April 2023

Computer Vision-based Approach for Rejecting Portraits and Mannequins

HP INC

Computer Vision-based Approach for Rejecting Portraits and Mannequins

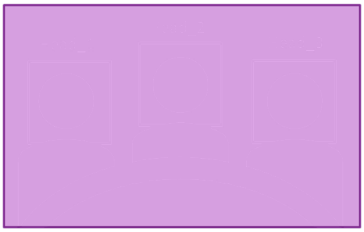
***Abstract*— In this innovation, we propose to detect Human faces that are stationary, and that way make sure that statues, mannequins, portraits containing people, and face false positives appearing on other static objects can be removed from the further processing through framing and tracking experiences. Training halls, seminar and board rooms, conference rooms, office huddle and personal workspaces often have statues, mannequins, or other human face like 3- dimensional physical figures or people in the portraits or personal photo collection that is coming in the field of view of camera. An ability to detect and remove from further processing of such static stationary human faces makes the framing and tracking experience much more stable for the end users; such framing and tracking is reactive only to real life human faces and not to stationary static faces found in statues, mannequins, and portraits.**

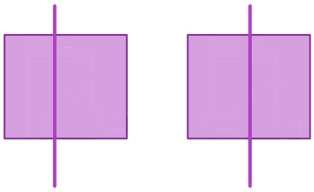
Keywords—computer vision, face detector, motion analysis, video calling, video camera, machine learning

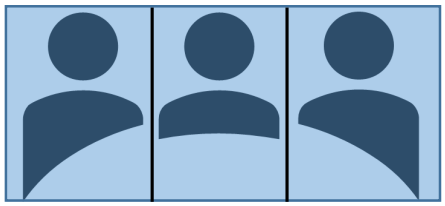
I. Introduction

Various applications in video conferencing systems need to detect people in the conference rooms so experiences like but not limited to frame a group (detect all the people in the conference room and frame them), frame active speakers (detect the active speakers and focus them for the far sight viewing), track presenters (detect active speakers and continuously track them), and frame people individually (detect each individual person the conference room and make a composite stream by assigning each in their own frame).

The examples are given below:







People Framing

However, training halls, seminar and board rooms, conference rooms, office huddle and personal workspaces often have statues, mannequins, or other human head like three dimensional physical figures or people in the portraits or personal photo collection that is coming in the field of view of camera. An ability to detect and remove from further processing of such static stationary human heads/faces makes the framing and tracking experience much more stable for the end users; such framing and tracking is reactive only to real life human heads and not to stationary static heads found in statues, mannequins, and portraits.

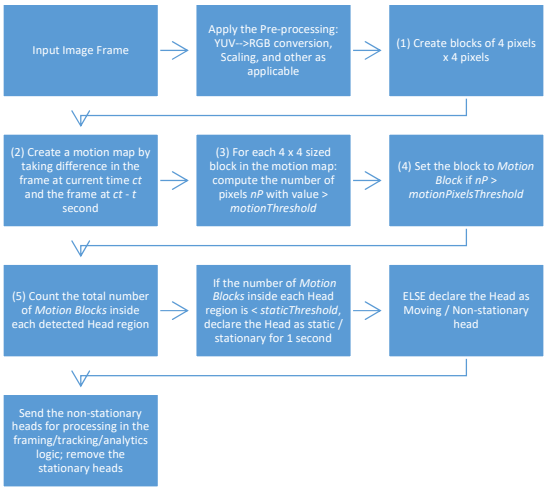
II. Key technology discussions

All the detected people using the AI detection algorithm are kept in the data structure that holds the coordinates of each Head/People along with their detection confidence.
The data structure looks like the following if *n* people are detected:

$$\begin{Bmatrix} x_1\ y_1\ Width_1\ Height_1\ Score_1 \\ x_2\ y_2\ Width_2\ Height_2\ Score_2 \\ \vdots\vdots \\ x_n\ y_n\ Width_n\ Height_n\ Score_n \end{Bmatrix}$$

Here {*x*, *y*, *width*, *height*} refers to the image plane coordinates of heads with their width and height information. *Score* is in the range (0, 100] and reflect confidence in % for each detected heads and *d* is distance in meters for each head. This data structure is then used as an input to various applications such as framing, tracking,

composing, recording, switching, Poly Lens reporting, encoding and as such.



The intuitive examples of components 1 through 5 in the above flow diagram are explained below:

1. Consider an input frame of 720P resolution. That is 1280 x 720. Create a new frame by considering 4 x 4 blocks, computing the mean (average) of each block, and considering that as a pixel value for the new frame that is $1/4^{th}$ the dimension of actual image frame. The new average frame is of dimension 180P. That is 320 x 180. Perform this operation for every frame that is 1 second apart. The timing can be changed depending on the application and the desired motion sensitivity.
2. Compute the difference between the average block frame at time *ct* and *ct-t* second. We recommend keeping *t* = 1 or half a second for this application. Call this a motion map and motion map are of dimension 180P.
3. In 180P dimension motion map, compute the number of pixels in each blocks that has the motion value < *motionThreshold.* Call this value *np.* For example a motion threshold of 25 could be considered.
4. Call this block a motion block if *np* > *motionPixelThreshold*. Motion pixel threshold is a constant and if number of pixels in the motion block is greater than this threshold, we declare the block as motion block, i.e., 1; otherwise stationary block, i.e., 0.
5. Inside each detected Head Region, count the total number of *motion* and *stationary* blocks. If *stationary* blocks count is greater that *motion* blocks, declare the Head as stationary for *t* seconds. Here *t* is 1 second for example.
6. Repeat the operations for set number of seconds, e.g., 30. If the Head is found to be stationary for that many seconds, declare the Head as stationary, statue, portrait, mannequin and discard from the downstream processing through framing and tracking logic.

III. Advantages

Fully software solution performing on AI-based head detectors. Can be applicable in detecting any other static objects on top of the object detector, i.e., static chair detector if the object detection model is detecting chairs.

Works with the existing camera installations that has the view of Humans from any angle.

The detection itself running locally on the embedded device, i.e., camera end point. All the cloud-based connectivity, software-based post-processing of the detection itself for image quality enhancements, and sharing, recording, framing, and streaming of Human Head can be done on top of the AI-based detections in the application layer.