object in an image. 2) The class label associated with each bounding box. 3) The probability/confidence score associated with each bounding box and class label.

V. Training

We first need to select a localization model to train. The main problem we faced was, there are so many options to we had to choose from. Every option vary in performance in terms of speed and accuracy. One have to choose the right model option for the right job. This paper is a good read for the one who wishes to learn more about the trade-off model options. SSDs are fast but they sometimes might fail to detect small objects with accuracy,

whereas Faster RCNNs are relatively slow and large, but have better accuracy. TensorFlow Object Detection API has given us a bunch of pre-trained models. It is highly beneficial to initialize training using a pre-trained model. It can heavily reduce the training time.

In this paper we are presenting a method We've utilized that inclused, use of fixed size sliding windows, which slide from left-to-right and top-to-bottom to localize objects at different locations. Also Use of An image pyramid to detect objects at varying scales And Classification via a pre-trained (classification) Convolutional Neural Network .

At each stop of the sliding window + image pyramid, we extract the ROI, feed it into a CNN, and obtain the output classification for the ROI. If the probability of classification of label L is higher than some threshold T, we mark the bounding box of the ROI as the label L. Repeating this process for every stop of the sliding window and image pyramid, we obtain the output object detectors. Finally, we applied non-maxima suppression to the bounding boxes yielding our final output detections.

VI. Conclusion

We are introducing a new attention based model for detection in multi-person videos. Our model can identify the most relevant key person without being explicitly trained with such annotations. Our method can generalize to any multi-person setting.We propose an extremely lightweight yet highly effective approach that builds upon the latest advancement in detection and video understanding.

Reference

[1]. A Krenker, J Bešter and A Kos," Introduction to the Artificial Neural Networks", Edited Kenji Suzuki, Published by InTech,, JanezaTrdine, Croatia, (2011), pp 3-18.

[2]. Er. P Kumar, Er.P Sharma, "ARTIFICIAL NEURAL NETWORKS-A Study", International Journal of Emerging Engineering Research and Technology, vol. 2, no. 2, (2014), pp. 143-148.

[3]. Vignesh Ramanathan1, Jonathan Huang2, Sami Abu-El-Haija2, Alexander Gorban2,Kevin Murphy2, and Li Fei-Fei, "Detecting events and key actors in multi-person videos" IEEE 2016.

[4]. [4]NouarAlDahoul, AznulQalid Md Sabri, and Ali Mohammed Mansoor, Research article,Computational Intelligence and Neuroscience ",Volume 2018, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia.

[5]. Athanesious, J. & Suresh, P., 2012. Systematic Survey on Object Tracking Methods in Video. Int. J. Adv. Res. Comput. Eng. Technol. 1, 242–247.

[6]. Avidan, S., 2004. Support vector is tracking. IEEE Trans. Pattern Anal. Mach. Intell. 26, 1064–1072. doi:10.1109/TPAMI.2004.53

[7]. Badrinarayanan, V., Perez, P., Le Clerc, F. &Oisel, L., 2007. Probabilistic Color and Adaptive MultiFeature Tracking with Dynamically Switched Priority Between Cues, in: 2007 IEEE 11th International Conference on Computer Vision. IEEE, pp. 1–8. doi:10.1109/ICCV.2007.4408955

[8]. Bagherpour, P., Cheraghi, S.A. & Bin MohdMokji, M., 2012. Upper body tracking using KLT and Kalman filter. Procedia Comput. Sci. 13, 185–191. doi:10.1016/j.procs.2012.09.127