International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 21 - 27
Received: 24/03/2018. Published: 06/04/2018

Monitoring Students using Deep Neural Networks and Clustering Algorithms

[1]Rajive Gandhi C,
[2]Abhishek Dabholkar, [3]Sruthik P, [4]Nishanth Ramesh and [5]Kevin Matthews
[1]Assistant Professor, [2,3,4,5] UG Scholars
Department of Computer Science and Engineering SRM Institute of Science and Technology
Vadapalani Campus, Chennai, India

Abstract—This paper provides a method to classify whether a student inside a classroom is attentive or not. Given a video file of a classroom filled with students as input, we first segregate the video into individual frames and run an Object Detection algorithm to identify the number of students present in each frame. The differences in appearances of each individual student are taken into account to successfully track the set of detected students as unique entities. Bounding boxes are created for the set of students and the coordinates of these bounding boxes are then stored for later inference within the network. We apply Jenks Natural Breaks clustering on the coordinates of the bounding boxes and classify students as attentive or not.
Keywords – clustering, action detection, attentiveness, neural networks

1. Introduction

The field of Computer Vision came a long way since it started under the guise of ”The Summer Vision Project” [1] at MIT’s Computer Science and Artificial Intelligence Lab in the late 1960’s. From very humble beginnings to formulating the benchmark algorithms of the present such as Edge detection, Optical flow and Motion estimation and finally overlapping with the field of Machine Learning by harnessing the power of Convolotional Neural Networks, the field has seen a rapid growth in not just popularity but even in the sheer amount of new algorithms and sub fields being explored. Object Detection and Activity Recognition is one such field where the introduction of constructs such as Neural Networks and Machine Learning algorithms has boosted the efficiency of algorithms and computing devices such as the GPU increased the speed of the older algorithms.

A subset of the Object Detection problem is Human Recognition which forms a very important part in our algorithm for classifying students as attentive or not. Initial work done in the field of Human Recognition was based on SIFT (Scale Invariant Feature Transform)[2] which worked on the basis of identifying interesting points on objects being classified to learn feature descriptors for that particular class. The SIFT algorithm was adept in handling edge cases such as varying scales between training and test object data and partial occlusion of objects in the test data.

An improved version of SIFT was introduced by Robert McConnell [3] in 1986 named Histogram of Oriented Gradients (HOG). While most of the technical details of the algorithm were similar to SIFT, HOG algorithms divided the input image into a collection of connected cells and computed the directed gradients of each cell to detect edges of objects. This increased the accuracy by a lot when compared to the original SIFT algorithm The current state of the art Computer Vision algorithms use Artificial Neural Networks[4] as a backbone. This helps immesly in terms of intuition and accuracy as Artificial Neural Networks succeed in emulation the Human brain’s way of thinking. More specifically, Convolutional Neural Networks [5] have

International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 21 - 27
Received: 24/03/2018. Published: 06/04/2018

been desiged specifically to tackle the overlapping fields of Machine Learning and Computer Vision.

1.1. Goals of the project

The purpose of this project is to apply a combination of Convolutional Neural networks and clustering algorithms to classify whether students in a classroom are paying attention to the teacher. Monitoring of students has always been a problem for teachers because of the sheer disproportion in the student to teacher ratio. Handling and observing multiple students while taking class simultaneously is a chore and we aim on reducing the stress on teachers by automating the process of monitoring students.

2. Defining Attentiveness

The notion of a person being attentive or paying attention is vague. People paying attention to a particular task at hand show some signs that can be inferred as signs of concentrating or paying attention. But these signs vary depending on the task at hand. A carpenter's attentiveness might be quantized into a few parameters which include the relative stillness of his eyes, a repeated pattern in the movement of his hands etc while a driver's signs of attention might be a lack of change in the spatial position of his/her body. Since our usecase pertains to the situation of students present inside a classroom or an exam hall, we will be limiting our definition of attentiveness and its related characteristics to such a scenario.

Figure 1. Action Detection Pipeline

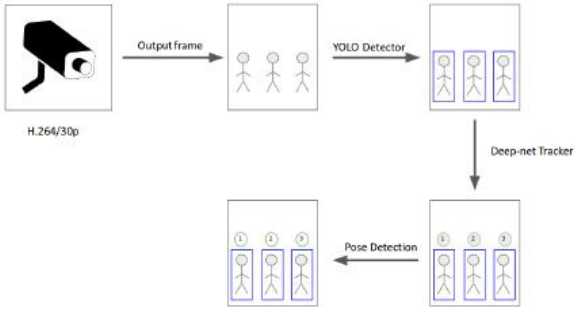
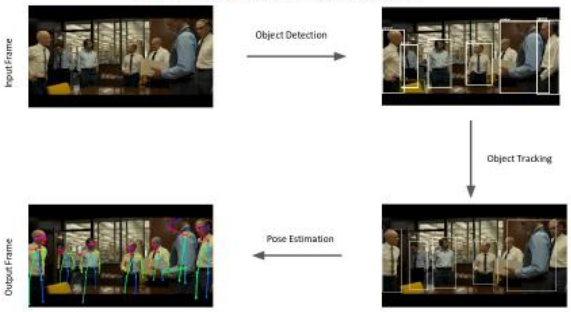


Figure 2. Stages of Action Detection



The algorithm will consider two base situations where it can be applied : A classroom and an examination hall. Inside a classroom, a student paying attention will exhibit a pattern of movements. These patterns may include activites such as moving the neck vertically to glance at the teacher and at the table. All such activites can be translated into a collection of coordinates that vary over time in a restricted space. We deploy clustering algorithms to detect such class of activities based on the movement pattern and classify based on the number of clusters detected and the relative variance between the collected coordinates.

3. Methodology

The algorithm we propose will use the output of various modules such as Object Detection and tracking systems coupled with Jenks Natural Breaks clustering algorithm in order to classify whether a particular student is attentive in the class.

3.1. Algorithm Pipeline

We use a set of modules to help the achieval of our goals. The first module is the Object Detection module which takes care of recognizing a set of objects in a given frame. We use the

International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 21 - 27
Received: 24/03/2018. Published: 06/04/2018

darknet Neural Network framework combined with YOLO [7] model store cognize people in a given frame. The second module is the tracking of distinct people. This module takes care of identifying unique people in the frame, assigning them a number and keeping a track of them throughout the lifetime of the context. The final module is the Pose detector which is concerned with tracking the pose and movements of the previously tracked people. We focus mainly on hand and spatial movements to keep track of a person's attentiveness.

3.2. Pipeline Stages

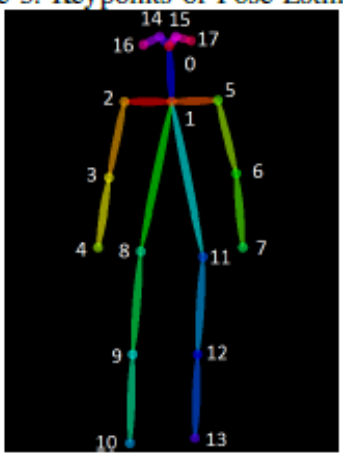
3.2.1. Object Detection Stage. The first stage involves applying object detection algorithms on the frames captured by a camera. We chose YOLO (You only look once) Object detection as it is fast and accurate. The model was trained on the PASCAL VOC datset. An advantage of YOLO was its ability to detect objects based on the global context of the image unlike other methods such as R-CNN (Regional Convolutional Neural Networks) [8]. The image is first divided into a 5∗5 grid which gives rise to a number of cells. If an object's center falls into one of these cells, it is the responsibility of that particular grid cell to detect the object. In a similar fashion, each and every grid cell predicts a number of objects and gives out confidence scores for all the classes they've predicted. These scores convey the relative confidence of an object belonging to that particular class and also whether the object is present at that particular grid cell The Neural Network model used to implement YOLO is a variant of Google's LeNet [9] containing 24 CONV layers with 2 Fully Connected layers at the end. Max pooling is done between successive convolution layers to reduce the resolutionoftheimagesandfocusonmorefinegraindetails. The convolution layers are pretrained on ImageNet's classes using the Darknet framework and for the error function we use Residual Sum of squares.

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Once predicted, bouding boxes are drawn to enclose the objects. For the purpose of detecting human beings only, the other classes are disabled and YOLO is forced to draw bounding boxes for objects that classify humans only. Along with the bounding boxes we get the (x,y) coordinates of the box with respect to the screen space coordinates.

3.2.2. Object Tracking Stage. The next stage deals with tracking the objects detected in the previous stage. In our case, the objects are just people and we use a SORT (Simple Online Realtime Tracking) [10] algorithm to track individual people.

Figure 3. Keypoints of Pose Estimation



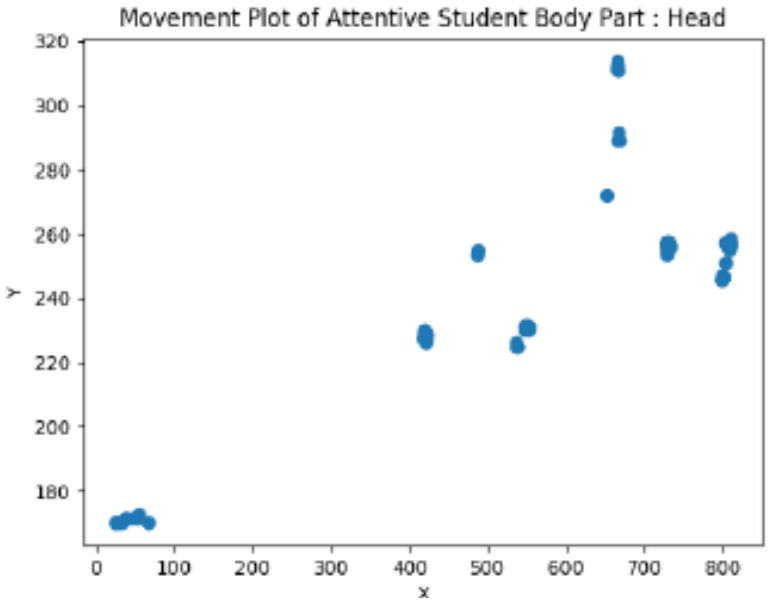
International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 21 - 27
Received: 24/03/2018. Published: 06/04/2018

More specifically, we use a SORT algorithm with a deep association metric colloquially named Deep Sort [11] which integrates the appearance information of people to get more accurate results. The algorithm works on the principle of Linear Quadratic Estimation which uses values measured over time to perform analysis and predicts estimates of unknown variables. The results are more accurate by estimating a joint probability distribution over the unknown variables for each frame across the timeframe There are two steps to Linear Quadratic Estimation - Step 1 is the prediction step where the algorithm first predicts the estimates for the state variables over a given frame. It then does the same for the next frame in the batch and using the results obtained in the later frame, it updates the older estimates by using a weighted average. Using Linear Quadratic Estimation in SORT algorithms promises real time performance and high accuracy, both parameters which are needed for real time monitoring of students. Each person once tracked is still under consideration even if we are unable to track him but once a person can't be tracked for a consecutive number of frames, it is assumed that the person has left the context.

3.2.3. Pose Estimation Stage.

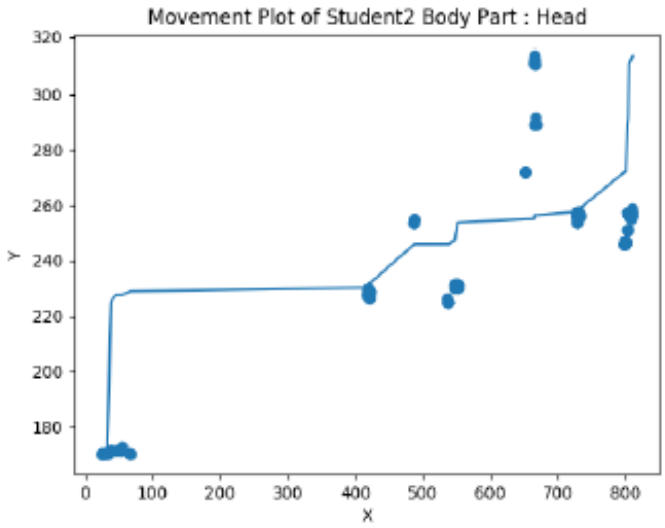
To identify the 2D pose of multiple people present in a frame we use the Open Pose set of tools released by CMU. Open Pose uses Part Affinity Fields [12] that help learn parts of body and associate them with the image. PAFs are a group of 2D vectors that encode the location of various joints of the human body over the context of the image. A split convolution neural network predicts a set of feature maps of the body parts and the part affinity vector fields present in a given image. Both the outputs are then parsed using greedy interference to output 2D keypoint data for every person detected in the video. An L2 loss function is used for both the branches of the CNN. Using Pose Estimation we are able to receive JSON data of the coordinates of the various joints present in people. More specifically, the data is formatted as : [x_1,y_1,c_1,x_2,y_2,c_2,...] where x_1,y_1 are the coordinates of the particular pose keypoint with respect to the image size coordinates and c1 is the confidence parameter of the resspective pose keypoint.

Figure 4. Scatter plot of Attentive Student recorded over 300 frames



International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 21 - 27
Received: 24/03/2018. Published: 06/04/2018

Figure 5. Plot of Attentive Student clustered via Jenks Natural Breaks



3.3. Classifying based on the Pose

After running the pose estimation, the output will contain x and y coordinates for each of the 18 keypoints. We mainly focus on the keypoints that refer to a persons head, neck and shoulders as those are the most important and used elements of a human body during a class. We plot a scatter graph of all the acquired data for each person and it becomes apparent that the people who are paying attention have a pattern of movements. Each of the patterns are clustered into specific regions that can be related to the student's activities performed during a class. Actions such as moving the neck to have a better look at the board result in the neck's keypoint moving in a specific pattern and this is repeated whenever the student performs the activity and forms a specfic pattern in the plot. An unattentive student's coordinate plot will have a lot of variation in the patterns of his movements and there will be numerous number of hotspots present in the scatter plot. To classfiy, we first collate the coordinates of a keypoint collected over a number of frames and sort them in ascending order. We then apply the Jenks Natural Breaks clustering algorithm to cluster the various coordinates collected. Since these coordinates represent the screen space coordinates, a person sitting closed to the camera will have different coordinates when compared to another person sitting far away even if both of them are attentive.

Figure 6. Frame from dataset showing detected people

International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 21 - 27
Received: 24/03/2018. Published: 06/04/2018

So it is important we consider each group of coordinates as independent entities. The reasoning behind choosing Jenks Natural Breaks algorithm was that the data we were using was One Dimensional (Seperate x and y data) unlike other multidimensional clustering algorithms such as k-means algorithm. We finally classify the students based on their attentiveness depending on the number of clusters found and the standard deviation across each cluster.

4. Results

Totesttheproposedalgorithmweacquireddatasetsfrom online YouTube videos that contained static footage of a class being held inside a classroom. The dataset collected was split into three smaller datasets each varying in duration to test the accuracy of the proposed methodology[13]. The object detection module was able to successfully detect students present closer to the camera and had worse luck in identifying students sitting in the last few rows. Overall out of 20 students, we were able to detect 15 resulting in an accuracy of 75%

5. Benchmarks

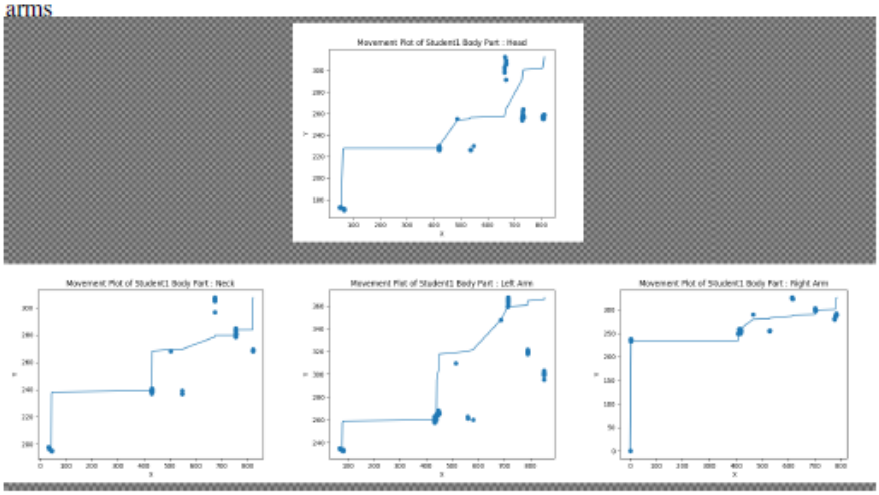
Benchmarks		
Device	Number of Students	Frames per second
CPU	1 10	2 FPS 0.1 FPS
GPU	1 10	30 FPS 10 FPS

The CPU used was an AMD Ryzen 5 1600 running at 3.8 GHz and the GPU was an Nvidia GTX 1060 6GB. The machine learning backend was Caffe with OpenBLAS as its BLAS library

6. Conclusion

A system to monitor students inside a classroom in order to track their attentiveness is proposed and presented in this paper. The accuracy of the system is calculated and estimated to be around 75%

Figure 7. Individual scatter plots of an attentive student's head, neck and arms



VOL 2 ISSUE 4 (2018) PAGES 21 - 27

Received: 24/03/2018. Published: 06/04/2018

References

[1] Papert, Seymour A. The Summer Vision Project, URL : http://hdl.handle.net/1721.1/6125
[2] David G. Lowe Distinctive Image Features from Scale-Invariant Keypoints, URL : https://www.cs.ubc.ca/ lowe/papers/ijcv04.pdf
[3] Robert K. McConnell, Method of and apparatus for pattern recognition, Patent US06400948 [4] Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain
[5] Yann LeCun et al. (1998). Gradient-Based Learning Applied to Document Recognition
[6] Steve R. Gunn Support Vector Machines for Classification and Regression
[7] Joseph Redmon et al. You Only Look Once: Unified, Real-Time Object Detection, arXiv:1506.02640
[8] Ross Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation, arXiv:1311.2524
[9] Christian Szegedy, Yangqing Jia et al. Going deeper with convolutions, arXiv:1409.4842
[10] Alex Bewley et al. Simple Online and Realtime Tracking, arXiv:1602.00763
[11] Nicolai Wojke, Alex Bewley and Dietrich Paulus Simple Online and Realtime Tracking with a Deep Association Metric, arXiv:1703.07402
[12] Zhe Cao and Tomas Simon and Shih-En Wei and Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, CVPR 2017
[13] Student Monitoring Dataset URL : https://goo.gl/AZV4AJ